Predicting Protective Linear B-cell Epitopes using Evolutionary Information

Yasser EL-Manzalawy[1], Drena Dobbs[2], Vasant Honavar[1]
Computer Science Department[1]
Department of Genetics and Cell Biology[2]
Iowa State University
Ames, Iowa, USA
{$yasser, ddobbs, honavar$} @iastate.edu

Abstract

Mapping B-cell epitopes plays an important role in vaccine design, immunodiagnostic tests, and antibody production. Because the experimental determination of B-cell epitopes is time-consuming and expensive, there is an urgent need for computational methods for reliable identification of putative B-cell epitopes from antigenic sequences. In this study, we explore the utility of evolutionary profiles derived from antigenic sequences in improving the performance of machine learning methods for protective linear B-cell epitope prediction. Specifically, we compare propensity scale based methods with a Naive Bayes classifier using three different representations of the classifier input: amino acid identities, position specific scoring matrix (PSSM) profiles, and dipeptide composition. We find that in predicting protective linear B-cell epitopes, a Naive Bayes classifier trained using PSSM profiles significantly outperforms the propensity scale based methods as well as the Naive Bayes classifiers trained using the amino acid identity or dipeptide composition representations of input data.

1. Introduction

B-cell epitopes are antigenic determinants that are recognized and bound by receptors (membrane-bound antibodies) on the surface of B lymphocytes [26]. The identification and characterization of B-cell epitopes plays a crucial role in vaccine design, immunodiagnostic tests, and antibody production. At present, several techniques are available for experimental identification of B-cell epitopes [19]. However, their high cost prohibits their use on a genomic scale. Hence, there is an urgent need for computational methods for reliable prediction of B-cell epitopes [14].

There are two types of B-cell epitopes: linear (continuous) and conformational (discontinuous). Linear epitopes are short peptides, corresponding to a contiguous amino acid sequence fragment of a protein [3, 17]. In contrast, conformational epitopes are composed of amino acids that are not contiguous in primary sequence, but are brought into close proximity within the folded protein structure. Although it is believed that a large majority of B-cell epitopes are discontinuous [34], experimental epitope identification has focused primarily on linear B-cell epitopes [12]. Several linear B-cell epitopes in B-cell epitope databases [25, 28] fail to produce neutralizing antibodies (and hence fail to offer protective immunity). This has led to efforts to compile well-characterized datasets of protective linear B-cell epitopes, i.e., those that offer protective immunity [31]. The primary focus of this paper is on predicting protective linear B-cell epitopes.

Classical methods of identifying potential linear B-cell epitopes from antigenic sequences typically rely on the use of amino acid propensity scales [23, 21, 15, 11, 24, 22, 1, 20, 29]. However, as shown by Blythe and Flower [5], the performance of such methods is only marginally better than that of random guessing. Hence, several methods based on machine learning and statistical approaches have been recently proposed for predicting linear B-cell epitopes [18, 30, 32, 8, 31, 10, 9].

Inspired by the analysis presented by Söllner et al. [31] and several studies [27, 6, 13] suggesting that conserved regions in antigens are good targets for developing vaccines, we explore the utility of evolutionary profiles features, e.g., position-specific scoring matrices (PSSM), to improve the performance of predicting protective linear B-cell epitopes. We compare propensity scale based methods with a Naive Bayes classifier using three different representations of the classifier input: amino acid identities, position specific scoring matrix (PSSM) profiles, and dipeptide composition. We compared these methods on two datasets: a dataset of linear B-cell epitopes derived from BciPep database [28]; and the dataset of protective linear B-cell epitopes introduced by Söllner et al. [31]. Our experimental results show that in predicting protective linear B-cell epitopes, a Naive Bayes

classifier trained using PSSM profiles significantly outperforms the propensity scale based methods as well as the Naive Bayes classifiers trained using the amino acid identity or dipeptide composition representations of the data.

2. Materials and Methods

2.1 Datasets

We used two datasets in this study:

1. **Protectivity dataset** [31], which is, to the best of our knowledge, the first and only available dataset of protective linear B-cell epitopes. This dataset is comprised of 57 non-redundant pathogen proteins extracted from IEDB database [25]. Each of these 57 antigens is annotated with a number of linear B-cell epitopes that are classified as "leading to biological activity". The resulting dataset of B-cell epitopes is believed to closely approximate a dataset of protective linear B-cell epitopes [31].

2. **BciPep dataset**, a dataset of 125 non-redundant antigens at 30% sequence similarity cutoff constructed from BciPep database [28]. Peptide-based methods for identifying linear epitopes utilize the target antigen for deciding on a set of overlapping epitopes to be synthesized on pins (PEPSCAN), on a cellulose membrane support (SPOT), or on micro-arrays [19]. The synthetic peptides are then being examined for antibody binding. Hence, the presence of a purified antigen for mapping of linear B-cell epitopes using peptide-based methods is not required. Based on this observation of the independence of antigen when identifying linear B-cell epitopes, we label the residues in the 125 antigen sequences as follows: First, we collect a set of 1230 unique B-cell epitopes included in BciPep database. Then, we compare each protein sequence against each epitope in the set of unique epitopes to find exact matches. For each hit, we assign positive labels to antigen residues included in that match. Thus, if a reported epitope sequence is repeated in an antigen sequence, all occurrences of that epitope will receive positive labels. For example, this procedure assigns positive labels to each of the 25 occurrences of the synthetic epitope "TPSTPA" in the repetitive shed acute-phase antigen (SAPA) from Trypanosoma cruzi. Moreover, if an epitope sequence x is reported to be in an antigen A but it happens that x also occurs in an antigen B and both A and B are in our dataset, then both occurrences of x receive positive labels.

2.2 Feature representation

In our setup, the classifier receives a nine amino-acid window as input. A label is assigned to the instance corresponding to the label of the residue at the center of the window. A positive label indicates that the target residue, the residue at the center of the window, is included in an epitope. A negative label denotes that the target residue is not included in any reported epitope.

We explore three alternative representations of the nine amino acid windows: (i) *Amino Acid Identity (ID) representation*: Each 9-mer window is represented by an ordered 9-tuple of amino acids from the 20-letter amino acid alphabet; (ii) *PSSM representation*: Each antigen sequence in the datasets is aligned against a non-redundant dataset of all currently known sequences using PSI-BLAST [2] with three iterations and cut-off at 10^{-3}. Each residue in the 9 amino acid window is then encoded using the (PSSM) matrix for that residue in the resulting PSSM profile. Thus, each 9-mer window is represented by 9×20 feature vector; (iii) *Dipeptide composition (DC) representation*: Dipeptide composition represent an amino acid sequence (of any length) using the observation frequency for each possible dipeptide in the given sequence. With 20 amino acid alphabet, each 9-mer window is represented by a feature vector of 400 dimensions which correspond to the frequencies of occurrence of each of the 20×20 possible dipeptides.

3. Results and Discussion

We compared the performance of the Naive Bayes classifiers using the sequence identity (NBID), PSSM profiles (NBPSSM), and dipeptide composition (NBDC) representations of the data with five propensity scale based methods [21, 15, 11, 24, 16] on the protectivity and BciPep datasets using 5-fold *sequence-based* cross-validation [7]. The predictive performance measured by the area under the Reciever Operating Characteristic (ROC) curve is summarized in Table 1. The ROC curves are shown in Figure 1.

In predicting protective linear B-cell epitopes, Parker's method [21] slightly outperforms the other four propensity scale based methods and even the Naive Bayes classifiers evaluated using sequence identity or dipeptide composition features; and Naive Bayes classifier evaluated using the PSSM representation of the data outperforms all other methods. NBPSSM ROC curve dominates the ROC curves for each of the other methods.

In predicting linear B-cell epitopes, we find that all of the five propensity scale based methods marginally outperform random guessing (AUC=0.5). This result is consistent with the results of Blythe and Flower's study on a smaller dataset of 50 proteins [5]. Perhaps more interesting is the finding that none of the three Naive Bayes classifiers offer improve-

ments over the propensity scale based methods. Thus, the ROC curves for all of the methods are close to a diagonal connecting points (0,0) and (1,1) which corresponds to a classifier that assigns labels by random guessing.

The superior performance of a Naive Bayes classifier evaluated using PSSM-based data representation in predicting protective linear B-cell epitopes, underscores the functional importance of sequence conservation (previously noted by several authors [27, 6, 13, 31]). It also suggests that conserved regions in antigenic sequences are good candidates to target for developing new vaccines. This possibility needs to be further explored by applying sequence variability analysis methods [4, 13]. It is also worth noting that highly variable sequence residues can be functionally important [13]. PSSM profiles contain information that can be used to distinguish highly conserved residues from those that are not conserved and from those that are only moderately conserved. Hence, combining PSSM profiles with machine learning methods provides a powerful tool for discovering useful patterns for predicting functionally important residues without the need for any a-priori assumptions regarding the conservation or variability of the functional residues.

Work in progress is aimed at further improving the performance of methods for protective B-cell epitope prediction by:

1. Incorporating feature selection, feature abstraction, and dimensionality reduction methods to minimize the deleterious effects of redundant and irrelevant features.

2. Exploring more sophisticated machine learning methods such as Support Vector Machine [33].

3. Exploring the utility of additional sequence-derived features, e.g., predicted solvent accessibility (since linear B-cell epitopes are believed to be exposed to the surface of the antigen).

References

[1] A. Alix. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*, 18:311–4, 1999.

[2] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3390–3402, 1997.

[3] D. Barlow, M. Edwards, J. Thornton, et al. Continuous and discontinuous protein antigenic determinants. *Nature*, 322:747–748, 1986.

[4] C. Berezin, F. Glaser, J. Rosenberg, I. Paz, T. Pupko, P. Fariselli, R. Casadio, and N. Ben-Tal. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, 20:1322–1324, 2004.

[5] M. Blythe and D. Flower. Benchmarking B cell epitope prediction: Underperformance of existing methods. *Protein Sci*, 14:246–248, 2005.

[6] H. Bui, J. Sidney, W. Li, N. Fusseder, and A. Sette. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics*, 8:361, 2007.

[7] C. Caragea, J. Sinapov, V. Honavar, and D. Dobbs. Assessing the Performance of Macromolecular Sequence Classifiers. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, pages 320–326, 2007.

[8] J. Chen, H. Liu, J. Yang, and K. Chou. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, 33:423–428, 2007.

[9] Y. El-Manzalawy, D. Dobbs, and V. Honavar. Predicting Flexible Length Linear B-cell Epitopes. *7th International Conference on Computational Systems Bioinformatics (CSB'08)*, pages 121–132, 2008.

[10] Y. El-Manzalawy, D. Dobbs, and V. Honavar. Predicting linear B-cell epitopes using string kernels. *J Mol Recognit*, 21:243–255, 2008.

[11] E. Emini, J. Hughes, D. Perlow, and J. Boger. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, 55:836–839, 1985.

[12] D. Flower. *Immunoinformatics: Predicting immunogenicity in silico*. Quantum distributor, 1st edition, 2007.

[13] M. Garcia-Boronat, C. Diez-Rivero, E. Reinherz, and P. Reche. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic Acids Res*, 36:W35, 2008.

[14] J. Greenbaum, P. Andersen, M. Blythe, H. Bui, R. Cachau, J. Crowe, M. Davies, A. Kolaskar, O. Lund, S. Morrison, et al. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit*, 20:75–82, 2007.

[15] P. Karplus and G. Schulz. Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen. *Naturwiss*, 72:21–213, 1985.

[16] A. Kolaskar and P. Tongaonkar. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276(1-2):172–4, 1990.

[17] J. Langeveld, J. martinez Torrecuadrada, R. boshuizen, R. Meloen, and C. Ignacio. Characterisation of a protective linear B cell epitope against feline parvoviruses. *Vaccine*, 19:2352–2360, 2001.

[18] J. Larsen, O. Lund, and M. Nielsen. Improved method for predicting linear B-cell epitopes. *Immunome Res*, 2:2, 2006.

[19] G. Morris. Epitope Mapping: B-cell Epitopes. *Encyclopedia of Life Sciences*, 2007.

[20] M. Odorico and J. Pellequer. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J Mol Recognit*, 16:20–22, 2003.

[21] J. Parker and H. R. Guo, D and. 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*, 25:5425–5432, 1986.

Table 1. AUC values for different prediction methods on protectivity and BciPep datasets using 5-fold sequence-based cross-validation tests.

METHOD	PROTECTIVITY	BCIPEP
PARKER HYDROPHILICITY	0.58	0.56
CHOU AND FASMAN BETA-TURN	0.54	0.56
EMINI SURFACE ACCESSIBILITY	0.57	0.55
KARPLUS AND SCHULZ FLEXIBILITY	0.56	0.54
KOLASKAR AND TONGAONKAR ANTIGENICITY	0.46	0.46
NBID	0.54	0.54
NBPSSM	0.61	0.55
NBDC	0.56	0.54

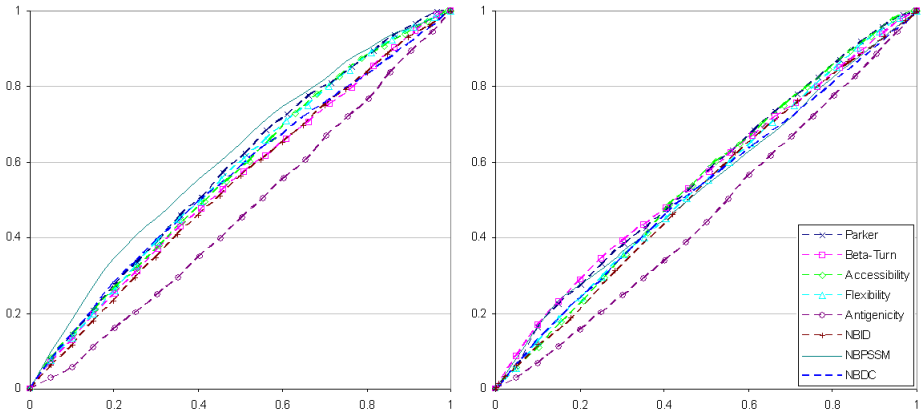

Figure 1. ROC curves for different methods on protectivity (left) and BciPep (right) datasets estimated using 5-fold sequence-based cross-validation.

[22] J. Pellequer and E. Westhof. PREDITOP: a program for antigenicity prediction. *J Mol Graph*, 11:204–210, 1993.

[23] J. Pellequer, E. Westhof, and M. Van Regenmortel. Predicting location of continuous epitopes in proteins from their primary structures. *Meth Enzymol*, 203:176–201, 1991.

[24] J. Pellequer, E. Westhof, and M. Van Regenmortel. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol Lett*, 36:83–99, 1993.

[25] B. Peters, J. Sidney, P. Bourne, H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, et al. The Immune Epitope Database and Analysis Resource: From Vision to Blueprint. *PLoS Biology*, 3:e91, 2005.

[26] G. Pier, J. Lyczak, and L. Wetzler. *Immunology, infection, and immunity*. ASM Press, 1st edition, 2004.

[27] P. Reche and E. Reinherz. Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. *J Mol Biol*, 331:623–641, 2003.

[28] S. Saha, M. Bhasin, and G. Raghava. Bcipep: a database of B-cell epitopes. *BMC Genomics*, 6:79, 2005.

[29] S. Saha and G. Raghava. BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physicochemical properties. *Artificial Immune Systems, Third International Conference (ICARIS 2004), LNCS*, 3239:197–204, 2004.

[30] S. Saha and G. Raghava. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*, 65:40–48, 2006.

[31] J. Sollner, R. Grohmann, R. Rapberger, P. Perco, A. Lukas, B. Mayer, and M. Blythe. Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins. *Immunome Res*, 2008:1–17, 2008.

[32] J. Söllner and B. Mayer. Machine learning approaches for prediction of linear B-cell epitopes on proteins. *J Mol Recognit*, 19:200–208, 2006.

[33] V. Vapnik. *The nature of statistical learning theory*. Springer, 2nd edition, 2000.

[34] G. Walter. Production and use of antibodies against synthetic peptides. *J Immunol Methods*, 88:149–61, 1986.