BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 12 2008, pages 1459–1460
doi:10.1093/bioinformatics/btn199

Structural bioinformatics

PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure

Michael J. Sweredoski[1,2] and Pierre Baldi[1,2,*]

[1]Department of Computer Science and [2]Institute for Genomics and Bioinformatics, University of California, Irvine, 92697-3435, California, USA

Received on March 3, 2008; revised on April 18, 2008; accepted on April 20, 2008
Advance Access publication April 28, 2008
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Accurate prediction of B-cell epitopes is an important goal of computational immunology. Up to 90% of B-cell epitopes are discontinuous in nature, yet most predictors focus on linear epitopes. Even when the tertiary structure of the antigen is available, the accurate prediction of B-cell epitopes remains challenging.

Results: Our predictor, PEPITO, uses a combination of amino-acid propensity scores and half sphere exposure values at multiple distances to achieve state-of-the-art performance. PEPITO achieves an area under the curve (AUC) of 75.4 on the Discotope dataset. Additionally, we benchmark PEPITO as well as the Discotope predictor on the more recent Epitome dataset, achieving AUCs of 68.3 and 66.0, respectively.

Availability: PEPITO is available as part of the SCRATCH suite of protein structure predictors via www.igb.uci.edu.

Contact: pfbaldi@ics.uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

B-cell epitope prediction is an important, but unsolved problem in bioinformatics. The ability to accurately predict B-cell epitopes would aid researchers in a variety of immunological applications.

Initial attempts at predicting B-cell epitopes involved the calculation of propensity scales (Hopp and Woods, 1981). While this information can be useful in predicting B-cell epitopes, Blythe and Flower (2005) showed that propensity scales alone are not enough to accurately predict epitopes.

Many of the previous predictors have focused on linear B-cell epitopes. Some of these methods include ABCpred (Saha and Raghava, 2006), BEPITOPE (Odorico and Pellequer, 2003), Bepipred (Larsen *et al*., 2006) and PEOPLE (Alix, 1999). However, past surveys have estimated that only 10% of the B-cell epitopes are continuous (van Regenmortel, 1996). Additionally, van Regenmortel (2006) noted that even linear epitopes adopt a conformational structure and therefore the distinction is somewhat blurred. Far fewer predictors have been developed for discontinuous B-cell epitopes. One of the first methods explicitly created for identification of discontinuous epitopes was conformational epitope predictor (CEP) (Kulkarni-Kale *et al*., 2005). Another method described by Rapberger *et al*. (2007) incorporates epitope–paratope shape complementarity to predict interaction sites. One of the most recent, state-of-the-art, predictors of discontinuous epitopes is Discotope (Andersen *et al*., 2006), which uses both contact numbers (i.e. the number of Cα atoms within a certain distance threshold) and an amino-acid propensity scale.

Our predictor, PEPITO, attempts to overcome some of the limitations of previous predictors by incorporating an amino-acid propensity scale along with side chain orientation and solvent accessibility information using half sphere exposure values (Hamelryck, 2005). To increase robustness, PEPITO uses propensity scales and half sphere exposure values at multiple distance thresholds from the target residue.

2 METHODS

2.1 Datasets

We obtained epitope datasets for benchmarking prediction methods from both the Discotope Supplementary Materials (Andersen *et al*., 2006) and Epitome (Schlessinger *et al*., 2006). The two datasets contain different sets of protein chains and differ in their epitope/non-epitope classification rules. The Discotope dataset, which consists of 75 protein chains, labels all residues in antigen chains within 4 Å of an antibody as epitopes. The Epitome dataset, which consists of 140 protein chains, seeks to eliminate incidental contacts by labeling residues in the antigen within 6 Å of the complementary determining regions of the antibody chains as epitopes.

We derived two additional datasets, C[Discotope] and C[Epitome], from the set of protein chains that are common to both the Epitome and Discotope datasets. The two datasets differ in the method used to identify epitope residues. Eight hundred and seventy-five of the residues in the derived datasets are defined as epitopes using both methods. Four hundred and seventy-one of the residues in the derived datasets are defined as epitopes using the Epitome method but not the Discotope method. One hundred and nine of the residues in the derived datasets are defined as epitopes using the Discotope method but not the Epitome method. The assertions by Schlessinger *et al*. (2006) would indicate that the 471 residues are integral to the antigen–antibody binding while the 109 residues result from incidental contacts.

Testing procedures require that the protein chains present in the datasets be clustered to prevent any one family from dominating the performance measures. Protein families were previously annotated for the Discotope dataset. UniqueProt (Mika and Rost, 2003) was used to identify protein families in the Epitome dataset and the two derived datasets.

[*]To whom correspondence should be addressed.

2.2 Prediction

For each residue r in the target protein chain, we calculate an epitope score E(r). Large values of E(r) indicate a higher likelihood that the residue r is an epitope residue. The score E(r) is calculated using a linear combination of terms. We also explored non-linear methods such as SVMs, ANNs and Gaussian Mixture Models, but they did not achieve higher performance levels. The score is given by:

$$\mathrm{E}(r) = \sum_{k\in\{8,10,\ldots,16\text{Å}\}} \alpha \cdot PS(r,k) + \beta \cdot \mathrm{HSEup}(r,k) + \gamma \cdot \mathrm{HSEdown}(r,k)$$

The first term PS(r, k) is the sum of the propensity scale scores, averaged over a linear window of nine residues, for all residues within k Å of residue r. The second half-sphere exposure term HSEup(r, k) is the number of Cα atoms in the *up* half sphere within k Å of residue r, and similarly for the third term using the *down* half sphere within k Å of residue r. Intuitively, the HSEup term encodes information on the relative orientation of the side chain—toward the center of the protein or toward the surface—and the side chain accessibility.

Currently, PEPITO uses the propensity scale described by Andersen *et al.* (2006). The coefficients ($\alpha = 1, \beta = -1/2, \gamma = -1/4$) are derived from those previously used by Andersen *et al.* (2006) and the correlations between half sphere exposures and contact number (Hamelryck, 2005).

The server version of PEPITO calculates the epitope score using all residues—only the antigen chain should be used, not the antibody–antigen complex. PEPITO returns a simplified PDB file with the epitope score, expressed as a Z-score, in the B-factor field of each atom. A Z-score threshold of 1.3 will produce a sensitivity >0.3 and specificity >0.9.

3 RESULTS

We benchmark PEPITO, as well as Discotope, on the Discotope dataset as well as the more recently curated Epitome dataset. Following the recommendations made by Greenbaum *et al.* (2007), we use the area under the curve of the receiver operator characteristic (ROC AUC) as the primary performance measure. We also calculate all the other standard performance measures and they are available online. To avoid skewing the performance measures by overrepresented protein families, the results are averaged over the mean performance within protein families.

On the Discotope dataset, PEPITO achieves an ROC AUC of 75.38. On the Epitome dataset, PEPITO achieves an ROC AUC of 68.31. In Table 1, we see that the average ROC AUC increase between PEPITO and Discotope for each protein family is between 2.31 and 3.51. Similar performance improvements were found for the other three datasets. Additionally, the bootstrap estimate of the 95% confidence interval (CI) of the increase in ROC AUC for each dataset shows that the improvements are statistically significant. Additional analyses show that the ROC AUC decreases to 74.42 if contact numbers are used instead of half sphere exposure values and the ROC AUC decreases to 73.50 if only the 10 Å threshold distance is used.

Table 1. ROC AUC for various methods and datasets

Dataset (# families)	Discotope	PEPITO	Mean Δ ROC AUC between families (with 95% CI)
Discotope (25)	72.60	75.38	2.79 (1.69, 4.05)
Epitome (35)	66.00	68.31	2.31 (0.78, 3.94)
C[Discotope] (21)	70.28	73.80	3.51 (1.91, 5.37)
C[Epitome] (21)	67.78	70.45	2.66 (0.78, 4.76)

CIs were determined using 100 000 bootstrap samples.

While there is an increase in sensitivity at 95% specificity on the Discotope dataset from 18.70 using the Discotope method to 20.87 using PEPITO, the bootstrap estimate of the 95% CI (−1.62, 6.90) shows the difference in sensitivities is not statistically significant.

By comparing the ROC AUCs on C[Discotope] and C[Epitome], the two datasets derived for direct comparison of epitope definition, we see that both Discotope and PEPITO perform better on C[Discotope]. This is most likely caused by the optimization of the propensity scores for the Discotope definition of epitopes. Recalculation of the propensity scores for the Epitome definition would likely improve the results for the Epitome and C[Epitome] datasets.

Thus PEPITO is a state-of-the-art B-cell epitope predictor which takes advantage of multiple distance thresholds and half sphere exposure. An online version of PEPITO is available as part of the SCRATCH server suite (Cheng *et al.*, 2005).

ACKNOWLEDGEMENTS

We thank P. Andersen for her assistance with details on Discotope.

Funding: This work was supported by NIH grant LM-07443-01, NSF grant EIA-0321390, and a Microsoft Faculty Research Award to P.B.

Conflict of Interest: none declared.

REFERENCES

Alix,A.J. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*, **18**, 311–314.

Andersen,P. *et al.* (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.*, **15**, 2558–2567.

Blythe,M.J. and Flower,D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.*, **14**, 246–248.

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, w72–w76.

Greenbaum,J.A. *et al.* (2007) Toward a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.*, **20**, 75–82.

Hamelryck,T. (2005) An amino acids has two sides: a new 2D measure provides a different view of solvent exposure. *Prot. Struct. Func. Bioinform.*, **59**, 38–48.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci.*, **78**, 3824–3828.

Kulkarni-Kale,U. *et al.* (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res.*, **33**, w168–w171.

Larsen,J.E. *et al.* (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res.*, **2**, 2.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein-sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

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

Rapberger,R. *et al.* (2007) Identification of discontinuous antigenic determinants on proteins based on shape complementarities. *J. Mol. Recognit.*, **20**, 113–121.

Saha,S. and Raghava,G.P.S. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural networks. *Prot. Struct. Funct. Bioinform.*, **65**, 40–48.

Schlessinger,A. *et al.* (2006) Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res.*, **34**, D777–D780.

van Regenmortel,M.H. (1996) Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods*, **9**, 465–472.

van Regenmortel,M.H. (2006) Immunoinformatics may lead to a reappraisal of the nature of B cell epitopes and of the feasibility of synthetic peptide vaccines. *J. Mol. Recognit.*, **19**, 183–187.