Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

B-Pred, a structure based B-cell epitopes prediction server

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
24 July 2012
Number of times this article has been viewed

Luciano Giacò[1]
Massimo Amicosante[2]
Maurizio Fraziano[1]
Pier Federico Gherardini[1]
Gabriele Ausiello[1]
Manuela Helmer-Citterich[1]
Vittorio Colizzi[1]
Andrea Cabibbo[1]

[1]Department of Biology, [2]Department of Internal Medicine, University of Rome "Tor Vergata", Rome, Italy

Abstract: The ability to predict immunogenic regions in selected proteins by in-silico methods has broad implications, such as allowing a quick selection of potential reagents to be used as diagnostics, vaccines, immunotherapeutics, or research tools in several branches of biological and biotechnological research. However, the prediction of antibody target sites in proteins using computational methodologies has proven to be a highly challenging task, which is likely due to the somewhat elusive nature of B-cell epitopes. This paper proposes a web-based platform for scoring potential immunological reagents based on the structures or 3D models of the proteins of interest. The method scores a protein's peptides set, which is derived from a sliding window, based on the average solvent exposure, with a filter on the average local model quality for each peptide. The platform was validated on a custom-assembled database of 1336 experimentally determined epitopes from 106 proteins for which a reliable 3D model could be obtained through standard modeling techniques. Despite showing poor sensitivity, this method can achieve a specificity of 0.70 and a positive predictive value of 0.29 by combining these two simple parameters. These values are slightly higher than those obtained with other established sequence-based or structure-based methods that have been evaluated using the same epitopes dataset. This method is implemented in a web server called B-Pred, which is accessible at http://immuno.bio.uniroma2.it/bpred. The server contains a number of original features that allow users to perform personalized reagent searches by manipulating the sliding window's width and sliding step, changing the exposure and model quality thresholds, and running sequential queries with different parameters. The B-Pred server should assist experimentalists in the rational selection of epitope antigens for a wide range of applications.

Keywords: B-cell epitopes, immunoinformatics, bioinformatics, web server, epitope prediction

Introduction

Several experimental techniques are currently available for the experimental mapping of B-cell epitopes.[1] However, the time and costs involved in comprehensive epitope mapping of even a single target protein make this approach unfeasible on a genomic scale. In contrast, in-silico B-cell epitope prediction methodologies are a manageable alternative that allow for virtual cost-effective, genome-wide scans in the search for molecules with diagnostic, vaccinal, or immunomodulatory potential. They also generally provide options for the selection of useful biotechnological reagents.

The problem of predicting potential epitopes for target proteins has been investigated by several groups using various methodologies.[2,3] Early methods, based on the analysis of linear protein sequences, attempted to correlate aminoacid propensity scales based on hydrophobicity, hydrophilicity, flexibility, accessibility, or secondary structure

Correspondence: Andrea Cabibbo
University of Rome "Tor Vergata", Department of Biology, Via Della Ricerca Scientifica 1, 00133, Rome, Italy
Tel +39 06 7259 4236
Fax +39 06 7259 4224
Email andrea.cabibbo@uniroma2.it

features with epitopic regions.[4–9] While these methods were widely used to assist in the design of peptides for immunological purposes, their validity has been questioned.[10] Some groups have proposed that combining several different linear prediction indexes might produce more reliable results. This is the basis of the Linear Epitope Prediction Database (LEPD) method.[11]

Among the wealth of parameters considered for optimal epitope prediction,[2,12] the role of residue solvent exposure in the determination of antigenic properties is still a matter of debate. Parameters correlated with exposure, such as hydrophilicity, have been used since the early era of in-silico B-cell epitope prediction methods,[6] based on the commonsense assumption that epitopic residues should be exposed and available for interaction with the antibody–antigen binding regions. The efficacy of such simple propensity scales has been extensively reviewed using updated datasets, and is proven to be extremely poor if not close to random.[10,13] Despite this, simple propensity scales can still be used in combination with other parameters, as shown by the Conformational B-cell Epitope Prediction (CBTOPE) server[14] that incorporates the Parker scale[6] among the aminoacid physico-chemical properties used in the prediction procedure.

By surveying an extensive epitopes dataset that includes structural information, Lollier et al concluded that "epitopic residues are not distributed among any specific Relative Surface Accessibility and Protrusion index values, and in some cases epitopes cover the entire antigenic sequence."[15]

Wang et al[12] propose a highly accurate prediction methodology, which is based on an evaluation, from the linear sequence, of a number of structural features, evolutionary information, disorders, and low complexity information. These data are then used in different combinations to identify the best prediction methodology. It is worth noting that the best predictor includes residue accessibility information as one of the parameters, while disorders and low complexity information do not contribute significantly to the method.

More recently, Ponomarenko et al[16] proposed a structure-based method called ElliPro that approximates protein surface patches as ellipsoids and then scores the residues belonging to the surface patches based on a protrusion index, which is a sophisticated way of selecting the most exposed residues in the structure.

In a recent paper, Zhang et al[17] presented a novel prediction methodology based on the concept of the "thick surface patch," in which both exposed residues and interior residues concur on the definition of the potential conformational B-cell epitopes.

The present study addresses the possibility of using solvent exposure, directly calculated from 3D data, as a simple parameter to be correlated with the probability that a given peptide sequence would contain a B-cell epitope. We find that the correlation improves if the local quality of the model/structure is also taken into account. The method, named B-Pred, is accessible on the Internet at http://immuno.bio.uniroma2.it/bpred. The method and server focus on the prediction of potential linear epitopes, as opposed to discontinuous epitopes. This makes the server particularly useful in the selection and design of immunological and biotechnological reagents, as the conversion of a linear epitope to a peptidic reagent is relatively straightforward, while the same does not hold true for discontinuous epitopes. A few novel features that are not presently available in other similar B-cell prediction web servers further facilitate the reagent-design aspect. These include the visualization of positive stretches directly on the structure, a customizable output for the sliding window results that also allows the user to quickly download the analysis results in different formats, and the computation of protein epitopic hotspots. The server also has a unique feature not found in other B-cell epitope prediction servers, which is the ability to compute and highlight interface residues when analyzing a molecule within a multi-chain structure. This has interesting implications when the server is used to design peptidic reagents, as it allows the user to select targets with the desired characteristics.

Materials and methods

Building the experimental epitopes reference dataset

The B-cell epitopes dataset was built using the Immune Epitope Database (IEDB) version 2.0[18] and the Bcipep database.[19] Only positive linear epitopes were used. Duplicated entries and epitopes that could not be mapped univocally on the protein linear sequence were removed. The dataset included 1336 experimentally determined epitopes derived from 106 proteins for which at least two different epitopes were known, and for which a model could be built with an average quality score ≥ 0.2 as determined by the Verify 3D (V3D) software.[20–22] This dataset is available at http://immuno.bio.uniroma2.it/bpred/table/table2.pl and provides the following information for each epitope: the epitope id linked to the original epitope entry (ie, IEDB epitopes), the epitope reference ID (ie, Bcipep epitopes), the PubMed ID linked to the corresponding publication record on PubMed, the peptide start and stop residues, sequence, length, and the (NCBI GI) number of the protein

linked to the corresponding protein record on the National Center for Biotechnology Information (NCBI) website.

Molecular modeling

The templates for the protein models were identified by HMM-HMM comparison using the pdb70_7Mar09 Hidden Markov Model (HMM) database[23] via the HHpred server.[23] Psiblast was used for the Multiple Sequence Alignment (MSA) generation method with three MSA generation iterations and a local alignment mode. The E-value threshold was set at 1E-3 for MSA generation and the minimum coverage threshold for MSA hits was set at 20%. Structures predictions were performed by the MODELLER software[24,25] using the optimal multiple template of the HHpred server.

Identification of interface residues (protein complexes only)

The Naccess software[26,27] was used to calculate the solvent exposure of each residue using the full quaternary structure of the protein (ie, including all protein chains) as input. These solvent exposure values were then compared with the values obtained when using single protein chains as input. The difference in solvent accessibility was calculated in these two cases for each residue. This difference is zero if the residue is not located within a protein-protein interface, and has a negative value otherwise. This simple method allows for the easy identification of the residues located at interfaces, which are subsequently highlighted in the output of the web server's analysis.

Dataset analysis with published methods

For the ElliPro method, the dataset was analyzed using the protein models with the default ElliPro parameters.[16] Of all the predicted epitopes, only the linear ones were considered in the analysis. For the LEPD[11] and the CBTOPE [14] methods, the analysis was performed using the default parameters.

For the purpose of this analysis, five different randomized datasets were generated from the main dataset, thereby generating five "test set/training set" pairs. In order to minimize possible biases in the randomization process, each pair was evaluated as an independent training and test set. The performance values were determined by the mean values of the five independent determinations.

Statistical analysis

Data are expressed using mean ± standard deviation of the mean, or the frequency of positives as appropriate. The groups were compared using the *t*-test for continuous data and the χ^2 test for categorical data, applying Yates' correction when appropriate. A *P*-value below 0.05 was considered significant. Receiving operator characteristic (ROC) analysis was used for the identification of the optimal B-Pred parameter conditions and cut-off values. All tests were performed using the GraphPad Prism 4.0 software package (GraphPad Software, San Diego, CA).

Web server and programming

B-Pred was developed on a Linux Ubuntu 10.04 server running PHP 5.2.3, Perl 5.10.1, Naccess version 2.1.1, and the current release of V3D (http://nihserver.mbi.ucla.edu/Verify_3D/). The web server interface, web forms, and data output were written using HTML4.01/CSS and PHP.

First, the user uploads or selects a model or structure to work with. After uploading the model or retrieving the structure from the Protein Data Bank (PDB), a job folder is created on the server by a PHP script where the model data and all subsequent analysis data are stored. A job password is then created and stored in a dedicated text file. Once the model file is stored, an interface with the analysis options (sliding window step, analysis thresholds) is shown to the user. After selecting the analysis parameters, the PHP script makes a call to a perl script (bpredmain.pl) that is in charge of all the analysis steps. The first task executed by the perl script is to call the Naccess application, and retrieve the Naccess solvent exposure data for the selected chain (computed alone) and the Naccess data for the full protein complex. Comparing the data for the chain alone and the data from the full complex identifies the interface residues. The structural data for the selected chain are then analyzed for local model quality by the perl script, which then calls the V3D software. Naccess data, V3D data, and interface determinations are then stored in a text file in csv format. Next, the PHP script reads this detailed analysis file to format the results and generate the server output. Figure 5 presents the server algorithm. With this data, a visualization of the structure is dynamically provided through the Jmol applet.[28] Finally, solvent exposure and model quality charts from the server output are dynamically generated using the JpGraph PHP software.[29]

Results

Method flowchart

The proposed B-cell epitope prediction method, B-Pred, assumes the availability of a 3D structure of the target protein, either as an experimentally determined structure or as a model. This structure is used to assign two scores

to each residue, one for solvent exposure and one for local structure quality. The solvent exposure values are computed for the selected chain alone and for the selected chain in the context of the whole protein complex. Comparing these two sets of values identifies and flags the residues involved in the protein–protein interface. The protein is then split into a number of overlapping peptides using a sliding window approach. A solvent exposure score and a quality score are then computed for each peptide, as the average of the scores of its constituent residues. If the values of the solvent exposure and structure quality are above the predefined thresholds, the peptide is flagged as potentially containing a B-cell epitope. The following two sections describe the computation of the default threshold values. Figure 1 provides a diagram of this flowchart, and Figure 5 describes the server algorithm underlying the flowchart.

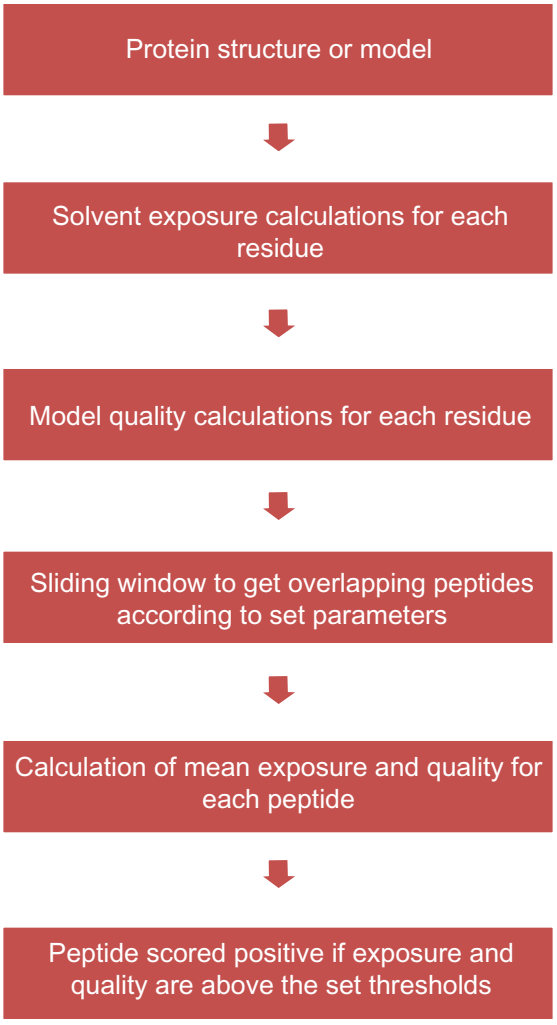

Figure 1 Flowchart of the B-Pred method.
Notes: Starting with a protein structure or model, the solvent exposure and the local structural quality are calculated for each protein residue. The protein is split into overlapping peptides using a sliding window approach according to the parameters selected by the user on the web interface. The mean exposure and structural quality are calculated for each peptide and the hotspot sequence stretches are detected. Peptides are flagged as positive if the scores are above the set threshold.

In order to test the predictive power of this method and determine the optimal thresholds for solvent exposure and structure quality, a dataset of 1336 experimentally determined epitopes derived from 106 proteins was built (see methods section). This dataset is available at http://immuno.bio.uniroma2.it/bpred/table/table2.pl.

Selection of default sliding window size and default sliding step for peptide generation and analysis

This study analyzed the distribution of epitope sizes in the experimental dataset. As expected, there is a clear bias toward round numbers, in particular 10 mers (12.4% of total epitopes), 15 mers (23.4%) and 20 mers (11.9%). There is a prevalence of 15 mers, which are probably a good compromise between the willingness of the researcher to include a number of residues higher than the expected size for an average B-cell epitope in the synthesized peptides and the cost of the peptide synthesis. Nevertheless, a significant number of 20 mers are also present in the dataset. Epitopes and peptides longer than 20 are poorly represented (9.6%).

The criteria for scoring a peptide as true positive are based on the inclusion of the experimental epitope in the considered peptide. Therefore, we selected 20 as the default sliding window size in order to include as many epitopes from the experimental dataset as possible, while retaining a good degree of resolution in the results. This allowed us to consider 90.4% of the experimental epitopes in our dataset.

In order to avoid the overselection of contiguous peptides in the same protein region that would be obtained by moving the sliding window by one residue each time, we introduced a default sliding step of three residues.

Identification of optimal thresholds for Naccess/V3D analysis

The experimental dataset was randomly partitioned into five different training (80% of the peptides) and test (20%) sets. The training sets were evaluated independently to determine the optimal thresholds for both the solvent exposure and structure quality scores. To this end, a 20-mer scan was performed on the proteins in each set. Each peptide was assigned a solvent exposure score (mean of individual Naccess scores of the 20 residues) and a local model quality score (mean of individual V3D scores of the 20 residues).

First, the ability of the Naccess score to identify the true epitopes at different V3D thresholds was assessed. Figure 2 shows the mean values of the area under the curve obtained with the ROC analysis at the different values of V3D for the five training sets. The training sets performed quite homogenously, and the best performances were obtained from the V3D values in the range of 0–0.2 ($P < 0.05$ all comparisons, χ^2 test).

In agreement with the ROC analysis, the sensitivity of the method slightly improved when the model quality, as assessed by V3D, was taken into account (Figure 3). In accordance with the original publication that describes the V3D software,[22] the best cutoff values for V3D were in the range of 0–0.2.

Interestingly, the five different training sets determined similar cut-off values at each level of V3D (data not shown). Among the five training sets, the worst performance was a sensitivity of 36.65%, which was associated with a specificity of 70% for the Naccess cut-off value of 43.05, and a V3D above 0.

Predictive power of the proposed method compared to other established methods

The 20 mers in the five test sets were also scored using the ElliPro, LEPD, and CBTOPE methods for comparison (see methods section).

Figure 4 presents the result from this comparative analysis in terms of sensitivity, specificity, and predictive values. The mean values obtained for each parameter for the test sets are also reported. Interestingly, B-Pred presented a significant

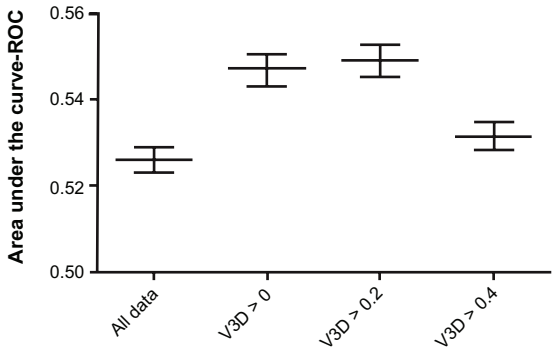

Figure 2 ROC analysis for the Naccess algorithm according to the different values of the V3D algorithm for the five training sets of peptide fragments analyzed.
Notes: The graph shows the mean value + SD of the different areas under the curves obtained by ROC analysis with the Naccess algorithm. These values are within the five training sets of data and are based upon the quality of the 3D model of the protein in the portion analyzed according to the V3D algorithm. The best performances of the Naccess algorithm were obtained with peptides with a V3D value in the range of 0–0.2, with a decline in performance of 0.4.
Abbreviations: ROC, receiving operator characteristic; SD, standard deviation; V3D, Verify3D software.

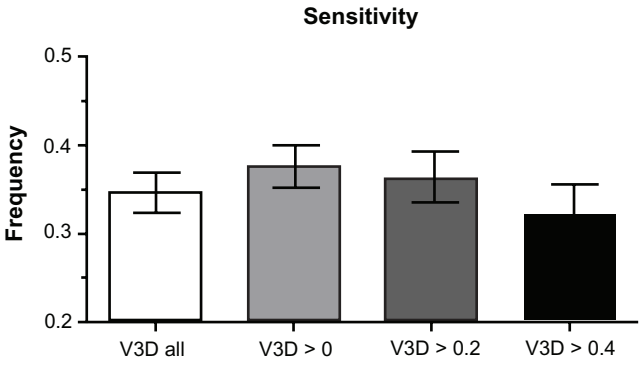

Figure 3 Sensitivity of the training sets at different V3D values.
Notes: The graph shows the mean value + SD of the sensitivity of all five training sets obtained by ROC analysis for a value corresponding to 0.7. The analysis was repeated at different V3D score thresholds (x-axis). The best performances were obtained with peptides with V3D values between 0–0.2.
Abbreviations: ROC, receiving operator characteristic; SD, standard deviation; V3D, Verify3D software.

improvement in positive predictive values ($P < 0.03$ all comparisons, χ^2 test) compared to the other three methods used.

Furthermore, the B-Pred thresholds were set up with the aim of achieving a high specificity and a similar approach is at the basis of the cut-off value used in the ElliPro method; therefore, specificities of the two systems are comparable (Figure 4). The sensitivity of B-Pred is equivalent to the sensitivity of ElliPro, which supports the notion that B-Pred performs as well, if not better than ElliPro. Indeed, the sensitivity of the B-Pred was worse than 0.35 for test set 4, which was the same as ElliPro ($P > 0.05$, χ^2 test) (Table 1).

On the other hand, B-Pred presented a significantly lower sensitivity than CBTOPE (0.37 versus 0.54, respectively, $P < 0.02$ all comparisons, χ^2 test) (Table 2). However, B-Pred presented a significantly higher specificity (0.69 versus 0.49, respectively, $P < 0.01$, χ^2 test). Therefore, B-Pred achieved a significantly higher positive predictive value (Figure 4).

Development of the B-Pred web server

After achieving these positive results, we incorporated the method described above in a web server, which is accessible at http://immuno.bio.uniroma2.it/bpred (Figure 5). The server can use a protein structure, a model uploaded by the user, or a PDB id as input.[30] Each job is assigned a random number that can be used to retrieve the results at a later time.

After the model file is uploaded, or the structure has been retrieved from the PDB, the website displays details on the structure including a 3D rendering with the Jmol applet. By default, the server proposes the parameter values (length of peptides, sliding window size, solvent exposure, and structure

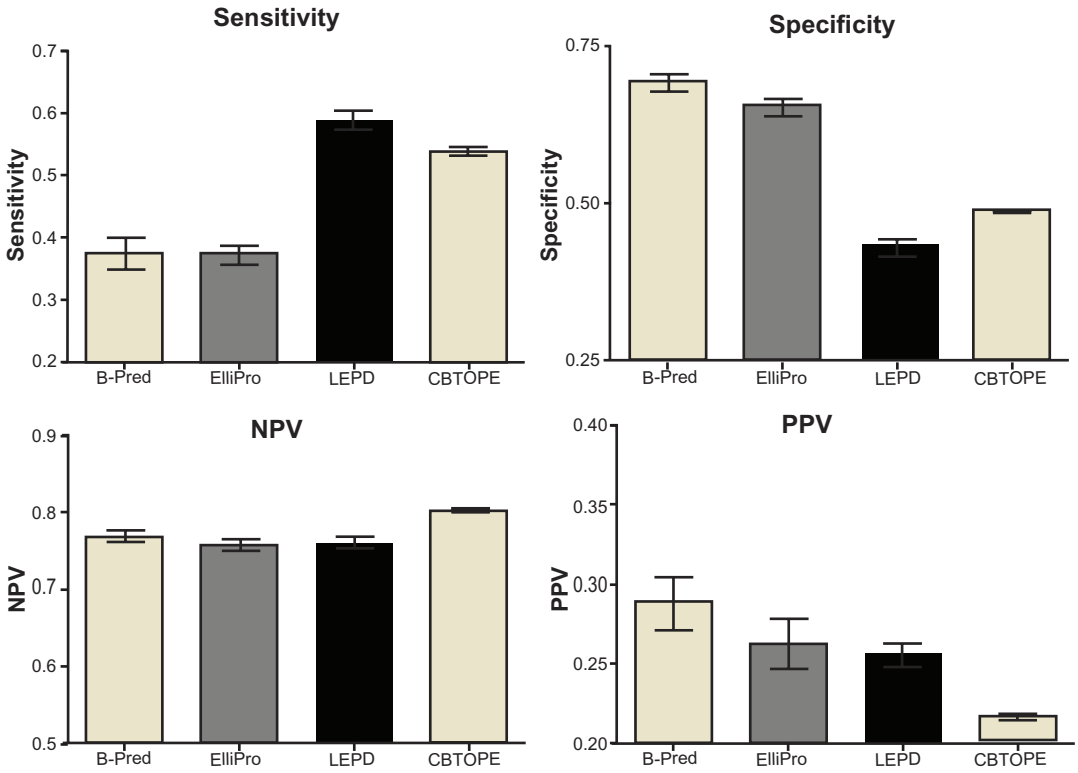

Figure 4 Performance of the different B-cell epitope prediction algorithms.
Notes: The graph shows the mean value + SD of the sensitivity, specificity, negative predicting value (NPV) and positive predicting value (PPV) for the different predicted algorithms within the five test sets.
Abbreviations: CBTOPE, Conformational B-cell Epitope Prediction; LEPD, Linear Epitope Prediction Database; SD, standard deviation.

quality thresholds) that have been optimized for the dataset presented in this paper. These values can be changed at any time, as the analysis can be reiterated with different parameters within the same work session.

If the uploaded structure or model comprises more than a single chain, the user can select the chain to be analyzed.

The detailed results of the analysis (Figure 6) are presented as follows:

1. Sequence overview: the full sequence is shown with the positive aminoacids (those with exposure and quality values above the selected thresholds) highlighted in red. Aminoacids in the linear sequence that are missing from the structure are highlighted in light grey, if the full sequence information is available; otherwise they are shown as Xs. If the uploaded structure comprises more than a single molecule, the interface residues in the selected chain are underlined. Details about the individual aminoacids (letter and position) are displayed in a tooltip.
2. Peptide scan results: each line provides the results of the analysis for an individual peptide, with the same formatting code as the previous view. Solvent exposure and structure quality values for a peptide correspond to the mean of the individual aminoacids that compose the peptide. If these mean values are above the selected thresholds and at least 70% of the peptide is composed of residues with known 3D coordinates, the peptide score is positive and a "1" flag is displayed at the end of the line. Raw peptide data can be downloaded by the user in csv or plain text format for later reference.
3. Local peaks/hotspots: single aminoacids or contiguous aminoacid stretches with positive scores are reported.

Table 1 Determination of sensitivity of the four tested methods in the five test sets

Sensitivity*	B-Pred	ElliPro	LEPD	CBTOPE
Test set 1	0.37	0.39	0.57	0.54
Test set 2	0.37	0.38	0.59	0.54
Test set 3	0.41	0.38	0.60	0.53
Test set 4	0.35	0.35	0.58	0.54
Test set 5	0.37	0.36	0.60	0.53
Mean	0.37**	0.37**	0.59	0.54
SD	0.025	0.015	0.015	0.006

Notes: *Data is expressed as frequency; **$P < 0.05$ versus LEDP and CBTOPE, χ^2 test.
Abbreviations: CBTOPE, Conformational B-cell Epitope Prediction; LEPD, Linear Epitope Prediction Database; SD, standard deviation.

Table 2 Sensitivity, specificity, and positive and negative predicting values for the four methods considered in this study

	B-Pred	ElliPro	LEPD	CBTOPE
Sensitivity*	0.37 ± 0.025**	0.37 ± 0.015**	0.59 ± 0.015	0.54 ± 0.006
Specificity*	0.69 ± 0.015**	0.65 ± 0.012**	0.43 ± 0.010	0.49 ± 0.001
Positive predicting value*	0.29 ± 0.016***	0.26 ± 0.015	0.25 ± 0.007	0.22 ± 0.001
Negative predicting value*	0.77 ± 0.007	0.76 ± 0.007	0.76 ± 0.009	0.80 ± 0.001

Notes: *Data are expressed as mean + SD of the frequency observed with the five test sets; **$P < 0.05$ versus LEDP and CBTOPE, χ^2 test; ***$P < 0.01$ versus LEDP and CBTOPE, χ^2 test.
Abbreviations: CBTOPE, Conformational B-cell Epitope Prediction; LEPD, Linear Epitope Prediction Database.

4. Quick Jmol view: the structure of the uploaded model can be explored interactively in the Jmol applet. Positive peptides or clusters of positive peptides are highlighted in red.
5. Solvent accessibility profile: plots the Naccess values along the linear sequence.
6. Model quality profile: plots the V3D values along the linear sequence.

Figure 3 shows part of the server output screen in which the sequence overview and the peptide scan results are shown for a sample PDB file.

Discussion

This paper investigates whether the solvent accessibility of a peptide in the context of a full protein structure can be used to identify potential epitopes. Lollier et al[15] recently questioned this relationship that is widely used in algorithms for B-cell epitope prediction tools based on linear sequence analysis[31] and protein structures.[2]

This paper presents a simple approach to analyzing 3D models or structures using Naccess and V3D algorithms to obtain values for solvent exposure and local model/structure quality, respectively. Selection occurs by scoring a sequence as "positive" when these values are above the defined threshold.

Several B-cell epitope prediction methods have recently been developed based on the linear protein sequence or on protein structure coordinates.[2,3] However, B-cell epitope prediction methods are often largely inaccurate

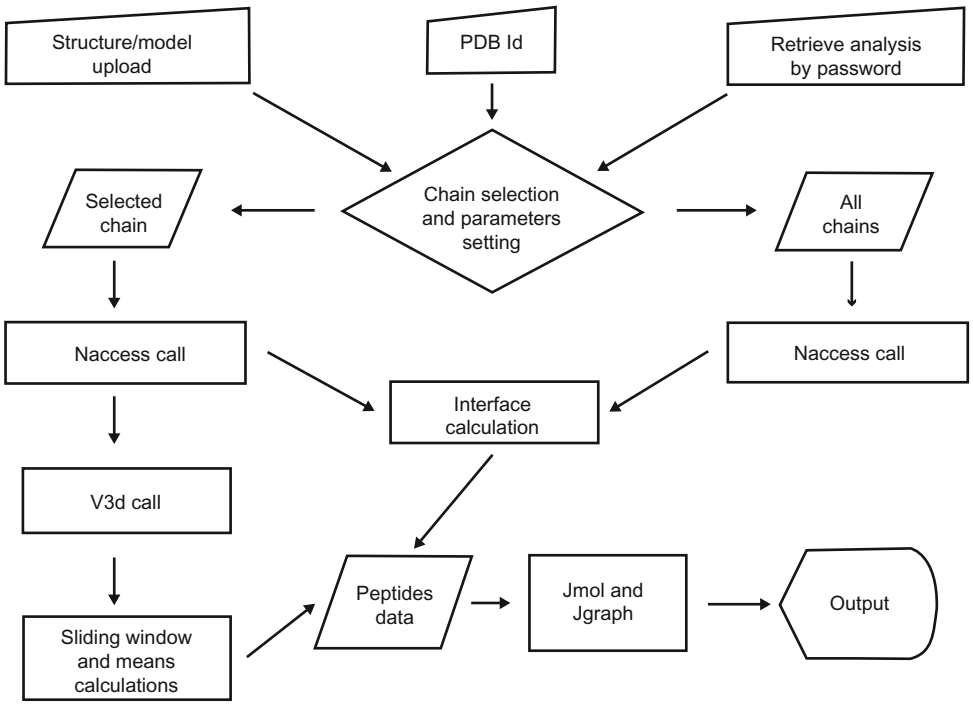

Figure 5 Algorithm of the B-Pred server.
Notes: The server accepts a model/structure as input data, a PDB Id entry, or the retrieval of a past analysis using a password. The user selects the analysis parameters such as a chain to analyze the solvent exposure, the quality structure thresholds, and the sliding window. For the selected chain, the solvent exposure and quality model calculations are performed by using Naccess and Verify3D software. Overlapping peptides are generated and the mean scores are calculated. The peptide data are stored with the interface information derived from comparing the solvent exposure measures of the selected chain with the same measures in the context of the whole protein complex (all chains). The graphic applet is called, the graphics are generated through Jgraph, and the output is displayed.

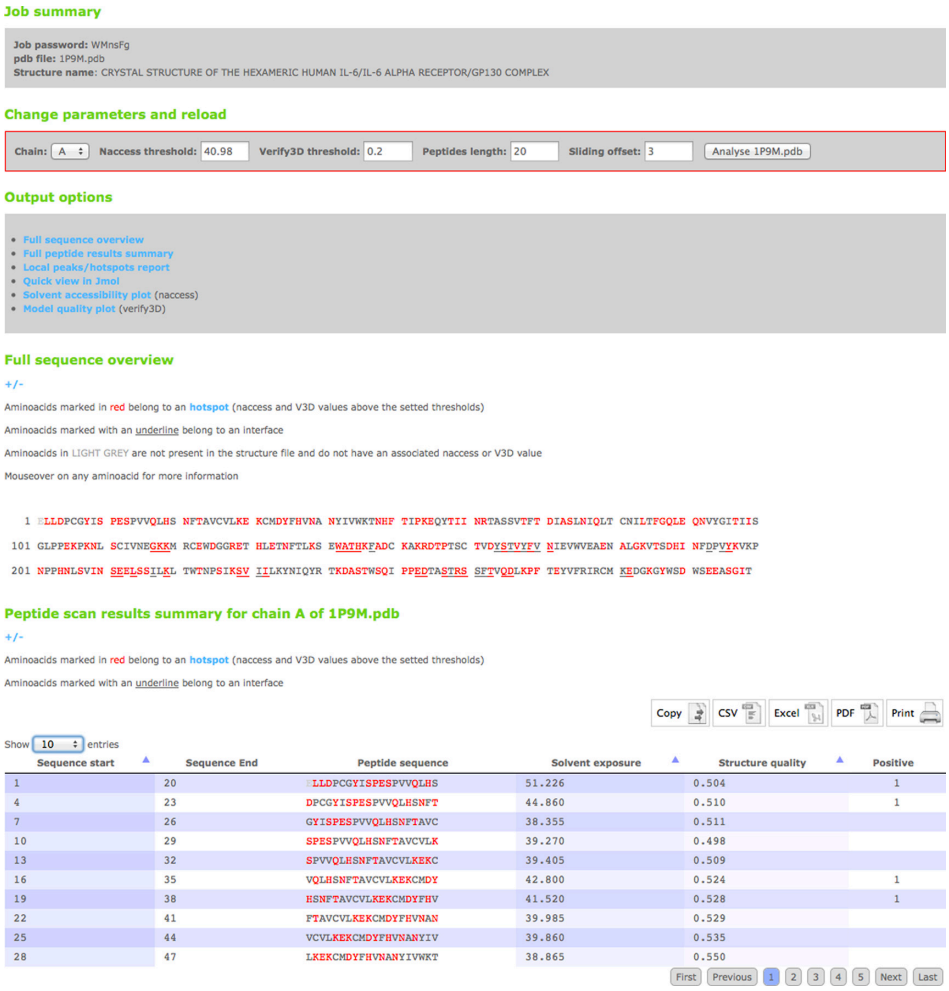

Figure 6 A screenshot of the results page from the B-Pred server.
Notes: In the "change parameters" section the user can set the selected chain, the Naccess and V3D thresholds, the size of the sliding window, and a sliding window offset. In the "full sequence overview" section, the full sequence is shown with residues above the threshold for both Naccess and V3D highlighted in red and, for structures with more than one chain, the interface residues are underlined. In the peptide scan results, the nacces and V3D values of each peptide are reported. If the mean Naccess and V3D values of the peptide are above the thresholds currently in use and at least 70% of the residues are present in the structure file, the peptide is predicted as containing a potential epitope. This is indicated by a "1" in the dedicated column.
Abbreviation: V3D, Verify3D software.

for several reasons.[10,13] The characteristics that render a sequence suitable for antibody binding are still poorly defined despite extensive research in this area, making even the prediction of linear epitopes a difficult task. In terms of prediction methods based on 3D structures, a paradigm is emerging that shows that a significant number of proteins (about 40% of all human proteins) contain at least one disordered segment of 30 aminoacids or more, while 25% of all human proteins are likely to be entirely disordered and might reach a defined structure only when interacting with a ligand.[32,33] Therefore, experimental crystal structures or structure models may not necessarily reflect the real conformation of proteins or protein complexes in a solution. Despite these limitations, a number of methods demonstrate significant predictive power when challenged with experimental datasets.

This report presents a rather simple structure-based method that only analyzes two different parameters (local solvent exposure and local structure quality). The software is called B-Pred and can be freely accessed through a web server located at http://immuno.bio.uniroma2.it/bpred. This method is aimed at predicting linear, continuous epitopes (as opposed to conformational/discontinuous epitopes). B-Pred showed a sensitivity of 0.37 at a specificity of 0.69 (Table 1), making the method comparable with other published methods that are based on linear protein sequences (LEPD), 3D coordinates (ElliPro), or SVM models (CBTOPE), with a slightly increased specificity, thereby minimizing the number of false-positive predictions. It should be noted that the conditions used to test LEPD, ElliPro, and CBTOPE are different from the ones reported in their original papers, as the epitopes

database is different and assumptions were made in order to score 20 mers with these methods.

This method was implemented and made publicly available by the development of a web server with a number of novel features that are not available in similar servers. This server is biased toward scoring potential immunological reagents (peptides) derived from protein sequences. B-Pred uses a sliding window to scan the sequence and identify potential epitopes. The parameters of the analysis can be modified during subsequent iterations in order to identify the reagents that are most suited to the specific needs of the user.

A unique feature of the B-Pred server is the identification of the residues located in protein-protein interaction surfaces. This information can be relevant in designing peptides for use in the production of antibodies/antisera with specific characteristics. In this context, it could be speculated that antibodies directed at protein-protein interfaces could display neutralizing activity by preventing or competing with the formation of active protein complexes. Conversely, antibodies targeted at areas not involved in complex formation, but still located on solvent exposed regions, could be suitable reagents for the immunoprecipitation of whole protein complexes. Of course, the B-Pred server is just a contribution toward this ambitious goal.

Although the B-Pred server considers conformational and structural information to determine solvent accessibility, it exclusively focuses on the prediction of continuous linear epitopes, as opposed to discontinuous epitopes that can be predicted by other B-cell prediction servers. While this limits the scope of the method, it allows for an immediate translation of the results into peptidic reagents for bench research, which is one of the main purposes of this system.

According to Lollier et al, surface and solvent exposure, as assessed by different methods (Relative Surface Accessibility and Protrusion Index) cannot be reliably correlated to antigenic propensity.[15] There are a number of reasons why an experimentally determined epitope can have poor solvent exposure in the context of the 3D structure of the full protein. For instance, a protein or allergen can be denatured or otherwise processed before or after being injected into an animal for immunization. Before the availability of prediction methods based on structure, peptides were selected from protein sequences using propensity scale indexes and were successfully used to raise antisera or monoclonal antibodies. It is common knowledge that monoclonals exist that will only work in western blots, and thereby recognize sequences that become exposed only after protein denaturation during electrophoresis and blotting procedures. However, other monoclonals are suitable for immunoprecipitation of the target protein from undenatured lysates, and thereby recognize solvent accessible surface sequence stretches that are either continuous or discontinuous with respect to the linear sequence. For these and other reasons, many sequences stored in databases as containing B-cell epitopes can have an overall poor surface exposure. The detailed information about the methods and experimental conditions used for their identification would be extremely useful for the development of more targeted prediction methodologies that would be able to take all of the above considerations into account.

It should be noted that, despite the report by Lollier et al,[15] in the present study we do observe a correlation between surface exposure and antigenic propensity. This could be due to a number of reasons that are worth investigating. For example, for intrinsic reasons, our epitopes dataset is entirely biased toward proteins for which a 3D structure is either available or for which a model can be reliably computed. Therefore, it is possible that our experimental epitopes dataset is biased toward peptides that were predicted for subsequent synthesis and experimental testing using existing structural prediction methods that most often incorporate surface exposure information in their algorithms. Since the methodology for peptide design is not readily available in epitope databases, it is not easy to verify this kind of hypothesis. Research is underway to address these issues.

Since the current B-Pred implementation is based on a single parameter (solvent exposure) that is filtered on local model quality, it is reasonable to assume that the method could be further improved by the inclusion of additional structural parameters[12,34] and/or by combining the solvent exposure, as directly determined from the structure, with classical linear propensity scales. Research is under way to investigate the possible inclusion of additional parameters to improve the prediction accuracy of the current method.

Among the possible applications of this method, the development of diagnostic reagents for serological analysis is worth mentioning. A protein encoded in the genome of a pathogen of interest can be analyzed for potential B-cell epitopes that could be targeted by the humoral host response. Peptides containing these epitopes have potential as diagnostic reagents for serological tests if the selected sequences are specific to the pathogen of interest. In order to optimize the discovery of B-cell epitopes with diagnostic potential by identifying amino acid sequences that are only present in a given pathogen strain (which is related to the concept of

"conservancy"[35]), an interesting development of this work will be to automatically link the B-Pred analysis with a BLAST analysis. Work is currently in progress to achieve this goal in a future version of the B-Pred web server.

In conclusion, this study provides a new, freely accessible online tool for the selection of candidate B-cell epitopes in proteins of interest, and is focused on the design of experimental reagents for a variety of biological and biotechnological applications.

Acknowledgments

This study was supported by the "Distretto Tecnologico delle Bioscienze del Lazio," FILAS 2009. We are grateful to Professor Gajendra PS Raghava for allowing us to access the full data of the Bcipep database and to Dr Emanuele Buonomo for providing informatics advice for building the epitopes dataset.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Van Regenmortel MH. Structural and functional approaches to the study of protein antigenicity. *Immunol Today*. 1989;10(8):266–272.
2. El-Manzalawy Y, Honavar V. Recent advances in B-cell epitope prediction methods. *Immunome Res*. 2010;6(Suppl 2):S2.
3. Yang X, Yu X. An introduction to epitope prediction methods and software. *Rev Med Virol*. 2009;19(2):77–96.
4. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Nat Acad Sci U S A*. 1981;78(6):3824–3828.
5. Welling GW, Weijer WJ, van der Zee R, Welling-Wester S. Prediction of sequential antigenic regions in proteins. *FEBS Lett*. 1985;188(2):215–218.
6. Parker JM, 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(19):5425–5432.
7. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*. 1990;276(1–2):172–174.
8. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157(1):105–132.
9. Pellequer JL, Westhof E, Vanregenmortel MHV. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol Lett*. 1993;36(1):83–100.
10. Blythe MJ, Flower DR. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci*. 2005;14(1):246–248.
11. Chang HT, Liu CH, Pai TW. Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches. *J Mol Recognit*. 2008;21(6):431–441.
12. Wang Y, Wu W, Negre NN, White KP, Li C, Shah PK. Determinants of antigenicity and specificity in immune response for protein sequences. *BMC Bioinformatics*. 2011;12:251.
13. Ponomarenko JV, Bourne PE. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct Biol*. 2007;7:64.
14. Ansari HR, Raghava GP. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Res*. 2010;6:6.
15. Lollier V, Denery-Papini S, Larré C, Tessier D. A generic approach to evaluate how B-cell epitopes are surface-exposed on protein structures. *Mol Immunol*. 2011;48(4):577–585.
16. Ponomarenko J, Bui HH, Li W, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics*. 2008;9:514.
17. Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. *BMC Bioinformatics*. 2011;12:341.
18. Vita R, Zarebski L, Greenbaum JA, et al. The immune epitope database 2.0. *Nucleic Acids Res*. 2010;38(Database issue):D854–D862.
19. Saha S, Bhasin M, Raghava GP. Bcipep: a database of B-cell epitopes. *BMC Genomics*. 2005;6:79.
20. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 1991;253(5016):164–170.
21. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature*. 1992;356(6364):83–85.
22. Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol*. 1997;277:396–404.
23. Söding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005;21(7):951–960.
24. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234(3):779–815.
25. Eswar N, Webb B, Marti-Renom MA, et al. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*. 2007;Chapter 2:Unit 2.9.
26. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 1971;55(3):379–400.
27. Naccess [computer program]. Version 2.1.1. Manchester, UK: Hubbard S, Thornton JM; 1996. Available from: http://www.bioinf.manchester.ac.uk/naccess/. Accessed March 14, 2011.
28. Jmol: an open-source Java viewer for chemical structures in 3D [computer program]. Version 11.8.24. Minnesota, MN: Jmol Development Team; 2010. Available from: http://jmol.sourceforge.net/. Accessed April 12, 2011.
29. JpGraph [computer program]. Version 3.0.7. Japan: Asial Corporation; 2011. Available from: http://jpgraph.net/. Accessed April 12, 2011.
30. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235–242.
31. Reimer U. Prediction of linear B-cell epitopes. *Methods Mol Biol*. 2009;524:335–344.
32. Chouard T. Structural biology: breaking the protein rules. *Nature*. 2011;471(7337):151–153.
33. Uversky VN, Dunker AK. Understanding protein non-folding. *Biochim Biophys Acta*. 2010;1804(6):1231–1264.
34. Su CH, Pal NR, Lin KL, Chung IF. Identification of amino acid propensities that are strong determinants of linear B-cell epitope using neural networks. *PLoS One*. 2012;7(2):e30617.
35. Bui HH, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics*. 2007;8:361.