Computational prediction of B cell epitopes from antigen sequences

Running title: Sequence-based prediction for B cell epitopes

Jianzhao Gao[1] and Lukasz Kurgan[2]*

[1] School of Mathematical Sciences, Nankai University, Tianjin, People's Republic of China,

[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

* lkurgan@ece.ualberta.ca

Summary

Computational identification of B-cell epitopes from antigen chains is a difficult and actively pursued research topic. Efforts towards the development of method for the prediction of linear epitopes span over the last three decades, while only recently several predictors of conformational epitopes were released. We review a comprehensive set of thirteen recent approaches that predict linear and four methods that predict conformational B-cell epitopes from the antigen sequences. We introduce several databases of B-cell epitopes, since the availability of the corresponding data is at the heart of the development and validation of computational predictors. We also offer practical insights concerning the use and availability of these B-cell epitope predictors, and motivate and discuss feature research in this area.

Key words: B cell epitope; linear epitope; conformational epitope; antigen; immunotherapeutic; vaccine; prediction; database

1. Introduction

One of the key aspects of an immune system is the antibody-mediated ability to identify foreign, infectious objects, such as bacteria and viruses. This is implemented through binding of the antibodies and antigens (e.g., proteins from the pathogenic entity) at sites known as B-cell epitopes. Ability to identify these binding areas in the antigen sequence or on its surface is important for the development of vaccines and immunotherapeutics (***1***). The B-cell epitopes are categorized into two classes: linear/continuous and conformational/discontinuous. The former B-cell epitope is a short segment in the corresponding amino acid sequence (**Figure 1A**). Majority of the B-cell epitopes are conformational, which means that they are distributed over multiple segments in the protein chain that are located in close proximity in the folded 3-dimensional structure (**Figure 1B**) (***2***).

Although several experimental techniques can be used to identify the B-cell epitopes (***3***), they are relatively time consuming and expensive, particularly when considering to do that on large, genomic scale. Computational methods are a viable alternative to provide a fast and cost-effective way to predict the B-cell epitopes (***4***). A fairly large number of computational B-cell epitope predictors, which are characterized by varying degrees of success and scope, have been developed over the last three decades (***4-7***). Although progress has been accomplished in the context of the development and applications of these computational methods, much remains to be done, particularly considering modest predictive performance of these approaches (*see* **Note 1**). In parallel, a few efforts to collect, annotate and deposit B-cell epitopes into publicly accessible databases are currently underway (***8-10***) and integrated resources that provide access to multiple tools for prediction and analysis of epitopes are available (***11,12***). Such efforts should make these technologies more accurate (more data allows for building more accurate predictive models) and more convenient (freely available and integrated) for the end users.

The algorithms that predict the B-cell epitopes are classified into sequence-based and structure-based. The structure-based methods use the 3-dimensional structure of the antigen to perform the prediction, while the sequence-based methods utilize only the sequence of the antigen. While the structure-based predictors usually provide higher predictive performance when directly compared with the sequence-based methods (***13-15***), they are constrained to a relatively small set of targets for which the structure is available. They also suffer from a limited availability of the annotated data. Recent years have witnessed a revival of the development of the sequence-based methods, which currently are capable of finding both linear and conformational epitopes. To this end, we overview major relevant databases and summarize a comprehensive set of 17 sequence-based predictors of the B-cell epitopes, which expands over the coverage of recent predictors offered by the prior reviews (***4, 6***).

2. Databases of B-cell epitopes

Several databases that store experimentally annotated B-cell epitopes were developed over the last decade. They differ in scope and sources of data. These databases provide data that are used to develop and evaluate new and improved predictors of B-cell epitopes (*see* **Note 2**). We briefly summarize, in chronological order, six publicly available databases.

2.1 AntiJen

This repository was developed in 2001 at the Edward Jenner Institute for Vaccine Research in UK (***16***). It was later updated to version 2.0 (***10, 17***). It stores experimental thermodynamic binding data concerning the interaction of peptides including B-cell receptors, T-cell receptors, Major Histocompatibility Complexes (MHCs), TAP transporters and immunological protein-protein interactions. The B-cell and T-cell epitopes are also included. As of January 2013, there were total of 24000 entries in this database, and according to (***17***) 816 entries were related to B-cell epitopes. Users can search for the relevant data utilizing BLAST (***18***) and a variety of specialized search

options that allow defining specific experimental conditions and molecules. Based on the Web of Knowledge as of June 2013, this resource accumulated 211 citations across the three publications.
Availability: http://www.ddg-pharmfac.net/antijen/

2.2 IEDB

IEDB (Immune Epitope DataBase) was established in 2004 at the La Jolla Institute of Allergy and Immunology in San Diego (***19,20***) and it was recently upgraded to version 2.0 (***8***). This comprehensive resource provides integrated access to experimentally characterized B-cell epitopes, T-cell epitopes and data on the MHC binding. The data are extracted from epitope-related articles available in PubMed and from direct submissions from scientists. The database includes epitope sequence and structure, source antigen and organism from which the epitope is derived, and details concerning experiments describing recognition of an epitope and related assays including MHC ligand elution assays and MHC binding assays. Users can conveniently query the database through a web interface utilizing a variety of criteria, such as the source antigen, source organism, epitope structure, immune recognition context, host organism, etc. Based on the Web of Knowledge as of June 2013, this database is highly cited with the combined number of citations for the three articles totaling to 332.
Availability: http://www.iedb.org

2.3 Bcipep

This resource was developed in 2004 at the Institute of Microbial Technology Chandigarh in India (***21***). It provides access to experimentally determined linear B-cell epitopes, which were extracted from literature in PubMed and collected from other publicly available databases. As of January 2013, it contained 3031 entries including 539 entries from bacteria, 2046 from viruses, 236 from protozoa, 53 from fungi, and 157 from other organisms. Users can search the database through a variety of options including keywords related to the relevant publications, sequence, entry number, source organism, etc., by utilizing sequence similarity with BLAST, and by scanning through the associated protein structures.

Availability: http://www.imtech.res.in/raghava/bcipep/

2.4 CED

CED (Conformational Epitope Database) was built in 2005 by Huang and Honda at the University of Electronic Science and Technology in China (***22***). This database focuses on the conformational epitopes. The entries were extracted from peer-reviewed journal articles collected from PubMed and ScienceDirect. CED provides the location of the epitope in the sequence and structure, immunological properties of the epitope, the source antigen, and corresponding antibody. The database can be browsed or searched using keywords through a website interface. As of January 2013, CED included 293 entries.

Availability: http://immunet.cn/ced/

2.5 Epitome

This database was established in 2005 by Rost group at the Columbia University (***23***). Epitome provides access to a collection of antigen-antibody complex structures, including annotation and visualization of residues that are involved in the interactions and information concerning certain structural characteristics of the binding regions. The entries were collected from Protein Data Bank (PDB) (***24***). User can search the database utilizing keywords with options to specify chain and certain structural properties of antigen and antibody, and also by finding similar sequence with BLAST. This resource contains 142 antigens from protein-antibody complexes (***23***).

Availability: http://www.rostlab.org/services/epitome/

2.6 SEDB

Structural Epitope Database (SEDB) was developed in 2011 at the Pondicherry University in India (***9***). It provides access to a comprehensive set of structures of B-cell, T-cell and MHC binding proteins. The data was collected from PDB, PDBsum (***25***), MHCBN (***26***), IMGT/3D (***27***), Bcipep, and IEDB databases. SEDB includes information concerning epitope sequence and position, antigen-antibody interacting residues, corresponding taxonomic identifiers, and is cross-linked to relevant databases such as IEDB, UniprotKB (***28***), PDB, and NCBI (***29***). The database can be either browsed or searched by finding, using BLAST, similar chains. It currently includes 614 entries with 273 B-cell epitopes.

Availability: http://sedb.bicpu.edu.in/

3. Sequence-based predictors of linear B-cell epitopes

Prediction of linear B-cell epitopes from the antigen sequences dates back to 1980s. The trailblazing methods were fairly simple and utilized a single propensity (flexibility, solvent accessibility, etc.) of the underlying chain or chain fragment (***2, 30-35***). A new generation of methods that combined multiple physicochemical propensities to predict B-cell epitopes has surfaced in 1990s. They include PREDITOP (***36***), PEOPLE (***37***), BEPITOPE (***38***), BcePred (***39***), and LEP-LP (***40***) predictors. Predictive quality of these approaches was questions in 2005 in a study by Blythe and Flower (***41***). They analyzed predictive performance of close to 500 amino acid propensity scales on 50 antigens and determined that these propensities performed only slightly better than random. Since than this field has observed a revival that resulted in the development of more sophisticated knowledge-based methods, particularly in the context of the predictive models that they utilize. The considered models included a neural network in ABCpred (***42***); hidden Markov model in BepiPred (***43***); and naïve Bayes that was used in Epitopia (***13, 14***). The dominant model used in recent years is the Support Vector Machine (SVM), which was applied in a wide range of methods, such as AAP (***44***), BCPred (***45***), FBCPred (***46***), COBEpro (***47***), BayesB method (***48***), BROracle (***49***), LEPS (***50***), SVMTriP (***51***), and LBtope (***52***). These approaches differ in the formulation and scope of information extracted from the input antigen sequence, in the size of data that were used to compute the SVM model, and in the type of SVM kernel function used. **Table 1** summarizes methods that were developed since 2005 and includes one representative older method, BEPITOPE (*see* **Note 3**). COBEpro can also predict conformational epitopes and thus it is discussed later in this chapter. Several predictors of linear B-cell epitopes are widely cited in the literature, relative to when they were published. Based on the Web of Knowledge as of June 2013, ABCpred and BepiPred that were published in 2006 were already cited 139 and 145, respectively. The AAP method that was published in 2007 was cited 106 times, and the newer articles for BCPred and Epitopia that were released in 2008 and 2008 already accumulated 54 and 47 (for the two publications combined) citations, respectively. Most of the abovementioned recent sequence-based linear B-cell epitope predictors, except BROracle, are available as convenient web servers that require the end user only to provide an input antigen sequence. Five methods, BepiPred, AAP, BCPred, FBCPred, and Epitopia can be also downloaded as standalone applications, which would appeal to the users who would like to incorporate such tools into their computational pipelines. Following, we summarize the 13 predictors from **Table 1** in the chronological order.

3.1 BEPITOPE

BEPITOPE was published in 2003 by Pellequer's group at the Centre de Marcoule at CEA in France (***38***). BEPITOPE utilizes a scoring function that combines information from over 30 selected physicochemical propensities including hydrophilicity, flexibility, propensity to form beta-turns, and surface accessibility. User can define sequence motifs to filter the predictions.

Inputs: Protein sequence in FASTA format or accession number.

Outputs: Numerical profile over the input chain where putative epitopes are indicated by peaks.

Architecture: Scoring function.

Availability: This program is available for free for academic use and has to be requested from the authors. User is required to sign a license agreement before receiving a copy of the software. Web server is not available.

3.2 ABCpred

ABCpred was developed in 2006 by Raghava's group at the Institute of Microbial Technology Chandigarh in India (***42***). This method was one of the first to use a more sophisticated, machine learning-based prediction model. This model is a recurrent neural network that has a single hidden layer with 35 neurons. It utilizes a segment of 16 consecutive residues to perform prediction.

Inputs: Amino acid sequence using since-letter encoding. User can also set values of several parameters including threshold to identify epitopes and segment length. Default values are used in case if user does not want to set parameter values.

Outputs: Starting position and numeric score for predicted epitope(s).

Architecture: Recurrent neural network.

Availability: Web server at http://www.imtech.res.in/raghava/abcpred/.

3.3 BepiPred

BepiPred was created in 2006 by Lund's group at the Technical University of Denmark (***43***). This is the first and so far the only method that utilizes hidden Markov model. This model combines multiple physicochemical propensities including antigenicity, hydrophilicity, hydrophobicity, solvent accessibility and secondary structure, which are pre-processed using a running mean window.

Inputs: Protein sequence or a set of sequences (up to 2000) in FASTA format. Each sequence has to have at least 10 and no more than 6000 amino acids. User can also set value of threshold to identify epitopes; default value (0.35) is used otherwise.

Outputs: Numeric score for each residue in the query protein sequence. The predicted epitope is composed of residues with scores higher than the threshold.

Architecture: Hidden Markov model.

Availability: Web server at http://www.cbs.dtu.dk/services/BepiPred/. Standalone version for UNIX platform is also available at this web site.

3.4 AAP

AAP (amino acid pair antigenicity) predictor was developed in 2007 at the Shanghai Jiaotong University in China (***44***). This is the first method that utilizes the SVM-based prediction model. The authors introduced antigenicity propensity scale, which was empirically shown to improve over previously used physicochemical propensities, that was utilized to convert the query sequence into numerical inputs for the SVM.

Inputs: Amino acid sequence using since-letter encoding. User can also select the length of the epitope to be predicted, with default value set to 20 and allowed values of 12, 14, 16, 18, 20, and 22.

Outputs: Predicted epitope segments with the predefined length.

Architecture: Support Vector Machine with RBF kernel.

Availability: The authors do not provide the software. However, a web server that is a part of BCPREDS platform can be found at http://ailab.cs.iastate.edu/bcpreds/. Standalone version is also available at this web site.

3.5 LEP-LP

LEP-LP was released in 2008 by Tun-Wen Pai's group at the National Taiwan Ocean University (***40***). The authors utilized mathematical morphology to extract local peaks from a numerical profile that implements combination of several weighted physicochemical propensity scales, such as hydrophilicity, solvent accessibility, polarity, flexibility, antigenicity, and secondary structure.

Inputs: Amino acid sequence using since-letter encoding.

Outputs: Ranked putative epitope segments with the associated numeric scores.

Architecture: Scoring function based on mathematical morphology

Availability: Web server at http://biotools.cs.ntou.edu.tw/lepd_antigenicity.php (currently unavailable).

3.6 BCPred

BCPred was published in 2008 at the Iowa State University (***45***). This is the second method that applied SVM-based prediction model, however this model is customized to use string kernel. The authors utilized a specific type of the string kernel, subsequence kernel, which considers a feature (input) space generated by a set of k-mer subsequences of the input chain.

Inputs: Amino acid sequence using since-letter encoding. User can also select the length of the epitope to be predicted, with default value set to 20 and allowed values of 12, 14, 16, 18, 20, and 22.

Outputs: Predicted epitope segments with the predefined length and with the associated numeric scores.

Architecture: Support Vector Machine with string kernel.

Availability: Web server at http://ailab.cs.iastate.edu/bcpreds/. Standalone version is also available at this web site.

3.7 FBCPred

FBCPred was developed in 2008 at the Iowa State University (***46***). Similar to BCPred, this method also uses SVM model with the subsequence kernel. FBCPred targets prediction of linear B-cell epitopes of variable length, in contrast to BCPred that assumes fixed (user-defined) length.

Inputs: Amino acid sequence using since-letter encoding. User can also select the length of the epitope to be predicted, with default value set to 14.

Outputs: Predicted epitope segments with the predefined length and with the associated numeric scores.

Architecture: Support Vector Machine with string kernel.

Availability: Web server at http://ailab.cs.iastate.edu/bcpreds/. Standalone version is also available at this web site.

3.8 Epitopia

This predictor was published in 2009 by Tal Pupko group at the Tel Aviv University in Israel (***13, 14***). Epitopia predicts linear B-cell epitopes from either a protein structure or sequence; here we focus on the sequence-based version. This method uses Naïve Bayes classifier by considering a small sliding window of 7 residues. The inputs for the classifier are generated from this window by using 14 physicochemical propensities including polarity, flexibility, antigenicity, hydrophilicity, solvent accessibility, secondary structure, and ratio between the frequency of selected amino acid in the window and the remaining part of the sequence.

Inputs: Protein sequence in FASTA format and an email address of the user.

Outputs: Numeric immunogenicity score and corresponding probability for each amino acid in the query protein sequence. The immunogenicity scores are used to derive a ranked list of epitope segments.

Architecture: Naïve Bayes classifier.

Availability: Web server at http://epitopia.tau.ac.il. Standalone version for LINUX platform is also available at this web site.

3.9 BayesB

This method was created in 2010 at the National University of Singapore (***48***). BayesB utilizes the SVM model and employs Bayes feature extraction that is based on differences in the frequency of occurrence of amino acid types at each position in a predefined (training) set of epitopes and non-epitope segments.

Inputs: Protein sequence in FASTA format or using since-letter encoding. User can also select the length of the epitope to be predicted, with default value set to 20.

Outputs: Predicted epitope segments with the predefined length.

Architecture: Support Vector Machine with RBF kernel.

Availability: Web server at http://www.immunopred.org/bayesb/.

3.10 BROracle

B-Cell Epitope Oracle (BROracle) method was developed in 2011 at the Dana-Farber Cancer Institute (***49***). This predictor is implemented using SVM model. The input to the model were generated from the sequence and a variety of sequence-derived characteristics including evolutionary information calculated from PSI-BLAST output (***53***), secondary structure predicted with PSI-PRED (***54***), solvent accessibility predicted with ACCpro (***55***), disorder predicted withVSL2 algorithm (***56***), and sequence complexity computed with SEG algorithm (***57***).

Inputs: Protein sequence.

Outputs: Unknown.

Architecture: SVM classifier with polynomial kernel.

Availability: Standalone program at https://sites.google.com/site/oracleclassifiers/ (currently unavailable). Web server is not available.

3.11 LEPS

LEPS (Linear Epitope prediction based on Propensities scale and SVM) was created in 2011 by Tun-Wen Pai's group at the National Taiwan Ocean University (***50***). This method extends the LEP-LP predictor by the same group. First, candidate epitopes are predicted with LEP-LP. Next, SVM model is used to remove less probable candidates utilizing their amino acid sequences.

Inputs: Protein sequence in FASTA format or using since-letter encoding. The user has an option to adjust 32 parameters related to the setup of the propensities considered in LEP-LP. Default parameter values are used in case if user does not want to set parameter values.

Outputs: Ranked list of predicted epitope segments.

Architecture: Support Vector Machine with RBF kernel.

Availability: Web server at http://leps.cs.ntou.edu.tw.

3.12 SVMTriP

SVMTriP was created in 2012 by Chi Zhang's group at the University of Nebraska, Lincoln (***51***). This predictor is based on SVM model that utilizes similarity, calculated with Blosum62 matrix, and frequency of tripeptides (3-mers) from the input antigen chain.

Inputs: Protein sequence in FASTA format or using since-letter encoding. User can select the length of the epitope to be predicted, with default value set to 20.

Outputs: Predicted epitope segments with the predefined length and with the associated numeric scores.

Architecture: Support Vector Machine with string kernel.

Availability: Web server at http://sysbio.unl.edu/SVMTriP.

3.13 LBtope

LBtope was published in 2013 by Raghava group at the Institute of Microbial Technology Chandigarh in India (***52***). This method converts the antigen chain into numerical features (descriptors) that are based on dipeptide (2-mer) profiles. These features are fed into the SVM model that predicts epitopes.

Inputs: Protein sequence or a set of sequences, in FASTA format. User can also select model type, using fixed size epitope fragments (20 residues long) or variable length epitopes (user-defined between 5 and 30); default value (variable length with 15 residues segment) is used otherwise.

Outputs: Predicted epitope segments with the predefined length and with the associated numeric scores.

Architecture: Support Vector Machine with undisclosed type of kernel.

Availability: Web server at http://crdd.osdd.net/raghava/lbtope/.

4. Sequence-based predictors of conformational B-cell epitopes

A few methods were recently developed to predict the conformational B-cell epitopes from protein chains. This is a challenging problem given the fact that the corresponding epitopic residues are potentially distributed over an entire protein chain, without necessarily being clustered into longer segments. The prediction methods score each amino acid in an input protein chain (using a numeric or binary value) to indicate whether it is part of an epitope. A drawback of this prediction is that these programs do not group the predicted epitopic residues into the corresponding epitopes, which could be an issue if a given chain contains more than one epitope. The sequence-based predictors of conformational epitopes, which are summarized in **Table 2**, include COBEpro that was designed to predict linear epitopes and extended to predict conformational epitopes (***47***), CBTOPE (***58***), BEST (***15***), and Bprediction (***59***) (*see* **Note 4**). The first three methods apply the SVM model, while the most recent Bprediction is based on the random forest model, which utilizes a set of decision trees. Based on the Web of Knowledge as of June 2013, the oldest sequence-based predictor of conformational B-cell epitopes, COBEpro, which was published in 2009, was already cited 30 times. The other methods are too recent to accumulate citations. COBEpro, CBTOPE, and Bprediction are available to the end users via web servers. Two of the methods, CBTOPE and BEST, are provided as standalone software that the end users would install and use on their computers. Next, we summarize these four predictors in the chronological order.

4.1 COBEpro

COBEpro was published in 2009 by Baldi's group at the University of California (***47***). COBEpro has a two-tier architecture where the first layer applies SVM to predict short segments (5 to 18 residues long) in the input chain utilizing information based on their similarity to epitopic segments in a training database, and secondary structure and solvent accessibility predicted with SSpro (***60, 61***) and ACCpro (***55***), respectively. The second layer is used to combine the above predictions to calculate epitopic propensity score for each amino acid. This allows COBEpro to be used for the prediction of discontinuous B-cell epitopes.

Inputs: Protein sequence or a set of sequences, using since-letter encoding, and an email address of the user.

Outputs: Ranked (according to propensity) list of most likely predicted epitopes, including their predicted secondary structure and solvent accessibility, and numeric propensity scores for each amino acid in the query protein sequence.

Architecture: Support Vector Machine with Gaussian kernel.

Availability: COBEpro is incorporated into the SCRATCH web server suite at http://scratch.proteomics.ics.uci.edu/.

4.2 CBTOPE

CBTOPE was released in 2010 by Raghava's group at the Institute of Microbial Technology Chandigarh in India (***58***). This method applies a sliding window (a segment of 19 residues that is moved along the input antigen sequence) to predict the epitopic score for the residues in the middle of a given window. CBTOPE computes amino acid composition, which is represented using a binary vector, of the residues in the window and these values are inputted into the SVM model that predicts epitopic propensity.

Inputs: Protein sequence in FASTA format or using since-letter encoding. User can select a threshold for the output scores from the predictor, with a default value set to -0.3. Residues with scores above the threshold are assumed to be epitopic.

Outputs: Numeric propensity scores for each amino acid in the query antigen chain. The scores are integers between 0 and 9, where higher value denotes a higher likelihood of a given residue to be in an epitope.

Architecture: Support vector machine with Gaussian kernel.

Availability: Web server at http://www.imtech.res.in/raghava/cbtope/. Standalone version for Windows operating system is also available at this web site.

4.3 BEST

BEST (B-cell Epitope prediction using Support vector machine Tool) was published in 2012 by Kurgan's group at the University of Alberta in Canada (***15***). This method utilizes SVM model and a comprehensive set of sequence-derived characteristics of the antigen chain. BEST is implemented using a two-layer architecture, see **Figure 2**. In the first layer, the input antigen sequence is processed using sliding widows of 20 amino acids. Each 20-mer segment is encoded by a numerical feature vector that utilizes sequence conservation computed based on Weighted Observation Percentage (WOP) matrix generated with PSI-BLAST (***53***), similarity to training epitopes based on measure proposed in (***47***), and secondary structure and relative solvent accessibility predicted with SPINE (***62, 63***). This vector is inputted into SVM model and the predictions from SVM are combined to generate the epitopic propensities in the second layer.

Inputs: Protein sequence or a set of sequences, in FASTA format.

Outputs: Numeric propensity scores for each amino acid in the query protein sequence.

Architecture: Support Vector Machine with RBF kernel

Availability: Standalone software for Linux platform is available at http://biomine.ece.ualberta.ca/BEST/. Web server is not available.

4.4 Bprediction

Bprediction was made available in 2012 by Zhang's group at the Wuhan University in China (***59***). This predictor has a two-level design and applies an ensemble of random forest models that take a set of numerical features computed from sliding windows of size 9 (9-mers) generated over the antigen chain as their inputs. The inputs are divided into nine set, where each set is utilized by a different random forest model, which include (1) physicochemical propensities including flexibility, hydrophilicity, solvent accessibility, polarity, and propensity for formation of beta-turns; (2) amino acid composition of the residues in the window represented using binary vectors and (3) real-valued vectors; (4) composition of amino acids sets defined based on their R-groups; (5) values from the Position Specific Scoring Matrix (PSSM) generated by PSI-BLAST (***53***); (6) composition of dipeptides (2-mers) in the window; and (7) secondary structure and (8) relative solvent accessibility predicted with SABLE (***64***). The second level generates the output propensity scores by computing weighted average of normalized, based on z-scores, values of predictions from these nine models, see **Figure 3**.

Inputs: Protein sequence using since-letter encoding and an email address of the user.

Outputs: Numeric propensity scores for each amino acid in the query protein sequence.

Architecture: Ensemble of random forests.

Availability: Web server at http://bcell.whu.edu.cn.

The overall architectures of the two most recent conformational B-cell epitope predictors, BEST and Bprediction, are relatively similar (**Figures 3** and **4**). Both utilize the two-layered design and use multiple sequence alignments computed with PSI-BLAST and predictions of secondary structure and solvent accessibility. The main differences are in the fact that they use different prediction models (SVM vs. ensemble of decision forests) and several different inputs (similarity scores vs. physicochemical propensities and various amino acid compositions). In spite of utilizing these relatively sophisticated architectures, the predictive performance of these and other predictors of conformational epitopes is at modest levels (*see* **Note 1**). This calls for more research towards the development of more accurate methods (*see* **Note 5**).

5. Notes

(1) We sampled recent publications that evaluated predictive performance of the current B-cell epitope predictors. For simplicity we concentrate on the area under the ROC curve (AUC)

measure (***4***). AUC values range between 0.5 and 1, with 0.5 denoting a random prediction and higher values corresponding to better predictive performance. Five methods that predict epitopes from antigen sequences were compared side-by-side in (***15***) and were shown to achieve AUC between 0.52 and 0.57 on a benchmark dataset consisting of 149 antigens. In another study, six and two methods that predict epitopes from antigen structures and sequences, respectively, were evaluated on a small dataset with 19 antigens; their AUC values were in the 0.57 to 0.63 range (***59***). A recent review of predictors that utilize antigen structure demonstrates that AUC values for the prediction of conformational epitopes range between 0.57 and 0.64 (***5***). Overall, these results reveal that further research is needed to improve the currently modest levels of predictive performance.

(2) One of reasons behind relatively low predictive performance of B-cell epitope predictors is a relatively small size of the currently available annotated data. Most of the current and more successful methods are knowledge-based, which means that they utilize annotated, with the location of epitopes, structures or sequences of antigens to calculate and optimize their predictive models. Availability of additional annotated data would likely results in an improved performance of predictors, as the data used to build them would be more representative of the complete population of epitopes.

(3) When testing sequence-based predictors of linear B-cell epitopes we found that two of them, LEP-LP and BROracle, were no longer available. The web server implementations of the remaining methods allow predictions for a single chain. In case a user wants to predict a set of chains, (s)he has to supply and predict them one at the time. The two exceptions are BepiPred and LBtope that simultaneously process prediction of multiple chains, with a limit of up to 2000 sequences for a single run of BepiPred. Moreover, the BayesB predictor cannot predict peptides shorter than 25 residues.

(4) Three sequence-based predictors of conformational B-cell epitopes are available to the end users as web servers and two as standalone applications. Two of them, COBEpro and

Bprediction, are limited in the sense that they can predict only one sequence at the time. The other two, BEST and CBTOPE, are capable of predicting multiple chains in a single run. A further limitation of COBEpro is that it can be used to predict chains shorter than 1500 residues.

(5) There are potentially many ways to pursue the development of more accurate predictors of the B-cell epitopes. One possibility is to utilize a consensus of different predictors. Although Bprediction already implements a consensus approach, it is limited to the same predictive models and the same prediction flow. Instead, the consensus should consider combining outputs of multiple methods that use different models and flows, say BEST, Bprediction, BCTOPE and COBEpro. Similar attempts were shown to be successful for related prediction tasks, such as prediction of MHC class II peptide binding (***65***) and T-cell epitopes (***66***). Another potential direction is to find new and useful sources of information that are helpful in identifying epitopic regions. Examples include predicted disordered regions and flexible residues, predicted regions involved in protein-protein interactions, and results generated through homology modeling.

Acknowledgement

This work was supported by National Science Foundation of China (NSFC) grants 31050110432 and 31150110577 to LK. JG was supported by the Fundamental Research Funds for the Central Universities grant 65011491.

References

1. Kringelum JV, Lundegaard C, Lund O, Nielsen M (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol, 8(12):e1002829

2. Pellequer JL, Westhof E, van Regenmortel MH (1991) Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol, 203:176-201.

3. Reineke U, Schutkowski M (2009) Epitope mapping protocols. Methods Mol Biol, vol. 524

4. El-Manzalawy Y, Honavar V (2010) Recent advances in B-cell epitope prediction methods. Immunome Res, 6(Suppl 2):S2

5. Yao B, Zheng D, Liang S, Zhang C (2013) Conformational B-cell epitope prediction on antigen protein structures: a review of current algorithms and comparison with common binding site prediction methods. PLoS One, 8(4):e62249

6. Ansari HR, Raghava GP (2013) In silico models for B-cell epitope recognition and signaling. Methods Mol Biol, 993:129-138

7. Yang X, Yu X (2009) An introduction to epitope prediction methods and software. Rev Med Virol, 19(2):77-96

8. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. Nucleic Acids Res, 38:D854-862

9. Sharma OP, Das AA, Krishna R, Kumar SM, Mathur PP (2012) Structural Epitope Database (SEDB): A Web-based database for the epitope, and its intermolecular interaction along with the tertiary structure information. J Proteomics Bioinform, 5: 84-89

10. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova AI, Guan P, Hattotuwagama CK, Flower DR (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res, 1:4

11. Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, Greenbaum J, Lundegaard C, Sette A, Lund O, Bourne PE, Nielsen M, Peters B (2012) Immune epitope database analysis resource. Nucleic Acids Res, 40:W525-530

12. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J, Lund O, Lundegaard C, Nielsen M, Ponomarenko J, Sette A, Zhu Z, Peters B (2008) Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res, 36:W513-518

13. Rubinstein ND, Mayrose I, Martz E, Pupko T (2009) Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics, 10:287

14. Rubinstein ND, Mayrose I, Pupko T (2009) A machine-learning approach for predicting B-cell epitopes. Mol Immunol, 46: 840-847

15. Gao J, Faraggi E, Zhou Y, Ruan J, Kurgan L (2012) BEST: improved prediction of B-cell epitopes from antigen sequences. PloS One, 7:e40104

16. Blythe MJ, Doytchinova IA, Flower DR (2002) JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics, 18:434-439

17. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. J Chem Inf Comput Sci, 43:1276-1287

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol, 215(3):403-410

19. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol, 3(3):e91

20. Peters B, Sette A (2007) Integrating epitope data into the emerging web of biomedical knowledge resources. Nat Rev Immunol, 7(6):485-490

21. Saha S, Bhasin M, Raghava GP (2005) Bcipep:A database of B-cell epitopes. BMC Genomics, 6(1):79

22. Huang J, Honda W (2006) CED: a conformational epitope database. BMC Immunol, 7:7

23. Schlessinger A, Ofran Y, Yachdav G, Rost B (2006) Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Res, 34:D777-780

24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res, 28(1):235-242

25. Laskowski RA (2001) PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res, 29:221-222

26. Bhasin M, Singh H, Raghava GP (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics, 19:665-666

27. Kaas Q, Ruiz M, Lefranc MP (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res, 32:D208-210

28. Magrane M, UniProt Consortium. (2011) UniProt Knowledgebase: a hub of integrated protein data. Database, bar009

29. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res, 40:D130-135

30. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci, 78:3824-3828

31. Welling GW, Weijer WJ, van der Zee R, Welling-Wester S (1985) Prediction of sequential antigenic regions in proteins. FEBS Lett, 188:215-218

32. Karplus PA, Schulz GE (1985) Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen. Naturwissenschaften, 72:212-213

33. Parker JM, Guo D, Hodges RS (1986) 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

34. Kolaskar AS, Tongaonkar PC (1990) A semi empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett, 276:172-174

35. Pellequer JL, Westhof E, van Regenmortel MH (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol Lett, 36(1):83-99

36. Pellequer JL, Westhof E (1993) PREDITOP: a program for antigenicity prediction. J Mol Graph, 11:191-202

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

38. Odorico M, Pellequer JL (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. J Mol Recognit, 16(1):20-22

39. Saha S, Raghava GP (2004) BcePred: prediction of continuous b-cell epitopes in antigenic sequences using physico-chemical properties. Third Intern Conf on Artificial Immune Systems, 197-204

40. Chang HT, Liu CH, Pai TW (2008) Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches. J Mol Recognit, 21(6):431-441

41. Blythe MJ, Flower D (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci, 14:246-248

42. Saha S, Raghava, G.P. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins, 65(1):40-48

43. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immunome Res, 24:2:2

44. Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids, 33(3):423-428

45. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. J Mol Recognit, 21(4):243-255

46. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting flexible length linear B-cell epitopes. Comput Syst Bioinformatics Conf, 7:121-132

47. Sweredoski MJ, Baldi P (2009) COBEpro: a novel system for predicting continuous B-cell epitopes. Protein Eng Des Sel, 22(3):113-120

48. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC (2010) SVM-based prediction of linear B-cell epitopes using Bayes feature extraction. BMC Genomics, 11(Suppl 4):S21

49. Wang Y, Wu W, Negre NN, White KP, Li C, Shah PK (2011) Determinants of antigenicity and specificity in immune response for protein sequences. BMC Bioinformatics, 12:251

50. Wang HW, Lin YC, Pai TW, Chang HT (2011) Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. J Biomed Biotechnol, 2011:432830

51. Yao B, Zhang L, Liang S, Zhang C (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PLoS One, 7(9):e45152

52. Singh H, Ansari HR, Raghava GP (2013) Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One, 8(5):e62216

53. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25(17):3389-3402

54. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, 292(2):195-202

55. Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res, 33:W72-76

56. Vucetic S, Brown CJ, Dunker AK, Obradovic Z (2003) Flavors of protein disorder. Proteins, 52:573-584

57. Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol, 266:554-571

58. Ansari HR, Raghava GP (2010) Identification of conformational B-cell Epitopes in an antigen from its primary sequence. Immunome Res, 6:6

59. Zhang W, Niu Y, Xiong Y, Zhao M, Yu R, Liu (2012) Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning. PLoS One, 7(8):e43575

60. Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins, 47(2):142-153

61. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins, 47(2):228-235

62. Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins, 74: 847-856

63. Dor O, Zhou Y (2007) Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins, 66: 838-845

64. Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59, 467-475

65. Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol, 4(4):e1000048

66. Moutaftsi M, Peters B, Pasquetto V, Tscharke DC, Sidney J, Bui HH, Grey H, Sette A (2006) A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. Nat Biotechnol, 24(7):817-819

Figures

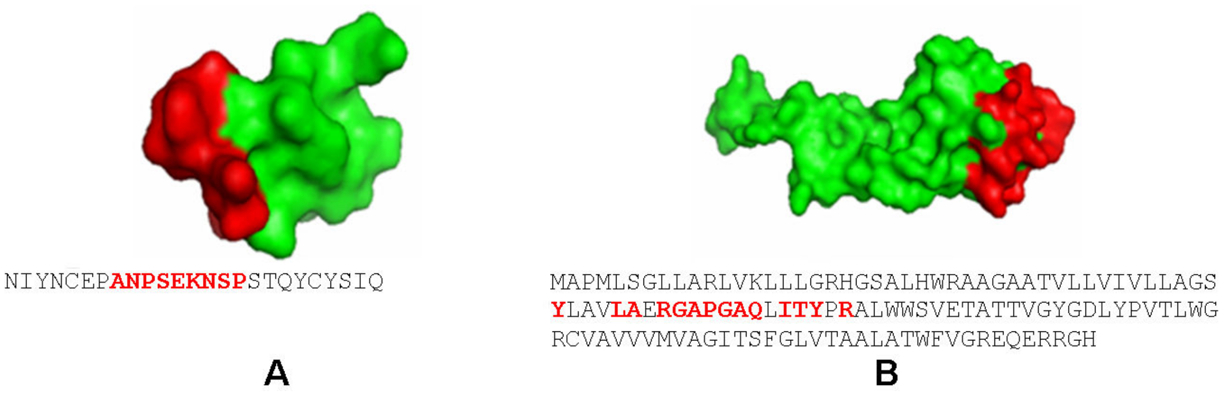
NIYNCEP**ANPSEKNSP**STQYCYSIQ
A
MAPMLSGLLARLVKLLLGRHGSALHWRAAGAATVLLVIVLLAGS
YLAV**LA**E**RGAPGAQ**L**ITY**P**R**ALWWSVETATTVGYGDLYPVTLWG
RCVAVVVMVAGITSFGLVTAALATWFVGREQERRGH
B

Figure 1. Example linear and conformational epitopes. Panel A shows linear epitope for the B-lymphocyte antigen CD20 from *H. sapiens* (IEDB ID: 1610831; PDB ID: 3PP4:P). Panel B gives conformational epitope for the voltage-gated potassium channel from *S. lividans* (IEDB ID: 142362; PDB id: 1K4D:C). Annotations of epitopes were extracted from the Immune Epitope DataBase (IEDB) (***8***) and the protein structures were collected from the Protein Data Bank PDB (***24***). Red color denotes localization of the B-cell epitope on the surface of the antigen protein and red and bold font shows the epitope in the corresponding sequence.

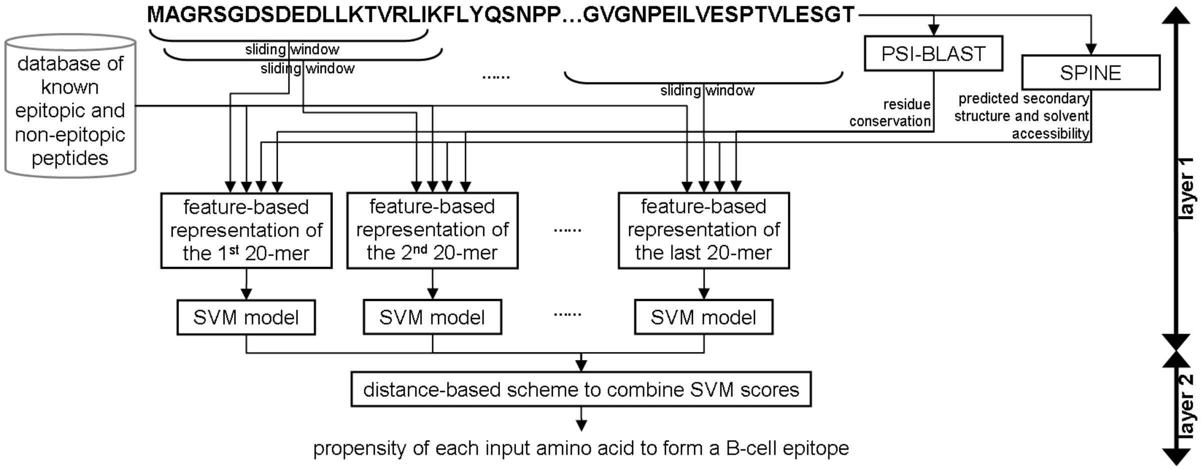
MAGRSGDSDEDLLKTVRLIKFLYQSNPP…GVGNPEILVESPTVLESGT
sliding window
sliding window
……
sliding window
database of known epitopic and non-epitopic peptides
PSI-BLAST
SPINE
residue conservation
predicted secondary structure and solvent accessibility
feature-based representation of the 1st 20-mer
feature-based representation of the 2nd 20-mer
……
feature-based representation of the last 20-mer
SVM model
SVM model
……
SVM model
distance-based scheme to combine SVM scores
propensity of each input amino acid to form a B-cell epitope
layer 1
layer 2

Figure 2. Architecture of the BEST predictor of conformational B-cell epitopes. SVM stands for support vector machine.

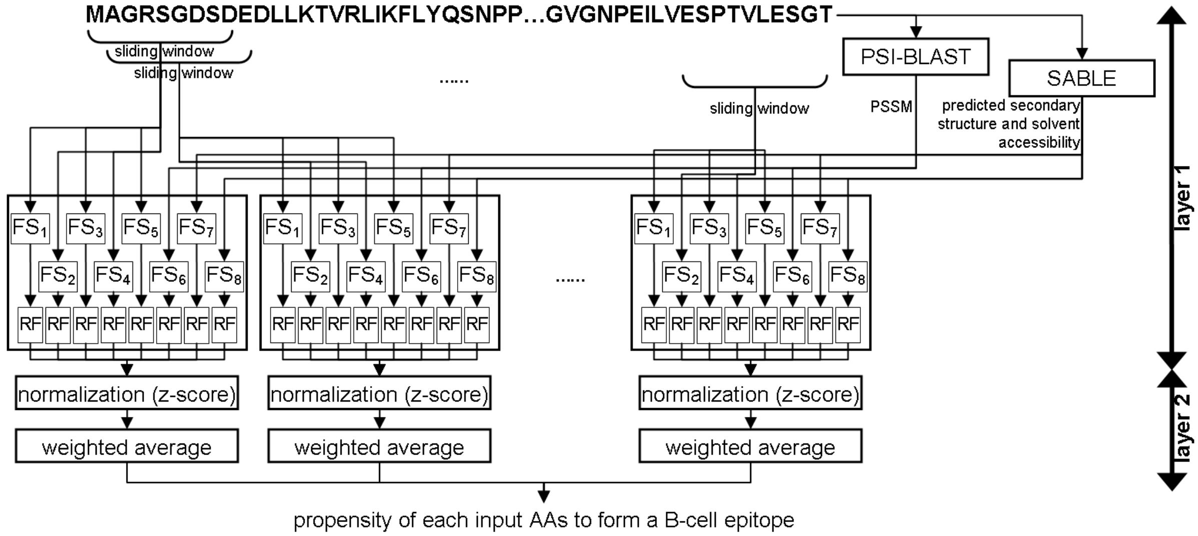
MAGRSGDSDEDLLKTVRLIKFLYQSNPP…GVGNPEILVESPTVLESGT
sliding window
sliding window
......
sliding window
PSI-BLAST
SABLE
PSSM
predicted secondary structure and solvent accessibility
FS1
FS3
FS5
FS7
FS2
FS4
FS6
FS8
RF
......
normalization (z-score)
weighted average
propensity of each input AAs to form a B-cell epitope
layer 1
layer 2

Figure 3. Architecture of the Bprediction method for the prediction of conformational B-cell epitopes. FS$_i$ refers to i^{th} feature set, where i = 1 (physicochemical propensities), 2 (binary amino acids composition), 3 (real-valued amino acids composition), 4 (composition of amino acids sets), 5 (composition of dipeptides), 6 (PSSM values), 7 (predicted secondary structure), 8 (predicted relative solvent accessibility). RF stands for random forest.

Table Captions

Table 1. Summary of sequence-based predictors of linear B-cell epitopes. The methods are sorted chronologically.

Method	Year	Model	Type[1]	Input[2]	Availability
BEPITOPE	2003	Scoring function	SP	SC	by contacting the authors
ABCpred	2006	Neural network	WS	SC	http://www.imtech.res.in/raghava/abcpred/
BepiPred	2006	Hidden markov model	WS+SP	MC	http://www.cbs.dtu.dk/services/BepiPred/
AAP	2007	Support vector machine	WS+SP	SC	http://ailab.cs.iastate.edu/bcpreds/
LEP-LP	2008	Scoring function	WS	unknown	http://biotools.cs.ntou.edu.tw/lepd_antigenicity.php[3]
BCPred	2008	Support vector machine	WS+SP	SC	http://ailab.cs.iastate.edu/bcpreds/
FBCPred	2008	Support vector machine	WS+SP	SC	http://ailab.cs.iastate.edu/bcpreds/
Epitopia	2009	Naïve Bayes	WS+SP	SC	http://epitopia.tau.ac.il
BayesB	2010	Support vector machine	WS	SC	http://www.immunopred.org/bayesb/
BROracle	2011	Support vector machine	SP	unknown	https://sites.google.com/site/oracleclassifiers/[3]
LEPS	2011	Support vector machine	WS	SC	http://leps.cs.ntou.edu.tw
SVMTriP	2012	Support vector machine	WS	SC	http://sysbio.unl.edu/SVMTriP
LBtope	2013	Support vector machine	WS	MC	http://crdd.osdd.net/raghava/lbtope/

[1] SP: standalone program; WS: web server

[2] SC: method predicts a single chain, i.e., prediction has to be restarted for each chain; MC: multiple chains can be predicted at the same time

[3] a given predictor is currently unavailable

Table 2. Summary of sequence-based predictors of conformational B-cell epitopes. The methods are sorted chronologically.

Method	Year	Model	Type[1]	Input[2]	Availability
COBEpro	2009	Support vector machine	WS	SC	http://scratch.proteomics.ics.uci.edu
CBTOPE	2010	Support vector machine	WS+SP	MC	http://www.imtech.res.in/raghava/cbtope/
BEST	2012	Support vector machine	SP	MC	http://biomine.ece.ualberta.ca/BEST/
Bprediction	2012	Random forest	WS	SC	http://bcell.whu.edu.cn

[1] SP: standalone program; WS: web server

[2] SC: method predicts a single chain, i.e., prediction has to be restarted for each chain; MC: multiple chains can be predicted at the same time