Chapter 10

Prediction of Conformational B-Cell Epitopes

Wen Zhang, Yanqing Niu, Yi Xiong, and Meng Ke

Abstract

Conformational B-cell epitopes play an important role in the epitope-based vaccine design. The increase of available data promotes the development of computational methods. Compared with the wet experiments, the computational methods are faster and more economic. In the past few years, a number of computational methods (especially the machine learning-based methods) have been developed to predict the conformational B-cell epitopes. In this chapter, we introduce important data resources and computational methods, which are publicly available. Moreover, we introduce our ensemble learning-based method that can predict the conformational epitopes from sequences. These promising methods may assist immunologists in identifying potential vaccine candidates.

Key words Conformational B-cell epitopes, Machine learning, Epitope-based vaccine design

1 Introduction

Antigen–antibody interaction is a critical event in the immune process, which may elucidate the underlying mechanism of immune recognition [1–4]. The sites on antigens recognized and bound by B cell-produced antibodies are well known as B-cell epitopes. The location of B-cell epitopes is useful for synthesizing peptides that can elicit the immune response with specific cross-reacting antibodies. For this reason, the identification of B-cell epitopes facilitates the design of the potentially safer peptide-based vaccines. B-cell epitopes can be classified into two categories: linear (continuous) epitopes and conformational (discontinuous) epitopes. Linear epitopes are formed by continuous amino acid sequences, while conformational epitopes consist of residues that are distantly separated in the sequences but spatially proximal.

In the last decade, the increase of available data promotes the development of computational methods, which may be fast and economic [5]. Although the majority of all epitopes (about 90 %) are conformational, the study began fairly late. In the prediction work, there are several definitions ever used for the conformational
epitopes inferred from the X-ray structures of antigen–antibody complexes. By definitions, the epitope residue is an antigen residue with area loss upon antibody binding more than a given threshold or an antigen residue separated from any antibody residue by a Euclidean distance less than 4 Å. The study revealed that these definitions do not make significant difference. Here, we must emphasize that the epitopes in the computational work are not functional but structural, and structural epitopes cannot definitely lead to the immune response. Currently, the prediction of functional epitopes is a tough task. Thereafter, the epitopes mean the structural epitopes.

Although some protein docking methods (such as Patch Dock [6] and ClusPro [7]) can be used to predict conformational epitopes, these methods are different from those which are specially designed for the conformational epitope prediction. The docking methods require the structures of both antigens and antibodies to make prediction, while the specially designed methods attempt to predict the epitopes from antigens in the absence of antibodies.

CEP is the pioneering method proposed for the prediction of conformational epitopes [8], which uses the residue solvent accessibility. DiscoTope [9] exploits the surface accessibility, spatial information, and amino acid statistics information to identify epitopes. PEPITO [10] combines amino acid propensities and half-sphere exposure values at multiple distances to make prediction. ElliPro [11] uses Thornton’s propensities and residue clustering to make prediction. In SEPPA [12], two concepts, “unit patch of residue triangle” and “clustering coefficient,” are introduced to describe the local spatial context and spatial compactness. EPITOPIA [13] combines structural and physicochemical features and then adopts naive Bayes classifier to make prediction. EPCES [14] uses the consensus score of several structural and physicochemical terms. EPSVR [15] uses support vector machine (SVM) and combines various features for prediction. EPMeta [16] is a meta method combining the predictions from several existing servers. Liu et al. [17] adopted the logistic regression to predict the conformational epitopes. We [18] proposed a random forest-based method by dealing with the imbalanced dataset and combining various features. Above methods construct the prediction models based on antigen structures.

Although a great number of structure-based methods have been developed, their application is undermined by the limited number of available structures, and the experimental techniques that determine structures are costly and time consuming. Instead of making predictions from structures, Ansari et al. made the first attempt on sequence-based conformational epitope prediction [19]. Gao et al. developed a method based on averaging selected scores generated from sliding 20-mers by SVMs [20]. Recently, we proposed an ensemble learning model using the antigen sequences [21].
2 Materials

2.1 Database

(a) Immune Epitope Database (IEDB) ([http://www.iedb.org/][22]) can provide a highly annotated set of B-cell epitopes curated from crystal structures of antigen–antibody complexes.

(b) Conformational Epitope Database (CED) ([http://immunet.cn/ced/][23]) collected the conformational epitopes thoroughly sourced from articles published in the peer-reviewed journals. Initially, references were obtained by exhaustive querying on PubMed and ScienceDirect. The references were further manually filtered to annotate conformational epitopes.

(c) AntiJen [24] is a database with the published experimentally determined conformational B-cell epitopes ([http://www.ddg-pharmfac.net/antijen/]).

2.2 Dataset

In the conformational epitope prediction, the antigen–antibody complexes are analyzed to annotate the binding sites (epitope residues) on the antigens, and then only the antigens (structures or sequences) are used to develop the prediction models.

Several datasets are widely used in the conformational epitope prediction. The structure datasets can be classified into two kinds: bound dataset and unbound dataset. A bound dataset consists of the antigen–antibody complex structures, and the epitopes on antigen are annotated according to the definition of the conformational epitope. Then, the structures of the antigens are directly extracted from the complexes for modeling. An unbound dataset consists of complex structures and unbound structures of antigens. Annotated epitope residues on complexes (calculated according to the definition) are aligned to the residues on unbound structures of antigens. Then, the unbound antigen structures are used for modeling. One popular bound dataset is published by Rubinstein, which consists of 66 non-redundant complex structures, available at [http://epitopia.tau.ac.il/trainData/](http://epitopia.tau.ac.il/trainData/). Liang’s unbound dataset including 48 complexes and the unbound structures of antigens are available at [http://sysbio.unl.edu/services/](http://sysbio.unl.edu/services/). The antigen sequences can be extracted from the antigen–antibody complexes for the sequence-based prediction. Ansari et al. published benchmark sequence datasets available at [http://www.imtech.res.in/raghava/cbtope/supple.php](http://www.imtech.res.in/raghava/cbtope/supple.php).

3 Method

In this section, we introduce widely used conformational epitope prediction methods and their public servers (see Note 1).
3.1 DiscoTope

DiscoTope [9] is a structure-based method for conformational epitope prediction. The method uses the amino acid propensity (Parker hydrophilicity scale), spatial information (contact numbers), and surface accessibility to make prediction.

Parker hydrophilicity scale is an amino acid propensity, which can be obtained from AAIndex database. The residue contact number is the number of Ca atoms in the antigen within a distance of 10 Å of the residue Ca atom. The relative solvent-accessible surface area per antigen residue is calculated using the NACCESS program with a probe radius of 1.4 Å.

Given an antigen–antibody complex structure, the contact number score and surface accessibility score of each antigen residue are calculated. Here, the Parker hydrophilicity score of each residue is calculated over a smoothing window of seven residues. For a candidate residue, the weighted sum of the Parker hydrophilicity score, contact number score, and surface accessibility score is used for prediction. According to a preset threshold, the residue is predicted as epitope or non-epitope.

The web server of DiscoTope is available at http://www.cbs.dtu.dk/services/DiscoTope/. The users can use the PDB IDs of antigen–antibody complexes or the PDB files as input, and the server will return the prediction results. Users can specify the threshold for epitope identification.

3.2 Epitopia

For a given structure, a patch of 20 amino acids is constructed around each solvent-accessible antigen residue. Rubinstein et al. statistically evaluated a wide range of amino acid physicochemical and structural-geometrical properties [13]. These properties are (1) the ratio between the frequencies of some amino acid types in the patch and the remaining antigen surface, (2) the ratio between the frequency of helix secondary structures in the patch and the remaining antigen surface, (3) the average relative accessibility of the patch to the solvent, (4) the average accessibility of the patch, (5) the average curvature of the patch atoms, and (6) several amino acid propensities.

Then, Rubinstein et al. use the feature selection technique to obtain the optimal property subset. Starting with all properties, one property for which the deletion had the least effect on prediction accuracy is removed at each iteration. Finally, the subset of properties with the highest number of successful predictions was selected as the optimal set. The optimal property subset is used to represent patches as feature vectors. Then, naïve Bayes is used as the classification engine to build prediction model. Thus, a server named “Epitopia” is constructed to predict conformational epitopes.

Epitopia is available at http://epitopia.tau.ac.il. Users can enter the PDB ID or upload the PDB file for prediction.
3.3 EPCES

In EPCES [14], a patch (with 20 residues) is formed around each candidate antigen residue. EPCES uses consensus score from six different scoring terms to make prediction. These scoring terms are residue epitope propensity, conservation score, side-chain energy score, contact number, surface planarity score, and secondary structure composition.

The residue epitope propensity was calculated as the product of the normalized solvent-accessible surface of the residue and the logarithm ratio of the epitopic area to the rest area. The conservation score was calculated by the position-specific substitution matrix generated from PSIBLAST and the diagonal element of BLOSUM62. The side-chain energy score was calculated from the side-chain energies of all possible rotamers. The contact number is as same as the introduction in Subheading 3.1. The planarity of each patch was calculated as the root mean squared deviation of all the Cα atoms in the patch from the least squares plane through the atoms. The secondary structure composition was the fraction of patch residues forming turns or loops in all 20 patch residues.

For each candidate residue, the residue epitope propensity, conservation score, and side-chain energy score were calculated at the residue level and distance-based averaged over all residues in the patch by following distance-based equation:

\[ E_{\text{patch}}(i) = \sum_{k=1}^{20} E_{\text{residue}}(K) \cdot e^{-\frac{d}{T}} \]

where \( E_{\text{residue}}(K) \) is the score of residue \( K \) in the patch, \( d \) is the distance between \( K \) and the central residue of the patch, and \( T \) is the parameter needed to be optimized.

Each scoring term can predict a candidate residue as epitope or non-epitope according to its score and a given threshold. For a residue, if more than five scoring terms yield the scores greater than a given threshold, it is finally predicted as the epitope residue.

A web-based EPCES application is available at http://sysbio.unl.edu/services/EPCES/. The PDB ID of an unbound structure or the PDB file is used as the input. The output will be displayed on this web page when the prediction is completed. The output includes the predicted antigen residue and its possibility of being an epitope residue.

3.4 EPSVR

EPSVR [15] uses a support vector regression (SVR) method to integrate six scoring terms ever used in the EPCES.

For each surface patch, the number of epitopic residues could be any integer value between 0 and the patch size (i.e., 20). Therefore, each patch is assigned a real value associated with the number of epitopic residues, and the prediction of conformational
epitopes is transformed as a problem of regression. Each surface patch had six SVR attributes, whose values were calculated with the six scoring terms: residue epitope propensity, conservation score, side-chain energy score, contact number, surface planarity score, and secondary structure composition. The six scores and the number of observed epitope residues in the patch were scaled to 0–1. Then, the SVR-based model is built to make prediction.

The web server of EPSVR is available at \url{http://sysbio.unl.edu/EPSVR/}. The input and output of EPSVR are same as those of EPCES.

3.5 **CBTOPE**

CBTOPE [19] is the first method of predicting conformational B-cell epitopes from antigen sequences. The fixed-length window is shifted over the antigen sequences to generate residue segments (peptides). According to the central residues (epitope or non-epitope), the peptides are labeled as positive or negative. Then, each peptide can be represented as a feature vector by several encoding schemes, including binary profile, physicochemical profile, and composition profile.

Binary profile represents each amino acid as a 21-dimensional vector. Physicochemical profile uses Grantham polarity, Karplus–Schulz flexibility, Kolaskar antigenicity, Parker hydrophobicity, and Ponnuswami polarity index to represent amino acids. Amino acid composition is the percentage of each amino acid type in a peptide. Three encoding schemes are used for peptide representation, and the prediction models are constructed by using SVM. Among all encoding schemes, the composition profile can produce the best results.

A web server CBTOPE has been developed to predict conformational epitopes, available at \url{http://www.imtech.res.in/raghava/cbtope/}. Users can enter antigen sequences for prediction.

4 **The Sequence-Based Ensemble Learning Method**

We follow the work pioneered by CBTOPE and focus on two aspects concerning the sequence-based prediction [21]. One is to explore more potential sequence-derived features relevant to conformational epitopes. The other is to effectively use various features which may share redundant information. In order to address these issues, we evaluate several sequence-derived features, which are ever used in the epitope prediction or similar tasks. Second, we consider the ensemble learning technique that can incorporate useful features, and the weighted scoring approach is adopted to build the prediction model.

4.1 **The Basic Idea of Ensemble Learning Method**

The overlapping residue segments (peptides) are generated from the antigen sequences by using a sliding window of the length \( L \). For simplifying, let \( L \) be an odd integer. For a sequence with \( N \) residues, a total of \( N - L + 1 \) peptides are extracted, and each peptide
is labeled as positive or negative according to the label of its central residue (epitope residue or non-epitope residue). The prediction of conformational epitopes from sequences is formulated as the problem of binary classification. We consider several sequence-derived features, which are described as follows.

Physicochemical propensities: These physicochemical propensities are flexibility scale, hydrophilicity scale, surface-exposed residue scale, polarity scale, beta-turn scale, and accessibility scale.

Sparse profile: Sparse profile is a widely used representation of amino acids. Each amino acid type (20 common types in all) can be represented by a 20-bit binary string, in which the value at one bit is 1 and others are 0.

Amino acid composition: According to the previous study, some amino acid types are significantly overrepresented in epitopes, and others are underrepresented; thus the amino acid composition can be used to differentiate epitope regions from non-epitope regions. Here, we use the amino acid composition of the residue segments (also called as sliding windows or samples) extracted from the whole sequences.

Amino acid function group: Since contacts between antibodies and the antigens are mostly determined through functional moieties of the R-groups, functional moieties can influence the location of antibody–antigen-binding sites. According to different R-groups, 20 amino acid types are classified into 13 classes. In order to take antigen–antibody interaction into consideration, we present a novel feature named “amino acid function group” and use 13-bit binary strings to represent 13 functional classes.

Amino acid functional composition: By incorporating both amino acid function group and amino acid composition, we present a novel feature “amino acid functional composition,” which represents the percentage of each amino acid functional type in a sequence.

Evolutionary profile: The evolutionary conservation is represented by the position-specific scoring matrix (PSSM), which is obtained by aligning the target sequence against NCBI non-redundant reference sequences with PSI-BLAST tool. For an amino acid sequence with \( L \) residues, the PSSM has \( L \) rows and 20 columns. PSSM values in each row are rescaled by the standard logistic function \( f(x) = 1 / (1 + e^{-x}) \). When using the evolutionary profile, a residue is represented by its corresponding 20-dimensional row vector in the matrix.

Amino acid pair profile: The amino acid pair profile is usually observed to be associated with the protein functions. Amino acid pair profile of a sequence represents the percentage of each amino acid pair type.

Although structural information cannot be directly obtained from antigen sequences, some state-of-the-art tools can help to predict it. Here, the SABLE program \cite{25} is adopted, for the
A stand-alone tool is publicly available. With the given sequences as input, the software can predict the secondary structures (SS) and relative accessible surface areas (RASA) of residues. The predicted SS of a residue is denoted as H, E, or C (helix, sheet, coil), and (1, 0, 0), (0, 1, 0), and (0, 0, 1) are, respectively, used to represent three types. The predicted RASA of a residue is a real value between 0 and 100, representing the percentage of exposed area of the residue over its full area.

The statistical study indicates that all features have the ability of differentiating epitope regions from non-epitope regions [21]. Since the amino acid functional composition incorporates both amino acid composition and amino acid group, seven groups of features including physicochemical propensities, evolutionary profile, amino acid functional composition, sparse profile, amino acid pair, sequence-predicted secondary structure, and sequence-predicted relative solvent accessibility are finally used for the development of prediction models.

Obviously, there are much more non-epitopes than epitopes, and the instances are seriously imbalanced. A strategy based on the data bootstrap is used to deal with the imbalanced data, and random forest [26] is used as the classification engine. Thus, a classification model which consists of multiple random forests is constructed (described in Fig. 1) and used as the base module for ensemble learning.

Since a peptide can be represented as different feature vectors by different descriptors (features), multiple base modules can be constructed. We adopt a simple ensemble strategy named weighted scoring [27] to integrate modules and develop the ensemble model.

**Fig. 1** The classification model based on the random forest and data bootstrap
Given an instance, each base module will produce a score, and then these scores are normalized. Further, a weight is assigned to the normalized score yielded by a base module, and the sum of weighted scores is adopted as the final prediction (see Note 2).

The web server is constructed by JavaScript and Tomcat. In order to calculate the conservation score, secondary structures, and relative accessible surface areas, we have to use some external tools (i.e., PSI-BLAST and SABLE program). PSI-BLAST is a Windows version executive program; SABLE [25] is written in Perl. The outputs from external tools are parsed to obtain feature values used for sequence representation.

We adopt the Weka package [28] to implement the machine learning methods. Weka is a collection of java code implementing machine learning algorithms, including data preprocessing, classification, regression, clustering, association rules, and visualization. Here, we use the random forest class in Weka to develop our ensemble learning model. The inputs of the model are the feature vectors representing sequences, and probability of being an epitope residue is returned for each residue. The server is available at http://bcell.whu.edu.cn.
In the web page of prediction (shown in Fig. 3), users can enter an antigen sequence and its information (sequence name and chain name). In addition, the e-mail address should be specified to receive the prediction result. A typical task (a sequences of 30 residues) takes about 15–20 min. The running time depends on the length of the submitted sequence. In the returned result (shown in Fig. 4), the first column is the residue id; the second column is the residue name; and the third column is the probability for the residue to be the epitope residue.

**Fig. 3** The web page of the server

**Fig. 4** An example of the returned result

<table>
<thead>
<tr>
<th>Residue</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.4933333333333334</td>
</tr>
<tr>
<td>V</td>
<td>0.6533333333333332</td>
</tr>
<tr>
<td>T</td>
<td>0.28666666666666667</td>
</tr>
<tr>
<td>T</td>
<td>0.24666666666666667</td>
</tr>
<tr>
<td>Y</td>
<td>0.263525434170393</td>
</tr>
<tr>
<td>K</td>
<td>0.31678109640819324</td>
</tr>
<tr>
<td>L</td>
<td>0.08668649400409438</td>
</tr>
</tbody>
</table>

In the web page of prediction (shown in Fig. 3), users can enter an antigen sequence and its information (sequence name and chain name). In addition, the e-mail address should be specified to receive the prediction result. A typical task (a sequences of 30 residues) takes about 15–20 min. The running time depends on the length of the submitted sequence. In the returned result (shown in Fig. 4), the first column is the residue id; the second column is the residue name; and the third column is the probability for the residue to be the epitope residue.

5 Conclusion

This chapter introduces the data resources and computational methods related with the conformational B-cell epitope prediction, especially our sequence-based conformational epitope prediction method and the public server. The above-discussed methods
have large potential for the practical use. The publicly available servers will assist immunologists in identifying potential vaccine candidates.

6 Notes

1. As far as we know, some structure-based methods are trained and evaluated on the bound dataset (DiscoTope, SEPPA, Epitopia), and others are constructed and tested on the unbound dataset (EPSVR, EPCES). CBTOPE and our ensemble method are developed by using antigen sequences.

2. The sequence-based ensemble learning method has some advantages. First, the ensemble model provides a flexible frame that incorporates individual feature-based classifiers. Second, the ensemble model can select the features by itself and integrate them based on the discriminative power. According to the optimal weights, we can approximately know the components of the ensemble model. Therefore, this ensemble model is easy to not only implement but also explain.

Acknowledgments

This work is supported by the National Science Foundation of China (61103126), the Ph.D. Programs Foundation of Ministry of Education of China (20100141120049), and Natural Science Foundation of Hubei Province (No. 2011CDB454).

References

25. Sable server available at: http://sable.cchmc.org/