

Chapter 14

Bioinformatics Methods to Predict Allergen Epitopes

Ying He and Ailin Tao

Abstract Epitopes are composed of contiguous or discontinuous specific amino acid residues on an antigen. Binding of these epitopes with their corresponding receptors on the surface of lymphocytes is a necessary step for an antigen to activate these cells and mount an immune response in the body. Therefore, precise determination of antigenic epitopes is very important in vaccine design research, modification of allergenicity, understanding immune tolerance mechanisms, and more. With the rapid development of bioinformatics, many free online servers with programs to predict B-cell or T-cell epitopes have emerged in recent years. These prediction methods are based on different modeling principles and algorithms, so their outcomes and accuracy varies. This chapter will introduce some typical programs for the prediction of B-cell and T-cell epitopes as well as their modeling principles and algorithms.

Keywords Bioinformatics · B-cell epitope · MHC binding · Prediction · T-cell epitope

Y. He · A. Tao (✉)

Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology,
The State Key Clinical Specialty in Allergy, The State Key Laboratory of Respiratory Disease,
The Second Affiliated Hospital of Guangzhou Medical University,
250# Changgang Road East, Guangzhou, China
e-mail: Aerobiologiao@163.com

Y. He

e-mail: heying0605@163.com

14.1 Introduction

Type I hypersensitivity can be divided into two stages, sensitization, and challenge. In the sensitization phase, an allergen induces B cells to produce IgE. With their Fc segment, IgE binds to the Fc ϵ RI on the surface of mast cells, basophils, and other target cells, and the body becomes sensitized. In the challenge stage, the same allergen is encountered again, specific IgE binds to the surface of the target cells, and anaphylaxis can occur. Mast cell and/or basophil degranulation causes an allergic reaction by releasing previously synthesized or newly synthesized inflammatory mediators such as histamine and leukotrienes. IgE plays a key role in this process and its production involves the differentiation of T_H2 cells and IgE class switching. The antigen-presenting cells process the allergen to antigenic peptides, forming a complex of antigenic peptide and MHC molecule that is recognized by the T-cell antigen receptor (TCR). The naive T cells then differentiate into T_H0 cells, which further differentiate into T_H2 cells under the influence of multiple factors, thereby promoting B-cell secretion of IgE. On the other hand, B-cell antigen receptors (BCR) also identify allergens, resulting in activation of STAT6, which can promote IgC ϵ exon germline gene transcription and promote an IgE class switch (Linehan et al. 1998). Thus, for an allergen to act as an antigen to induce an immune response, it must first be processed and the antigenic peptides presented, with recognition by T cells and B cells being a key step. Currently, methods to predict T-cell epitopes are mainly achieved by predicting the binding affinity of antigen peptides to MHC molecules. However, rarely, bioinformatics tools are available to help identify how the TCR binds to the peptide-MHC complex and which amino acid residues on the peptide bind directly to the TCR.

The antigenic epitope refers to an antigen-specific molecule determined by specific chemical groups. Antigens bind through their antigenic epitopes to corresponding lymphocyte surface antigen receptors leading to the activation of the lymphocytes and eliciting an immune response. The character, number, and spatial configuration of the antigenic epitopes determine the specificity of the antigen, thus influencing the production of the corresponding antibodies or sensitized lymphocytes. Epitopes can be divided into B or T-cell epitopes according to which cell type they bind to. B-cell epitopes are linear fragments or have a conformation that can bind to the BCR or specific antibody. Linear antigen epitopes contain a contiguous amino acid sequence that is at the core of the interaction with the receptor or antibody. Most of the current predictions using bioinformatics are based on linear epitopes. However, 90 % of B-cell epitopes are conformational, not linear epitopes. Conformational epitopes contain some noncontiguous amino acids in the primary protein structure that become adjacent to each other in the folded tertiary structure. T-cell epitopes are linear epitopes that are recognized by a T-cell receptor (TCR). Using bioinformatics to predict antigenic epitopes will help us to study the modification of allergenicity, vaccine design, and immune tolerance mechanisms.

14.2 Prediction of Linear B-Cell Epitopes

14.2.1 Prediction Scheme

Methods to predict linear B-cell epitopes are based on the primary structure of the protein together with a comprehensive measure of physicochemical properties, structural features, and other characteristics of the protein antigen. The most commonly evaluated parameters include hydrophilicity, surface probability, antigenic index, flexibility, charge distribution scheme, and secondary structure.

- **Hydrophilicity**

Amino acid residues of the protein antigen are divided into hydrophilic residues and hydrophobic residues. Generally, hydrophobic residues are buried inside the protein, whereas hydrophilic residues are on the surface and are therefore closely linked to the antigen epitopes. Previous investigations have demonstrated that charged, polar, and hydrophilic amino acid side chains are common features of antigenic determinants. In 1981, a method to predict the antigenic determinants by analyzing amino acid sequences was proposed (Hopp and Woods 1981). In this prediction method, each amino acid residue is assigned a hydrophilicity value and then repetitively averaging these values in a hexapeptide along the peptide chain. The point of highest local average hydrophilicity would invariably lie within or be immediately adjacent to an antigenic determinant. Since the time this method was developed, the Kyte–Doolittle scale (Kyte and Doolittle 1982), HPLC scale (Parker et al. 1986), and other hydrophilic prediction schemes have been proposed.

- **Surface probability**

Surface probability refers to the possibility of amino acid residues contacting surface solvent molecules. It reflects the distribution of the outer layers and inside of the residues, which indirectly reflect the capacity for antibody binding. The Emini scale (Emini et al. 1985) is a typical method for predicting surface probability.

- **Flexibility**

Conformational protein antigens are not rigid and contain a polypeptide chain skeleton that has a certain degree of flexibility. The strong active amino acid residue is the site of a large flexibility locus. Since the antigen–antibody binding is a lock-and-key process and protein conformations change, “flexible” amino acid residues that are prone to twist and fold are most likely to become antigenic epitopes. Karplus and Schulz developed a predictive method for protein fragment activity based on the atomic temperature factor of 31 proteins with a known three-dimensional structure (Karplus and Schulz 1985).

- **Charge distribution**

The charge of specific antibodies produced in response to alkaline antigens tends to be acidic, whereas the charge of specific antibodies to acidic antigens tends to be alkaline (Gershoni et al. 1997).

- **Antigenic index**

Welling et al. studied 606 amino acids from 20 proteins of which the antigenic regions have been well researched in order to determine the average composition of an antigenic region (Welling et al. 1985). Each amino acid is described by the frequency that it occurs in the antigenic region which is then divided by the frequency of each amino acid in the total protein to generate a scale value. Therefore, this method is based on the percentage of each amino acid present in known antigenic determinants compared with the percentage of that amino acid in the total protein. Studies have shown that some antigenic regions predicted by Welling scale are consistent with the results from the hydrophilic method of Hopp and Woods (1981). But the disadvantage of this method is that the database is limited. Another antigenic index prediction method was proposed in 1988, which predicts the topological features of a protein directly from its primary amino acid sequence. The computer program generates values for surface accessibility parameters and combines these values with those obtained for regional backbone flexibility and predicted secondary structure (Jameson and Wolf 1988).

- **Secondary structure**

Secondary structure prediction programs show that a protruding β turn structure is more likely to become an epitope as it often appears on the surface of protein antigens and is easy for antibodies to bind to unlike α -helix and β -sheet structures that are generally not epitopes because they are not flexible and are difficult for antibodies to bind to. Protein secondary structure prediction methods include the Chou–Fasman (Chou and Fasman 1978), Garnier–Robson (Garnier–Robson 1978), Cohen (Colloc'h and Cohen 1991), and other methods, but the success rates of these methods for predicting secondary protein structures are below 65 %. Cohen has a high prediction rate for β turn with an accuracy rate as high as 95 % for the known types of protein folding ($\alpha\alpha$ type, $\beta\beta$ type, and α/β type). In the case of a protein with unknown structure, prediction methods mentioned above for three types secondary structures are also available. In addition, the presence of a protruding β turn that is consistent in three prediction results is helpful in identifying epitopes.

The various methods for predicting correct epitopes are not that successful. Generally speaking, B-cell epitopes should be easily located or able to move to the cell surface in order to bind with their specific antibodies. In addition, a certain amount of flexibility is required because protein conformations will change when the antigen and antibody bind together. Typically, protein epitopes are present in the β turn. Therefore, the amino acid residues that are often present in the β turn also often appear in the antigenic epitope. Furthermore, the antigenic epitope generally consists of the amino acid residues with high hydrophilicity. However, current computer software programs cannot guarantee accurate predictions so further experiments are needed to verify.

14.2.2 Prediction Software

With the intensive study of the physical properties of amino acid residues, computer programs are now available, which predict B-cell linear epitopes including PEOPLE (Alix 1999), BEPITOPE (Odorico and Pellequer 2003), BcePred (Saha and Raghava 2004), and Protean of DNASTAR.

BcePred (<http://www.imtech.res.in/raghava/bcepred/>) is an online software program that predicts linear B-cell epitopes based on their physical and chemical properties (Saha and Raghava 2004). This dataset contains 1029 nonredundant B-cell epitope peptides from the Bcipep database, which were verified in that the same number of non-epitopic peptides was extracted from the SWISS-PROT database. The accuracy of this software based on the different prediction schemes was between 52.92 % and 57.53 %. After combining four properties (hydrophilicity, flexibility, polarity, and exposed surface) of the amino acid, the prediction accuracy increases to 58.70 % and the threshold value is 2.38. This server allows users to select any combination of amino acids and different physical and chemical characteristics including hydrophilicity, flexibility, accessibility, polarity, antigenic propensity, exposed surface and turns to predict the B-cell epitopes. The users can also choose any threshold value [−3 to +3] for the different parameters with the result being that when the threshold value is increased the specificity of the prediction can be improved but the sensitivity is decreased. The default threshold values for the different parameters are set based on the optimal sensitivity and specificity. BcePred prediction results can be presented in graphical or tabular form. The graphical format displays the score for every parameter of each amino acid residue along the protein backbone in different colors, helping users to easily observe the B-cell epitopes. Amino acid residues with values higher than the threshold value (default is 2.38) are considered to be B-cell epitopes.

14.3 Prediction of Conformational B-Cell Epitopes

Due to the limited number of known crystal structures of antigen–antibody complexes, the mechanism of antigen–antibody binding is not yet clear and, because of the difficulty in designing prediction algorithms as well as for other reasons, progress in predicting conformational B-cell epitopes has been relatively slow. In recent years, with the development of bioinformatics, some software programs that can predict conformational B-cell epitopes have been created.

The prediction of conformational B-cell epitopes usually involves the identification of surface and epitope residues, construction of benchmark datasets, algorithm design, and performance evaluation. Currently, prediction methods can be divided into four types: (i) machine learning methods; (ii) non-machine learning methods; (iii) methods based on phage display experiments; and (iv) protein–protein interface methods. Most of these methods are based on the protein structure,

and some are based on amino acid sequence data or are combined with the phage display data. Some methods use training data that have been validated and some methods use nonvalidated data. These prediction methods generally need to take the three-dimensional structure of the antigen into consideration, and a fully automated protein structure homology-modeling server such as SWISS-MODEL is available for evaluating antigens with an unknown three-dimensional structure. The result of the epitope prediction includes the predicted epitopic residue, position and number of epitopic residues. Different methods have various forms of output. Some software programs give a score for each surface residue that reflects the probability of it being a protein epitope. Some software programs predict whether the surface patch, consisting of surface residues, is an epitope. Other software combines the above two functions. Some even provide a graphical visualization of three-dimensional structures to display the epitope mapping. Most of the software programs are able to predict both linear and conformational epitopes.

14.3.1 Machine Learning Method

In 2005, an online conformational B-cell epitope prediction software program CEP (<http://bioinfo.ernet.in/cep.htm>) was published based on the machine learning prediction method (Kulkarni-Kale et al. 2005). It is the first free online server that uses the Protein Data Bank (PDB) file of the antigen protein tertiary structure as input and a prediction of the conformational epitope is its output. It can also predict linear epitopes, and the results are visualized graphically. This algorithm employs a structure-based bioinformatics method and solvent accessibility of amino acids in a specialized manner. By calculating the relative surface area of each amino acid in the protein structure, this software is able to confirm which regions are sufficiently exposed to be an antigenic epitope. The antigenic residues with ≥ 25 percent accessible surface area are identified. Antigenic determinants are identified if at least three contiguous accessible residues are present. If the spatial distance between the antigenic residues is below a specified cut-off value, it is predicted to be a conformational epitope. The prediction accuracy of this software is as much as 75 %. This server was the first approach available for the prediction of conformational B-cell epitopes, which is of groundbreaking significance, but its disadvantage is that it has fewer feature selections compared with subsequently developed similar software.

ElliPro (Ponomarenko et al. 2008) and many other prediction software programs based on the machine learning method have now been created. Machine learning methods can reveal complex nonlinear relationships in the dataset and are suitable for this field with their huge amount of data, but there is noise and lack of a unified theory. In order to overcome the deficiencies that include poor prediction accuracy, these methods are usually combined with additional features, and sometimes data verified by experiments are used to make predictions. However, further

improvements in the performance of the system depend on the growing number of benchmark datasets and other more valuable features yet to be found.

14.3.2 Non-machine Learning Method

Use of non-machine learning methods to predict B-cell epitopes refers to the use of geometric features of the protein antigenic structure, or the joint use of the geometric characteristics together with the physical and chemical characteristics, to make predictions directly through calculations and without the use of machine learning methods. Such methods are intuitively clear and trusted by some researchers. Representative software include DiscTope (Haste Andersen et al. 2006) and PEPITO (Sweredoski and Baldi 2008).

DiscTope (<http://www.cbs.dtu.dk/service>) is the first online server created just for the prediction of non-contiguous B-cell epitopes (Haste Andersen et al. 2006). The algorithm is currently amassing a dataset that includes the crystal structures of known antigen–antibody complex object data, conformational epitopes characterized by extensive statistical measurements, and analyses of the spatial characteristics and calculations related to the potential for surface location all of which contribute to the prediction of noncontiguous B-cell epitopes. DiscTope is trained on a compiled dataset of noncontiguous epitopes from 76 antigen–antibody complexes with X-ray diffraction structures. The DiscTope program allows for adjustments to the epitope threshold value in order to adjust the specificity and sensitivity of the prediction. DiscTope can detect 15.5 % of residues located in discontinuous epitopes with a specificity of 95 %, while the other linear epitope prediction method can only detect about 11 % of the residues with the same 95 % specificity. When software developers set a specificity of 90 % and a sensitivity of 24 % to predict the discontinuous epitopes of malaria protein apical membrane antigen 1 (AMA1), the results showed 43 of 311 residues in AMA1 were predicted as epitope residues.

14.3.3 Phage Display Method

After years of development and improvement, phage display technology provides an easy method for epitope research with high throughput, low costs, and a practical platform. The phage display prediction method involves finding mimetic peptides with antibody affinity obtained by antibody screening phage display peptide libraries, using them to obtain probe motifs, and then searching for the best matching sequence on the surface of the antigen's tertiary structure. Typical software programs include Findmap (Mumey et al. 2003), SiteLight (Halperin et al. 2003), 3DEX (Schreiber et al. 2005), Mapitope (Bublil et al. 2007), PepSurf (Mayrose

et al. 2007b), Pep-3D-Search (Huang et al. 2008), MIMOX (Huang et al. 2006), MIMOP (Moreau et al. 2006), and Pepitope (Mayrose et al. 2007a).

The Pep-3D-Search (<http://kyc.nenu.edu.cn/Pep3DSearch/>) program is based on mimotope analysis to predict conformational B-cell epitopes (Huang et al. 2008). The algorithm flow is as follows: First, input a three-dimensional structure of the antigen (a PDB file) and a set of mimotopes or a motif. Pep-3D-Search identifies all surface residues of the query antigen and creates a graph of surface residues. Second, either of two modes can be used: the mimotope mode and motif mode. In the mimotope mode, the ACO (ant colony optimization) algorithm is employed to search for matching paths on the surface of the antigen. Each matching path is scored to the corresponding mimotope. Finally, a *P*-value calculation algorithm and DFS (depth-first search) algorithm are used to screen and cluster these paths and the candidate epitopes are revealed. In the motif mode, the motifs are located on the antigen surface using the ACO algorithm and the paths with a high score are directly output as candidate epitopes.

Phage display peptide library technology combined with computer modeling for the prediction of conformational epitopes has a more solid experimental foundation and therefore, predictions are more reliable. However, its limitation is obvious. It can only predict the epitopes of antigens for which a monoclonal antibody is available, thus, it is difficult to carry out more extensive applications for epitope prediction.

14.3.4 Protein–Protein Interface Method

The interaction between antigen and antibody is a type of protein–protein interaction, therefore, some prediction tools based on protein–protein interactions can also be used to predict B-cell epitopes. The predictions using these methods depend on the identification of the characteristics of amino acid residues in the protein complex interface. In recent years, there has been a rapid growth of data on the three-dimensional structures of proteins, and a number of protein–protein interface databases have been created for use. These databases can be used to analyze the characteristics of amino acid residues in the protein–protein interface, the features of sequence conservation, secondary structure, solvent accessibility, and side-chain conformational entropy that all make a prominent contribution in epitope prediction. The recently created servers for predicting protein–protein interfaces include the following: ClusPro (Comeau et al. 2004) (<http://structure.bu.edu>) and PatchDock (Schneidman-Duhovn et al. 2003) (<http://bioinfo3d.cs.tau.ac.il/sources.html>), both of which are based on protein–protein docking; PPI-PRED (Bradford and Westhead 2005) (<http://bioinformatics.leeds.ac.uk/ppi-pred>), which combines a support vector machine approach with surface patch analysis to predict protein–protein binding sites; PIER (Kufareva et al. 2007) (<http://abagyan.ucsd.edu/PIER/>), which is based on local statistical properties of the protein surface derived at the level of atomic groups for predicting interfaces.

Prediction of conformational B-cell epitopes is generally required to determine the tertiary structure of the antigen and computer programs are able to predict the tertiary structures for those antigens with unknown tertiary structures. The currently published tools for the prediction of conformational B-cell epitopes have some limitations, but computer prediction programs combined with biological experiments, or integrated with the induced fit mechanism of antigen–antibody recognition, will all contribute to improve the accuracy of future successful epitope prediction.

14.4 Prediction of T-Cell Epitopes

T cells usually do not directly recognize the epitopes of native protein antigens. In order for the antigen to be recognized by the TCR, it must first be processed by an antigen-presenting cell (APC). First, after entering the cell by endocytosis, the native antigen is degraded into antigen peptides that then bind to the MHC molecule and form a peptide-MHC complex, which expressed on the surface of the APC and recognized by TCR. Thus, the primary goal in predicting T-cell epitopes is to determine the binding affinity of antigenic peptides and MHC molecules. There are two types of MHC molecules involved in antigen presentation, class I and class II. MHC-I molecules bind an endogenous antigen peptide that is then expressed on the surface of CD8⁺ cytotoxic T cells for recognition by TCR. In contrast, MHC-II molecules bind an exogenous antigen for recognition by the TCR on CD4⁺ helper T cells. Thus, bioinformatics methods to predict T-cell epitopes include predictions for both cytotoxic T-cell (CTL) epitopes and helper T-cell (T_H) epitopes.

In the adaptive immune system, the binding of MHC-II molecules to antigenic peptides plays an important role in the immune defense response. Exogenous proteins are enclosed in an endosome after entering the APC by endocytosis. This endosome merges with the primary lysosome to form a phagolysosome, where the exogenous proteins are degraded into peptides. The processed peptides bind to MHC-II molecules and form complexes that are transported to the APC surface for recognition by T helper cells. Since allergens are generally exogenous antigens, the main way to predict allergenic T-cell epitopes is to predict the binding affinity of the antigenic peptide with MHC-II molecules, but this approach is not able to predict the binding affinity of the antigen peptide with TCR.

An empty MHC molecule contains a groove that the peptides fit into. The ends of the groove of MHC-I molecules are closed, so a protein fragment must have a relatively fixed length to fit this groove. However, unlike MHC-I molecules, the ends of the groove of MHC-II molecules are open, allowing for some extension of the amino terminal and carboxyl terminal of the peptide (Nielsen et al. 2010b). This increases the complexity of predicting peptide binding to MHC-II molecules, where changes in the length of T_H-cell epitopes are large, usually composed of 9–25 amino acid residues, and sometimes more than 30 amino acids residues.

In addition, genes encoding MHC molecules are extremely diverse. Human MHC molecules are termed human leukocyte antigen (HLA), and the HLA gene is located on human chromosome VI. The HLA class I genotype mainly includes HLA-A, HLA-B, and HLA-C. The HLA class II genotype mainly includes HLA-DP, HLA-DQ, and HLA-DR. So far, more than 3000 HLA II alleles have been found. When analyzing the MHC binding affinity with antigen peptide to predict T-cell epitopes, sufficient consideration should be given to the polymorphisms and geographical distribution differences of HLA alleles.

14.4.1 Modeling Methods and Software

Since the late 1990s, researchers have put a lot of effort into computer-based methods for predicting T-cell epitopes, and there have been many prediction programs based on bioinformatics, so choosing a prediction tool with the highest accuracy is very important. Li et al. (2010) selected 309 known epitope peptides, and comprehensively evaluated the accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of seven popular epitope prediction software programs that include ARB Matrix(Bui et al. 2005), MHCPre (Guan et al. 2003), NetMHCII (Nielsen et al. 2007), NetMHCIIpan (Nielsen et al. 2010a), Propred (Singh and Raghava 2001), RANKPEP(Reche et al. 2004), and SYFPEITHI (Rammensee et al. 1999). The results show the existence of differences in the reliability of the prediction methods, of which NetMHCIIpan and NetMHCII have the highest accuracy (see Table 14.1). We can combine the results of these two software programs to determine the epitopes with relatively high reliability, which can provide important information for epitope modification and vaccine design.

Table 14.1 Statistical results of epitope prediction software (Li et al. 2010)

Software	Accuracy	Sensitivity	Specificity	Positive predictive values	Negative predictive values
ARB	0.77	0.80	0.71	0.87	0.37
RANKPEPE	0.74	0.68	0.89	0.94	0.34
SYFPET	0.84	0.86	0.84	0.93	0.40
Propred	0.81	0.75	0.96	0.98	0.37
NetMHCIIpan	0.88	0.83	0.85	0.93	0.40
NetMHCII	0.88	0.83	0.87	0.94	0.40
MHCPre	0.61	0.54	0.05	0.76	0.11

According to the different modeling methods that are based on peptide sequences, the prediction methods can be divided into the binding motif method, quantitative matrix method and machine learning method as described below:

- **Binding motif method**

This early modeling method is based on binding motifs, and its principle is that when a linear polypeptide binds to MHC molecules, only few amino acids contribute to the binding with their side chains and the tightest binding occurs at conserved amino acids located at both ends of the peptide. This binding site is called an “anchor site” and the corresponding amino acid residues at anchor sites are called “anchor residues.” Some of these anchor residues together form a fixed pattern, commonly referred to as a “motif.” This modeling method works by evaluating whether the peptide fragment has known motifs in order to determine whether there will be binding between peptide and MHC-II molecule. Thus, the more verified MHC-II binding motifs there are, the better the accuracy of T_H-cell epitope prediction.

SYFPEITHI (<http://www.syfpeithi.de/>) is a representative motif modeling method. Its database is based on T-cell epitopes and MHC ligands that have been verified in the literature and covers 7000 peptides that bind to MHC-I and MHC-II molecules (Rammensee et al. 1999). Each peptide has been cited in the literature and is directly linked to the EMBL and PubMed databases. Thus, the number of peptides is constantly updated. This prediction method is based on published motifs (pool sequencing and natural ligands) and takes into account the amino acids at anchor positions and auxiliary anchor positions, as well as other frequently occurring amino acids. A score for the amino acid residue is calculated from a set of criteria that looks at whether or not they are anchor residues, auxiliary anchor residues or preferred residues. Ideal anchor residues are given 10 points, unusual anchor residues 6–8 points, auxiliary anchor residues 4–6 and preferred residues 1–4 points. Amino acid residues that are considered to have a negative effect on the binding ability is given a score between –1 and –3. Recent studies on anchor residues have shown that, due to great variability in the binding groove of MHC-II molecules, the restriction of anchor residues is not as strong as that of MHC-I. In addition, some peptides that bind to MHC-II molecules are not required to exactly match the anchor site. Thus, using binding motif methods to predict T-cell epitopes is not very accurate and the reliability of SYFPEITHI prediction method is only about 50 %.

- **Quantitative matrix method**

The development of quantitative matrix modeling methods to identify MHC-II binding peptides heralded a new stage in epitope prediction. The principle of quantitative matrix modeling is that the contribution of each amino acid residue in a peptide to binding to MHC-II molecules is independent, and the binding values can be added linearly. In the quantitative matrix, a protein sequence is fragmented into 9-mer peptides in a sliding window, each amino acid in the 9-mer peptides is assigned a binding value based on the type of amino acid and its position, and

then the values of each amino acid in the 9-mer peptide are added to obtain a total score. This score is compared with a threshold score to determine the likelihood that a peptide will bind to MHC molecules. In general, the prediction accuracy of the quantitative matrix approach is higher than the binding motif method.

TEPITOPE (<http://www.imtech.res.in/raghava/propred/>) is a typical quantitative matrix method that has been widely used to study a variety of subunit vaccines (Sturniolo et al. 1999). The MHC-II type alleles have enormous polymorphisms (there are more than 690 kinds of known HLA-DR alleles) to allow for binding to different peptides, so prediction of peptide molecule binding to MHC-II is complex. TEPITOPE covers more than 50 HLA-DR alleles. Using TEPITOPE as its foundation, Raghava et al. established the online server Propred that has an improved graphical display so that a variety of specific HLA-DR binding peptides can be displayed overlapping with each other at a glance, and these overlapping regions may be the heterozygous T-cell epitopes.

• Machine learning method

The machine learning method is an advanced method developed in recent years to predict helper T-cell epitopes and overcomes the disadvantages of the motif method and the matrix method. It can reveal the complicated nonlinear relationship in dataset with high specificity, accuracy, and adaptability, so it has been widely used. The main mode of the machine learning method is to first use machine learning approaches to determine the binding center, and then to establish prediction model based on the recognized binding center. A typical modeling method of machine learning was proposed in 1998 (Brusic et al. 1998), which was a MHC-II binding peptides prediction method based on an evolutionary algorithm (EA) and an artificial neural network (ANN). The main steps of this type of modeling are: (1) Binding data extraction—peptide sequences and their binding affinities are collected from a variety of sources; (2) Peptide alignment—an EA is used to generate alignment matrices that are then used to find and align putative nonamer cores of known binders; (3) ANN training and classification—the ANN training set comprising aligned binding nonamer cores and nonbinding nonamers is used to train ANNs to predict the binding affinity of query peptides. Machine learning algorithms include artificial neural networks, evolutionary algorithms, hidden Markov chains, Gibbs sampling, and support vector machines. NN-align and NetMHCIIpan are representative servers that are based on artificial neural networks to predict T-cell epitopes.

NetMHCIIpan 2.0 (<http://www.cbs.dtu.dk/services/NetMHCIIpan-2.0/>) is an improved pan-HLA-DR binding prediction method that uses concurrent alignments and a weight optimization training program (Nielsen et al. 2010a). This method is a pan-specific version of the earlier published allele-specific NN-align algorithm. It does not require pre-alignment of the input data, which means that it does not require a certain amount of data related to MHC molecules in the training data to generate an accurate result of specific alleles. Thus, this method is useful for making predictions for alleles with limited known binding data. Using large and diverse benchmark data to evaluate the predictive accuracy of NetMHCIIpan

2.0, the results showed that the method was significantly better than the earlier MHC-II prediction methods and was particularly improved for alleles with very limited binding data where the conventional methods often have poor prediction accuracy.

14.4.2 Application of Software Prediction

With advances in bioinformatics, a number of T-cell epitope prediction tools have been developed but the accuracy of these predictions must be experimentally verified. Ailin Tao et al. from Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology have worked extensively in this field. They used the NetMHCII 2.2 online server to predict the T_H-cell epitopes for nearly 20 kinds of representative allergens including mites (Der f 1, Der f 2, Der f 3, Lep d 2, and Sar s 14), cat (Fel d 1), house (Equ c 1), fungi (Mal f 1), *Parietaria judaica* (Par j 1), birch (Bet v 1), ryegrass (Lol p 5b), oysters (Cra g 1), mackerel (Sco j 1), soybean (Gly m 4), peanut (Ara h 2 and Ara h 8), lentils (Len c 1), and so on. Then, in order to reduce their allergenicity, the strong binding sites of the allergen were modified by site-directed mutagenesis PCR to attenuate their MHC-II binding affinity. Finally, the allergenicity was evaluated using a human dendritic cell model as well as with sensitized mice models to verify the prediction accuracy of the epitope modification.

References

- Alix A. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*. 1999;18(3–4):311–4.
- Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*. 2005;21(8):1487–94.
- Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*. 1998;14(2):121–30.
- Bublil EM, Freund NT, Mayrose I, Penn O, Roitburd-Berman A, Rubinstein ND, Pupko T, Gershoni JM. Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. *Proteins*. 2007;68(1):294–304.
- Bui H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton K, Mothe B, Chisari F, Watkins D, Sette A. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*. 2005;57(5):304–14.
- Chou P, Fasman G. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*. 1978;47:45–148.
- Colloc'h N, Cohen FE. Beta-breakers: an aperiodic secondary structure. *J Mol Biol*. 1991;221(2):603–13.
- Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*. 2004;20(1):45–50.

- Emini E, Schleif W, Colonna R, Wimmer E. Antigenic conservation and divergence between the viral-specific proteins of poliovirus type 1 and various picornaviruses. *Virology*. 1985;140(1):13–20.
- Garnier J, Osguthorpe D, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*. 1978;120(1):97–120.
- Gershoni JM, Stern B, Denisova G. Combinatorial libraries, epitope structure and the prediction of protein conformations. *Immunol Today*. 1997;18(3):108–10.
- Guan P, Doytchinova IA, Zygouri C, Flower DR. MHCPre: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res*. 2003;31(13):3621–4.
- Halperin I, Wolfson H, Nussinov R. SiteLight: binding-site prediction using phage display libraries. *Protein Sci*. 2003;12(7):1344–59.
- Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci*. 2006;15(11):2558–67.
- Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *PNAS*. 1981;78(6):3824–8.
- Huang J, Gutteridge A, Honda W, Kanehisa M. MIMOX: a web tool for phage display based epitope mapping. *BMC Bioinform*. 2006;7:451.
- Huang Y, Bao Y, Guo S, Wang Y, Zhou C, Li Y. Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinform*. 2008;9:538.
- Jameson BA, Wolf H. The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput. Appl. Biosci*. 1988;4(1):181–6.
- Karplus P, Schulz G. Prediction of chain flexibility in proteins. *Naturwissenschaften*. 1985;72(4):212–3.
- Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. PIER: protein interface recognition for structural proteomics. *Proteins*. 2007;67(2):400–17.
- Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. *Nucleic acids research*. 2005;33(Web Server issue):W168–71.
- Kyte J, Doolittle R. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157(1):105–32.
- Li S, Liu C, Zou Z, Tao A. Comparative Assessment of Different Softwares for MHC - II Antigen Epitope Prediction (in Chinese with English Abstract). *J Biomed Eng Res*. 2010;29(2):128–32.
- Linehan LA, Warren WD, Thompson PA, Grusby MJ, Berton MT. STAT6 is required for IL-4-Induced germline Ig gene transcription and switch recombination. *J Immunol*. 1998;161(1):302–10.
- Mayrose I, Penn O, Erez E, Rubinstein ND, Shlomi T, Freund NT, Bublil EM, Ruppin E, Sharan R, Gershoni JM, et al. Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics*. 2007a;23(23):3244–6.
- Mayrose I, Shlomi T, Rubinstein ND, Gershoni JM, Ruppin E, Sharan R, Pupko T. Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res*. 2007b;35(1):69–78.
- Moreau V, Granier C, Villard S, Laune D, Molina F. Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics*. 2006;22(9):1088–95.
- Mumey BM, Bailey BW, Kirkpatrick B, Jesaitis AJ, Angel T, Dratz EA. A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins. *J Comput Biol J Comput Mol Cell Biol*. 2003;10(3–4):555–67.
- Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinform*. 2007;8:238.
- Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S. NetMHCIIpan-2.0—Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res*. 2010a;6:9.

- Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. *Immunology*. 2010b;130(3):319–28.
- Odorico M, Pellequer JL. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J Mol Recogn JMR*. 2003;16(1):20–2.
- 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):5425–32.
- Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, Peters B. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinform*. 2008;9:514.
- Rammensee H, Bachmann J, Emmerich N, Bachor O, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999;50(3–4):213–9.
- Reche P, Glutting J, Zhang H, Reinherz E. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*. 2004;56(6):405–19.
- Saha S, Raghava G. BcePred: prediction of continuous B-Cell epitopes in antigenic sequences using physico-chemical properties. *Lect Notes Comput Sci*. 2004;3239:197–204.
- Schneidman-Duhovni D, Inbar Y, Polak V, Shatsky M, Halperin I, Benyamini H, Barzilay A, Dror O, Haspel N, Nussinov R, et al. Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins*. 2003;52(1):107–12.
- Schreiber A, Humbert M, Benz A, Dietrich U. 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J Comput Chem*. 2005;26(9):879–87.
- Singh H, Raghava GPS. ProPred: prediction of HLA-DR binding sites. *Bioinformatics*. 2001;17(12):1236–7.
- Sturniolo T, Bono E, Ding J, Radrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti M, Sinigaglia F, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol*. 1999;17(6):555–61.
- Sweredoski MJ, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics*. 2008;24(12):1459–60.
- Welling G, Weijer W, van der Zee R, Welling-Wester S. Prediction of sequential antigenic regions in proteins. *FEBS Lett*. 1985;188(2):215–8.

Author's Biography



Ying He is an Assistant Researcher working at the Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology, The State Key Clinical Specialty in Allergy, The State Key Laboratory of Respiratory Disease, the Second Affiliated Hospital of Guangzhou Medical University. Email: heying0605@163.com.

Ying He received her Master's Degree from Jinan University with a major in Genetics in 2009 and working on the anti-fibrosis effect of the protein drugs FGFR2 and its mechanisms. Currently, she is working in allergy research. Her main research interests focus on the prediction of allergen epitopes using bioinformatics tools, the cloning and purification of recombinant allergens, and the evaluation of allergenicity and modification of recombinant proteins. She has participated in several research

programs of National Natural Science Foundation of China.



Ailin Tao is a Professor at Guangzhou Medical University, Director of Guangdong Provincial Key Laboratory of Allergy and Clinical Immunology, Principal Investigator of the State Key Laboratory of Respiratory Disease, Deputy Director of the State Key Clinical Specialty in Allergy of the Second Affiliated Hospital of Guangzhou Medical University, Member of the State Committee for Transgenic Safety Assessment, Standing Committee Member of Allergy Branch of Guangdong Medical Association, Member of Guangdong Provincial Committee for Transgenic Safety Assessment and Master Tutor.

Professor Ailin Tao earned his doctorate degree from the State Key Laboratory of Crop Genetic Improvement of Huazhong Agricultural University in 2002, followed by a post-doctoral training at Postdoctoral Station of Basic Medicine

in Shantou University Medical College, majoring in allergen proteins. His most recent research has been on allergy bioinformatics, allergy and clinical immunology, and disease models such as allergic asthma, allergic rhinitis, infection and inflammation induced by allergy, inflammatory, and protracted diseases caused by antigens or superantigens. He has gained experience in the field of allergology including the mechanisms of immune tolerance, allergy triggering factors, chronic inflammation pathways, and allergenicity evaluation and modification for food and drugs. He proposed some new concepts including "Representative Major Allergens," "Allergenicity Attenuation" of immunotoxin and allergens, "broad-spectrum immunomodulator" as well as the theoretical hypothesis of "Balanced Stimulation by Whole Antigens." Professor TAO's laboratory focuses on the diagnosis of allergic disease and the medical evaluation of food and drug allergenicity and its modification. Professor TAO has now constructed a system for the prediction, quantitative assessment, and simultaneous modification of epitope allergenicity, which has been applied to more than 20 allergens, and he also developed a bioinformatics software program for allergen epitope prediction, SORTALLER (<http://sortaller.gzhmu.edu.cn>), which performed significantly better than other existing software, reaching a perfect balance of high specificity (98.4 %) and sensitivity (98.6 %) for discriminating allergenic proteins from several independent datasets of protein sequences of diverse sources. Furthermore, this program has a Matthews correlation coefficient as high as 0.970, a fast running speed and can rapidly predict a set of amino acid sequences with a single click. The software has been frequently used by researchers from many institutions in China and over 30 countries worldwide, thus becoming the number one allergen epitope prediction software program. Professor TAO has set up an allergen database ALLERGENIA (<http://ALLERGENIA.gzhmu.edu.cn>) that has several advantages over other databases such as a wide selection of nonredundant allergens, excellent stringency and accuracy, and friendly and usable analytical functions.