

An Overview of Bioinformatics Tools and Resources in Allergy

Zhiyan Fu* and Jing Lin*

Abstract

The rapidly increasing number of characterized allergens has created huge demands for advanced information storage, retrieval, and analysis. Bioinformatics and machine learning approaches provide useful tools for the study of allergens and epitopes prediction, which greatly complement traditional laboratory techniques. The specific applications mainly include identification of B- and T-cell epitopes, and assessment of allergenicity and cross-reactivity. In order to facilitate the work of clinical and basic researchers who are not familiar with bioinformatics, we review in this chapter the most important databases, bioinformatic tools, and methods with relevance to the study of allergens.

Key words Food allergens, Allergy, Bioinformatics, Machine learning, Epitope prediction

1 Introduction

Allergy has become an emerging public health problem. It is an immunologically mediated adverse reaction, and the triggering agents are known as allergens that are otherwise not harmful. Allergic diseases are mainly mediated by immunoglobulin E (IgE) antibodies through the following mechanisms (Fig. 1): Exogenous allergens are processed by endosomal proteases in antigen presenting cells (APCs) to peptides/epitopes, which are presented to T cells by major histocompatibility complex (MHC) class II molecules on cell surface. CD4⁺ T helper cells are activated upon recognition of these peptide-MHC complexes by T-cell receptor (TCR) molecules and start the production and secretion of cytokines (i.e., interleukin-4 (IL-4) and IL-13), which induce class switching and IgE production in stimulated B cells. Allergen-specific IgE (sIgE) antibodies play a central role in allergic reactions. Cross-linking of membrane-bound IgE antibodies by the specific allergens results in activation

*These authors contributed equally to this article.

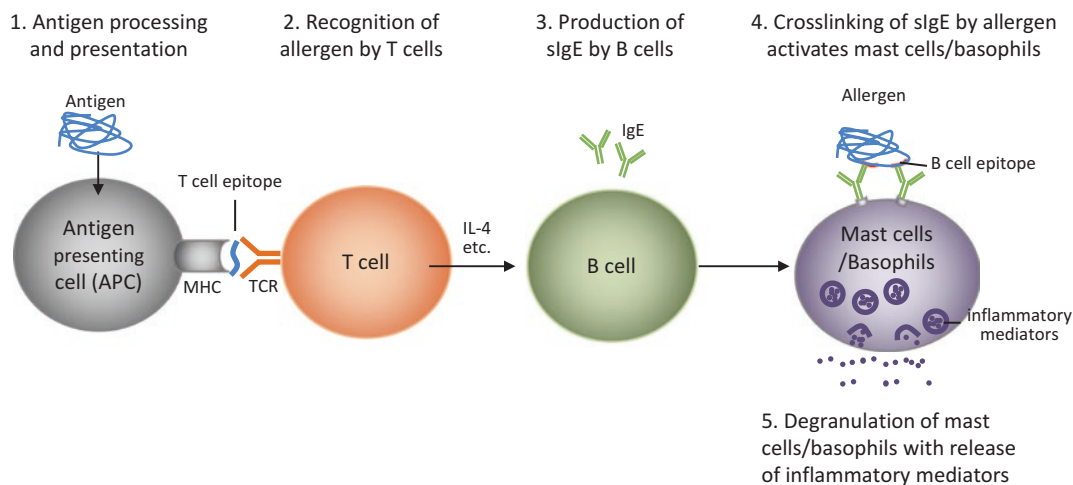


Fig. 1 A schematic diagram representing the IgE-mediated allergic reaction

and degranulation of effector cells (mast cells or basophils) with release of histamine and other inflammatory mediators, which together lead to the symptoms of allergic diseases [1]. In addition to IgE-mediated allergic diseases, other types of allergy also exist. For example, delayed allergic reactions to certain small therapeutic drugs (e.g., abacavir, nevirapine) are mediated by T cells and have been shown to be associated with specific MHC class I and/or II alleles [2].

Epitopes are the groups of amino acids within allergens that bind to antigen-specific receptors on lymphocytes or to secreted antibodies. There are two types of epitopes, continuous (or linear) and discontinuous (or conformational) in allergens. Continuous epitopes are short linear peptide fragments that are usually sequences of contiguous amino acids, while conformational epitopes are comprised of amino acids that line up because of the tertiary structure of an allergen. Studies of epitopes are important to both clinical and basic researchers in allergy or immunology in general as they hold huge potential for vaccine design, disease prevention, diagnosis, and treatment. For example, the current trends in allergy immunotherapy are to modify B-cell epitopes to prevent IgE binding, while preserving T-cell epitopes to retain the capacity for immunotherapy, or peptide-based immunotherapy containing major T-cell epitopes of the relevant allergen as an alternative strategy [3].

Epitopes can be categorized into B- and T-cell epitopes based on their respective receptors. B-cell epitopes are also known as IgE epitopes and can be either linear or conformational. The majority of T-cell epitopes are linear because antigens must be processed into peptide fragments or epitopes by proteasomes prior to MHC binding and presentation to Th cells.

Traditionally, studies of food allergens and allergenic epitopes depend exclusively on biochemical and immunological experiments, such as ELISA, protein/peptide microarray [4, 5], immunofluorescence, radioimmunoassay, Western blotting, immunohistochemistry, molecular cloning, NMR, and X-ray crystallography studies of antibody/antigen structure, most of which are expensive and time-consuming. In the post-genome era where massive amount of data of protein sequences and structures are available online, bioinformatics, which combine biomedical and computer science together, have progressed and developed vast varieties of methods and tools. These bioinformatics tools are now widely used for sequence analyses [6–8], protein structure modeling, and function annotation, reducing the time and effort spending on wet lab experiments [9].

In the last few years, machine learning technique, one of the computer science branches, advanced quickly and was widely applied in bioinformatics area [10]. For example, in addition to the traditional heuristic algorithms such as BLAST and FASTA used to detect sequence homology in large databases, better accuracy has been achieved using machine learning algorithms, such as support vector machine (SVM). An extensive number of customized kernels (functions that quantify the similarity between sequences) have been proposed and are more effective than general-purpose kernels (e.g., polynomial and radial basis function kernels) [11]. Classifiers developed to explicitly optimize the area under ROC curve (AUC) have been shown to be more useful than to optimize the classification error when the AUC is the performance metric of interest [12], and ensemble learning methods (e.g., boosting and bagging) have been shown effective in improving the performance of a single learning algorithm [13].

Now, with the aid of bioinformatics and machine learning, we are able to narrow protein of our interest, predict epitope using computation, increase allergy diagnostic accuracy, and drastically reduce the number of wet lab experiments. For example, we have applied bioinformatics tools in combination with peptide microarray experiment to identify epitope biomarkers that are capable of predicting clinical allergy with over 90% accuracy [14].

Currently, many bioinformatics resources with user-friendly web interface have been developed, which can facilitate the work of clinical and basic researchers who are not familiar with bioinformatics. However, the average researchers, who have no much background in computer science, can be easily overwhelmed by such a broad array of bioinformatics tools that are highly specific in use, not well understood or defined, and tested on limited data sets [15]. Our goal in writing this review is to introduce the essential updated bioinformatics resources in allergy field to clinicians and allergists and help them make informed selection among these tools.

2 Bioinformatic Prediction Methods

Current bioinformatics research in allergy field are primarily focused on assessment of protein allergenic potential, and prediction of B- and T-cell epitopes. Predicting the allergenic potential of a protein, particularly in novel foods, focuses on two main aspects: inherent protein allergenicity and cross-reactivity [16]. Inherent allergenicity refers to the potential of a protein to induce T-cell response and production of specific IgE antibodies, whereas cross-reactivity refers to the reactivity of a protein with pre-existing IgE antibodies that were previously induced by another allergen. The molecular basis of allergenicity of a protein can be studied through the analysis of its sequence, structure, and B- or T-cell epitopes. Online allergen databases are being used as reference for finding the sequence similarity for the purpose of allergenicity evaluation [17]. As allergen-specific IgE antibodies are the main effector molecules inducing allergic symptoms, prediction of allergen is closely related and often involves IgE epitope prediction. For example, in some of the allergen prediction methods, a protein is considered an allergen if it has a region or peptides identical to a known IgE epitope [18].

Allergen/epitope prediction dates back to 1981 when the first B-cell epitope prediction method was developed by Hopp and Woods [19]. Since then many methods have been developed or adapted from other computational tools. Despite the early start, however, prediction for B-cell epitopes is still in their infancy due to the difficulty in predicting discontinuous epitopes. For T-cell epitope prediction, MHC-I-binding predictions are now very strong by integration with predictions of proteasomal cleavage and TAP (transporter associated with antigen processing) binding sites. MHC-II-binding predictions are not as well developed as MHC-I-binding predictions, but are progressing at a fast pace. Based on the type of information being used, prediction methods can be categorized into sequence-based methods, structure-based methods, and hybrid methods.

2.1 *Sequence-Based Prediction*

Sequence-based method is based on the concept that protein sequence determines its structure and identical structures may lead to identical functions. Epitopes may have a common sequence pattern or motif, as well as MHC allele specificity determining patterns. BLAST and FASTA are common sequence–sequence comparison methods used to find sequence homologues and assign amino acid scores by multiple alignment of query protein sequence with template sequences in databases. However, experiments have shown that proteins with high sequence identity may possess differential functions due to folding, interactions among variable regions in different cellular compartmentization [20]. The methods using only sequence information are limited to the prediction of continuous epitopes.

Physico-chemical properties are also used to predict epitopes, such as charge, hydrophilicity, flexibility, number of proline residues, the distance/proximity to the N or C terminus, etc. Due to the enormous number of physico-chemical properties that may be associated with epitopes, simple quantitative descriptors of amino acid properties can be used for computation. The classical amino acid scale-based methods of identifying potential linear epitopes from sequences typically rely on the use of amino acid propensity scale that is to compute the scores of a residue i in a given protein sequence. The $i - (n - 1)/2$ neighboring residues on each side of the residue i are used to compute the score for residue i in a window size of n amino acids. The final score for residue i is the average of the scale values for n amino acids in the window. Pellequer [21] has compared several propensity scale methods using a dataset of 14 epitope-annotated proteins and found that the classical scores developed by Parker et al. [22], Chou and Fasman [23], Levitt [24], and Emini [25] provide better results than the other scores tested [26].

2.2 Structure-Based Prediction

Structure-based epitope prediction was first used by Kolaskaret al. in 1999 [27]. It is based on the 3D protein structural similarity between query protein and template proteins, and it also considers the related propensity scales, including geometric attributes and structure-based physico-chemical properties. With the development of various omics and bioinformatics, experimental data of conformational B-cell epitopes has been increasing rapidly in recent years. Current structure-based epitope identifying techniques mainly include mutagenesis, competition experiments, free energy scoring function, quantitative structure–activity relationship (QSAR) analysis [28], protein threading [29], rigid/flexible docking [30], and homology modeling [31]. Atomistic molecular dynamics simulations have also been applied but not broadly used as it is computational intensive.

The structure-based epitope prediction approaches generally give more accurate prediction due to their several advantages over the sequence-based one. For example, discontinuous epitopes can only be predicted based on structure. In addition, as protein functions can be predicted by its structure, the structure-based prediction can be used to predict epitopes for alleles/genes that have not been extensively studied, where sequence-based approaches have failed. However, the structure-based approach has been greatly restricted due to high computational cost, development complexity, and limited availability of protein 3D structures.

2.3 Hybrid Prediction: Combining Sequential with Structural Information

Due to the limitations of epitope predictors based only on either sequence or structure analysis, predictive methods taking advantage of both sequential and structural information have been developed. For example, a method that integrates 3D protein structure with properties of amino acids using Hidden Markov Model (HMM) improved the prediction precision to a significant degree [26].

In addition to the above three categories, prediction methods can also be categorized based on the techniques or algorithms, such as binding motifs [32], quantitative matrices (QM) [33], machine learning algorithms [34], linear programming [35], etc. They all have their relative advantages and disadvantages [36]. For example, in a comparative study, Yu et al. found that methods based on binding motifs gave the most accurate MHC-binding peptides predictions (potential T-cell epitopes) with a limited dataset, but as the data volume increased, machine learning predictions considering both structure and sequence information became more reliable [37]. Among machine learning methods, Bahsin and Ragahava have compared SVM (support vector machine) and ANN (artificial neural network) techniques, and found that SVM techniques performed better [38]. SVM is particularly attractive to biological analysis, due to its ability to filter noise, large input spaces, and good performance [10].

Ensemble methods (also known as consensus methods) combine the predictions of several predictors and often outperform individual predictors in many biomolecular sequence and structure classification studies. Several strategies for combining a set of predictors, S , into a single consensus or meta-predictor exist: (1) majority voting, (2) weighted linear combination, and (3) meta-learning [39]. A large number of nearest neighbor- and decision tree-based classifiers are trained using different sets of training data features for developing an ensemble of linear B-cell epitope classifiers [40].

2.4 Performance Evaluation of Prediction Methods

To evaluate and compare the performance of various prediction methods and software, ROC analysis is commonly used. An ROC analysis includes positive and negative predictive values (PPV, NPV), area under the curve (AUC), specificity, sensitivity, and accuracy. A ROC curve is obtained by plotting true positives on the y -axis against false positives on the x -axis. AUC that is the area under the rectangular hyperbola ROC curve represents the accuracy of prediction. AUC equals 0.5 (ROC plotted as a straight diagonal line) means the test is no better than random chance alone, while AUC equals 1.0 means the test has perfect accuracy. In order to make evaluation more accurate, cross-validation method is usually used [14]. Cross validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. Usually, multiple rounds of cross-validation are performed using different partitions and the evaluation (i.e., AUC) obtained from cross validation is the average over rounds.

3 Allergen Prediction Tools and Resources

According to the guideline proposed by World Health Organization (WHO) and Food and Agriculture Organization (FAO) for evaluating protein allergenicity (FAO/WHO 2001) [41], a protein that has at least six to eight consecutive identical amino acids or >35% identity within any window of 80 amino acids when compared with any known allergens is considered an allergen. This guideline has been widely accepted for allergen prediction and implemented into some web tools, such as SDAP [42]. However, prediction methods based on the guideline are sequence-based and tend to produce results with high sensitivity but very low specificity. It has been reported that only 1 of 200 “positive matches” represents a true allergen when using the FAO/WHO guidelines [43].

Several allergenicity prediction methods have been developed which showed improvement over the FAO/WHO guideline. For example, Stadler developed a motif-based method to achieve higher specificity by generating a minimal set of motif sequences from a database of known allergens. A protein was classified as a potential allergen if it contained at least one of the motifs [43]. Bjorklund et al. compared the sequence of >570 allergens and 700 non-allergens and identified a dataset of allergen-representative peptides (ARP). Their data showed that the ARP-based allergen prediction method outperformed the methods based on the FAO/WHO guideline [44]. Epitope-based methods have also been reported which are based on similarity search against IgE epitope database [18]. ARP and epitope-based methods are restricted by the number of known epitopes and allergenic domains. More recently, Wang et al. [45] have tried to integrate the FAO/WHO guideline-based method with motif-based and SVM-based methods and achieve >90% accuracy in allergen prediction.

Online allergen sequence databases are essential for evaluating the allergenic potential of protein as they not only serve as reference for finding the sequence/structure similarity with known allergens, but also provide useful web tools for structural and physio-chemical property analysis. Here, we will briefly introduce a few important and currently online available allergen prediction databases and tools. These resources are summarized in Table 1.

The allergen nomenclature database (www.allergen.org) is the official site for the systematic allergen nomenclature that is approved by the World Health Organization and International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature Sub-committee and has been constantly updated [46].

The Allergome website (www.allergome.org) [47] has been designed to supply information on allergens. The primary data source

Table 1
Resources for allergen prediction

Names	URLs
Database of IUIS	http://www.allergen.org
Database of Allergome	http://www.allergome.org
SDAP	http://www.fermi.utmb.edu/SDAP/
AlgPred	http://www.imtech.res.in/raghava/algpred
proAP	http://gmobl.sjtu.edu.cn/proAP/main.html

for Allergome is the literature published since the early 1960s. It has a total of 2484 allergen sources and 33,097 uploaded bibliographic references (updated in March 2016). It contains allergens that are not in the IUIS list but these allergens are added after careful characterization under stringent criteria. The Allergome platform has developed several modules aiming to support specific purposes. For example, “Real Time Monitoring of IgE sensitization” (ReTiME) module allows uploading of raw data from both in vivo and in vitro experiments.

SDAP (Structural database of Allergenic Proteins) (<http://fermi.utmb.edu/SDAP/>) [42] is a web server that integrates a database of allergenic proteins with various computational tools that can assist structural biology studies related to allergens. In its latest update, it has 1526 allergens and isoallergens. SDAP provides links to retrieve related information from the most common databases of protein sequences and structures (e.g., SwissProt, PIR, NCBI, PDB), and tools to assess allergenic potential of proteins using the FAO/WHO guideline.

Two web tools, AlgPred [18] and proAP [45], allow allergen prediction using several different approaches, such as motif-based using MEME/MAST [43], ARP-based, or SVM-based methods, or a hybrid option to predict allergen using combined approach (SVMc + IgE epitope + ARPs BLAST + MAST).

One problem facing the development of allergenicity prediction methods is the lack of a standard dataset for training and testing, including a good negative dataset of non-allergens. Allergen-related data are currently scattered across various publicly accessible databases. The lack of a standard dataset has made comparison between the various methods difficult and is hindering progress.

4 Epitope Prediction Tools and Resources

4.1 B-Cell Epitopes

B-cell epitopes are either continuous/linear or discontinuous/conformational [26]. So far various kinds of B-cell epitope sequences and structures have been identified. However, experiments

designed to identify B-cell epitopes are mostly based on linear epitopes [4, 5]. Because many of the B-cell epitopes are discontinuous, prediction of which requires 3D protein structure and notoriously difficult to predict, their predictive software and databases are generally with lower accuracy and based on a smaller scale of dataset. Currently, the top performing B-cell epitope prediction methods have average AUC values ranging between 0.6 and 0.7, depending on the evaluation dataset used [48–50]. Here, we will briefly introduce the most common software/tools and databases on B-cell epitopes (summarized in Table 2), all of which are currently available online.

4.1.1 Conformational B-Cell Epitope Prediction Software

As proteins can be recognized by B cells in their native forms, the input for conformational/discontinuous B-cell epitope prediction software is usually 3D protein structures and the output can be viewed graphically as 3D structures. Anderson et al. presented a method called DiscoTope [51], which used a combination of amino acid statistics, spatial information, and surface exposure, and could detect 15.5% of residues located in discontinuous epitopes with a specificity of 95%.

BEPro, formerly known as PEPITO, was initially developed as an alternative method to Discotope. Similar to Discotope, BEPro combines propensity scales with surface exposure information. BEPro utilizes the Discotope amino acid propensity scale in its algorithm, along with side-chain orientation and solvent accessibility data [48].

One of the most complete tools in this field is ElliPro [31], which is available on the IEDB analysis resource website. This tool predicts both linear and discontinuous epitopes based on protein's 3-D structure. ElliPro associates each predicted epitope with a structural related score, defined as a PI (Protrusion Index) value. In ElliPro, the input is protein sequence. Based on the structural template selected by the user, a 3-D structure will be predicted for the input protein sequence by homology modeling, and linear and discontinuous epitopes will be computed based on the predicted structure.

Three web servers SEPPA [52], Epitopia [49], and EPCES [53] were published in 2009. SEPPA employs two concepts, “unit patch of residue triangle” and “clustering coefficient” to describe the local spatial context and spatial compactness, and combines them to predict epitopes. Epitopia uses several structural and physiochemical features to represent the surface patch and applies machine-learning algorithm (Naive Bayes classifier) for epitope prediction. EPCES introduced a consensus scoring method for epitope prediction based on several structural and physicochemical features. In 2011, two server applications EPSVR and EPMeta were presented by the same author of EPCES [54]. EPSVR adopted the SVM regression method to integrate the scoring terms as EPCES, while EPMeta is a meta-server that combines EPSVR, EPCES, Epitopia, SEPPA, PEPITO, and Discotope1.2.

Table 2
List of B-cell epitope prediction servers

Method	Online service (program)	Predict discontinuous/ continuous epitopes	Brief notes (features used in prediction)
DiscoTope	http://www.cbs.dtu.dk/services/DiscoTope/	Discontinuous	Amino acid statistics, spatial context, and surface accessibility
ElliPro	http://tools.iedb.org/elliPro/	Both	Prominent index
PEPITO/ BEPro	http://pepito.proteomics.ics.uci.edu/	Discontinuous	Half sphere exposure values
SEPPA	http://lifecenter.sgst.cn/seppa/index.php	Discontinuous	Unit patch of residue triangle
Epitopia	http://epitopia.tau.ac.il/	Both	Based on Naive Bayes classifier with physicochemical and structural geometrical properties
EPCES	http://sysbio.unl.edu/services/EPCES/	Discontinuous	Consensus score by six functions Shinji Soga Not stated Antibody-specific epitope propensity index
EPSVR &EPMeta	http://sysbio.unl.edu/services/	Discontinuous	Based on SVR and meta-analysis
MEPS	http://www.caspur.it/meps	Discontinuous	Surface mimicking peptides
MIMOX	http://immunet.cn/mimox	Discontinuous	The first free Mimotop-based web tool
Pepitope	http://pepitope.tau.ac.il	Both	The first method based on amino acid pairs
EpiSearch	http://curie.utmb.edu/episearch.html	Discontinuous	An automated sequence analysis based on sequence and 3D profiles
Pep-3D-Search	http://kyc.nenu.edu.cn/Pep3DSearch	Discontinuous	Ant colony optimization algorithm
MimoPro	http://informatics.nenu.edu.cn/MimoPro	Discontinuous	Based on both patch and graph searching
Bcepred	http://www.imtech.res.in/raghava/bcepred/	Continuous	Linear B-cell epitopes prediction by physico-chemical properties
BepiPred	http://www.cbs.dtu.dk/services/BepiPred/	Continuous	Linear B-cell epitope prediction
ABCPred	http://www.imtech.res.in/raghava/abcpred/	Continuous	Artificial neural network linear B-cell epitope predictor

The above B-cell epitopes prediction methods are based on the structure features of antigen. Most of them are available as web servers or can be downloaded as a standalone version (Table 2) with user-friendly interfaces. Researchers just need to enter the PDB ID (*see* Subheading 4.1.4) or upload the local file in PDB format, enter the Chain id for the protein chain of interest, and/or specify a threshold (may be changeable on the output view). Yao et al. [55] evaluated the performance of prediction methods available at that time and showed that overall the accuracy of EPMeta is the highest value by all conditions and methods.

4.1.2 Mimotope-Based Prediction Methods

One unique type of method to predict conformational B-cell epitope is Mimotope-based prediction, which is a combinatorial method that requires both a list of affinity-selected peptides and the protein 3D structure as input. To obtain the affinity-selected peptides, random peptides are generated and displayed on the surface of bacteriophages, and screened with immobilized antibodies (e.g., IgE from patients). The unbound phages are removed and the bound ones are then eluted, amplified, and used for further rounds of selection. After several rounds, phage clones displaying peptides with higher affinity binding to antibodies are highly enriched. These affinity-selected peptides are defined as mimotopes that have the similar functionality to and a high sequential similarity with the native epitope [56]. Mapping these mimotopes back to the source antigen can help in finding the native epitopes. Mimotope-based prediction methods are essential to map mimotopes back to the surface of a source antigen to locate the best alignment sequences and predict possible epitope regions. Currently available mimotope-based web servers (programs), algorithms, and brief notes are listed in Table 2.

These include MIMOP [56], which predicts linear and conformational epitopes based on MimAlign and MimCons methods, MIMOX [57], Pepitope [58] which uses Pepsurf [59] and Mapitope [60] methods, Pep-3D-Search [61], and MimoPro [62]. These methods can be classified into two categories: sequence-sequence alignment category and sequence-structure alignment category [63]. MimAlign method of MIMOP belongs to the sequence-sequence alignment category that aligns mimotope sequences with the antigen sequence. The other methods belong to the sequence-structure alignment category that aligns mimotope sequences and the antigen 3D structure. In the sequence-structure alignment category, there are motif-based methods (i.e., MIMOX, MimCons method of MIMOP), pairs-based methods (i.e., Mapitope), patch-based methods (i.e., EpiSearch) [64], and graph-based methods (i.e., Pep-3D-Search and Pepsurf). To improve the performances of existing programs, hybrid methods such as MimoPro [62] and meta-servers such as Pepitope [58] were also developed.

4.1.3 Continuous B-Cell Epitope Prediction Software

Continuous B-cell epitopes are easier to predict than discontinuous ones, and share many common characteristics with T-cell epitopes that are mostly continuous. The input format for continuous B-cell epitope prediction software can be either sequence or structure. Methods for predicting continuous B-cell epitopes are generally sequence-based methods, amino acid propensity scale-based methods, or machine-learning methods. Common predictive software on continuous B-cell epitope includes ABCpred, BcePred, BepiPred, and the prediction tool provided on IEDB.

ABCpred is based on standard feed-forward (FNN) and recurrent neural network (RNN) for predicting B-cell epitopes in an antigenic sequence with an estimated accuracy of 65.9% [65].

BcePred server predicts linear B-cell epitopes with 53–58% accuracy [66] based on the physico-chemical properties (i.e., hydrophilicity, flexibility/mobility, accessibility, polarity, exposed surface, and turns) or combination of properties. It was developed based on a dataset of 1029 B-cell epitopes obtained from Bcipep database and an equal number of non-epitopes obtained randomly from Swiss-Prot database.

Bepipred [67] predicts linear B-cell epitopes using a combination of machine learning technique-hidden Markov model (HMM), and one of the two propensity scale methods, Parker [22] and Levitt [24]. When tested on the validation data set constructed from three datasets of linear B-cell epitopes: annotated proteins from literature, AntiJendatabase [68], and Los Alamos HIV database (<http://www.hiv.lanl.gov>), it achieves ~60–66% of accuracy for continuous epitope prediction.

4.1.4 B-Cell Epitope Databases

Experimental data plays an essential role in B-cell epitope prediction. The 3D structure of allergens or antigen-antibody complexes can be found in the Protein Data Bank (PDB) database [69], which is a database for the 3D structural data of proteins and nucleic acids. Its data are typically obtained by X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy. One can search structure according to the PDB-id at the home page and then view or download the structure in several formats.

Bcipep [70] provides comprehensive information about experimentally verified B-cell epitopes and tools for mapping these epitopes on an antigen sequence.

IEDB database (immune epitope databases and analysis resources) [71] is the most commonly used and most authoritative database containing a large number of epitopes (237,460 peptidic epitopes as in May 2016), and displays relatively complete information about each epitope as well as cross linked to other databases. IEDB covers a wide range of epitopes, including both B- and T-cell epitopes, and for a spectrum of host organisms, including human. It can be browsed and searched by different categories, such as epitope type (linear, discontinuous, etc.), source (organism or antigen name),

or diseases (allergic, infectious, autoimmune, etc.). It predicts continuous B-cell epitopes with the option of various methods, including six amino acid property-based methods, Bepipred Linear Epitope Prediction, etc. For prediction of discontinuous B-cell epitopes, it also provides methods based on antigen structure, such as ElliPro and DiscoTope, which have been mentioned in earlier sections.

Conformational epitope database (CED) [72] provides a collection of conformational epitopes and related information (e.g., location, immunological property, corresponding antibody) of the epitope. All epitopes in the database have been manually curated from the published literature. Most epitopes in CED can also be viewed interactively in the context of their 3D structures. In addition, the entries are also hyperlinked to other databases such as Swiss-Prot, PDB, KEGG, and PubMed, providing wide background information.

For mimotope-based methods, it needs both the antigen structure and the sequence data of mimotopes. Since the 3D structure can be obtained from PDB or by computational homology modeling, the small number of mimotope sequences derived from phage display becomes a limitation for the development of mimotope-based conformational B-cell epitope prediction. In recent years, several databases have been released, which integrated the structure data, the mimotopes data, and other associated information. One such example is ASPD, a curated database that incorporated data on peptides and proteins mainly from phage display experiments [73]. The current version released in 2001 has 195 entries. ASPD has a user-friendly interface, and researchers can search the needed information by means of simple search system or SRS (Sequence Retrieval System).

PepBank is a database of peptides based on sequential text mining and public peptide data sources and it has incorporated the short peptide sequences of the ASPD database [74]. This database only stores peptides with available sequences and 20 amino acids or shorter in length. PepBank has a web-based user interface with a simple, Google-like search function, advanced text search, and BLAST and Smith-Waterman search capabilities.

BDB (Biopanning Data Bank, formerly known as MimoDB) is an information portal to biopanning results of random libraries [75]. It is the latest and largest database for mimotopes. In its latest release (April 2016), it has >25,000 peptides grouped into ~3000 sets and collected from >1300 papers. For each entry, the target, template, library, and structure information are provided. In addition, BDB provides tools for simple and advanced search, structure visualization, BLAST, and alignment view on the fly.

These databases are listed in Table 3. They are important resources for B-cell epitope prediction, and provide a basis for computational biologists to derive benchmark and customize

Table 3
Databases for B-cell epitopes

Databases	Websites
PDB	http://www.rcsb.org/pdb/home/home.do
BCIPEP	http://www.imtech.res.in/raghava/bcipep
CED	http://immunet.cn/ced/
IEDB	http://www.iedb.org/
ASPD	http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd
PEPBANK	http://pepbank.mgh.harvard.edu
BDB	http://immunet.cn/bdb

datasets for tool evaluation and new algorithm development. However, limitations exist due to limited number/types of epitope entries in databases, limited number of discontinuous epitope databases, incomplete epitope information, and rare integration of databases and epitope prediction software.

4.2 T-Cell Epitopes

A functional T-cell response requires binding of antigenic peptides with MHC class I or class II molecules (also known as HLA, human leukocyte antigen) and a proper interaction of the peptide-MHC complex with TCR on T cells. Therefore, currently T-cell epitopes are largely predicted indirectly by predicting peptide binding to MHC molecules.

MHC-binding prediction consists of MHC-I and II-binding prediction, which are distinct from each other. This is because the peptide-binding groove of MHC class I molecules is closed, while the corresponding groove on class II is open at both ends that can accommodate peptides of variable length, typically 13–25 amino acids long [76]. As a result, MHC-II-binding prediction has much lower accuracy rate than that of MHC I. T-cell epitope prediction methods and algorithms can be roughly grouped into five distinct types: quantitative matrix, machine learning techniques (ANN, SVM, etc.), MHC peptide threading, 3DQSAR (three-dimensional quantitative structure and activity relations), and molecular modeling [77]. Their performance in predicting T-cell epitopes is compared in Table 4. The number of MHC-binding predictive software is expanding rapidly and achieved impressive prediction accuracies of 70–99% sensitivity but only 40–80% specificity with limited range of MHC allele coverage and for peptide with fixed length [78]. A few commonly used T-cell epitope predictive software with publicly available web interfaces are listed in Table 5. Most of these software and databases are user friendly with straightforward output, providing online tutorials for those who are not familiar in this field to select the suitable tools.

Table 4
Comparison of T-cell epitope prediction methods

Method	Training dataset	Prediction dataset	Accuracy	AUC	Correlation coefficient
3D QSAR	A*0201	A*0201			0.679
EpiDock	A*0201	A*0201, B*2705	75–85%		
Threading	10 alleles		71%		–0.61 to 0.07
HMM-SMM	DRB1*0101	DRB1*0101		0.89	
ANN + QM	30 alleles		93.6%		
SVM	30 alleles	30 alleles	86–99%		

However, T-cell epitope prediction based only on MHC-binding prediction has its limitations. Binding to MHC is necessary but not sufficient for recognition by T cells. Many factors are involved in epitope recognition and not every peptide with high affinity to MHC proteins are epitopes. Therefore, combining the prediction results of multiple well-selected tools can achieve more accurate prediction than any single tool. For example, to predict T-cell epitopes from an antigenic protein sequence, an integrated prediction of proteasomal cleavage, TAP transport efficiency, and MHC class I binding affinity will yield better results than a single-prediction software [79]. A study by Trost et al. has shown that a consensus method combining the predictions of 16 MHC-I-binding peptide prediction servers performed better than any individual server [80].

4.2.1 Software for MHC-I-Binding Prediction

Although MHC class I molecules are not directly associated with antigen presentation to T helper cells and B-cell stimulation, bindings of certain drugs with specific MHC class I molecules have been shown to associate with drug allergy and prediction of these bindings can help with drug allergy prediction.

MHC I epitope prediction is quite reliable and currently provided in most software. Software exclusively focused on MHC I epitopes includes BIMAS, CTLpred, ProPred-1, various tools provided on IEDB, etc. BIMAS allows users to locate and rank potential 8- to 10-mer peptides based on a predicted half-time of dissociation to MHC class I molecules. The analysis is based on coefficient tables deduced from the paper published by Parker et al. [81]. BIMAS and SYFPEITHI [82] used to be two of the most commonly used T-cell epitope prediction tools.

MMBPred [83] server predicts mutated promiscuous and high-affinity MHC class I-binding peptide from protein sequence. It uses the matrix data in a linear prediction model, in which peptide conformation is ignored.

NetMHC server [84] uses artificial neural network (ANN) and weight matrices. It has been trained on data from 55 MHC

peptides (43 human and 12 nonhuman) and position-specific scoring matrices (PSSMs) for additional 67 HLA alleles.

Similar to NetMHC, NetMHCpan [85] is also based on neural-network method. It generates a quantitative prediction of the affinity of any peptide-MHC class I interaction, covering HLA-A and HLA-B for humans as well as chimpanzee, macaque, gorilla, cow, pig, and mouse, which very few databases can achieve. Both NetMHC (ANN) and NetMHCpan tools are currently provided by the IEDB Analysis Resource website for T-cell epitope prediction.

In addition to MHC I binder prediction, the web server EpiJen [77] has tried to model the entire process of the MHC I degradation pathway, and includes proteasome cleavage and TAP binding into its prediction algorithm. It is a multi-step algorithm for T-cell epitope prediction, which belongs to the next generation of in silico T-cell epitope identification methods. Similar to EpiJen, NetCTL [86], and WAPP (Whole Antigen Processing Pathway) [87] web servers, based on ANN and SVM, respectively, also integrate predictions of proteasomal cleavage, TAP transport efficiency, and MHC class I binding affinity for T-cell epitope prediction. NetCTLpan [79], a pan-specific MHC class I epitope predictor, is an improved and extended version of NetCTL.

4.2.2 Software for MHC-II-binding prediction

As mentioned earlier, MHC-II-binding peptides and epitopes are more difficult to predict than MHC I epitopes, owing to the inherent physical and chemical properties of MHC II molecules. A few web-based tools (listed in Table 5), built specifically for MHC-II-binding prediction, are currently available, such as ProPred, MHC2Pred, and MHC-II-binding tools provided by IEDB.

Propred [33] is a graphical web tool for predicting MHC class II binding regions in antigenic protein sequences, using quantitative matrices derived from the published literature by Sturniolo et al. [88]. MHC2Pred is a SVM-based method for the prediction of promiscuous MHC class II binder. The average accuracy of the SVM-based method for 42 alleles is 80%. The performance of the method was lower for some alleles due to a smaller size of dataset [33].

The MHC-II-binding prediction server on IEDB provides various tools for predicting MHC II binders. The input format can be space separated sequences, one continuous sequence, or in FASTA format. When choosing the prediction method, by default the consensus method that combines several other tools is selected, as it has the highest overall performance ranking than any single tool.

4.2.3 Software for MHC-I and II-Binding Prediction

Software that can be used to predict both MHC I and II binders are usually hybrid method, or have combined several methods. SVMHC [34] is a software based on SVM for both MHC I and II epitope prediction. The server can be used to find the most likely binders in a protein sequence and to investigate the effects

Table 5
List of T-cell epitope prediction servers

Server name	Link	Predictive server for	Predictive method
BIMAS	http://www-bimas.cit.nih.gov/molbio/hla_bind/	MHC I	Published coefficient tables
ProPred I	http://www.imtech.res.in/raghava/propred1/	MHC I	QM
MMBPred	http://www.imtech.res.in/raghava/mmbpred/	MHC I	QM
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	MHC I	ANN
NetMHCpan	http://www.cbs.dtu.dk/services/NetMHCpan-3.0/	MHC I	ANN
IEDB binding	http://tools.immuneepitope.org/analyze/html/mhc_processing.html	MHC I	SVM + ANN
EpiJen	http://www.ddg-pharmfac.net/cpijen/EpiJen/EpiJen.htm	MHC I	Multi-step algorithm
NetCTL	http://www.cbs.dtu.dk/services/NetCTL	MHC I &TAP	ANN-regression
NetCTLpan	http://www.cbs.dtu.dk/services/NetCTLpan/	MHC I &TAP	ANN-regression
MHCPred	http://www.ddg-pharmfac.net/mhcpred/MHCPred/	MHC I &II	Additive method
SYFPEITHI	http://www.syfpeithi.de/bin/MHCServer.dll/EpitopePrediction.htm	MHC I &II	Published motifs
RANKPEP	http://bio.dfci.harvard.edu/RANKPEP/	MHC I &II	PSSM
SVMHC	http://abi.inf.uni-tuebingen.de/Services/SVMHC	MHC I &II	SVM
ProPred	http://www.imtech.res.in/raghava/propred/	MHC II	QM
MHC2Pred	http://www.imtech.res.in/raghava/mhc2pred/	MHC II	SVM
CTLDPred	http://www.imtech.res.in/raghava/ctlpred/	MHC I &CTL	SVM + ANN
TAP Pred	http://www.imtech.res.in/raghava/tappred/	TAP	SVM
WAPP	http://abi.inf.uni-tuebingen.de/Services/WAPP/information	TAP	SVM
PAProC	http://www.paproc.de	Proteasomal cleavages	
FRAGPREDICT	http://www.mpiib-berlin.mpg.de/MADPP/cleavage.html	Proteasomal cleavages	
NetChop	http://www.cbs.dtu.dk/services/NetChop	Proteasomal cleavages	ANN
Pcleavage	http://www.imtech.res.in/raghava/pcleavage	Proteasomal cleavages	SVM

QM quantitative matrix, SVM support vector machine, ANN artificial neural network, PSSM position-specific scoring matrices

of SNPs (single nucleotide polymorphisms) in terms of MHC-peptide binding.

RANKPEP predicts peptide binders to MHC I and MHC II molecules from protein sequences or sequence alignments using Position Specific Scoring Matrices (PSSMs), which derived from a set of aligned peptides known to bind to a given MHC molecule. The binding potential of any peptide sequence (query) to the MHC molecule is determined by comparing the query to the PSSM. In addition, it also determines whether the C-terminus of the predicted MHC I peptide binders are the result of proteasomal cleavage [89].

MHCPred [90] uses the additive method to predict the binding affinity to MHC I and II molecules and also to the TAP. It is composed of a number of allele-specific Quantitative Structure Activity Relationship (QSAR) models created using PLS, a robust multivariate statistical method.

4.2.4 Software for Proteasomal Cleavage Sites Prediction

Predictions of proteasomal cleavage sites provide an alternative way for T-cell epitope prediction, although their prediction sensitivity is not very high. This is expected because proteasomal cleavage is only the first stage in antigen degradation and MHC presentation process, and the majority of peptides may not bind to MHC molecules during later stages. However, they may complement predictive software based on other mechanisms. A few commonly used software with publicly available web interfaces for proteasomal cleavage site prediction, including PProC, NetChop, FRAGPREDICT, and Pcleavage, are listed in Table 5.

PProC [91] is a prediction tool that predicts cleavages by human and yeast 20S proteasomes, based on experimental cleavage data. Pcleavage [92] is a SVM-based software for the prediction of constitutive and immunoproteasome cleavage sites in antigenic sequences. It uses SVM, Parallel Exemplarbased Learning (PEBLs), and Waikato Environment for Knowledge Analysis (Weka). Pcleavage and NetChop [93] that are based on neural network prediction are considered to be the best methods for proteasomal cleavage site prediction.

4.2.5 Software for TAP-Binding Prediction

The transporter associated with antigen processing (TAP) delivers the peptide fragments of the proteolyzed antigen to the endoplasmic reticulum, where they bind to MHC class I molecules. Therefore, prediction of TAP-binding peptides is helpful in identifying the MHC class I-restricted T-cell epitopes and has been implemented into several T-cell epitope prediction software (i.e., EpiJen, NetCTL, and WAPP). But as TAP binding alone is not very informative, there are only very few software exclusively for TAP-binding prediction. For example, TAPPred [94] is an online tool to predict binding affinity of peptides toward the TAP transporter. The server allows the user to predict TAP binders using a simple SVM model or cascade SVM model, and to customize the display/output.

Table 6
T-cell epitope related databases

Database	URL	Description
IEDB	http://epitope2.immuneepitope.org/home.do	Biothreat pathogen T-cell epitope database
IMGT/TR	http://imgt.cines.fr/textes/IMGTrepertoire	Aligned and annotated T-cell receptor sequences
IPD-IMGT/HLA	http://www.ebi.ac.uk/ipd/imgt/hla/	Aligned and annotated HLA sequences by WHO nomenclature
SYFPEITHI	http://www.syfpeithi.de	MHC binders, MHC-specific anchor and auxiliary motifs

4.2.6 T-Cell Epitope Databases

There are more databases containing larger amount of data (usually tens of thousands of epitopes) of T-cell epitopes than B-cell epitope databases as discontinuous epitopes are more difficult to analyze. A few commonly used T-cell epitope databases are listed in Table 6. These databases usually provide user-friendly search interface to search or browse for epitopes. Among these databases, IEDB, which has been mentioned earlier in B-cell epitope databases, is one of the several complete databases containing a large number of epitopes, including both T- and B-cell epitopes. When searching for epitopes, it allows specifying MHC restriction to limit the search by one or more MHC classes or specific alleles.

Syfpeithi [32] is a database comprising more than 7000 peptide sequences known to be MHC class I and II ligands and peptide motifs of humans and other species, such as apes, cattle, and mouse. The database is based on previous publications on T-cell epitopes and MHC ligands. Syfpeithi also allows prediction of T-cell epitopes based on published motifs by calculating a score based on amino acids in the anchor, auxiliary anchor positions, and other frequent amino acids. However, the Syfpeithi website has not been updated since August 2012.

IMGT® [95] (the international ImMunoGeneTics information system®) has a wide collection of immunoglobulins, T-cell receptors, MHC, and related proteins of the immune system of human and other vertebrates. It consists of sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB, etc.), genome database (IMGT/GENE-DB), structure database (IMGT/3Dstructure-DB), monoclonal antibodies database (IMGT/mAb-DB), Web resources (IMGT Repertoire, etc.), and interactive tools (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/StructuralQuery, etc.).

As part of the IMGT, the IPD-IMGT/HLA Database [96] provides a specialist database for sequences of the human MHC

(currently containing 14,473 allele sequences) and includes the official sequences named by the WHO Nomenclature Committee for Factors of the HLA System.

5 Conclusions

Bioinformatics methods, software, and databases related to the studies of allergens and allergenic epitopes are developing and expanding quickly. This review is intended to provide an introduction and overview of the publicly accessible bioinformatics tools and databases related to allergy field. We hope to assist researchers in locating and leveraging on these tools and to provide insights into future development in this field. Although current prediction methods and tools are not perfect, especially for allergen and B-cell discontinuous epitopes discovery, a good selection and combination of the proper methods can still save researchers a considerable amount of work and effort. As this field progresses, more data sources and improvements on the methods are expected to make prediction more reliable and discovery even less expensive and time consuming.

References

1. Larche M, Akdis CA, Valenta R (2006) Immunological mechanisms of allergen-specific immunotherapy. *Nat Rev Immunol* 6(10): 761–771
2. Bharadwaj M, Illing P, Theodossis A et al (2012) Drug hypersensitivity and human leukocyte antigens of the major histocompatibility complex. *Annu Rev Pharmacol Toxicol* 52:401–401
3. Tanabe S (2007) Epitope peptides and immunotherapy. *Curr Protein Pept Sci* 8(1):109–118
4. Lin J, Bardina L, Shreffler WG et al (2009) Development of a novel peptide microarray for large-scale epitope mapping of food allergens. *J Allergy Clin Immunol* 124(2):315–322 322 e311–313
5. Ayuso R, Sanchez-Garcia S, Pascal M et al (2012) Is epitope recognition of shrimp allergens useful to predict clinical reactivity? *Clin Exp Allergy* 42(2):293–304
6. Mou C, Zhang S, Lin J et al (2002) EST analysis of mRNAs expressed in neurula of Chinese amphioxus. *Biochem Biophys Res Commun* 299(1):74–84
7. Fu Z, Lu Y, Ye L et al (2002) Construction and application of a large scale cDNA sequences analysis system based on Unix. *Acta Sci Nat Univ Sunyatseni* 41(5):60–63
8. Liu Z, Lin J, Fu Z et al (2001) A sequence-based typing method for HLA-DPB1 genotyping. *Acta Sci Nat Univ Sunyatseni* 40(5): 83–87
9. Fu Z (2008) Post-structural genomics: correlate protein sequence, structure, and function. ProQuest, Ann Arbor, MI
10. Fu Z, Lin J (2015) Applying supervised machine learning in bioinformatics analysis. In: Clary TS (ed) *Horizons in computer science research*, vol 12. Nova, New York, NY
11. Saigo H, Vert JP, Ueda N et al (2004) Protein homology detection using string alignment kernels. *Bioinformatics* 20(11):1682–1689
12. Hakenberg J, Bickel S, Plake C et al (2005) Systematic feature evaluation for gene name recognition. *BMC Bioinformatics* 6(Suppl 1):S9
13. El-manzalawy YM (2008) Machine learning approaches for epitope prediction. Graduate theses and dissertations, Iowa State University
14. Lin J, Bruni FM, Fu Z et al (2012) A bioinformatics approach to identify patients with symptomatic peanut allergy using peptide microarray immunoassay. *J Allergy Clin Immunol* 129(5):1321–1328 e1325
15. De Groot AS (2006) Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov Today* 11(5–6):203–209

16. Aalberse RC (2000) Structural biology of allergens. *J Allergy Clin Immunol* 106(2):228–238
17. Brusci V, Petrovsky N, Gendel SM et al (2003) Computational tools for the study of allergens. *Allergy* 58(11):1083–1092
18. Saha S, Raghava GP (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 34(Web Server issue):W202–W209
19. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino-acid-sequences. *Proc Natl Acad Sci U S A* 78(6):3824–3828
20. Rosenberg HF, Dyer KD (1995) Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J Biol Chem* 270(50):30234
21. Pellequer JL, Westhof E, Van Regenmortel MH (1991) Predicting location of continuous epitopes in proteins from their primary structures. *Methods Enzymol* 203:176–201
22. 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(19):5425–5432
23. Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 47:45–148
24. Levitt M (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry* 17(20):4277–4285
25. Emini EA, Hughes JV, Perlow DS et al (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* 55(3):836–839
26. Greenbaum JA, Andersen PH, Blythe M et al (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit* 20(2):75–82
27. Kolaskar AS, Kulkarni-Kale U (1999) Prediction of three-dimensional structure and mapping of conformational epitopes of envelope glycoprotein of Japanese encephalitis virus. *Virology* 261(1):31–42
28. Doytchinova IA, Flower DR (2001) Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J Med Chem* 44(22):3572–3581
29. Schueler-Furman O, Altuvia Y, Sette A et al (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci* 9(9):1838–1846
30. Logean A, Rognan D (2002) Recovery of known T-cell epitopes by computational scanning of a viral genome. *J Comput Aided Mol Des* 16(4):229–243
31. Ponomarenko J, Bui HH, Li W et al (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* 9:514
32. Rammensee H, Bachmann J, Emmerich NP et al (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50(3–4):213–219
33. Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17(12):1236–1237
34. Donnes P, Kohlbacher O (2006) SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res* 34(Web Server issue):W194–W197.
35. Murugan N, Dai Y (2005) Prediction of MHC class II binding peptides based on an iterative learning model. *Immunol Res* 1:6
36. Doytchinova IA, Taylor P, Flower DR (2003) Proteomics in vaccinology and immunobiology: an informatics perspective of the immunone. *J Biomed Biotechnol* 2003(5):267–290
37. Yu K, Petrovsky N, Schonbach C et al (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med* 8(3):137–148
38. Bhasin M, Raghava GP (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* 22(23–24):3195–3204
39. El-Manzalawy Y, Honavar V (2010) Recent advances in B-cell epitope prediction methods. *Immunol Res* 6(Suppl 2):S2
40. Sollner J (2006) Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. *J Mol Recognit* 19(3):209–214
41. FAO/WHO (2001) Allergenicity of genetically modified foods. Report of a joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. FAO/WHO, Rome
42. Ivanciuc O, Schein CH, Braun W (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* 31(1):359–362
43. Stadler MB, Stadler BM (2003) Allergenicity prediction by protein sequence. *FASEB J* 17(9):1141–1143
44. Bjorklund AK, Soeria-Atmadja D, Zorzet A et al (2005) Supervised identification of allergen-representative peptides for in silico

- detection of potentially allergenic proteins. *Bioinformatics* 21(1):39–50
45. Wang J, Yu Y, Zhao Y et al (2013) Evaluation and integration of existing methods for computational prediction of allergens. *BMC Bioinformatics* 14(Suppl 4):S1
 46. Radauer C, Nandy A, Ferreira F et al (2014) Update of the WHO/IUIS allergen nomenclature database based on analysis of allergen sequences. *Allergy* 69(4):413–419
 47. Mari A, Scala E, Palazzo P et al (2006) *Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model.* *Cell Immunol* 244(2):97–100
 48. Sweredoski MJ, Baldi P (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics* 24(12):1459–1460
 49. Rubinstein ND, Mayrose I, Martz E et al (2009) Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinformatics* 10:287
 50. Kringelum JV, Lundegaard C, Lund O et al (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* 8(12):e1002829
 51. Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci* 15(11):2558–2567
 52. Qi T, Qiu T, Zhang Q et al (2014) SEPPA 2.0—more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen. *Nucleic Acids Res* 42((Web Server issue)):59–63
 53. Liang S, Zheng D, Zhang C et al (2009) Prediction of antigenic epitopes on protein surfaces by consensus scoring. *BMC Bioinformatics* 10:302
 54. Liang S, Zheng D, Standley DM et al (2010) EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. *BMC Bioinformatics* 11:381
 55. Yao B, Zheng D, Liang S et al (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
 56. Moreau V, Granier C, Villard S et al (2006) Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics* 22(9):1088–1095
 57. Huang J, Gutteridge A, Honda W et al (2006) MIMOX: a web tool for phage display based epitope mapping. *BMC Bioinformatics* 7:451
 58. Mayrose I, Penn O, Erez E et al (2007) Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics* 23(23):3244–3246
 59. Mayrose I, Shlomi T, Rubinstein ND et al (2007) Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res* 35(1):69–78
 60. Bublil EM, Freund NT, Mayrose I et al (2007) Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. *Proteins* 68(1):294–304
 61. Huang YX, Bao YL, Guo SY et al (2008) Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinformatics* 9:538
 62. Chen WH, Sun PP, Lu Y et al (2011) MimoPro: a more efficient Web-based tool for epitope prediction using phage display libraries. *BMC Bioinformatics* 12:199
 63. Huang J, Ru B, Dai P (2011) *Bioinformatics resources and tools for phage display.* *Molecules* 16(1):694–709
 64. Negi SS, Braun W (2009) Automated detection of conformational epitopes using phage display Peptide sequences. *Bioinf Biol Insights* 3:71–81
 65. Sun P, Ju H, Liu Z et al (2013) *Bioinformatics resources and tools for conformational B-cell epitope prediction.* *Comput Math Methods Med* 2013:943636
 66. Saha S, Raghava GP (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 65(1):40–48
 67. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. *Immunol Res* 2:2
 68. Toseland CP, Clayton DJ, McSparron H et al (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunol Res* 1(1):4
 69. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242
 70. Saha S, Bhasin M, Raghava GPS (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics* 6:79–79
 71. Vita R, Overton JA, Greenbaum JA et al (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 43((Database issue)):D405–D412
 72. Kulkarni-Kale U, Bhosle S, Kolaskar AS (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res* 33((Web Server issue)):W168–W171
 73. Valuev VP, Afonnikov DA, Ponomarenko MP et al (2002) ASPD (Artificially selected proteins/

- peptides database): a database of proteins and peptides evolved in vitro. *Nucleic Acids Res* 30(1):200–202
74. Shtatland T, Guettler D, Kossodo M et al (2007) PepBank—a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics* 8:280
75. Huang J, Ru B, Zhu P et al (2012) MimiDB 2.0: a mimotope database and beyond. *Nucleic Acids Res* 40((Database issue)):D271–D277
76. Meydan C, Otu HH, Sezerman OU (2013) Prediction of peptides binding to MHC class I and II alleles by temporal motif mining. *BMC Genomics* 14((Suppl. 2)):S13
77. Doytchinova IA, Guan P, Flower DR (2006) EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinformatics* 7:131
78. Cui J, Han LY, Lin HH et al (2007) Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol Immunol* 44(5):866–877
79. Stranzl T, Larsen MV, Lundegaard C et al (2010) NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62(6):357–368
80. Trost B, Bickis M, Kusalik A (2007) Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. *Immunol Res* 3:5–5
81. Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* 152(1):163–175
82. Schuler MM, Nastke MD, Stevanovic S (2007) SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol* 409:75–93
83. Bhasin M, Raghava GP (2003) Prediction of promiscuous and high-affinity mutated MHC binders. *Hybrid Hybridomics* 22(4):229–234
84. Lundegaard C, Lamberth K, Harndahl M et al (2008) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res* 36((Web Server issue)):W509–W512
85. Hoof I, Peters B, Sidney J et al (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 61(1):1–13
86. Larsen MV, Lundegaard C, Lamberth K et al (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* 8:424
87. Donnes P, Kohlbacher O (2005) Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci* 14(8):2132–2140
88. Sturniolo T, Bono E, Ding J et al (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 17(6):555–561
89. Reche PA, Reinherz EL (2007) Prediction of peptide-MHC binding using profiles. *Methods Mol Biol* 409:185–200
90. Guan P, Hattotuwigama CK, Doytchinova IA et al (2006) MHCpred 2.0: an updated quantitative T-cell epitope prediction server. *Appl Bioinformatics* 5(1):55–61
91. Nussbaum AK, Kuttler C, Haderl KP et al (2001) PAPROC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics* 53(2):87–94
92. Bhasin M, Raghava GP (2005) Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res* 33(Web Server issue):W202–W207.
93. Nielsen M, Lundegaard C, Lund O et al (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57(1–2):33–41
94. Bhasin M, Lata S, Raghava GP (2007) TAPPred prediction of TAP-binding peptides in antigens. *Methods Mol Biol* 409:381–386
95. Lefranc MP, Giudicelli V, Ginestoux C et al (2009) IMGT, the international ImmunoGeneTics information system. *Nucleic Acids Res* 37((Database issue)):D1006–D1012
96. Robinson J, Waller MJ, Fail SC et al (2009) The IMGT/HLA database. *Nucleic Acids Res* 37((Database issue)):1013–1017