Bioinformatics Tools for Epitope Prediction

4

Mohini Jaiswal, Shafaque Zahra, and Shailesh Kumar

Abstract

Immunological protection is conferred by immune cells, i.e., B and T cells, which can efficiently develop pathogen-specific memory and thus involved in adaptive immunity. More specifically, these immune cells can recognize a specific portion of their respective antigens termed as epitopes which possess their own significant values. There is a noble reason to identify the antigenic region of an antigen as it is having a great empirical cause, which includes exploration of disease etiology, the advancement of diagnosis assays, immune monitoring, and to design epitope-based vaccines. It requires detection and prediction of epitopes which is a considerable concern in the preparation of a peptide-based vaccine that is the centralized issue of immunoinformatics. Experimental screening is involved for large arrays of probable epitope candidates; thereby it is pricey and tedious. There is a requirement of more-advanced immunoinformatics tools as a prodigious amount of information has accumulated because of the onset of next-generation sequencing approaches for collection, analysis, and interpretation of data. Further, development of in silico epitope prediction methods has substantially reduced the difficulties related to epitope mapping by shortening potential epitope candidates list for experimental testing. These software tools have diverse applications in diagnosis of infectious diseases and allergies, understanding immune system function, vaccine designing, and prognosis of cancer. This chapter presents an outlook on how these tools are capable to predict epitopes of various antigens.

Keywords

B-cell and T-cell epitopes · Immunoinformatics · Immunological protection · Vaccine designing

M. Jaiswal · S. Zahra · S. Kumar (✉)
Data Sciences Laboratory, National Institute of Plant Genome Research (NIPGR),
New Delhi, India
e-mail: shailesh@nipgr.ac.in

S. Singh (ed.), *Systems and Synthetic Immunology*,
https://doi.org/10.1007/978-981-15-3350-1_4

4.1 Introduction

The adaptive immune system is also termed as acquired immune system as it is acquired during the lifetime rather than the inherited one and is considered as a subsystem of the global immune system whose constituents are highly specialized systemic cells and processes that help out in elimination of pathogens as well as in their growth prevention. Due to the existence of acquired immunity, immunological memory creates an initial response for each specific pathogen which results in a strong anamnestic response at the time of subsequent exposure to that particular pathogen. Vaccination is based on this particular feature of acquired immunity. B and T cells are involved in adaptive immunity which is responsive for humoral- and cell-mediated immunity, respectively. They recognize a specific portion of protein residing on the surface of pathogen rather than pathogens as a whole and that protein is termed as an antigen. Distinct receptors residing on the surface of B and T cells designated as B-cell and T-cell receptors (BCR & TCR) consist of membrane-bound immunoglobulins helping in the recognition of the solvent-exposed antigens. There is a remarkable difference between perceptions by B and T cells [30]. Different functions are triggered from antibodies released by B cells upon binding with their respective antigens. As a result, toxins and pathogens get neutralized and labeled as for destruction [20].

Besides this, cell surface-residing T-cell receptor (TCR) presented by T cells assist recognition of antigen-presenting cells (APCs) displayed antigens bounded with major histocompatibility complex (MHC) molecules. MHC I and II molecules are involved in T-cell epitopes presentation. Co-receptor CD4 expressed by helper T cells assists in the perception of antigen in the context of MHC class II, while antigen displayed by MHC class I molecules is acknowledged by cytotoxic CD8$^+$ T cells as per the immunological dogma. Subsequently, CD8 and CD4 T-cell epitopes exist. Meanwhile, CD4 T cells can act as a helper or regulatory T cells [20]. The immune response is amplified by helper T cells which are divided into three major subclasses that include Th1 involved in cell-mediated immunity against intracellular pathogens, Th2 involved in antibody-mediated immunity, and Th17 showing inflammatory response as well as defense across extracellular bacteria [37].

Along with the advancement in recombinant DNA technology, bioinformatics tools development and information of host immune response that acts as the genetic background of pathogen has led to the advancement of new vaccines which are more efficient, secure, and inexpensive in contrast to conventional vaccines. Conservation of chosen epitopes in a vaccine is a prerequisite event across distinct stages of pathogen and its variants. Intracellular antigen processing is required for cytotoxic T-cell-intervened response for which linear epitopes act as a prevailing target. In this respect, the binding affinity of selected epitopes should be with more than one major histocompatibility complex allele for a particular vaccine.

To identify B-cell and T-cell epitopes for vaccine designing is a decisive step as it requires to construct overlapping peptides based on experimental scanning result of epitope-active regions that span complete sequence of a protein antigen, and it is again a pricey and tedious job. Therefore, to elicit an immune response, in silico techniques are a perfect substitute to identify protein domains out of thousands of plausible candidates [29]. This chapter gives an insight regarding some of the commonly used bioinformatics tools developed for B-cell and T-cell epitope prediction.

4.2 Tools for B-Cell Epitopes Prediction

B-cell epitope anticipation tools aim to contribute to the detection of the specific antigenic peptide (epitope), and thus it has a significant purpose as it acts as a substitute of antigen for antibody production.

However, linear and conformational epitopes are the two groups based on B-cell epitopes classification. Sequential residues in primary sequence constitute a segment of linear epitope, whereas a cluster of antigen residues placed at a distance from each other in their primary sequence is regarded as conformational epitope that is brought to spatial vicinity because of polypeptide folding [1]. Thereby, linear and conformational B-cell epitopes are equally termed as continuous and discontinuous B-cell epitopes, respectively. This means that denatured antigens can be identified by antibodies which are used to identify linear B-cell epitopes, while in case of conformational B-cell epitopes, denaturation leads to recognizance failure. Unlike linear epitopes, conformational epitopes prediction depends on the three-dimensional structure of the protein. Linear B-cell epitopes are possessed by only a few of the native antigens; otherwise, approximately 90% of them are conformational [26].

4.2.1 Linear B-Cell Epitopes Anticipation

In spite of being a trivial one, linear B-cell epitopes can act as a substitute for immunization and antibody production. Thus, their anticipation received major attention. It has been predicted via methods based on a sequence from the primary sequence of antigens. Earlier computational methods were rooted on propensity scales of simplified amino acids featuring physicochemical characteristics for B-cell epitopes. For example, residue hydrophilicity calculations were implemented by Hopp and Wood to predict B-cell epitopes [11, 12] on the basis of the hypothesis that hydrophilic regions preferentially reside on the protein surface and are probably antigenic. For developing diverse prediction tools datasets, algorithms and training features used to differ.

Currently, accessible linear B-cell epitopes envision tools involve BcePred indulged in anticipation of linear B-cell epitopes as per their physicochemical attributes. Another one is Lbtope based on Immune Epitope Database (IEDB)-derived data of experimentally approved non-B-cell epitopes [39]. Analogous positive data of B-cell epitopes is required for training of artificial neural networks (ANNs) algorithm that has been implemented in Lbtope yet vary on negative data of non-B-cell epitopes.

Another one is BepiPred, which involves random forests algorithm-based training of B-cell epitopes derived from the three-dimensional architecture of antigen-antibody complexes. It is involved in the prediction of both varieties of B-cell epitopes [14]. On the whole, B-cell epitope prediction methods implementing machine learning algorithm outperformed other methods rooted on the basis of amino acid propencities.

4.2.2 Conformational B-Cell Epitopes Anticipation

It has been already mentioned that preferentially B-cell epitopes are conformational, even though linear B-cell epitopes anticipation is ahead of them, for that two major empirical approaches exist. Firstly, the requirement of conformational B-cell epitopes prediction is whole information of protein 3D structure which is available only for a few proteins [31]. The second one is the complicated task of discontinuous B-cell epitopes isolation from their corresponding protein frame to formulate a particular antibody. Its necessity is suitable scaffolds for epitope grafting. In spite of these difficulties, various mechanisms exist to envisage conformational B-cell epitopes.

One of them is CBTOPE which relies on Support Vector Machine (SVM) algorithm. Physicochemical characteristics and sequence-derived attributes are utilized for training of conformational B-cell epitopes, and a benchmark dataset of conformational epitopes derived from 3D structures of antibody-protein complexes is used for their assessment along with 86.59% accuracy from cross-validation experiments [1]. This tool is involved in predicting discontinuous B-cell epitope of an antigen based on its primary sequence by overcoming the first difficulty.

Another one is ElliPro that depends on the geometrical properties of protein structure. In addition to CBTOPE, ElliPro also assessed on the same benchmark dataset derivative of 3D structures of antibody-protein complexes [24].

There is a significant role of bioinformatics tools for each of the B-cell epitopes envision in peptide-based vaccine designing and disease identification [9, 22].

Although there are various tools for each of the B-cell epitope prediction, the five most commonly highly utilized tools are described in Table 4.1.

Table 4.1 Some freely accessible B-cell epitope anticipation tools

B-cell types	Tools	Method	Server (URL)	References
Continuous	BcePred	Physicochemical properties	http://www.imtech.res.in/raghava/bcepred/	[28]
	Lbtope	ML (ANN)	http://www.imtech.res.in/raghava/lbtope/	[35]
Discontinuous	ElliPro	Structure-based method (geometrical properties)	http://tools.iedb.org/ellipro/	[24]
	CBTOPE	Sequence based (SVM)	http://www.imtech.res.in/raghava/cbtope/submit.php	[1]
Both	BepiPred-2.0	ML (DT)	http://www.cbs.dtu.dk/services/BepiPred/	[14]

4.2.3 Description of Various Tools and Their Overall Performance Enlisted in Table 4.1

4.2.3.1 BcePred Server

BcePred server assists in envision of linear B-cell epitope rooted on physicochemical characteristics of amino acids. These properties comprised of mobility, turns, flexibility, exposed surface, accessibility, hydrophilicity, polarity, and antigenicity of any particular antigen. To quantify these properties, attributes value is allocated to all of the 20 natural amino acids. The user can opt for any combination of physicochemical attributes for epitopes prediction.

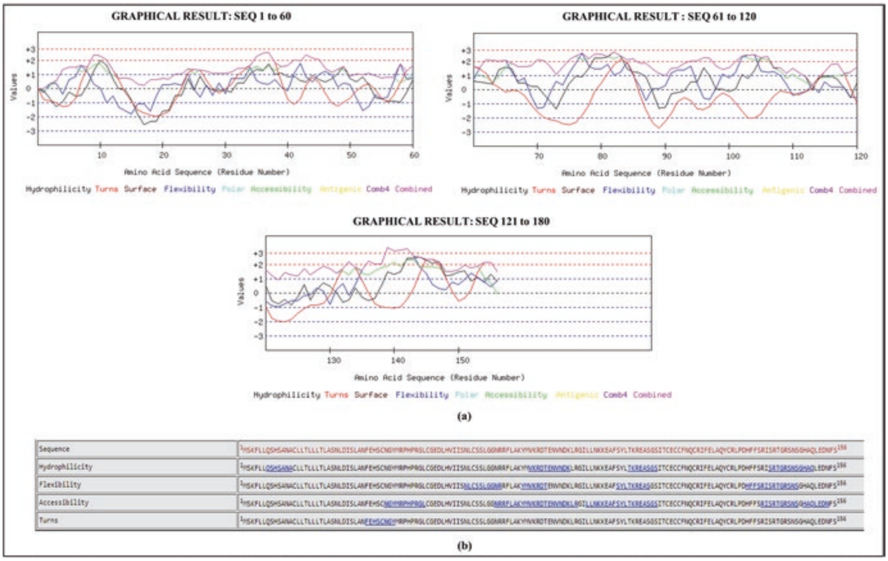
PERL version 5.03 is used for writing a common gateway interface (CGI) script. Sun Server (420E) with a UNIX (Solaris 7) environment is used for their installation.

Submission Form Using the Following Steps for BcePred Server

- Input data is in the form of sequence that should be written in submission form by using one-letter amino acid code: "acdefghiklm-npqrstvwy" or "ACDEFGHIKLMNPQRSTVWY." Other letters get transformed into "X" which were reviewed as obscure amino acids.
- Threshold values lie in the range of −3 to +3. As per the outstanding sensitivity and specificity value gained, default thresholds for various parameters have been opted.
- After pressing "Submit sequence" button, a WWW page will return as a result that delivers summarized information about entered query sequence in graphical (Fig. 4.1a) as well as in tabular and in overlap display format (Fig. 4.1b). The tabular format provides a normalized score of opted attributes with the respective amino acid residue of a protein as well as minimum, maximum, and average values of integrated methods opted.
- Quick picturing of B-cell epitope on protein is achieved when residue properties are plotted along protein backbone. A particular amino acid residue will be reviewed as expected B-cell epitope when their peak is having value above threshold (default value is 2.38 in the combined approach).

Pros and Cons

- By using BcePred server, prediction of B-cell epitopes can be made based on two or more physicochemical properties at a time. So it would be more accurate.
- However, there is no autonomous assessment or benchmarking of prevailing procedures in this server; thereby, the decision of much better residue property or method is a difficult task.


GRAPHICAL RESULT: SEQ 1 to 60
GRAPHICAL RESULT : SEQ 61 to 120
GRAPHICAL RESULT: SEQ 121 to 180
Values
+3
+2
+1
0
-1
-2
-3
10 20 30 40 50 60
70 80 90 100 110 120
130 140 150
Amino Acid Sequence (Residue Number)
Hydrophilicity Turns Surface Flexibility Polar Accessibility Antigenic Comb4 Combined
(a)
(b)

Sequence	^{1}MSKFLLQSHSANACLLTLLLTLASNLDISLANFEHSCNDYMRPHPRGLCGEDLHVIISNLCSSLGGNRRFLAKYMVKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISRTGRSNSGHAQLEDNFS^{156}
Hydrophilicity	^{1}MSKFLLQSHSANACLLTLLLTLASNLDISLANFEHSCNDYMRPHPRGLCGEDLHVIISNLCSSLGGNRRFLAKYMVKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISRTGRSNSGHAQLEDNFS^{156}
Flexibility	^{1}MSKFLLQSHSANACLLTLLLTLASNLDISLANFEHSCNDYMRPHPRGLCGEDLHVIISNLCSSLGGNRRFLAKYMVKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISRTGRSNSGHAQLEDNFS^{156}
Accessibility	^{1}MSKFLLQSHSANACLLTLLLTLASNLDISLANFEHSCNDYMRPHPRGLCGEDLHVIISNLCSSLGGNRRFLAKYMVKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISRTGRSNSGHAQLEDNFS^{156}
Turns	^{1}MSKFLLQSHSANACLLTLLLTLASNLDISLANFEHSCNDYMRPHPRGLCGEDLHVIISNLCSSLGGNRRFLAKYMVKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISRTGRSNSGHAQLEDNFS^{156}

Fig. 4.1 BcePred server showing B-cell epitope regions in insulin precursor sequence (length is 156 aa) of *Aplysia californica*. (**a**) Graphical result. (**b**) Overlap display in which selected programs are hydrophilicity, flexibility, accessibility, and turns having threshold value as 1.9, 2.0, 1.9, and 2.4, respectively. Predicted B-cell epitopes are shown in blue color and are underlined

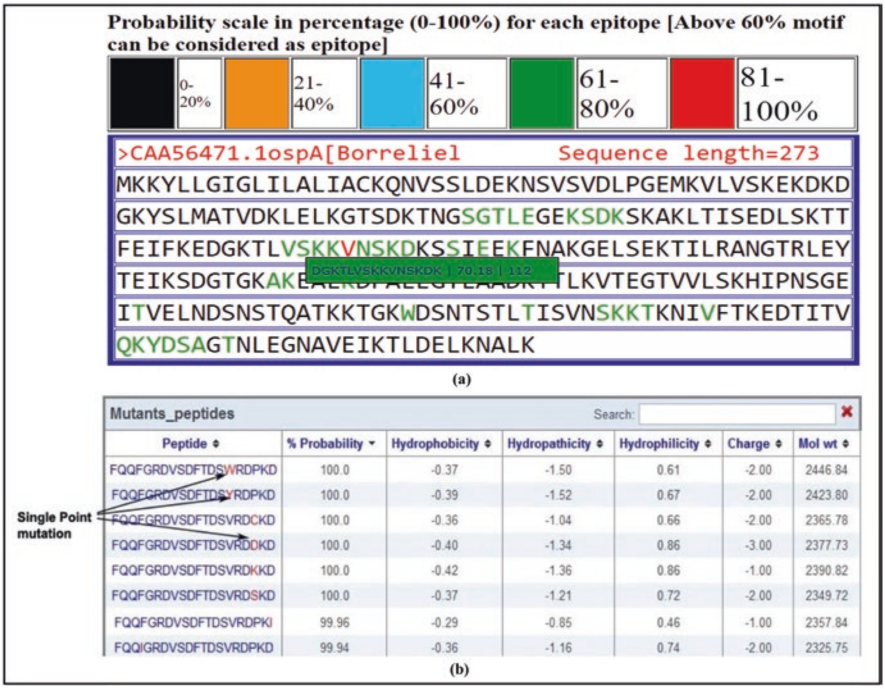
4.2.3.2 Lbtope

Lbtope is a tool designed to predict linear B-cell epitope. PHP 5.2.9, HTML, and JavaScript have been used to develop its front end. Further, Red Hat Enterprise Linux 6 server environment has been utilized for its installation. Along with experimentally certified B-cell epitopes, non-B-cell epitopes can be also retrieved from Immune Epitope Database (IEDB) which include five datasets termed as Lbtope_Fixed, Lbtope_Fixed_non_redundant, Lbtope_Variable, Lbtope_Variable_non_redundant, and Lbtope_Confirm dataset. Various models have been developed based on these datasets to discriminate B-cell epitopes from non-epitopes.

In Lbtope, SVM^{light} package is used for implementing SVM technique in association with Weka implemented Ibk.

Working Steps

I. Input data is the primary amino acid sequences in fasta format (Fig. 4.2a).
II. Overlapping peptides containing 20 amino acids and 5–30 amino acids are developed for Lbtope fixed dataset model and for variable datasets, respectively, for prediction of linear epitopes. Due to the very high specificity, nonredundant model is introduced as well.
III. Antigen sequences profiled with B-cell epitopes having probability scale of 20–80% comes as an output data (Fig. 4.2a).
IV. A higher score is meant for a higher possibility of a peptide to behave as B-cell epitope.



Probability scale in percentage (0-100%) for each epitope [Above 60% motif can be considered as epitope]

■	0-20%	🟧	21-40%	🟦	41-60%	🟩	61-80%	🟥	81-100%

```
>CAA56471.1ospA[Borreliel        Sequence length=273
MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEMKVLVSKEKDKD
GKYSLMATVDKLELKGTSDKTNGSGTLEGEKSDKSKAKLTISEDLSKTT
FEIFKEDGKTLVSKKVNSKDKSSIEEKFNAKGELSEKTILRANGTRLEY
TEIKSDGTGKAKE[illegible]TLKVTEGTVVLSKHIPNSGE
ITVELNDSNSTQATKKTGKWDSNTSTLTISVNSKKTKNIVFTKEDTITV
QKYDSAGTNLEGNAVEIKTLDELKNALK
```

DGKTLVSKKVNSKDK | 70.18 | 112

(a)

Mutants_peptides

Search:

Peptide	% Probability	Hydrophobicity	Hydropathicity	Hydrophilicity	Charge	Mol wt
FQQFGRDVSDFTDSWRDPKD	100.0	-0.37	-1.50	0.61	-2.00	2446.84
FQQFGRDVSDFTDSYRDPKD	100.0	-0.39	-1.52	0.67	-2.00	2423.80
FQQFGRDVSDFTDSVRDCKD	100.0	-0.36	-1.04	0.66	-2.00	2365.78
FQQFGRDVSDFTDSVRDDKD	100.0	-0.40	-1.34	0.86	-3.00	2377.73
FQQFGRDVSDFTDSVRDKKD	100.0	-0.42	-1.36	0.86	-1.00	2390.82
FQQFGRDVSDFTDSVRDSKD	100.0	-0.37	-1.21	0.72	-2.00	2349.72
FQQFGRDVSDFTDSVRDPKI	99.96	-0.29	-0.85	0.46	-1.00	2357.84
FQQIGRDVSDFTDSVRDPKD	99.94	-0.36	-1.16	0.74	-2.00	2325.75

(b)

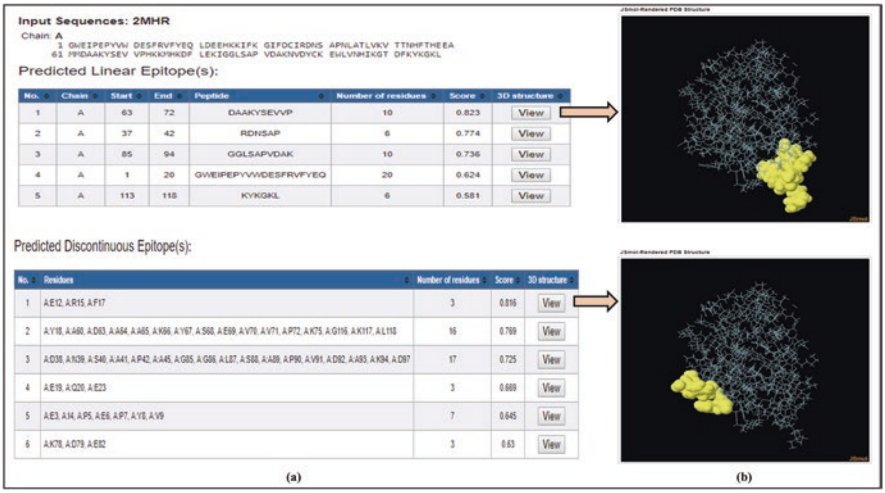
Fig. 4.2 (**a**) Sequence of OspA from *Borrelia burgdorferi* taken as input showing highlighted text as the predicted B-cell epitope along with probability scale. (**b**) Output data from peptide submission and mutant generation

Pros and Cons

- In addition to B-cell epitope prediction, this server exhibits a peptide mutation tool. It helps to create all plausible single-point mutations of a given peptide (Fig. 4.2b) and to predict its other properties. The further probability score is calculated based on a particular algorithm. Thereby, mutation tool is useful in the creation of peptide mutants and examination of its epitopic and other desired probability as well.
- Model based on Lbtope_Confirm dataset executed in an improved way as a comparison to mock-up established on Lbtope_Variable dataset. However, these model's activity decreased on nonredundant datasets.

4.2.3.3 ElliPro

ElliPro is a Web server obtained from Ellipsoid and Protrusion, that executes a modified version of Thornton's method according to which identification of continuous epitopes from protruding regions of protein globular surface becomes possible [38]. In addition to a residue clustering algorithm, the MODELLER program [8] and a Jmol viewer (Fig. 4.3b) are implemented in ElliPro as well. Due to this



Input Sequences: 2MHR

Chain: A

```
 1 GWEIPEPYVW DESFRVFYEQ LDEEHKKIFK GIFDCIRDNS APNLATLVKV TTNHFTHEEA
61 MMDAAKYSEV VPHKKMHKDF LEKIGGLSAP VDAKNVDYCK EWLVNHIKGT DFKYKGKL
```

Predicted Linear Epitope(s):

No.	Chain	Start	End	Peptide	Number of residues	Score	3D structure
1	A	63	72	DAAKYSEVVP	10	0.823	View
2	A	37	42	RDNSAP	6	0.774	View
3	A	85	94	GGLSAPVDAK	10	0.736	View
4	A	1	20	GWEIPEPYVWDESFRVFYEQ	20	0.624	View
5	A	113	118	KYKGKL	6	0.581	View

Predicted Discontinuous Epitope(s):

No.	Residues	Number of residues	Score	3D structure
1	A:E12, A:R15, A:F17	3	0.816	View
2	A:Y18, A:A60, A:D63, A:A64, A:A65, A:K66, A:Y67, A:S68, A:E69, A:V70, A:V71, A:P72, A:K75, A:G116, A:K117, A:L118	16	0.769	View
3	A:D38, A:N39, A:S40, A:A41, A:P42, A:A45, A:G85, A:G86, A:L87, A:S88, A:A89, A:P90, A:V91, A:D92, A:A93, A:K94, A:D97	17	0.725	View
4	A:E19, A:Q20, A:E23	3	0.669	View
5	A:E3, A:I4, A:P5, A:E6, A:P7, A:Y8, A:V9	7	0.645	View
6	A:K78, A:D79, A:E82	3	0.63	View

Fig. 4.3 (**a**) ElliPro prediction result for myohemerythin as an input sequence having sequence ID as 2MHR. (**b**) Epitope 3D structures for 2MHR via Jmol viewer program

implementation, envision of antibody epitopes as well as its visualization becomes possible in protein sequences as well as in structures. From 3D structures of antibody-protein complexes, a benchmark dataset of epitopes has been derived which is used to train ElliPro having the Area Under the ROC Curve (AUC) value as 0.732 [23].

Three algorithms are introduced in ElliPro to perform some major objectives that include an understanding of protein shape as an ellipsoid, estimation of residue protrusion index (PI), and grouping of neighboring residues as per their PI values.

Working Steps

I. Input data is either a protein structure or its primary amino acid sequence.

II. The sequence in fasta format or single-letter codes or their SwissProt/UniProt ID can be entered as a query in case the only sequence is available. To design a 3D structure of the submitted sequence, the selection of both a threshold for BLAST e-value and structural templates from PDB are required.

III. In case of structure, either a four-character PDB ID is entered in required space or a PDB file in PDB format can be uploaded (Fig. 4.3a). If submitted framework possesses more than one protein chain, then a specific chain has to be selected by the user on which calculation would be based.

IV. Threshold values are changeable based on parameters utilized by server to predict epitope, like minimum residue score (protrusion index), referred as S, that ranges in between 0.5 and 1.0 and maximum distance, termed as R, that ranges from 4 to 8 Å.

Pros and Cons

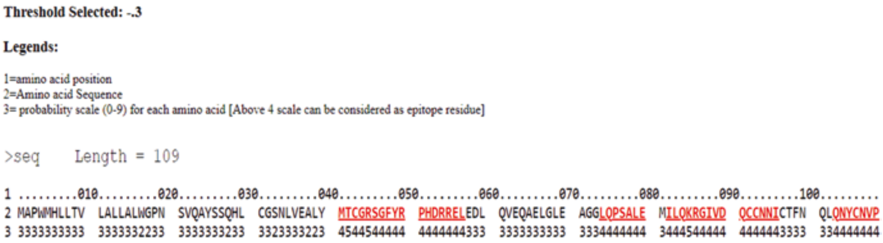
- ElliPro proves to be a helpful server for recognition of antibody epitopes from protein antigens and is helpful in identifying protein-protein interactions.
- A procedure that relies on geometrical attributes of protein structure has been introduced in this server which doesn't require training as well, so it is unable to properly differentiate between epitopes and non-epitopes.

4.2.3.4 CBTOPE

CBTOPE is a user-friendly Web server. It is established to anticipate conformational B-cell epitopes from antigen's amino acid sequence rather than based on their tertiary structure. A CGI script is written in Perl and HTML. Sun Server (420E) is used for installation under UNIX (Solaris 7) environment [1]. Development of this server is evident for envisioning of antigen's conformational B-cell epitope in which their primary amino acid sequences play a possible role.

Methodology

(a) For prediction via CBTOPE, main dataset is created by obtaining 526 antigenic sequences in combination with IEDB database as well as benchmark dataset [23] which is comprised of 161 protein chains derived from 144 antigen-antibody complex structures.
(b) Sequence redundancy is excluded by using program CD-HIT [16] at 40% cutoff.
(c) Finally, a nonredundant set of 187 antigens is gained. This set is devoid of sequences with the sequence identity of more than 40%.
(d) A different pattern is created. Standard procedure for assigning patterns is that if there would be any interaction between central residues and antibody, a positive value is assigned otherwise defined as negative (Fig. 4.4).
(e) By using patterns like the binary profile of pattern (BPP) and physiochemical profile of patterns (PPP), several models have been developed by using SVM as a classifier. It gained a maximal value of MCC as 0.22 and 0.17, respectively.
(f) Conventional characteristics of binary and physicochemical profiles are used and further assessed via fivefold cross-validation.



Threshold Selected: -.3

Legends:

1=amino acid position
2=Amino acid Sequence
3= probability scale (0-9) for each amino acid [Above 4 scale can be considered as epitope residue]

```
>seq    Length = 109

1 .........010.........020.........030.........040.........050.........060.........070.........080.........090.........100.........
2 MAPWMHLLTV  LALLALWGPN  SVQAYSSQHL  CGSNLVEALY  MTCGRSGFYR  PHDRRELEDL  QVEQAELGLE  AGGLQPSALE  MILQKRGIVD  QCCNNICTFN  QLQNYCNVP
3 3333333333  3333332233  3333333233  3323333223  4544544444  4444444333  3333333333  3334444444  3444544444  4444443333  334444444
```

Fig. 4.4 CBTOPE prediction result for insulin sequence of *Octodon degus* as an input. Predicted B-cell epitope is shown in red color

(g) The number of non-redundant protein chains is 187 comprising of 2261 antibody-interacting B-cell epitope residues that are used for training and assessment of all SVM models.

Working Steps

I. Input data is amino acid sequences in fasta format.
II. Total of 19 window patterns for each of the submitted sequences is created via server. The further amino acid composition is calculated to predict residues interacting with the antibody.
III. Amino acid sequence mapped with probability scale that ranges in between zero and nine comes as an output data for all amino acids where zero signifies the unusual possibility of residue to be a part of B-cell epitope and nine is the most plausible one (Fig. 4.4).
IV. For extraordinary precision (high-confidence) prediction, higher threshold value should be selected as per suggestion along with compromising the sensitivity of prediction. Nonetheless, lower threshold value should opt for maximum prediction of antibody-interacting residues.
V. The default threshold value is fixed at −0.3 as sensitivity and specificity are found to be equivalent at this value during CBTOPE development.

Pros and Cons

- Structure determination of a protein via techniques like X-ray crystallography proves to be costly, prolix, and time-consuming. Due to development of CBTOPE, one can predict conformational B-cell epitopes of antigens with ease which is lacking their tertiary structures with better sensitivity and AUC than other structure-based methods on same benchmark dataset as CPP composition-based SVM model is used in this server which outperformed others.
- Limitation of CBTOPE is its ineptitude for determination of number and distance required to obtain an epitope segment from antigen sequence.

4.2.3.5 BepiPred-2.0

BepiPred-2.0 is a Web server based on random forest algorithm for estimation of B-cell epitope, and annotated epitopes extracted from a dataset are used for its training which is composed of 649 antigen-antibody crystal structures and is derived from Protein Data Bank (PDB). Antibody molecules of each complex are recognized via HMM models.

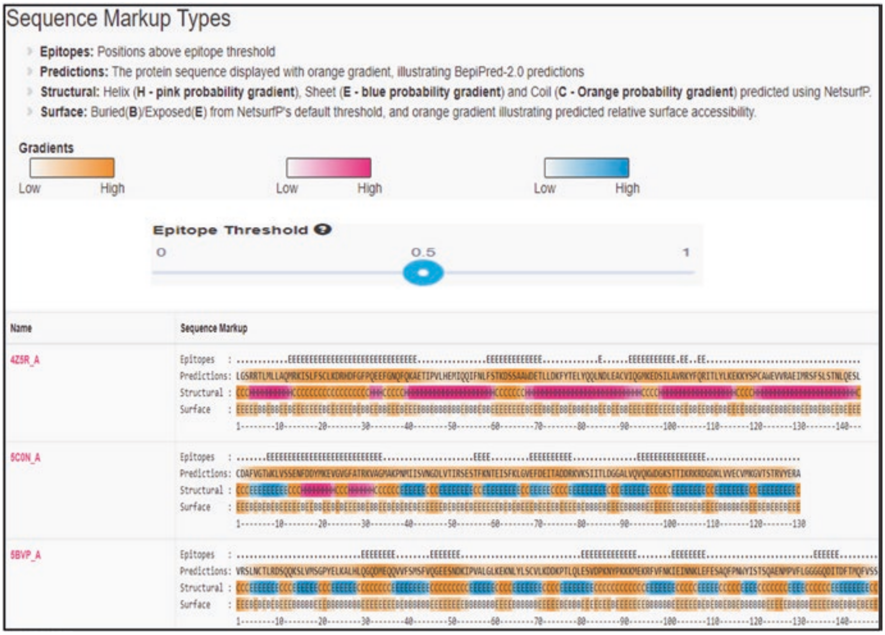
Methodology

(a) Random Forest Regression (RF) algorithm is assessed on a dataset to determine the plausibility of a given antigen residue so that it can be a part of an epitope with the usage of the fivefold cross-validation strategy.

(b) All of the residues is encrypted with the help of its polarity, hydrophobicity, computed volume along with secondary structure (SS), and relative surface accessibility (RSA) as anticipated by NetSurfP [21].
(c) The overall volume of antigen is gained via the addition of respective volumes of entire antigen's residues for almost 46 variables.
(d) Rolling average of window 9 is implemented on RF output to acquire concluding BepiPred-2.0 predictions.

Working Steps

I. Input data is protein sequences of interest having size more than 10 amino acids and lesser than 6000 in fasta format that can be entered into textbox either by pasting them or via uploading as a single file.
II. When predictions get completed, the user is automatically redirected to output page (Fig. 4.5) that has a navigation bar containing distinct tabs like "Summary" showing the result of each of the individual sequence in horizontal as well as in the form of a vertical table. Optionally, an email address can be given by the user so that after the job gets finished, result page link will be emailed.
III. "E" in "Epitopes" line is indicated as predictions higher than the user-defined threshold which is by default 0.5 above itself the protein sequence and is used to select the background color for protein sequences. Epitope classifications are alterable as per desire with the usage of "Epitope Threshold" slider.



Sequence Markup Types

- **Epitopes:** Positions above epitope threshold
- **Predictions:** The protein sequence displayed with orange gradient, illustrating BepiPred-2.0 predictions
- **Structural:** Helix (**H - pink probability gradient**), Sheet (**E - blue probability gradient**) and Coil (**C - Orange probability gradient**) predicted using NetsurfP.
- **Surface:** Buried(**B**)/Exposed(**E**) from NetsurfP's default threshold, and orange gradient illustrating predicted relative surface accessibility.

```
Name      Sequence Markup

4Z5R_A    Epitopes   : ...........EEEEEEEEEEEEEEEEEEEEEEEEEEEEEE..............EEEEEEEEEEEE............E......EEEEEEEEEEE.EE..EE..................................
          Predictions: LGSRRTLMLLAQMRKISLFSCLKDRHDFGFPQEEFGNQFQKAETIPVLHEMIQQIFNLFSTKDSSAAWDETLLDKFYTELYQQLNDLEACVIQGMKEDSILAVRKYFQRITLYLKEKKYSPCAWEVVRAEIMRSFSLSTNLQESL
          Structural : [illegible]
          Surface    : [illegible]
                       1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------140---

5CON_A    Epitopes   : .......EEEEEEEEEEEEEEEEEEEEEEEEEE....................EEEE.........EEEEEEEEEE..............EEEEEEEEEEEEEEEE......................
          Predictions: CDAFVGTWKLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVTIRSESTFKNTEISFKLGVEFDEITADDRKVKSIITLDGGALVQVQKWDGKSTTIKRKRDGDKLVVECVMKGVTSTRVYERA
          Structural : [illegible]
          Surface    : [illegible]
                       1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130

5BVP_A    Epitopes   : ..............................EEEEEEE.......EEEEEEE..........................EEEEEEEEEEEEE.......EEEEEEEE.........................EEEEEE........
          Predictions: VRSLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEESNDKIPVALGLKEKNLYLSCVLKDDKPTLQLESVDPKNYPKKKMEKRFVFNKIEINNKLEFESAQFPNWYISTSQAENMPVFLGGGGQDITDFTMQFVSS
          Structural : [illegible]
          Surface    : [illegible]
                       1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------140-------
```

Fig. 4.5 Sequence markup table of epitope predictions for three antigenic sequences to visualize the predictions on sequences in advanced output mode

IV. Predictions result are downloadable as JSON or CSV format via dropdown tab "Downloads." Besides this, by clicking the "All Downloads" tab, a short descriptive file can be found as well.

Pros and Cons

- BepiPred-2.0 attains a considerably better positive predictive value (PPV) and a moderately better true positive rate (TPR) in comparison to other methods. Also, it outperforms other available tools like BepiPred-1.0 and Lbtope for sequence-based epitope prediction relies on dataset retrieved from solved 3D structures or of a large collection of linear epitopes downloadable from IEDB database.
- The result format is informative as well as convenient.
- Limitation of BepiPred-2.0 is that it doesn't respond to nucleic acid sequences.

4.3 Tools for T-Cell Epitopes Prediction

Recognition of shortest peptides within an antigen is the main objective of T-cell epitope prediction which possesses immunogenicity, meaning capable to incite either CD4 or CD8 T cells. Immunogenicity is mainly based on three essential events which include processing of antigen and its binding with MHC molecules and acceptance from its respective TCR.

Amid all steps, MHC-peptide binding is the most discerning to delineate T-cell epitopes [13, 15]. Subsequently, the peptide-MHC binding prediction is the substantive baseline for prediction of T-cell epitopes.

4.3.1 Peptide-MHC Binding Anticipation

For peptide-MHC binding prediction, there should be an overview of already known peptide sequences that adhere with MHC molecules such as the existence of specified epitope databases, for instance, antigen [32], EPIMHC [18], and IEDB [39].

At the level of 3D structures of groove-resided bound peptides, resemblance exists between MHC I and II molecules, even though there is a major distinction between their binding grooves. For MHC I molecules, its peptide binding cleft consists of a single α chain; thereby, it is closed due to which their binding peptide length is reduced to 9 to 11 amino acid residues whose N- and C-terminal ends continue to stick by means of a linkage of hydrogen bonds with preserved residues of MHC I molecules [17, 36]. Tight physicochemical preferences also exist in addition to deep binding pockets in their peptide-binding groove that assist binding predictions. Alternative binding pockets exist for the same MHC I molecule which is often used by peptides of distinct sizes. Hence, there is a requirement of a fixed peptide length for the prophecy of MHC I-binding peptides. As mostly ligands have 9–11 residues, it can be the desired length.

On the contrary, open peptide-binding cleft is found in MHC II molecules, that allows expansion of peptide's N- and C-terminal ends beyond its binding groove [17, 36] which results in diversification of their peptide-binding length (9–22 residues). However, peptide-binding cleft allows to reside merely a core of nine residues, termed as peptide-binding core, into them. Consequently, the target of peptide-MHC II binding anticipation tools is to recognize peptide-binding cores mainly. The reason behind this imprecise forecasting of peptides that bind with MHC II molecule is their shallower and less demanding binding pockets than that of MHC I molecules [30].

Apart from this, peptide antigens derived from endogenous and exogenous pathway are offered by MHC I and MHC II molecules, respectively. Endosomal compartments are used for degradation and loading endocytosed antigens onto MHC II molecule [7], while antigens degraded via cytosolic pathway are transported via TAP to the endoplasmic reticulum and further loaded onto MHC I molecules. Before loading, peptides mostly go for trimming with the aid of ERAAP N-terminal aminopeptidases [10].

Along with MHC I and II-peptide binding anticipation tools, various tools are there to envisage even TAP binding that has been designed by training distinct algorithms on peptides having a significant affinity with TAP [3].

Consistently occurring amino acids are present in peptides at particular positions that bind with MHC molecules, termed as anchor residues thought to be liable for its binding with MHC molecule. However, later, it has been shown that along with anchor residues, peptide binding to a given MHC molecule is facilitated by non-anchor residues as well [27]. Accordingly, development of motif matrices (MM) helps in the assessment of input for each and all peptide positions of MHC molecule binding [19, 25].

Several ML algorithm has been used to solve mainly two distinct problems which are trained on datasets having peptides of known kinship to MHC molecules. First and foremost is the discernment of MHC binders from non-binders, and the second one is to envisage peptides binding affinity with MHC molecules.

MHC polymorphism is the major challenge in T-cell epitopes prediction. Human leukocyte antigen (HLA) is a term for MHC molecules in case of humans, and hundreds of their allelic variants exist which bind to peptide variants that need distinctive models to predict peptide-MHC binding. These variants are expressed at immensely diverse frequencies due to which HLA polymorphism creates hindrance in the advancement of T-cell epitope-based vaccines for distinct ethnic groups. In spite of all obstruction, there are various tools accessible for prediction of peptide-MHC binding. Some of them are described in Table 4.2.

Table 4.2 Some freely accessible T-cell epitope anticipation tools

MHC class	Tools	Method	Server (URL)	References
MHC I	nHLAPred	ANN	http://www.imtech.res.in/raghava/nhlapred/	[4]
	ProPred1	QAM	http://www.imtech.res.in/raghava/propred1/	[34]
	TAPPred	SVM	http://www.imtech.res.in/raghava/tappred	[6]
MHC II	ProPred	QAM	http://www.imtech.res.in/raghava/propred/	[33]
	EpiDOCK	SB	http://epidock.ddg-pharmfac.net	[2]

4.3.2 Description of Various Tools for T-Cell Epitope Prediction Enlisted in Table 4.2

4.3.2.1 nHLAPred

nHLAPred is a hybrid approach-based Web server which includes, firstly, a quantitative matrix (QM)-rooted technique in which involvement of each residue has been taken into consideration rather than just anchor residues and is formulated for 47 MHC class I alleles for which minimal 15 binders are accessible from MHCBN version 1.1 [5]. Secondly, an artificial neural network (ANN)-based method is implemented for 30 alleles out of 47 MHC alleles featuring at least 40 binders approachable from the database. Mutual approach (ANN and QM) has been used for the anticipation of 30 MHC alleles (Fig. 4.6), while the prediction of the remaining 37 alleles relies on QM [4]. The average accuracy of prediction is 92.8% that has ameliorated by 6% compared to each individual means with the development of this amalgam approach.

Sun Server 420R is used for installation under the Solaris environment. There is a partitioning of server in two substantial parts, ComPred and ANNPred, amid which ComPred enables for estimation of binders for 67 MHC class I alleles. Along with that, proteasomal matrices have been utilized by both parts to anticipate proteasomal cleavage site possessing MHC binders at C-terminal.

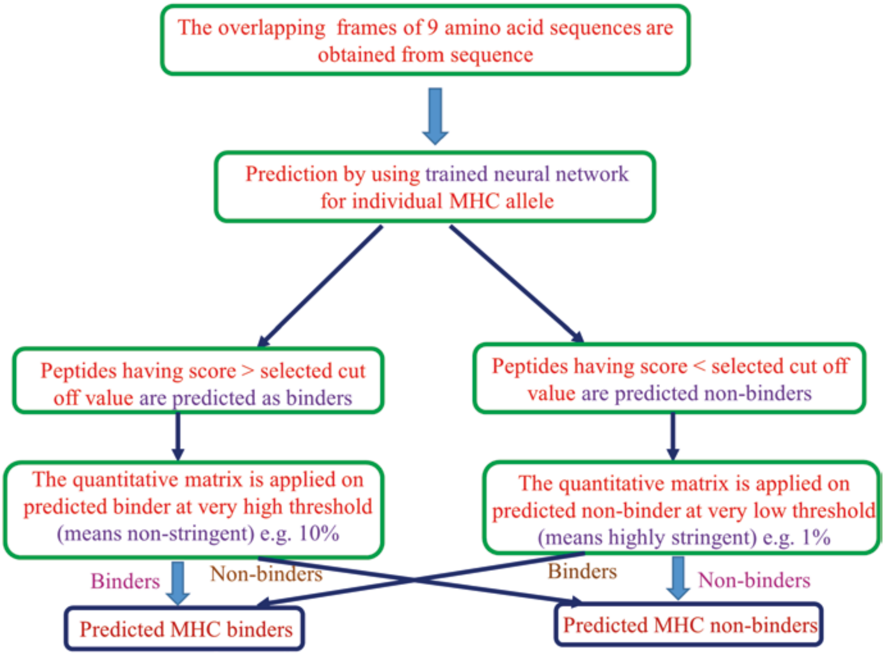

The overlapping frames of 9 amino acid sequences are obtained from sequence
Prediction by using trained neural network for individual MHC allele
Peptides having score > selected cut off value are predicted as binders
Peptides having score < selected cut off value are predicted non-binders
The quantitative matrix is applied on predicted binder at very high threshold (means non-stringent) e.g. 10%
The quantitative matrix is applied on predicted non-binder at very low threshold (means highly stringent) e.g. 1%
Binders
Non-binders
Binders
Non-binders
Predicted MHC binders
Predicted MHC non-binders

Fig. 4.6 Diagrammatic representation of combining ANNs and QM

Working Steps

I. ReadSeq developed by Dr. Don Gilbert has been implemented in the server, so input data can be the protein sequence query of any standard format.
II. For 47 MHC class I alleles, quantitative matrices are developed that are further assessed via jackknife validation test.
III. For each amino acid from point one to nine, coefficient value has been calculated via allocating the possibility of an amino acid at an exact point in binders as well as in non-binders.
IV. For prophecy of proteasomal cleavage sites which befall at the midpoint of 12mer peptides mainly six amino acids away from N-terminal, proteasomal and immunoproteasomal matrices are acquired from ProPred I server [34].

Pros and Cons

- The server is user-friendly, and its outcome demonstration format (HTML-II) is helpful in tracing promiscuous MHC-binding regions as of antigenic sequence with fair accuracy.
- However, certain limitations are also there like the incapability to handle nonlinearity in data because of significant confinement of quantitative matrix-based method. Also, the ANN-based method requires a large dataset for training.
- Proteasome cleavage site prediction procedures are less authentic due to extensive specificity of the proteasome in comparison of MHC-peptide binding specificity. Proteasome digested data are present in limited amount as well. Moreover, cleavage specificity depends on cleavage site-residing residues as well as on neighboring residues equally.

4.3.2.2 ProPred1

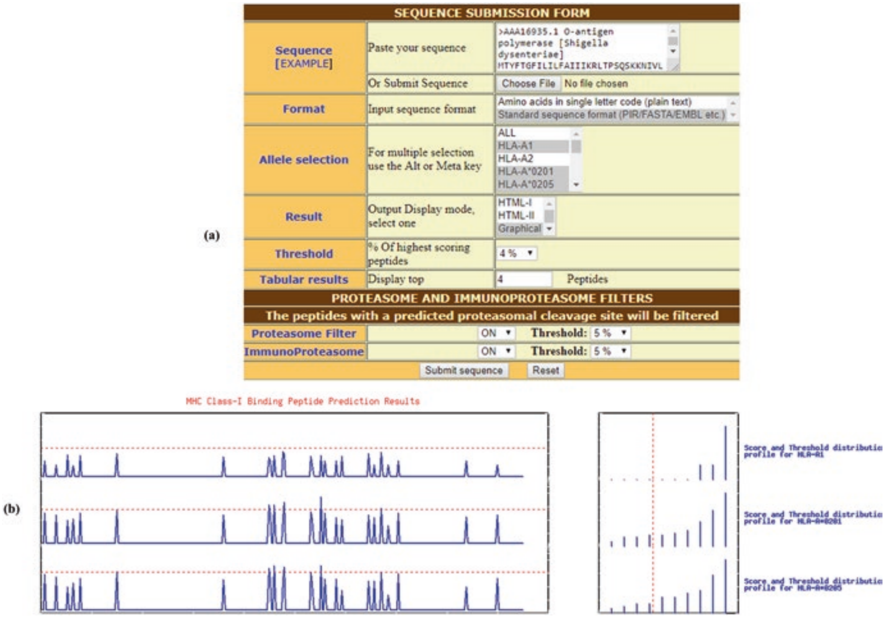
ProPred1 is an online matrix-based Web server in order to predict peptide binding to 47 MHC class I alleles. Matrices implemented have been acquired from BIMAS server as well as from literature. Results are in a user-friendly format that helps out users to identify promiscuous MHC binders in an antigen sequence.

The server enables users to predict MHC binders in an antigenic sequence along with their usual proteasome and immunoproteasome cleavage sites at C terminus simultaneously which results in identifying T-cell epitope with high potency.

PERL is used for writing a common gateway interface (CGI) script and is launched via Apache Web server. Further, Sun Server (420E) with a UNIX (Solaris 7) environment is used for installation.

Working Steps

I. Input data is the primary amino acid sequence of protein query in any frequently used sequence formats as the server uses ReadSeq to analyze input sequence (Fig. 4.7a).



SEQUENCE SUBMISSION FORM		
Sequence [EXAMPLE]	Paste your sequence	>AAA16935.1 O-antigen polymerase [Shigella dysenteriae] MTYFTGFILILFAIIIKRLTPSQSKKNIVL
	Or Submit Sequence	Choose File No file chosen
Format	Input sequence format	Amino acids in single letter code (plain text) Standard sequence format (PIR/FASTA/EMBL etc.)
Allele selection	For multiple selection use the Alt or Meta key	ALL HLA-A1 HLA-A2 HLA-A*0201 HLA-A*0205
Result	Output Display mode, select one	HTML-I HTML-II Graphical
Threshold	% Of highest scoring peptides	4 %
Tabular results	Display top	4 Peptides
PROTEASOME AND IMMUNOPROTEASOME FILTERS		
The peptides with a predicted proteasomal cleavage site will be filtered		
Proteasome Filter	ON	Threshold: 5 %
ImmunoProteasome	ON	Threshold: 5 %
Submit sequence	Reset	

(a)
(b)
MHC Class-I Binding Peptide Prediction Results
Score and Threshold distributio profile for HLA-A1
Score and Threshold distributio profile for HLA-A*0201
Score and Threshold distributio profile for HLA-A*0205

Fig. 4.7 (**a**) Sequence submission form of ProPred1 server showing protein sequence of O-antigen polymerase of *Shigella dysenteriae* as an input. (**b**) Prediction result in graphical format

II. There is an independency to select a threshold value for prediction.

III. Representation of output data in graphical (Fig. 4.7b) or text form provides assistance to the user in appropriate recognition of promiscuous MHC-binding domains in their query sequence.

IV. Firstly, for a given antigen sequence, all probable overlying 9mer peptides are produced followed by a quantitative matrix-based score calculation of selected MHC alleles. A peptide is designated as predicted binder if their score would be superior to a particular threshold value (e.g., at 4%) for selected MHC allele.

V. In an effort to forecast proteasome cleavage sites in an antigenic sequence, overlying 12mer peptides were developed for sequence followed by their score calculation with the usage of weight matrix of the proteasome.

VI. Further peptides having score superior to a certain threshold value (e.g., at 5%) are deemed as peptides featuring proteasome cleavage site at their midpoint positions (6-position left and 6-position right) as per prediction.

VII. Prediction of the immunoproteasome cleavage site of peptides shares analogy with proteasome cleavage site prediction.

VIII. Concurrent anticipation of MHC binders and proteasome cleavage sites results in removal of MHC binders not retaining cleavage site at C terminus.

Pros and Cons

- Purpose of ProPred1 development is to efficaciously attenuate wet lab experiments number indulged in to identify effective T-cell epitopes and thereby develop relevant vaccines.
- However, due to lack of sufficient data for MHC non-binders, calculation of threshold value is little bit crucial.

4.3.2.3 TAPPred

TAPPred is a user-friendly, support vector machine (SVM)-based Web server designed to predict TAP-binding affinity as well as translocation efficiency of the peptide. The server is initiated via public domain software package Apache on Sun server 420R in Solaris background. HTML is used for writing all the Web pages, while PERL and JavaScript are used for inscription of CGI scripts. By utilizing freely downloadable software, SVM^{light}, SVM has been implemented.

Working Steps

I. Input data is protein sequence as a single-letter amino acid code whose minimum length should be nine that is uploaded as a local sequence file or is pasted in required space, in any of the standard formats because of integration of ReadSeq.

II. Before running prediction sequence, uploaded format must be chosen by the user that it is in either plain or formatted form as server acknowledges both formatted and unformatted raw antigenic sequences which results in erroneous prediction if the selected format is false.

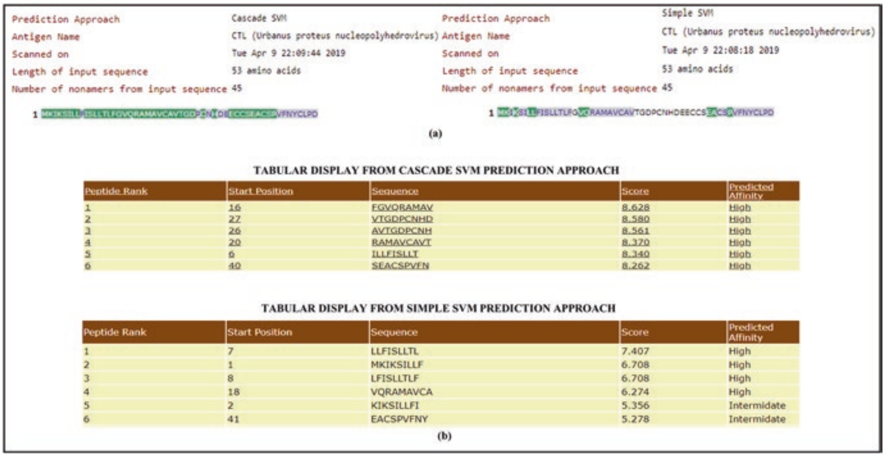
III. Prediction of binding affinity of the peptide has given permission by the server on the basis of two variants of SVM. Simple SVM involves prediction relied on sequential knowledge of peptides and is quicker than cascade SVM which includes characteristics of amino acids along with its sequential knowledge.

IV. Two tiers exist for prediction. Initially via joining characteristics of amino acids with sequential information, preliminary results are gained. Later on, the results of the first tier are further filtered. Despite having a slower rate of prediction, cascade SVM is more trustworthy as compared to simple SVM. Only a single approach can be selected for prediction at a time.

V. Results are depicted in two user-friendly formats. In the first format, the result is presented by coloring the residues. N-terminal is demarcated by the green color background of residues. Rest of the residues are represented with the violet-blue background (Fig. 4.8a).

VI. Type of peptides can be chosen to be displayed in the result.

VII. Tabular format display (Fig. 4.8b) has four alternatives. Only one output display can be selected at a time.



```
Prediction Approach                        Cascade SVM
Antigen Name                               CTL (Urbanus proteus nucleopolyhedrovirus)
Scanned on                                 Tue Apr 9 22:09:44 2019
Length of input sequence                   53 amino acids
Number of nonamers from input sequence     45

    1 MKIKSILLFISLLTLFGVQRAMAVCAVTGDPCNHDEECCSEACSPVFNYCLPD
```

```
Prediction Approach                        Simple SVM
Antigen Name                               CTL (Urbanus proteus nucleopolyhedrovirus)
Scanned on                                 Tue Apr 9 22:08:18 2019
Length of input sequence                   53 amino acids
Number of nonamers from input sequence     45

    1 MKIKSILLFISLLTLFGVQRAMAVCAVTGDPCNHDEECCSEACSPVFNYCLPD
```

(a)

TABULAR DISPLAY FROM CASCADE SVM PREDICTION APPROACH

Peptide Rank	Start Position	Sequence	Score	Predicted Affinity
1	16	FGVQRAMAV	8.628	High
2	27	VTGDPCNHD	8.580	High
3	26	AVTGDPCNH	8.561	High
4	20	RAMAVCAVT	8.370	High
5	6	ILLFISLLT	8.340	High
6	40	SEACSPVFN	8.262	High

TABULAR DISPLAY FROM SIMPLE SVM PREDICTION APPROACH

Peptide Rank	Start Position	Sequence	Score	Predicted Affinity
1	7	LLFISLLTL	7.407	High
2	1	MKIKSILLF	6.708	High
3	8	LFISLLTLF	6.708	High
4	18	VQRAMAVCA	6.274	High
5	2	KIKSILLFI	5.356	Intermidate
6	41	EACSPVFNY	5.278	Intermidate

(b)

Fig. 4.8 Prediction results from TAPPred server for CTL as an input sequence. (**a**) Displaying result in the form of colors. (**b**) Tabular display format

VIII. Only one output display can be selected by the user at a time that includes primarily a header and has data about the length of the peptide sequence, about nonamers obtained, as well as the date of prediction.

Pros and Cons

- The user can select parameters of their choice in this server.
- However, due to insufficient data for TAP-binding peptides, limited algorithms are there. Also, the minimum length of the query sequence should be nine; otherwise, it won't be accepted for prediction.

4.3.2.4 ProPred

ProPred is a graphics-based Web server in which matrix-based prediction algorithm has been deployed along with the implementation of amino acid or position coefficient table inferred from literature in order to foretell binding domain for MHC class II in antigenic sequences. Either as peaks in graphical interface or as colored residues in HTML interface, predicted binders can be envisioned. It has been developed mainly for 51 HLA-DR alleles whose matrices have been extracted from a pocket profile database defined by Sturniolo et al. in 1999 [33].

Working Steps

I. Input data is protein sequences in fasta or PIR format which are generally used as standard sequence formats and can be uploaded as a file.
II. In order to attain desirable results, selection of alleles, threshold, and other parameters are customizable.

III. An output as text or graphics is generated from the analysis of sequence data in which two choices have been provided by text display: the first choice in which binding regions of antigenic sequences are displayed by different colors thus providing easier detection. An option of representing binding score in a commonly used tabular format is also there that has been calculated from the matrix.
IV. The second choice involves the representation of coinciding regions independently on discrete lines; thus, delineation of specific regions from display becomes easier.
V. GDPlot library established by Lincoln D. Stein is used for graphics formulation in GIF format. HLA-DR-binding tendency laterally with the primary structure of a protein is represented as an output along with their binding strength. Consequently, it has an advantage over text presentation.
VI. Besides this, an alternative method is there for plotting threshold versus binding peptides, i.e., threshold profile, which renders assistance in the selection of a reasonable threshold value for finding promiscuous binders.

Pros and Cons

- All HLA-DR alleles are evaluated by server independently, and output is posted on a single screen that helps out the user in rapid visualization of promiscuous binders. Henceforth, it can be considered as a useful tool.
- Binding strength for all peptide frames in an opted subsequence can be computed by this server.

However, it is less expressive in representing overlapping binding regions.

4.3.2.5 EpiDOCK

EpiDOCK is the first structure-based server for prediction of peptide binding to 23 utmost common human MHC class II proteins which include 5 HLA-DP, 6 HLA-DQ, and 12 HLA-DR proteins. These alleles are the composition of more than 95% of the human population. The server is implicated to identify 90% of true binders as well as 76% of true non-binders, with a global precision of 83%.

Working Steps

I. Input data is protein sequence in fasta format. Multi-fasta protein format is likely reinforced.
II. Selection of HLA class II protein of concern is the next step that can be a single protein or all proteins.
III. Peptide-binding core is composed of nine adjacent residues due to which a collection of overlapping nonamers is formed as a result of input sequence conversion. A docking score-based quantitative matrix (DS-QM) is used for assessment of all nonamers retrieved for certain HLA class II protein and allotted a specific score.

IV. For any DS-QM, thresholds are defined with utmost certainty. Peptides having higher scores than the threshold or equal to them are expected to be binders, else considered as non-binders.

V. After that, if prophesied nonamer binder is a portion of recognized binder sequence, only then it will be categorized as an accurately foretold binder, else referred to as a false binder. Data is reported either in xls or csv formats.

VI. To validate anticipations, a test set of 7050 identified binders to HLA-DR, HLA-DQ, and HLA-DP proteins is implicated that originates from 1195 proteins, which is collected from Immune Epitope Database.

VII. Assigned values for specificity, sensitivity, accuracy, and AUC are 0.759, 0.903, 0.831, and 0.892, respectively.

Pros and Cons

- Structure-based approaches require information about peptide-MHC protein complex centered on their X-ray structure only rather than extensive preexisting experimental data.
- It is authentic and credible.
- Because of high resource implications of experimental testing at the time of scanning large proteome, a number of false positives can be more in contrast to a large number of false negatives which is a major problem to be dealt with.
- Amino acids having negative coefficients decrease the affinity of peptides for HLA-DRB1.

References

1. Ansari HR, Raghava GP (2010) Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res 6:6. https://doi.org/10.1186/1745-7580-6-6
2. Atanasova M, Patronov A, Dimitrov I et al (2013) EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. Protein Eng Des Sel 26:631–634. https://doi.org/10.1093/protein/gzt018
3. Bhasin M, Raghava GPS (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci 13:596–607. https://doi.org/10.1110/ps.03373104
4. Bhasin M, Raghava GPS (2007) A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. J Biosci 32:31–42. https://doi.org/10.1007/s12038-007-0004-5
5. Bhasin M, Singh H, Raghava GPS (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics (Oxford, UK) 19:665–666
6. Bhasin M, Lata S, Raghava GPS (2007) TAPPred prediction of TAP-binding peptides in antigens. Methods in molecular biology (Clifton, NJ):381–386
7. Blum JS, Wearsch PA, Cresswell P (2013) Pathways of antigen processing. Annu Rev Immunol 31:443–473. https://doi.org/10.1146/annurev-immunol-032712-095910
8. Eswar N, Webb B, Marti-Renom MA et al (2006) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics 15:5.6.1–5.6.30. https://doi.org/10.1002/0471250953.bi0506s15

9. Gershoni JM, Roitburd-Berman A, Siman-Tov DD et al (2007) Epitope mapping. BioDrugs 21:145–156. https://doi.org/10.2165/00063030-200721030-00002
10. Hammer GE, Gonzalez F, Champsaur M et al (2006) The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules. Nat Immunol 7:103–112. https://doi.org/10.1038/ni1286
11. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 78:3824–3828
12. Hopp TP, Woods KR (1983) A computer program for predicting protein antigenic determinants. Mol Immunol 20:483–489. https://doi.org/10.1016/0161-5890(83)90029-9
13. Jensen PE (2007) Recent advances in antigen processing and presentation. Nat Immunol 8:1041–1048. https://doi.org/10.1038/ni1516
14. Jespersen MC, Peters B, Nielsen M, Marcatili P (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 45:W24–W29. https://doi.org/10.1093/nar/gkx346
15. Lafuente E, Reche P (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr Pharm Des 15:3209–3220. https://doi.org/10.2174/138161209789105162
16. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158
17. Madden DR (1995) The three-dimensional structure of peptide-MHC complexes. Annu Rev Immunol 13:587–622. https://doi.org/10.1146/annurev.iy.13.040195.003103
18. Molero-Abraham M, Lafuente EM, Reche P (2014) Customized predictions of peptide–MHC binding and T-cell epitopes using EPIMHC. Methods in molecular biology (Clifton, NJ):319–332
19. Nielsen M, Lundegaard C, Worning P et al (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics 20:1388–1397. https://doi.org/10.1093/bioinformatics/bth100
20. Paul WE (2013) Fundamental immunology. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia
21. Petersen B, Petersen T, Andersen P et al (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 9:51. https://doi.org/10.1186/1472-6807-9-51
22. Pomés A (2010) Relevant B cell epitopes in allergic disease. Int Arch Allergy Immunol 152:1–11. https://doi.org/10.1159/000260078
23. Ponomarenko JV, Bourne PE (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol 7:64. https://doi.org/10.1186/1472-6807-7-64
24. Ponomarenko J, Bui H-H, Li W et al (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 9:514. https://doi.org/10.1186/1471-2105-9-514
25. Reche PA, Reinherz EL (2007) Definition of MHC supertypes through clustering of MHC peptide-binding repertoires. Methods in molecular biology (Clifton, NJ):163–173
26. Regenmortel MHV (2009) What is a B-cell epitope? Methods in molecular biology (Clifton, NJ):3–20
27. Ruppert J, Sidney J, Celis E et al (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. Cell 74:929–937. https://doi.org/10.1016/0092-8674(93)90472-3
28. Saha S, Raghava GPS (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. Springer, Berlin/Heidelberg, pp 197–204
29. Saha S, Raghava GPS (2007) Prediction methods for B-cell epitopes. Humana Press, pp 387–394
30. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA (2017) Fundamentals and methods for T- and B-cell epitope prediction. J Immunol Res 2017. https://doi.org/10.1155/2017/2680160
31. Scaiewicz A, Levitt M (2015) The language of the protein universe. Curr Opin Genet Dev 35:50–56. https://doi.org/10.1016/j.gde.2015.08.010

32. Singh SP, Mishra BN (2016) Major histocompatibility complex linked databases and prediction tools for designing vaccines. Hum Immunol 77:295–306. https://doi.org/10.1016/j.humimm.2015.11.012
33. Singh H, Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17:1236–1237. https://doi.org/10.1093/bioinformatics/17.12.1236
34. Singh H, Raghava GPS (2003) ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics 19:1009–1014. https://doi.org/10.1093/bioinformatics/btg108
35. Singh H, Ansari HR, Raghava GPS (2013) Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One 8:e62216. https://doi.org/10.1371/journal.pone.0062216
36. Stern LJ, Wiley DC (1994) Antigenic peptide binding by class I and class II histocompatibility proteins. Structure (London, UK: 1993) 2:245–251. https://doi.org/10.1016/S0969-2126(00)00026-5
37. Sun B, Zhang Y (2014) Overview of orchestration of CD4+ T cell subsets in immune responses. Adv Exp Med Biol 841:1–13
38. Thornton JM, Edwards MS, Taylor WR, Barlow DJ (1986) Location of "continuous" antigenic determinants in the protruding regions of proteins. EMBO J 5:409–413
39. Vita R, Overton JA, Greenbaum JA et al (2015) The immune epitope database (IEDB) 3.0. Nucleic Acids Res 43:D405–D412. https://doi.org/10.1093/nar/gku938