Epitope scanning using virtual matrix-based algorithms

Laura Raddrizzani and Juergen Hammer
Date received in revised form: 18th February 2000

Abstract
Protein sequence and expression databases (transcriptomes) contain the information required to identify epitopes capable of generating protective immune responses in humans. A key event in the initiation of an immune response against disease is the presentation of antigenic peptide epitopes to T cells by human leukocyte antigen (HLA) molecules. Computational filtering tools that allow the prediction of HLA/epitope interaction can be applied to sequence databases to select for candidate epitopes, thus minimising the subsequent amount of laboratory work. Here, the basic principles of epitope prediction and a summary of the available prediction approaches are presented, with a particular emphasis on the use of algorithms based on virtual HLA-II quantitative matrices, capable of predicting promiscuous HLA-II ligands.

INTRODUCTION
T cells are the key mediators of specific immune responses against diseases such as allergies, infectious diseases and cancer. A crucial event in T-cell activation is the presentation of peptides derived from disease-related protein antigens in the context of major histocompatibility complex (MHC) molecules, in humans termed human leukocyte antigens (HLA). This event is accomplished by the intracellular fragmentation of disease-specific protein antigens, followed by binding of the derived peptide epitopes to HLA molecules and subsequent presentation on the cell surface of antigen presenting cells (APCs), for the recognition by T-cell specific receptors.1

A crucial step in the design of subunit vaccines is the identification of T-cell epitopes in sets of disease-specific gene products. The unprecedented availability of genome-scale DNA sequence information, coupled with the use of genome-wide expression monitoring tools, such as DNA microarrays, has dramatically increased the number of possible disease-specific antigens. Experimental approaches to identify T-cell epitopes within these antigens, i.e. the synthesis and assay of overlapping peptides from proteins of interest, are not applicable to a large number of protein sequences. However, computer models capable of simulating and predicting the biological process of antigen presentation can be used to minimise the number of experiments, thus enabling a systematic scanning for candidate T cell epitopes (Figure 1).

Advances in understanding antigen presentation at the molecular level accelerated the development of computer models capable of predicting T-cell epitopes. Although some of the molecular aspects of antigen presentation are insufficiently defined to be of value for epitope prediction models, other models, such as the interaction of peptide fragments with HLA molecules, were characterised thoroughly. As a consequence, most of the currently available epitope prediction models are based on HLA peptide binding data. This approximation is supported by the observation that HLA peptide binding is a major bottleneck in the selection of epitopes, as indicated by the finding that most peptide sequences lack the capacity to interact with HLA molecules.
Figure 1: The combined use of genome-wide expression screening and prediction filtering tools for the creation of disease-specific candidate epitope databases. (a) Scheme of the application of subsequent genomic and computational tools for the identification of T-cell epitope sequences suitable for subunit vaccine design. (b) In this example (previously published in ref. 2), Affymetrix GeneChips™ were used to identify gene transcripts upregulated in tissue samples derived from colon cancer patients. The initial number of protein sequences to be screened for epitope identification was thus reduced from ~19,000 (number of gene sequences and contigs represented on the DNA microarrays used for this analysis) to 34 gene products. That is, those found to be upregulated in at least 30 per cent of the primary colon cancer tissues analysed. Analysis of this set of protein sequences using the HLA-II virtual matrix-based algorithm TEPITOPE identified 130 promiscuous candidate T-cell epitope sequences, which represent a manageable amount of data for subsequent laboratory testing. The left and middle columns in the histogram show an estimate of the number of peptides corresponding to the gene/protein sequences analysed.

HLA-BASED FILTERING TOOLS FOR T-CELL EPITOPE PREDICTION

The human MHC locus encodes two major classes of highly polymorphic molecules, designated HLA class I (HLA-I) and class II (HLA-II). Extensive crystallographic and binding studies revealed similar conformations of HLA ligands despite enormous differences in primary structure. HLA-I molecules interact with the peptide N- and C-termini, imposing a typical bulge on the peptide chain and allowing only peptides of a defined length (8–10 amino acid residues) to be accepted for binding. The binding groove in HLA-II molecules has open ends and peptides are bound in an extended conformation, allowing the peptides to be more variable in length (typically, 12–28 amino acids).

Polymorphic pockets and depressions in the HLA binding groove interact with several peptide side chains of the peptide core sequence, explaining the specificity of HLA/peptide interaction.

Several approaches have been developed for HLA-based epitope prediction, reflecting both the different characteristics of HLA-I/II peptide interaction, and the increasing structural and functional information that became available over the last decade. The main epitope prediction methods are listed below:

- Sequence similarity models are based on sequence homology search algorithms, ie the sequence alignment of a query protein sequence to known HLA-binding peptides. Publicly available programs such as BLAST can be used (Table 1). This method has proved useful in the identification of candidate epitopes for molecular mimicry, but has very poor generalisation properties.
- **Molecular models** are visualisation tools based on calculations that describe molecular interactions at the atomic level. In few cases, they can provide detailed explanation of peptide binding,\(^\text{13}\) but are subject to cumulative errors and cannot be used for large-scale prediction efforts.

- **Binding motifs** describe general position-based patterns of recurrent amino acids favourable for HLA peptide binding. They indicate which positions and which amino acids are important for binding;\(^\text{14-17}\) however, prediction models based on binding motifs are mostly all-or-nothing algorithms with very high false negative rates.\(^\text{5}\) Prediction algorithms based on HLA-I and II binding motifs are accessible on the Web (Table 1).

- **Artificial Neural Networks** (ANNs) are self-training systems that are able to extract and retain patterns present in the submitted data and subsequently recognize them in a previously unseen input. ANNs can 'learn', i.e., they are capable of automatic refinement as new sets of data become available.\(^\text{18,19}\) ANN-based models have proven very effective for the prediction of HLA-I ligands.\(^\text{20,21}\) Also for HLA-II ligands, the prediction accuracy is slightly higher than, for example, with matrix-based systems (see below). ANN models are limited by their requirement of large amounts of data (not available for most HLA alleles), and by their need for alignment matrices,\(^\text{18}\) which are required to train the HLA-II prediction systems. In addition, promiscuous prediction is not feasible since data for one allele

---

**Table 1: Epitope prediction algorithms accessible on the World-Wide Web**

<table>
<thead>
<tr>
<th>Type of algorithm</th>
<th>Name</th>
<th>Alleles/subtypes covered</th>
<th>Analysis output</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence similarity-based</td>
<td>BLAST</td>
<td>HLA-I, HLA-II</td>
<td>Alignment results are listed with a relative score to matching protein sequences</td>
<td><a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a></td>
</tr>
<tr>
<td>Binding motif-based</td>
<td>SYFPEITHI</td>
<td>11, 3</td>
<td>Peptides are listed separately for each selected HLA allele, ranked according to a relative score and aligned based on the main anchor positions</td>
<td>access via: <a href="http://www.uni-tuebingen.de/uni/kdd/">http://www.uni-tuebingen.de/uni/kdd/</a></td>
</tr>
<tr>
<td>EpiMatrix, EpiMer</td>
<td>na(^*), na(^+)</td>
<td></td>
<td>Single-allele or clustered predictions available(^\text{6})</td>
<td><a href="http://www.brown.edu/Research/TB-HIV_Lab/epimatrix/">http://www.brown.edu/Research/TB-HIV_Lab/epimatrix/</a> epimatrix.html</td>
</tr>
<tr>
<td>ANN-based</td>
<td>PREDICT</td>
<td>na(^+), na(^+)</td>
<td></td>
<td><a href="http://sdmc.krdl.org.sg:8080/predict/">http://sdmc.krdl.org.sg:8080/predict/</a></td>
</tr>
<tr>
<td>Quantitative matrix-based</td>
<td>BIMAS-HLA</td>
<td>33</td>
<td>For a single selected allele, peptides are listed and ranked according to a relative score; a summary of the analysis parameters is displayed</td>
<td><a href="http://bimas.dcterh.ch/gov:80/molbio/hla_bind/">http://bimas.dcterh.ch/gov:80/molbio/hla_bind/</a></td>
</tr>
<tr>
<td>Peptide Binding Predictions</td>
<td>TEPITOPE</td>
<td>25</td>
<td>Results are displayed simultaneously for all the selected alleles (see Figure 4) and included in a report file with a summary of the analysis parameters</td>
<td><a href="http://www.vaccinome.com%5C(%5E%5Cddagger%5C)">http://www.vaccinome.com\(^\ddagger\)</a></td>
</tr>
</tbody>
</table>

The limitations of the methods used by the above applications are mentioned in the section 'HLA-based filtering tools for T-cell epitope prediction'; additional information may be available on the respective web sites.

na: information not available.

*Both HLA-I and HLA-II alleles are covered by this algorithm; complete information can be obtained by requesting an analysis.

\(^\ddagger\) The PREDICT application will soon be available.

\(^\ddagger\) The TEPITOPE software can be obtained from the authors.
Figure 2: Quantitative matrix data can be split into single pocket profiles for the assembly of virtual matrices. (a) Schematic representation of the building process of quantitative matrices (see text). (b) Once a substantial pocket profile database is generated, virtual matrices can be assembled via HLA pocket homology\(^2\) (see text). Profiles for peptide positions 5 and 8 are not considered owing to their minimal effects on HLA-II binding.

cannot be generalised to other alleles. ANN-based prediction tools will soon be available on the Web (Table 1).

- **Hidden Markov models** (HMMs) are highly effective systems capable of generating multiple alignments and detecting conserved patterns in families of sequences; HMMs were recently reported for the prediction of HLA-II ligands.\(^2\)

- **Quantitative matrices** provide very detailed models in which the contribution to binding of each amino acid at each position within a binding core of peptide is quantified (Figure 2a). The position-specific amino acid values reflect the structural properties of HLA alleles, therefore constituting a sort of "fingerprint" of HLA binding domains. Quantitative matrix-based prediction systems are linear models that are easy to implement, and that result in a binding score for each query peptide.\(^6,23,24\) Their predictive accuracy is considerable, but they have less capacity to encode non-linear dependencies.\(^25\) Quantitative matrix-based prediction tools are
Epitope scanning

available on the Web for the prediction of HLA-I ligands (Table 1).

- **Virtual matrices**, like quantitative matrices, provide a detailed model in which the contribution to binding of each amino acid (ligand) with each pocket/position (HLA binding cleft) is quantified (Figure 2b). However, while quantitative matrices are determined individually for any given HLA allele, virtual matrices are formed by assigning and combining pocket-specific quantitative binding values derived from one HLA allele to other alleles via HLA sequence comparison (Figure 2; see also below). The advantage over quantitative matrices is that virtual matrices address the problem of HLA polymorphism and enable the systematic prediction of peptide ligands for a broad range of HLA binding specificity (= promiscuous peptides). The prediction of promiscuous binding ligands is considered to be a prerequisite for most subunit vaccine design strategies. Virtual matrix-based prediction models have been validated for HLA-II in several retrospective studies.\(^{25}\) Furthermore, they have been successfully applied to predict T-cell epitopes in the context of oncology,\(^{26}\) allergy\(^{27}\) and autoimmune diseases.\(^{12}\) The only virtual matrix-based epitope prediction tool available to date, TEPITOPE (Table 1), can be obtained from the authors. Virtual matrix-based prediction models for HLA-II molecules are described in detail in the following paragraphs.

THE CONCEPT OF HLA-II MATRICES

HLA-II molecules are extremely polymorphic. This generates a high degree of structural variability that profoundly affects HLA-II allele binding specificity. In recent years, however, extensive structural and functional studies improved our understanding of HLA-II/peptide interaction and led to the definition of a number of rules:

- The core sequence of the peptide interacting with the HLA-II binding cleft is nine amino acids long, and allows a flexible number of flanking peptide residues.
- The peptide positions interacting with HLA-II molecules are regularly spaced and exhibit higher frequencies of particular amino acid side chains, termed anchor residues.
- HLA-II polymorphic residues are mainly localised in the peptide binding cleft and often contribute to form the pockets and depressions that accommodate parts of the peptide ligand. This causes such pockets to have different size and chemical characteristics in different allelic variants, resulting in the preference for different ligand side chains. Some of the ligand peptide side chains that favourably interact with the residues forming these pockets increase the overall ligand binding affinity and specificity (anchor residues); other ligand side chains may interfere with the pocket architecture, thus reducing binding (inhibitory residues).\(^{38}\) With the opportune experimental design, these effects can be quantified.

- The effect of anchor residues is HLA allele-specific and depends on their relative position within the peptide binding frame, rather than on neighbouring residues, leading to the approximation of the independence of single amino acid effects on the overall peptide binding affinity.

- All these features allow a considerable variability in the peptide ligand sequences.

Quantitative matrices (see also above) provide a numerical description of the
interaction of all the amino acid side chains in a 9-mer peptide string, by assigning a position-specific coefficient to the effect of each side chain on HLA-II binding (Figure 2a). For each HLA-II allele, a quantitative matrix can be determined and represents a valuable tool for candidate epitope peptide prediction. However, the measurement of all effects of amino acids (PS) at each peptide position implies a huge experimental effort, involving multiple binding assays of hundreds of single-substituted peptides for each HLA-II allelic variant analysed. HLA polymorphism is so extensive that, in order to cover the majority of the human HLA-II binding specificity, the quantitative matrices for tens to hundreds of alleles should be determined, even if only focusing on the main HLA-II isotype, HLA-DR. A shortcut to the solution of this problem came from these further observations:

- Owing to the approximation of the independence of the effects of amino acid side chains from the effects of the neighbouring residues, a quantitative matrix can be split into its single components, or ‘pocket profiles’. Pocket profiles constitute the quantitative representation of the interaction of all natural amino acid residue with a single given HLA-II pocket (Figure 2). Pocket profile data are determined in competitive HLA-II binding experiments using sets of frame-fixed single-substituted designer peptides.

- A pocket profile can be associated to the respective HLA-II residue pocket composition; comparative studies of known pocket profiles demonstrated that pockets having the same residue composition show similar profiles.2

The consequence of these observations is that, once a pocket profile has been determined in vitro, it can be shared among other HLA-DR alleles, as long as their amino acid residues contributing to the pocket are identical. Profiles for identical pockets are therefore recycled from a pool rather than determined repeatedly for each allele. Thus, a relatively small number of profiles can be used to build a large number of HLA-DR matrices, named virtual matrices (Figure 2b). Extensive experimental work allowed the definition of pocket profiles for tens of different polymorphic pockets that were combined to build > 50 HLA-DR virtual matrices.2 These virtual matrices form the basis of the epitope prediction software TEPITOPE.

**BASIC PARAMETERS USED BY THE TEPITOPE SOFTWARE**

The TEPITOPE software utilises virtual matrix data in a linear prediction model,23 in which peptide conformation is entirely neglected. The only parameters describing peptide binding are peptide frame (PF), peptide side chain effects on binding (PS) and peptide length (PL). The basic analysis performed by TEPITOPE occurs in the following steps (Figure 3). First, all possible PF are extracted from the input protein sequence. Second, PS values are assigned to each residue within the PF. The position-specific and peptide side chain-specific PS values are derived from the virtual matrices (Figure 2b). Third, the linear combinations of all PS values for each PF from relative positions 1 to 9 result in a numerical value, or peptide score (PSC). The calculated PSC has been shown to correlate with the
Figure 3: Schematic representation of virtual matrix-based algorithms for the identification of candidate T-cell epitopes (see 'Basic parameters used by the TEPITOPE software'). If the score of the analysed PF is equal or higher than the score values belonging to the percentage threshold value initially selected by the user (e.g. 1 percent in this diagram), the output of the algorithm is the prediction of peptides containing such PF as candidate HLA-II epitopes.

- PF is a string variable, that is PF data consist of strings nine amino acids long (single letter code) corresponding to the nonamer core peptide frame of HLA-II ligands. Every PF within a given protein sequence will invariably start from amino acids having a hydrophobic side chain (i.e. aliphatic or aromatic), because this is a necessary requirement for the relative position 1 (P1) of an HLA-DR binding peptide. Any protein sequence

binding affinity of HLA-DR ligands, thus representing a measure for the likelihood of a peptide containing a given PF to bind to HLA-II molecules.6

None of the above parameters is visible or requires interference by the user when running the TEPITOPE software. However, since a basic understanding of these parameters improves the interpretation of the prediction results, additional information is provided for each one of these parameters.
Percentage threshold definition

A stretch that is longer than nine amino acid residues can contain multiple overlapping PFs. For simplification, TEPITOPE initially considers only one PF at a time, and integrates the results of multiple PFs via a visual user interface (Figure 4b). Thus, the capability of a peptide frame to bind to HLA-II molecules rather than that of the whole peptide sequence is predicted.

Peptide length

- **PL** affects binding of an HLA-II ligand, as previously shown by *in vitro* peptide truncation studies. These studies also demonstrated that PL could be neglected in a predictive model, when the nonamer–peptide core sequence is flanked by at least two amino acid residues on each side. This is the case for most peptide frames within a protein sequence, except for frames at the beginning or end of the query peptide or protein.

Peptide side chain effects

- **PS** values are derived from the corresponding virtual matrices and are assigned based on the letter and position within a given PF (Figure 2). They are relative binding data, i.e., they were calculated by normalizing experimental IC50 data with IC50 values obtained for alanine at the same peptide position. The data are expressed as the logarithm of alanine-normalised data. Consequently, the linear combination of PS values for PSC calculation is the addition of PS values (Figure 3).

Percentage threshold selection

A parameter that needs to be selected by the user is the 'percentage threshold'. The threshold is defined as the 'percentage of best scoring natural peptides'. For example, a threshold of 1 per cent would predict peptides in any given protein sequence which belong to the 1 per cent best-scoring natural peptides (Figure 3). The threshold values are derived as follows: (i) PSC for all valid PF in a representative database of natural protein sequences are calculated for any given virtual matrix. (ii) The distribution of PSC values within this database is checked; a database of sufficient size should generate a Gaussian-like distribution. (iii) PF are sorted based on their PSC, and the PSC values corresponding to the 1, 2, 3 per cent, etc. best-scoring peptides are determined (Figure 3). Thus, the threshold correlates with the PSC value and is therefore an indicator for the likelihood that the predicted peptides are capable of binding to a given HLA molecule.

The percentage threshold parameter allows the user to select for different stringency levels, in order to modulate the prediction results: a lower threshold corresponds to a higher stringency prediction, i.e., to a lower rate of false positives and to a higher rate of false negatives. In contrast, a higher threshold value (low stringency) corresponds to a higher rate of false positives and a lower rate of false negatives. In short, from the same protein sequence input, a threshold setting of 1 per cent will predict a lower number of peptide sequences and for a lower number of HLA-II alleles, compared with 2 per cent or higher thresholds (see example in Figure 4c). However, this will also ensure a higher likelihood of positive downstream experimental results. In different cases, the choice of the percentage threshold value may vary based on the overall context. These include:

- The type of analysed sequence (e.g., protein subunit vs. whole viral genome).
- The downstream epitope validation capabilities including economic considerations (e.g., number of *in vitro* stimulation assays that can be performed).
- The specific question asked (e.g., search for allele-specific or promiscuous epitopes).
Some guidelines are also provided by previously published experimental validation data. Normally, at least for a first round of screening, threshold values higher than 3 per cent are not desirable, since the rate of false positives can increase the size of the predicted repertoire to an amount unacceptable for later experimental testing.

**EPITOPE PREDICTION USING TEPITOPE**

An example of the identification of candidate T cell epitopes using TEPITOPE is illustrated in Figure 4. The sequence of the human myelin basic protein (MBP), a putative autoantigen involved in the pathogenesis of multiple sclerosis, is imported from a protein database source such as SWISS-PROT and TrEMBL (accessible via http://www.expasy.ch/) and displayed in the TEPITOPE visual user interface. After selection of the desired HLA-DR virtual matrices (listed on the left side of the screen Figure 4a), and the percentage threshold value (top left corner of the screen), scanning of the MBP sequence reveals an HLA-DR promiscuous area (MBP 87-103, underlined; Figure 4b). The results are simultaneously shown for each of the selected alleles by scrolling back and forth along the protein sequence. Interestingly, the DRB1*15 subtypes, specifically associated with susceptibility to develop multiple
sclerosis, are predicted to bind a peptide frame (MBP 87-96) which is slightly different from the one recognized by most other alleles; this finding is documented in numerous publications (for a review, see ref. 5). The quantitative analysis of the selected sequence (for any length 9 amino acids; Figure 4c) displays the threshold levels at which this peptide would be predicted for each HLA-DR virtual matrix: here, at the most stringent threshold (1 per cent), TEPITOPE would identify MBP 87-96 as a potential binder only for the DRB1*15 subtypes. For each of the selected peptides, a reporting function included in the software automatically generates a quantitative analysis together with a record of the parameters used for the prediction analysis. The experienced user will require approximately 1 - 5 minutes to scan a 1000 amino acid long protein sequence.

A qualitative analysis (Figure 4d) can also be performed by TEPITOPE for each PF contained in the selected sequence (only for core length = 9 amino acids). This function displays a 'summary at a glance' of all the anchor (A) or inhibitory (I) residues present at each position of the PF listed for each of the HLA-II alleles, together with eventual additional information on the type of hydrophobic residue at P1. A and I residues are defined as residues that have at least a 10 fold positive or negative effect on binding, respectively, based on the alanine baseline in virtual matrices. In this case, the specific peptide frame MBP 87-96 contains I residues at different positions for every HLA-DR allele, except for the DRB1*15 subtypes.

CONCLUSION

The application of computer models is increasingly changing the way biological knowledge is generated, providing tools to explore large biological databases in cases in which suitable laboratory experiments cannot be performed. In this paper, we analysed how the biochemical characterisation of HLA/peptide interaction constituted the basis for the design of candidate T-cell epitope prediction models. This resulted in the generation of a number of powerful bioinformatic tools that can lead to improved efficacy in the process of subunit vaccine design. The most innovative feature of virtual matrix-based prediction approaches is the possibility of generalising the prediction model by combining pre-existing pocket profile data to generate new virtual matrices. The only virtual matrix-based software available to date, TEPITOPE, is able to perform a simultaneous analysis for all the included HLA-II matrices. This has two important advantages: the antigenic potential of any protein sequence can be more rapidly predicted, and the promiscuity of the selected epitopes, a key factor in vaccine design, can be immediately evaluated 'at a glance' for all the HLA-II alleles of interest.

References

7. Madden, D. R. et al. (1991), 'The structure of HLA-B27 reveals nonamer self-peptides...


