MHC-Peptide binding prediction for epitope based vaccine design

Bing Zhao¹, Kishore R. Sakharkar²³, Chu Sing Lim³⁴, Pandjassarame Kangueane⁵, Meena K. Sakharkar³⁴*¹

¹ Department of Biochemistry, University of Western Ontario, Canada
² National University Medical Institutes, National University of Singapore, Singapore
³ Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore
⁴ Biopharmaceutical Engineering Cluster, BMERC, Singapore
⁵ Biomedical-Informatics, India

Submitted: 23 Aug. 2007; Accepted: 15 Oct. 2007

Abstract
The identification of MHC binding peptides for consideration as potential T-cell epitopes has application in peptide vaccine design and immunotherapy. Determination of MHCp binding using experimental measures is time-consuming and expensive. Therefore, efficient prediction models are required that facilitate systematic computational scanning of microbial genome for candidate T cell epitopes. These prediction models are either sequence or 3D structure based. This review provides a comparative analysis on the available prediction models with specific emphasis on the salient features of each model, its prediction efficiency, and merits and demerits.

Keywords: MHC-peptide, Binding, Prediction, Models, HLA-peptide.

INTRODUCTION

In the past century, medical research has improved health and increased life expectancy largely because of success in preventing and treating infectious diseases. Vaccines, in particular, offer protection against infectious diseases (Ada 2007). However, new threats to health continually emerge as bacteria and viruses evolve, are transported to new environments, or develop resistance to drugs and vaccines (Sawyer et al., 2006). Also, the current threat of the potential use of viruses and bacteria as biological weapons demands efficient methodologies that analyze genetic information related to potential viral and bacterial threats, identify potential vaccine target, and then develop diagnostic tests and vaccines for them.

With the advances in genomics, understanding of pathogen variability as well as of the diversity of the human immune system, has greatly improved. The availability of human genome data and various microbial genome sequences provides opportunity to establish computational methodologies to meet the demands. To date, genome sequences are available for approximately 500 viruses and bacteria, including those currently listed on the NIH and CDC priority pathogen list (Benson et al., 2007). With the advancement in our understanding of bacteria infections, virus replication, pathogenicity, and virus-host interactions and host-microbial interactions, on the basis of individual protein domains, individual genes, and whole genomes, a new trend in vaccine development has emerged, which focuses on the design of epitope-based vaccines. Compared to traditional vaccines, epitope based vaccines are more specific, safe, and easy to produce.

The key to success of these approaches is the prediction models for rapid scanning of pathogen genomes to identify effective T-cell epitopes.

*Corresponding author:
Meena K. Sakharkar, Ph.D.
Biopharmaceutical Engineering Cluster, BMERC,
Nanyang Technological University,
Singapore
Email: mmeena@ntu.edu.sg
The major histocompatibility complex (MHC) plays a pivotal role in regulating immune response (Viret and Janeway 1999). During T cell activation, peptides derived from protein antigens are presented by MHC molecules. This process is accomplished by the intracellular fragmentation of protein antigens, followed by binding of derived peptides to HLA molecules and subsequent presentation on antigen presenting cell (APC) surface, for recognition by T cell receptors (TCR) (Pamer and Cresswell 1998). Only a small fraction of the derived peptides are involved in eliciting host immune response (Viret and Janeway 1999). More than 1,500 HLA alleles are known today (Robinson et al., 2003). The polymorphic HLA molecules not only present peptides for cell death and cytokine release, but are also involved in T-cell repertoire selection and self versus non-self discrimination (Viret and Janeway 1999). The short antigenic peptides, derived from parent molecules have been proposed as interesting targets of vaccine design (Ishioka et al., 1999). Thus, a new generation of high efficiency, low toxicity epitope-based vaccines is in development.

Accessing subdominant specificities might be of particular value in the case of tumor antigens, where self-tolerance might have inactivated T cells recognizing the most dominant specificities (Disis et al., 1996). This approach also allows elicitation of immune responses against multiple conserved epitopes, a factor of crucial importance in the case of rapidly mutating pathogens, such as HIV and HCV (Cooper et al., 1999). It is also possible to overcome limitations in recombinant protein delivery by combining epitopes from multiple antigens into a single immunogen (Kawashima et al., 1998). They can also increase potency and break tolerance (Sarobe et al., 1998). T cells epitopes are being investigated as potential immuno therapeutics in tumor treatment (Iwasaki and Barber 1998; Morgan et al., 1998).

The primary objective in these prediction methodologies is the calculation of MHCp binding because high affinity binding often correlates with immunogenicity (Kubo et al., 1994; Sette et al., 1994) and this depends on the stability of MHCp complexes (van der Burg et al., 1996). A crucial step towards the rational design of peptide vaccines is the identification of T-cell epitopes from disease causing antigen proteins. As experimental determination of MHCp binding is expensive, considerable effort has been made in the development of MHCp binding prediction models (Meister et al., 1995). Such predictions are generally non-trivial due to extensive MHC polymorphism and peptide diversity in the pathogen proteome. Prediction models are either based on sequence data or structure data (Table-1). Thus, prediction efficiency and generality (scope and spectrum) varies depending on the training data used in model development. In this review, we discuss the available prediction models for MHCp binding prediction and their scope and limitations.

**Sequence Based Predictions**

Sequence-based methods rely on the primary sequence of peptides that are known to bind specific MHC allele using binding assays. The information on sequence anchors that are deterministic in binding are encoded into a binding motif (D'Amaro et al., 1995; Meister et al., 1995; Rammensee et al., 1999; Rammensee et al., 1995; Reche et al., 2002), a position-dependent matrix (Gulukota et al., 1997; Sette et al., 1989; Udaka et al., 2000), Hidden Markov Model (HMM) (Brusic et al., 2002; Noguchi et al., 2002), Support Vector Machine (SVM) (Bhasin and Raghava 2004; Cui et al., 2006; Cui et al., 2007; Donnes and Elofsson 2002; Donnes and Kohlbacher 2006; Liu et al., 2006), stepwise discriminant analysis (Mallios 1999), or an Artificial Neural Network (ANN) (Adams and Koziol 1995; Brusic et al., 1998; Milik et al., 1998). These models are developed using peptide data in large databases derived from naturally bound peptides (Rammensee et al., 1999; Rammensee et al., 1995) or synthetic peptide libraries (Parker et al., 1994; Stryhn et al., 1996).

In these approaches, the binding affinity of each peptide residue is scored using a matrix taking into account the relative contribution of residues at other positions. The sum of contributions by all the residues gives predicted binding value. It should be noted that such matrices are to be generated for each known HLA alleles. Sequence based methods have two main limitations: (1) they assumes that different peptide positions contribute in an additive manner to the overall binding affinity and overlook the known interplay between different peptide side chains; (2) their real predictive power is directly dependent on the amount of experimental data used to interpolate MHC binding properties. These limitations are reflected in their low prediction efficiency. The sequence based
methods can be classified into few groups based on the utilization of statistical ‘techniques’ or ‘procedures’ in their development which are discussed below.

**Binding Motif Based Methods**

Binding motif based statistical methods attempt to work around the problem of handling non-linearity by using statistical or experimental methods to discover correlations between amino acids in different positions (Mallios 1999). It was shown that peptides binding to MHC class I molecules are generally 8-11 residues long and they are restricted at two positions of the sequence called anchor points (second or fifth and last position). Based on these anchor positions, simple binding motifs have been defined for specific MHC alleles (Rammensee et al., 1999), and general position based patterns of recurrent amino acids for specific MHC peptide binding have been generated. Based on pool sequencing method, expanded motifs are developed from a large number of naturally presented peptides that bind to a specific MHC molecule (Mallios 1999). The allele-specific motifs have been used for building predictive methods using weighted motif scoring system (Disis et al., 1996; Meister et al., 1995).

Binding motifs are simple to implement and easy to understand, but they are of modest accuracy. They are particularly useful for MHC alleles where not much experimental data are available. However, the compliance of a peptide sequence to such a binding motif is neither sufficient nor necessary to ensure binding (Brusic et al., 2004). Even more detailed binding motifs, derived from large-scale binding experiments, do not cover the full repertoire of peptides capable of binding to a certain MHC molecule. Moreover, prediction models based on binding motifs are mostly all-or-nothing algorithms with very high false negative rates. Considerations on the basis of the sequence alone are also not sufficient enough to predict binding of a peptide to MHC molecule.

Based on the consideration that factors other than size and anchor residues also determine peptide binding (Ruppert et al., 1993), an improved definition of MHC class I binding motifs that defined favorable binding residues at particular positions, as well as those that have negative effect was proposed. These refined motif definitions, together with weighting schemes of anchor positions, brought the concept of binding motifs closer to definition of quantitative matrices.

**Quantitative Matrix Based Methods**

Quantitative matrices provide a detailed linear model with easy to implement capabilities. In this method, the contribution to binding from each amino acid at each peptide position within the binding groove is quantified (Parker et al., 1994). It is assumed that: (a) each position within the peptide contributes independently in binding to an MHC molecule, and (b) a residue located at a given peptide position contributes an equal amount to binding, even within different peptides. This method involves producing a matrix in which every entry (X, Y) represents a score associated with amino acid residue X at position Y. The position-specific amino acid values reflect the structural properties of HLA alleles, therefore representing a fingerprint for HLA binding domains. Summing the scores for every residue in a given peptide can be used with appropriate formulae to yield a predicted binding score.

The quantitative matrix method represents an extension of binding motifs, and is simple and efficient to implement. The matrix method enables prediction for a wide pool of peptides in a high-throughput manner unlike motif based approaches. A serious limitation is in the generation of binding coefficient matrix for each MHC allele that requires the experimental testing of hundreds of peptides. Another limitation of this method is that it assumes that every amino acid residue in a certain position influences binding independently of its neighbors, and ignores the contribution of overall peptide structure to binding. Matrix methods can give quick predictions on the basis of simple patterns, but they have little ability to encode non-linear dependencies. This limitation does not render matrix-based methods completely inadequate, as there is certain simple generalization about amino acid preferences at specific positions influencing binding. Quantitative matrices’ ability to over-fit data and their bias towards the sets of peptides used to derive the matrix coefficients results in lowering the generalization properties of such methods. Quantitative matrices are thus reported as more accurate predictors than binding motifs (Brusic et al., 1998).
MHC-Peptide binding prediction for epitope based vaccine design

**Virtual Matrix Based Methods**

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 (Raddrizzani and Hammer 2000; Sturniolo et al., 1999). However, while quantitative matrices are determined individually for any given HLA allele, virtual matrices are obtained by assigning and combining pocket-specific quantitative binding values derived from one HLA allele to other alleles via HLA sequence comparison. 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 class II HLA alleles in several retrospective studies. Furthermore, they have been successfully applied to predict T cell epitopes in the context of oncology, allergy and autoimmune diseases (Cochlovius et al., 2000; Gross et al., 1998; Hammer et al., 1995; Stassar et al., 2001).

**Neural Network Based Methods**

Artificial Neural Networks (ANNs) are complex, nonlinear and self-training systems that are able to extract and retain patterns present in the training data and subsequently recognize them in a previously unseen input (Adams and Koziol 1995; Brusic et al., 1998; Milik et al., 1998). ANNs can capture complex relationships in the data sets and are not based on simplifying assumptions such as quantitative matrices. ANNs can tolerate a degree of erroneous data, and are excellent at classifying nonlinear data, which makes them highly suitable for processing noisy biological information. ANN-based models have proven very effective for the prediction of class-I (A*0201) (Adams and Koziol 1995; Milik et al., 1998) and class II HLA ligands (DRB1*0401) (Brusic et al., 1998). The prediction accuracy of ANN-based methods is reported to be close to 80% sensitivity and 80% specificity (Brusic et al., 1998, Milik et al., 1998).

The advantages of ANNs are that they are adaptive and can self improve, they generalize well, are effective with nonlinear problems, and are tolerant to a certain level of erroneous data. Because they have larger number of parameters to be determined ANNs require larger amounts of binding data than simpler prediction methods. This is a bottleneck in the development of such systems. With increasing amount of peptide data, ANN-based models will become more trained and will further improve in prediction accuracy. In addition, promiscuous prediction is not feasible since data for one allele cannot be extrapolated to other alleles. Nonetheless, ANNs are reported as more accurate predictors than quantitative matrices (Borras-Cuesta et al., 2000).

An interesting algorithm described is a combination of neural networks and evolutionary algorithm. This approach achieved a correct classification percentage for both binders and non-binders in excess of 80% (Brusic et al., 1998). It should be noted that the authors also used an evolutionary algorithm to evolve a scoring matrix during preprocessing and the combined approach is shown to perform better.

**Hidden Markov Models**

HMMs are statistical models that can capture complex relationships in data sets. A HMM is a finite state machine that governs transition of a system-process between states using associated probability distributions. They can model sets of variable length sequences; and this makes them,
suitable for modeling biological sequences. HMMs are reported as high accuracy predictors of MHC-binding peptides (Mamitsuka 1998; Noguchi et al., 2002; Zhang et al., 2006). HMMs have been reported to be used for modeling MHC–peptide interactions for multiple alleles of a HLA-A2 super-type of MHC class I peptides (Brusic et al., 2002).

HMMs perform with a similar accuracy as ANNs and require larger data sets for training than quantitative matrices. The main advantage of a HMM is that it does not require pre-processing (alignment) of peptides before training.

Other Methods

Antigenic peptides have different length. Moreover, peptide binding groove of class II MHC is open at both ends, which causes ambiguity in positional alignment between the groove and peptide and uncertainty in modeling methods. Several approaches have been applied in recent years to handle the variability in peptide length and binding position. These methods, such as an iterative Partial Least Squares method (Doytchinova and Flower 2003), an Ant Colony search (Karpenko et al., 2002), a Gibbs sampling algorithm (Nielsen et al., 2004) and a kernel based approach (Salomon and Flower 2006), have significantly improved the prediction performance of simple conventional approaches reviewed above.

Other statistical methods that utilize binding data as training set in model development include SDS (Mallios 1999), HMM (Brusic et al., 2002), HMM + SSS (Noguchi et al., 2002), SVM (Bhasin and Raghava 2004; Cui et al., 2006; Cui et al., 2007; Donnes and Elofsson 2002; Donnes and Kohlbacher 2006; Liu et al., 2006), positional scanning (Udaka et al., 2000), profile motifs (Reche et al., 2002), quantitative matrix + ANN (Bhasin and Raghava 2007) and other hybrid approaches based on different classification algorithms (Antes et al., 2006; Bhasin and Raghava 2004; Trost et al., 2007). HLA alleles used for training and prediction in some typical methods are summarized in Table 1. The application of these models is restricted to either one or few HLA alleles depending on the availability of training set. As shown in Table 1, these models have been tested for DR1 (Mallios 1999), HLA-A2 (Brusic et al., 2002), DRB1*0101 (Noguchi et al., 2002), H-Kα, (Reche et al., 2002; Udaka et al., 2000), H-Dβ (Reche et al., 2002; Udaka et al., 2000), H-Lδ (Udaka et al., 2000) and another 26 alleles (Donnes and Elofsson 2002). It has been shown that their prediction accuracy varies from 90-100% for these datasets. It should be noted that the size of the training set is different in different cases. It is worth mentioning that the size of negative and positive data used in such developments is also variable and these factors play an important role in the overall estimation of their accuracy and predictive power. It will be interesting to compare the usefulness of these techniques either individually or combined and check their prediction efficiencies using the same set of training and prediction dataset.

Structure Based Predictions

The second categories of prediction methods use known three-dimensional structures of MHCp complexes. The current release of Protein Data Bank contains a number of unique MHC structures (Berman et al., 2002). These structures enable a better understanding of the structural principles governing peptide recognition by MHC molecules (Batalia and Collins 1997) (Figure 1). MHC molecules bind peptides of diverse sequences with high affinity and long half-life. Most peptides selected by class-I molecules are 8 to 10 residues long and conserved amino acids bind the invariant portions of the peptides, presenting anchoring backbone atoms at positions 2 and C, N termini (Guo et al., 1993; Madden et al., 1993). Auxiliary anchors at P1 and P3 usually fine tune peptide recognition (Madden et al., 1993; Ruppert et al., 1993). Each anchoring side chain interacts with one of the 6 polymorphic MHC pockets (Guo et al., 1993; Saper et al., 1991), whose structural fold is conserved in evolution with physicochemical diversity for allele specificity (Falk et al., 1990). Through a set of hydrogen bonds to the main chain of the peptide, the termini of the peptide are oriented into specific pockets that are designed to accommodate the chemical nature of the peptide residues. Thus, the orientation (amino to carboxyl) of the antigenic peptide is fixed for all MHC molecules (Figure 2). However, this arrangement is affected by peptide length. Longer peptides may zigzag (Madden et al., 1993) or bulge (Collins et al., 1995; Guo et al., 1992) to allow peptides of greater length to maintain the relative position of the termini. In addition, longer peptides may bind and maintain
**Figure 2:** Virtual Pockets are the basic functional unit of HLA molecules.

**a.** This illustration shows the peptide binding pockets. A HLA molecule binds and displays a peptide and the T-cell receptor recognizes the two polymorphic residues of the HLA molecules and one residue of the peptide. Anchor residues of peptide are accommodated in the pockets formed by the polymorphic residues in the binding groove of the HLA molecule.

**b.** Six binding pockets, denoted A through F, are formed by the polymorphic amino acid residues within the binding groove of HLA molecule. Pocket profiles is a detailed description of peptide binding environments in the pockets, which are determined by the HLA polymorphic side-chains delineate the geometry and chemical properties of these structural pockets. In this work, pocket profiles are defined as those residues in a fixed neighborhood of the peptide residues in known crystal structure complexes. Pockets are the basic functional unit of HLA molecules. The amino acid residues that form the pocket determine the antigen peptides that would preferentially be bound by the HLA allotype, accounting for the differential ability of different alleles to bind a variety of peptides. HLA alleles sharing same pocket profiles will preferentially bind the same kind of peptide anchor residues. The peptide-binding groove of the HLA is essentially an exchange or shuffling of pockets between different HLA allotypes.

The first approach based on molecular dynamics simulation (MDS) of MHCp complexes (Rognan et al., 1994) allows a crude discrimination of binders from non-binders. The approach reported...
change in free energy during simulation of the HLA-B*2705 complex in AMBER force field. Though efficient, this method is not suitable for high-throughput predictions due to extensive computational requirements in capturing the simulation trajectory. Concomitantly, various methods calculating binding free energy of MHCp complexes, based on different energy scoring functions have been developed. Since binding free energy of MHC-peptide complex is related to the affinity of MHC-peptide binding, binders and non-binders can be discriminated and the approach produces absolute or relative peptide binding affinity. Free energy calculation is either based on statistical pair-wise potentials tables or free energy scoring functions. An approach based on free energy involving threading of the peptides using known templates followed by evaluation of their binding by statistical pair wise potentials was presented (Altuvia et al., 1997; Jojic et al., 2006; Schueler-Furman et al., 2000; Schueler-Furman et al., 1998). However, unlike methods capable of calculating the absolute binding free energies from three-dimensional homology models, it does not allow the direct prediction of binding affinity values. This is explicable as it is difficult to develop a universal free energy function for MHCp binding, though several attempts have been made towards this goal (Rognan et al., 1999).

The backbone conformations of bound peptides are generally not conserved in the binding groove. The bound peptides are thus flexible and the middle part of the peptides usually bulges out the binding groove. This bulging part allows the backbones to take different patterns at the groove. The generic peptide structure determination methods using Monte Carlo, molecular dynamics simulations, dynamic programming, free energy mapping, or threading are suitable for binding free energy calculations, but they all have limitations in predicting the conformation of the peptide in the groove. Computational combinatorial ligand design (CCLD) (Zeng et al., 2001) has been used for placing amino acids in specific pockets. A method based on threedimensional quantitative structure affinity relationship (3-D QSAR) of MHCp complexes (Doytchinova and Flower 2001) has also been developed. Most approaches, except for threading (Altuvia et al., 1995; Jojic et al., 2006), are not generally suitable for systematic high-throughput genome scanning. The structure based methods can be classified into few groups which are discussed below.

**Knowledge Based Free Energy Scoring Methods**

The knowledge based scoring method based on pair-wise contacts use solved or modeled structure for MHCp binding calculations. The physical, chemical compatibility between peptide and MHC groove is estimated using pair-wise potential matrix. The binding score is obtained by adding all pair-wise values for residues in the pocket with the corresponding peptide residues at every position. This enables the ranking of peptides for MHCp binding (Altuvia et al., 1997; Schueler-Furman et al., 1998). This structure based approach is based on two main features: (1) the availability of appropriate peptide structural template; and (2) the choice of a pair-wise potential table.

Various knowledge based pair-wise potentials have been derived from known protein structures (Jernigan and Bahr 1996; Jones and Thornton 1996; Skolnick et al., 1997). A basic approximation underlying these potentials is that total “free energy” of a protein can be expressed as a sum of independent pair-wise interactions. The frequencies of residue pairs in the structures are assumed to represent the interaction preference between different types of residues. This interaction preference between two amino acids is expressed by its comparison with their affinity to a “reference state.” Various matrices have been published using distinct reference states (Skolnick et al., 1997). Miyazawa and Jernigan used solvent as reference a state and developed a matrix with emphasis on hydrophobic interactions (Miyazawa and Jernigan 1985; Miyazawa and Jernigan 1996). Betancourt and Thirumalai (Betancourt and Thirumalai 1999) modified the table by changing the reference state from solvent to a defined, single solvent-like molecule. The resulting matrix represents hydrophilic interactions. Altuvia et al. used the potential tables of Miyazawa and Jernigan to rank modeled MHCp structures (Altuvia et al., 1995; Altuvia et al., 1997). However, the procedure failed to predict hydrophilic interactions (Altuvia et al., 1997). The pair wise potential table of Betancourt and Thirumalai (Betancourt and Thirumalai 1999) has been reported to have successfully selected hydrophilic interactions (Schueler-Furman et al., 2000).
been used to calculate binding free energy of complexes. The FRESNO scoring function has been used to calculate binding free energy of MHCp interactions. The approach allows for the prediction of absolute binding affinities in a high throughput manner. An extension to this work is EpiDock (Logean and Rognan 2002) which is shown to predict potential T-cell epitopes from viral proteomes and can roughly predict unknown peptide binding motifs for novel class I MHC alleles. However, the critical issue in this approach is the identification of peptide templates for structure prediction.

**Free Energy Scoring Function Based Methods**

These methods aim is to determine the ligands capable of binding from a series of candidate ligands by calculating binding free energy. They generally do not require predetermined experimental data for model development and can produce relatively accurate binding affinity. It must be noted that MHCp binding or non-binding data is not enough to predict whether the peptide can induce immune response. Moreover, accurate calculation of MHCp binding free energy difference using 3D structures by simple free energy scoring functions is CPU intensive.

In recent years, a number of free energy scoring functions have been developed for different purposes and these functions have been used for MHCp binding predictions. One of the recent approaches for class I MHCp binding prediction is a tailor-made free energy scoring function (FRESNO) combined with homology modeling (Logean et al., 2001; Rognan et al., 1999). Starting from the primary sequence of the protein antigen, individual 3D structures of all possible class I MHC-peptide (8-, 9- and 10-mers) complexes are constructed by homology modeling. The FRESNO scoring function has been used to calculate binding free energy of MHCp interactions. The approach allows for the prediction of absolute binding affinities in a high throughput manner. An extension to this work is EpiDock (Logean and Rognan 2002) which is shown to predict potential T-cell epitopes from viral proteomes and can roughly predict unknown peptide binding motifs for novel class I MHC alleles. However, the critical issue in this approach is the identification of peptide templates for structure prediction.

**Molecular modelling**

Molecular modeling utilizes detailed knowledge of the crystal structure of MHC molecules (Rognan et al., 1994) and of protein–peptide interactions (Altuvia et al., 1997). Comparative modeling is used if known crystal structures and protein–peptide interactions are available as templates for building 3-D models. On the other hand, *Ab initio* modeling uses atomic simulations. Residue statistics is used when initial structural data are not available. Molecular modeling is a complementary approach to other data-driven approaches. Molecular modeling provides a detailed insight into specific 3-D structures and interactions. However, it is computationally demanding and therefore less suitable for large-scale screening. Molecular modeling can provide...
### Table 1: Comparison of Typical MHC-peptide Binding Prediction Methods

<table>
<thead>
<tr>
<th>Principle</th>
<th>Author</th>
<th>Year</th>
<th>Training Dataset</th>
<th>Prediction Dataset</th>
<th>ROC Analysis</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIMAS</td>
<td>Parker</td>
<td>1994</td>
<td>HLA-A2</td>
<td>-</td>
<td>154</td>
<td>-</td>
</tr>
<tr>
<td>ANN</td>
<td>Adams</td>
<td>1996</td>
<td>A*0201</td>
<td>-</td>
<td>1,038</td>
<td>A*0201</td>
</tr>
<tr>
<td>EA + ANN</td>
<td>Brusic</td>
<td>1998</td>
<td>DRB1*0401</td>
<td>-</td>
<td>336</td>
<td>312</td>
</tr>
<tr>
<td>ANN + QM</td>
<td>Bhaskar</td>
<td>2007</td>
<td>30 alleles</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SGS</td>
<td>Mailos</td>
<td>1999</td>
<td>DRI</td>
<td>-</td>
<td>562</td>
<td>98</td>
</tr>
<tr>
<td>HMM</td>
<td>Brusic</td>
<td>2002</td>
<td>HLA-A2</td>
<td>-</td>
<td>-</td>
<td>4 alleles</td>
</tr>
<tr>
<td>HMM + SSS</td>
<td>Kobayashi</td>
<td>2002</td>
<td>DRB1*0401</td>
<td>-</td>
<td>305</td>
<td>472</td>
</tr>
<tr>
<td>SVM/H</td>
<td>Elofsson</td>
<td>2002</td>
<td>26 alleles</td>
<td>-</td>
<td>-</td>
<td>6 alleles</td>
</tr>
<tr>
<td>SVM</td>
<td>Cul</td>
<td>2007</td>
<td>30 alleles</td>
<td>-</td>
<td>7460</td>
<td>1591</td>
</tr>
<tr>
<td>hybrid</td>
<td>Trost</td>
<td>2007</td>
<td>4 alleles</td>
<td>-</td>
<td>1473</td>
<td>2654</td>
</tr>
<tr>
<td>Mouse Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positional scan</td>
<td>Usaka</td>
<td>2000</td>
<td>H-κ, D, L</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Profile motifs</td>
<td>Reinhart</td>
<td>2002</td>
<td>H-κ, D</td>
<td>-</td>
<td>71</td>
<td>H-κ, D</td>
</tr>
<tr>
<td>Mostly Human Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHC</td>
<td>Rognan</td>
<td>1994</td>
<td>B'2705</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>DynaPred</td>
<td>Antes</td>
<td>2006</td>
<td>A*0201</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Threading</td>
<td>Margalit</td>
<td>1998</td>
<td>9 alleles</td>
<td>23</td>
<td>-</td>
<td>23</td>
</tr>
<tr>
<td>FESF</td>
<td>Rognan</td>
<td>1999</td>
<td>A*0201, H-2</td>
<td>23</td>
<td>A*0201, H-2</td>
<td>24, 16</td>
</tr>
<tr>
<td>Pockets</td>
<td>Hammer</td>
<td>1999</td>
<td>11 HLA-DR</td>
<td>-</td>
<td>51 alleles</td>
<td>223</td>
</tr>
<tr>
<td>Threading</td>
<td>Margalit</td>
<td>2000</td>
<td>10 alleles</td>
<td>37</td>
<td>123</td>
<td>-</td>
</tr>
<tr>
<td>CCLD</td>
<td>Zeng</td>
<td>2001</td>
<td>A*0201</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Structure-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D QSAR</td>
<td>Flower</td>
<td>2001</td>
<td>A*0201</td>
<td>1</td>
<td>102</td>
<td>84</td>
</tr>
<tr>
<td>EpiDock</td>
<td>Rognan</td>
<td>2002</td>
<td>A*0201</td>
<td>25/63</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3D Additive</td>
<td>Kanganeue</td>
<td>2003</td>
<td>8 alleles</td>
<td>29</td>
<td>-</td>
<td>&gt;50</td>
</tr>
</tbody>
</table>
information for building complex data-driven methods, i.e., for prediction of promiscuous MHC-binding peptides using HMMs (Brusic et al., 2002).

Molecular dynamics has been used to model binding of peptides to HLA-B*2705 (Rognan et al., 1994). This work has been extended to the prediction of peptide binding affinities using free energy scoring functions (Schueler-Furman et al., 2000). The crystal structures of 23 solved peptide–MHC structures are used for the development of a modeling algorithm (Schueler-Furman et al., 2000). The known peptide structures in the groove were used as a template for threading peptide candidates and their binding potentials were analyzed by statistical pair-wise potential for a broad range of MHC class I alleles (Schueler-Furman et al., 2000). Quantitative structure–function relationship studies were performed to identify physicochemical requirements of peptide binding to MHC molecules using large numbers of peptides (Altuvia et al., 1995; Antes et al., 2006; Doytchinova and Flower 2001; Jojic et al., 2006).

Virtual Pockets

A MHCp binding method has recently been developed by our group using structural information gathered from class-I MHCp crystal structures (Zhao et al., 2003). In this method, nine virtual pockets are defined and the binding affinity between MHC and peptide is given by the sum of residue-residue compatibility between peptide residues and corresponding virtual pockets (Zhao et al., 2003) (Figure 3).

Based on the analysis of HLAp molecular interactions using crystal structures, it is found that six binding pockets are formed by the polymorphic amino acid residues within the binding grooves of HLA molecules. Anchor residues of peptide are accommodated in the pockets. The binding environments of anchor residue in the pockets, or pocket profile, are determined by the HLA polymorphic side-chains delineate the geometry and chemical properties of these structural pockets. Pockets are the basic functional unit of HLA molecules. The amino acid residues that form the pocket determine the properties, such as size, shape, hydrophobicity, and electronic charge, and other properties, of the pocket and thus determine the antigen peptides that would be preferentially bound by the HLA allotype. This accounts for the differential ability of different alleles to bind a variety of peptides. HLA alleles sharing same pocket profiles will preferentially bind same kind of peptide anchor residues. Thus, the HLA peptide-binding groove is essentially an exchange or shuffling of pockets between different HLA allotypes.

By establishing the geometry and chemical compatibilities between individual pockets and peptide anchor residues, the binding strength of peptide anchor residues was quantified. Thereby, a peptide binding motif was identified and the accumulated binding strength between peptide and HLA molecule was roughly determined (Zhao et al., 2003). However, to accurately assess the binding affinity of a whole peptide, it is not enough to solely consider individual anchor residues that interact with binding pockets, because the peptides residues other than anchor residues are also contributing to the peptide binding affinity. To address this issue, the concept of structural pocket was expanded and the key functional residues for peptide binding in the HLA binding grooves were identified (Figure 1). A correlation between the peptide residue positions and the key functional HLA residue positions is established and the key functional HLA residues that directly interact with a certain peptide residue were designated as “virtual pocket”. This implies that their residue compositions are based on structural pockets, but their positional composition will change when they interact with different individual peptide residue.

The HLA peptide binding prediction model discussed in this section is developed based on the virtual pocket concept. The quantification of the interaction between the MHCp residue pair is calculated by the application of the Q matrix, which quantified the interaction between the 20 amino acids based on 237 physio-chemical properties (Mathura and Braun, 2001). The prediction method has been intensively verified by large quantity of MHCp binding data. The method has high efficiency (on average 60%); good sensitivity (50-73%) and specificity (52-58%). Although the prediction accuracy is moderate (60%), the method is simple, effective and most important applicable to all MHC allele whose sequence is clearly defined.
A limitation of this method is that the peptide length is fixed and needs to be predefined for a set of peptides with specific length because the virtual pockets are defined based on relationships among peptide residue positions. Since the majority of peptides in the available crystal structures of class I HLA peptide complexes are nonamers (9-mers), the model at present is solely based on the structural data of HLA-nonamer peptide complexes, and thus can predict the binding of nonamer peptides. However, the model can be easily extended in future to the peptides of different length when substantial structural data are available for them.

CONCLUSION

The MHCp binding prediction models are suited for high throughput scanning of a pathogen proteome with high sensitivity capable of covering maximum number of MHC alleles. The method that requires very few experiments in the identification of vaccine candidates is technologically advantageous. In this report the merits and demerits of several tools and techniques such as BIMAS, HLA BIND, ProPred1, SYFPEITHI, EPIMATRIX, EPIPREDICT, PREDICT, MDS, CCLD, 3D-QSAR, TEPITOPE, FRESNO, EpiDock and Virtual pockets for MHCp binding predictions is discussed. This review provides a comparison of different available methods. We highlight that the choice of the tools and its mode of development are critical to their application in immunology and users of such tools should be aware of such limitations (Table 1). Other coupled parameters such as peptide processing, transport, loading, TCR repertoires and subsequent immune elucidation factors have to be clearly modeled for appropriate application of MHCp binding prediction models in immunotherapeutics and vaccine design. The next few years promise many such prediction tools for use in immunobiology and these will be a step forward towards in silico vaccine design.

Acknowledgement


Altuvia Y, Schueler O and Margalit H (1995) Ranking potential binding peptides to MHC molecules by a computational threading approach, J. Mol. Biol. 249, 244-250.


Batalia MA and Collins EJ (1997) Peptide binding by class I and class II MHC molecules, Biopolymers 43, 281-302.


MHC-Peptide binding prediction for epitope based vaccine design


MHC-Peptide binding prediction for epitope based vaccine design

of the conformations of five viral peptides presented by HLA-A2, Cell 75, 693-708.


