BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 2. 96–108
Advance Access publication October 31, 2006
doi:10.1093/bib/bbl038

Methods and protocols for prediction of immunogenic epitopes

Joo Chuan Tong, Tin Wee Tan and Shoba Ranganathan

Submitted: 2nd June 2006; Received (in revised form): 14th September 2006

Abstract

T-cell recognition of peptide/major histocompatibility complex (MHC) is a prerequisite for cellular immunity. Recently, there has been an influx of bioinformatics tools to facilitate the identification of T-cell epitopes to specific MHC alleles. This article examines existing computational strategies for the study of peptide/MHC interactions. The most important bioinformatics tools and methods with relevance to the study of peptide/MHC interactions have been reviewed. We have also provided guidelines for predicting antigenic peptides based on the availability of existing experimental data.

Keywords: *MHC; antigens/peptides/epitopes; prediction*

INTRODUCTION

T-cell recognition of peptide/major histocompatibility complex (MHC) is a prerequisite for cellular immunity. The peptides that bind to specific MHC triggering T-cell recognition (T-cell epitopes) are targets for vaccine and immunotherapy development because they are the minimal essential peptide subunits that stimulate cellular immune responses. Precise identification of peptides binding to specific MHC alleles is important for the diagnosis and treatment of infectious [1], allergic [2], autoimmune [3] and neoplastic diseases [4]. The MHC genes in human, called human leukocyte antigen (HLA) are the most polymorphic human genes known [5]. By August 2006, 1964 protein-coding HLA alleles had been identified (http://www.anthonynolan.org.uk/HIG/). Binding studies show that each HLA allele recognizes a restricted set of peptides. Experimental approaches to determine HLA binding specificities is an expensive, laborious and time consuming process; and not applicable for studies involving large numbers of protein sequences.

Bioinformatics tools modeling the immune system network have played an instrumental role in advancing peptide vaccine discovery, with promising results in melanoma [6], multiple sclerosis [7], malaria [8] and anti-tumor vaccines [9]. T-cell epitope prediction tools help researchers identify allele-specific binding peptides, thus reducing the number of peptides to be synthesized and assayed. Tools for MHC supertype (superfamily) classification facilitate the identification of alleles with similar structural features and/or peptide specificities. More sophisticated bioinformatics tools enables the systematic scanning for candidate T-cell epitopes from larger sets of protein antigens, such as those encoded by complete viral genomes. These tools help researchers to identify regions with high concentrations of T-cell epitopes or immunological 'hot spots' and focus upon relevant experiments.

Corresponding author. Shoba Ranganathan, Department of Chemistry and Biomolecular Sciences & Biotechnology Research Institute, Macquarie University, NSW 2109, Australia. Tel: +61-2-9850-6262; Fax: +61-2-9850-8313; E-mail: shoba.ranganathan@mq.edu.au

Joo Chuan Tong is a researcher at the Institute for Infocomm Research, Singapore. His research focuses on T-cell epitope prediction, classification methods, and protein structure prediction.

Tin Wee Tan is an Associate Professor of Biochemistry at the Yong Loo Lin School of Medicine, National University of Singapore, working on bioinformatics methods and applications, including grid computing.

Shoba Ranganathan is the Chair Professor of Bioinformatics at Macquarie University and Adjunct Professor, National University of Singapore, working on computational structural biology, immunoinformatics and comparative genome sequence analysisis.

The aim of this article is to provide an overview of bioinformatics tools available for the study of peptide/MHC interactions.

PREDICTION OF MHC-BINDING PEPTIDES

Two main categories of specialized bioinformatics tools are available for prediction of MHC binding peptides—methods based on identifying patterns in sequences of binding peptides, and those that employ three-dimensional (3D) structures to model peptide/MHC interactions. The first group includes procedures based on binding motifs, quantitative matrices, decision trees, artificial neural networks (ANNs), hidden Markov models (HMMs) and support vector machines (SVMs). In contrast, the second category corresponds to techniques with distinct theoretical lineage and includes the use of homology modeling, docking and 3D threading techniques. An unequal amount and variety of techniques have explored for the two categories in the published reports, far fewer for structure-based approach due to higher complexity in development and longer computational time.

Simple sequence motifs

Discovery of anchor residues and sequence motifs

The earliest attempt to predict MHC-binding peptides started with the discovery that peptides binding to specific MHC alleles are functionally related and share residues with similar properties at various positions of their primary sequences. Class I and class II binding peptides contain residues with side-chains that fit into polymorphic cavities (or 'pockets') and bind to complementary residues of specific MHC alleles. These residues are called anchor residues because they 'anchor' the peptides firmly at various positions in the MHC binding cleft [10–13] and contribute to most of the binding interactions. This led to the definition of 'peptide motif' [10, 11, 14] for an array of class I and class II alleles. Numerous research groups, including Zhang *et al.* [15], Lipford *et al.* [16], Sette *et al.* [17], Sidney *et al.* [18], Parker *et al.* [19], Hammer *et al.* [20], Rammensee *et al.* [21], Meister *et al.* [22], D'Amaro *et al.* [23] and Rajapakse *et al.* [24] developed computational tools that scan peptides that fit these motifs. SYFPEITHI [25], a database for MHC ligands and binding motifs, was developed. As of August 2006 (last update in May 2006), SYFPEITHI [25] comprises more than 4500 peptide sequences known to bind class I and class II alleles from published reports (http://www.syfpeithi.de/).

It was later discovered that residues along other positions of a peptide also play a vital role in binding, and sequence motifs alone are inadequate to account for the comprehensive binding ability of a candidate peptide [26–28]. Immunodominant peptides without the required binding motifs were identified, and not all motif-conforming peptides do bind to the respective MHC alleles [29]. In an attempt to investigate the role of motifs in binding, Ruppert *et al.* [28] performed binding assays on peptides to HLA-A*0201 and found that only about 30% of motif-conforming peptides were actual binders. In practice, simple motif models have proven to be both nonsensitive and nonspecific [29]. This approach fails to detect binders not conforming to existing motifs and includes nonbinding sequences that fit the required patterns [22]. However, despite these limitations, this approach is still a useful alternative to random guessing or use of a complete overlapping set of peptides for selection of candidate binders [30].

Binding matrices

Binding matrices represent an enhancement of simple motif models by correlating peptide residue positions to binding. This approach employs the use of tables containing $l \times 20$ coefficients where l corresponds to the length of the binding motif and 20 for each amino acid symbol [31, 32]. Consensus scores are obtained by summing, multiplying or averaging the matrix coefficients and compared against a predetermined threshold. In general, matrices are constructed using amino acid frequencies at different position of known binders or quantitative MHC-binding data. The former indicates the binding likelihood of a peptide sequence to the MHC molecule, while the later provides means of quantifying the peptide binding affinity. Examples of matrices derived from simple counting of amino acid frequencies at different position of known peptide binders include EpiMatrix [33] and SYFPEITHI [25], while BIMAS [19] was developed by fitting of MHC-binding data.

More complex forms of matrix-based models have been developed to detect weak binding patterns and to account for noisy and collinear data. Reche *et al.* [34] employed the use of position-specific scoring matrices from a set of aligned binding

peptides to predict binders to an array of MHC class I and class II molecules. Peters *et al*. [35] introduced the use of stabilized matrix method (SMM) as predictor for HLA-A2 binding peptides. Nielsen *et al*. [36] applied a Gibbs sampler to detect weak sequence motifs in class I and class II binding peptides. Rajapakse *et al*. [37] utilized a multi-objective evolutional algorithm to identify a consensus motif for I-A^{g7}. Guan *et al*. [38, 39] and Doytchinova *et al*. [40] employed the use of multivariate statistics to improve the predictive performance of their matrices. An additive equation was formulated to account for individual amino acid contributions at each position and interactions with neighboring amino acids. The matrix was subsequently solved through the use of partial least square regression.

Decision trees

Decision trees are rule-based models that classify patterns using a sequence of well-defined rules [41]. Position-specific binding motifs are converted into rules and embedded within the nodes of a decision tree. The resulting tree structure indicates amino acid properties that are strongly correlated with physicochemical properties of binding peptides. Peptide sequences are threaded through a series of nodes and the result of all node-to-node transitions are used to determine the outcome of prediction. Because of its capability to elucidate both linear and nonlinear problems, this approach has been adopted by several groups to identify higher-level rules for binding. Savoie *et al*. [42] constructed a decision tree using the BONSAI program to investigate T-cell preference and adverse motifs for HLA-A*0201 binding peptides. Segal *et al*. [43] adopted a similar tree-structured technique to predict peptides binding to H2-K^b. An example of a decision tree network is shown in Figure 1.

Artificial neural networks

ANNs are connectionist models particularly well suited to perform classification and complex pattern recognition tasks [44]. ANNs can encode nonlinear data and have been used extensively for prediction of peptide binding to both class I and class II alleles [32, 45–50]. Peptide features are represented by amino acid descriptors such as composition, hydrophobicity, volume and charge. The descriptors are used to train an ANN for classifying peptides into binders and nonbinders. An example of ANN architecture is illustrated in Figure 2. An investigation on the predictive performance of ANNs revealed that this approach gradually outperforms motifs, matrices and HMMs with increasing peptide data [30]. A major drawback of ANN is the requirement of a fixed input length. As such, a given ANN model can only predict binding peptides that are of the same length as those in the training data set. This constraint restricts the ability of ANN to predict epitopes with length that differ from those used in the trained network.

Various groups have developed hybrid versions of ANN for peptide/MHC prediction. Nielsen *et al*. [50] described a combination of a series of neural networks using several sequence coding strategies including an HMM encoding to improve the predictive power of the system. Brusic *et al*. [46] integrates the strength of matrix models and evolutionary algorithm (EA) for processing ANN training set. New alignment matrices were selected by EA based on evolutionary principles. Each parent (matrix) produces two children consisting of an exact copy of itself and a mutant copy, and passes the child with the higher fitness value to the next generation. The highest scoring alignments from the final generation matrices were subsequently fed into ANN for training.

Hidden Markov models

HMM belongs to a type of probabilistic graphical models that have been successfully applied to a wide range of applications in statistical pattern recognition and classification [51]. In order to overcome the potential limitations of ANNs, HMMs have been applied to predict peptides binding to MHC [52]. Similar to decision trees and ANNs, HMMs have the ability to cope with nonlinear data and are suitable for representing time-series sequences having flexible lengths. Associated with each HMM is a series of discrete-state, time-homologous, first-order Markov chain (MC) with suitable transition probabilities between states and an initial distribution. Each state consists of a discrete or continuous distribution over possible emissions or outputs. These outputs are generated when particular state is visited or during transition from state to state. Transitions between states follow a set of transition and emission probabilities. The transition probability is the probability of moving from one state to another via a connected edge, and the emission probability is the probability of emitting a particular symbol at a state. The sequences of states underlying MC are hidden and cannot be observed, hence the name hidden

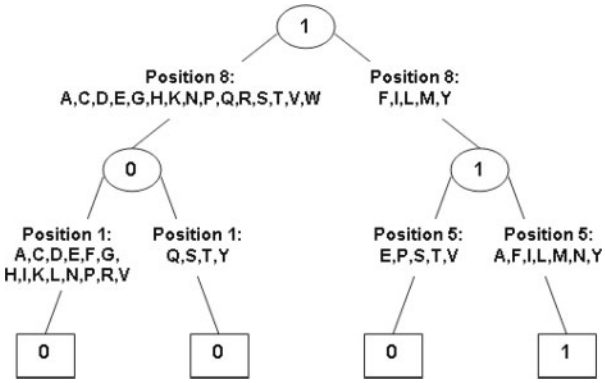

Figure 1: Subset of decision tree network employed by Segal *et al*. [43]. Each node represents grouping of preferential/nonpreferential amino acid residues at various positions of H2-K^b binding peptides. Predicted class at each node (ellipses—internal; rectangles—terminal) is given by the 0 (nonbinding) or 1 (binding) within each node.

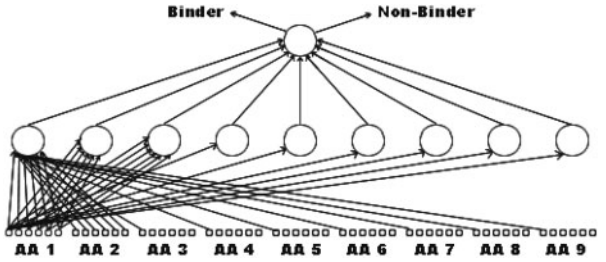

Figure 2: A three-layer ANN for predicting class I binding peptides by Brusic *et al*. [45]. The first layer represents input nodes with the number of nodes corresponding to the length of input peptide; the number of second (hidden) layer nodes equals to the ideal length of binding peptides; and a single output node predicts binding versus nonbinding.

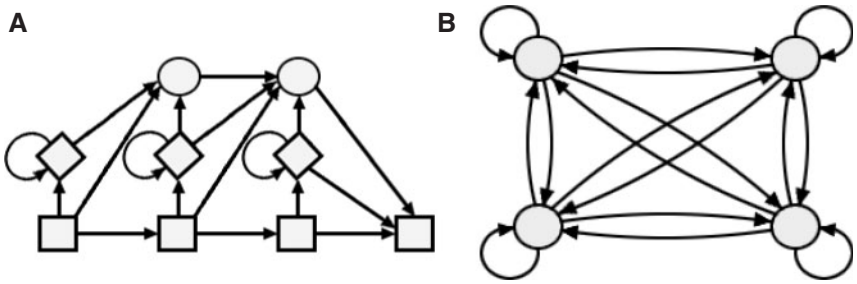

Figure 3: HMM topologies adopted for peptide/MHC prediction by Mamitsuka [52]. (**A**) A profile HMM, (**B**) a fully connected HMM.

Markov model. The probability of any sequence, given the model, is computed by multiplying the emission and transition probabilities along the path.

The use of HMM for peptide/MHC prediction was first reported in the literature [52] using two different HMM topologies: profile HMM and fully connected HMM. Profile HMMs (Figure 3A) are linear left-right models where the underlying directed graph is acyclic, with the exception of loops, hence supporting a partial order of the states. The profile HMM architecture [53] consists of three classes of states: the match state, the insert state and the delete state; and two sets of parameters: transition probabilities, and emission probabilities. The match and insert states always emit a symbol, whereas the delete states are silent states without emission probabilities. A fully connected HMM (Figure 3B) consists of states that are pairwise connected such that the underlying digraph is complete. There are no distinguished starting and terminating states and the transition matrix does not contain any zero entries with the exception of diagonal entries, which correspond to loops or self-transitions. Because there is no constraint on the structure of a fully connected HMM, this model permits the representation of more than one sequence pattern concealed in the training data.

Support vector machines

SVMs are statistical learning methods based on the structural risk minimization principle [54]. Similar to decision tree, ANN and HMM, it has the ability to handle both linear and nonlinear data. Every peptide sequence is represented by specific feature vector assembled from encoded representations of residue properties such as amino acid composition, hydrophobicity, polarity, charge, bulkiness and solvent accessibility. Parameters are trained by mapping input vectors into a high-dimensional feature space and maximizing the margin between the binders and nonbinders with an optimal separating hyperplane. SVM outperforms ANN and decision tree in the

absence of large training data set [55] and has been embraced by several groups including Dönnes and Elofsson [56], Bhasin and Raghava [57] and Bozic *et al*. [58] for predicting class I and class II binding peptides. Hybrid models based on ANN and SVM have also been developed by Bhasin and Raghava [59] for consensus and combine prediction of T-cell epitopes.

Structure-based approach

Protein threading

Protein threading [60] or side-chain conformational search [61] involves computing an alignment between a target amino acid sequence and the spatial positions of a 3D structure. In the context of peptide/MHC modeling, this involves substituting the backbone coordinates of a source peptide (P_1, $P_2, \ldots, P_n$) that is bound to a MHC molecule of interest with the target peptide sequence (S_1, $S_2, \ldots, S_n$) by replacing P_i with S_i. A search for the best side-chain conformations is usually performed, and a scoring scheme is subsequently applied to discriminate the binders from nonbinders.

Altuvia *et al*. [62] demonstrated the use of protein threading to detect binding peptides not conforming to HLA-A*0201 binding motifs using the statistical pairwise potential table of Miyazawa and Jernigan [63, 64]. This was subsequently extended to the analysis of peptides binding to an array of class I alleles [65, 66]. This approach successfully identified peptides binding to MHC molecules with hydrophobic binding pockets but not to MHC molecules with hydrophilic, charged pockets. In order to circumvent the problem, Kangueane *et al*. [67] introduced the use of knowledge-based rules to discriminate binders from nonbinders based on the number of observed atomic clashes between the MHC and its bound peptide, and the number of solvent exposed hydrophobic residues on the modeled peptide. The problem was later solved by Schueler-Furman *et al*. [68] through the use of a different pairwise potential table [69] that described hydrophilic interactions more appropriately. A hybrid technique that combines MHC class I sequences and peptide/MHC binding affinities was also proposed [70]. In an attempt to improve the accuracy of threading algorithms, Bui *et al*. [71] incorporated explicit water molecules at the peptide/MHC interface. An alternative discrimination scheme was also introduced by Doytchinova and Flower [72] that employ similarity indices from 3D quantitative structure-affinity relationship (QSAR) studies.

Homology modeling

Homology modeling [73, 74] employs the use of available homologous protein structure(s) to predict the unknown structure of a related amino acid sequence. In the context of peptide/MHC prediction, the aim is to model the bound conformation of a peptide sequence with an unknown structure given the 3D structure of other bound peptides to homologous MHC molecules. Hammer *et al*. [75] constructed a series of synthetic peptide/HLA-DRB1*0402 models from HA peptide/HLA-DRB1*0101 crystallographic structure to identify specific patterns of peptide binding. Rognan *et al*. [76] and Logean *et al*. [77] applied a similar two-step approach to construct the bound conformation of peptides to an array of class I alleles. Their modeling procedure begins by selecting peptide termini residues based on homology to the most similar MHC-bound peptide with available crystallographic structure. The remaining residues were subsequently constructed by satisfaction of spatial restraints using a knowledge-based loop search procedure. Several attempts to characterize TCR/peptide/MHC interaction were also reported. Michielin *et al*. [78] successfully developed a model of T1 T-cell receptor (TCR)/PbCS/H2-K^d complex based on its homology with the 2C TCR, the A6 TCR/Tax/HLA-A2 complex, the 1934.4 TCR Vα chain, the 14.3.d TCR Vβ chain, and the H2-K^b ovalbumine peptide. Buoyant by the excellent results, Michielin *et al*. [79] applied the methodology to identify critical residues of the A6 TCR that interacts with peptide/HLA-A2 complex. Almagro *et al*. [80] constructed a model of 5C.C7/MCC 93-103/I-E^k to study the structural role for TCR α1, α2, β1 and β2, in MHC interaction. A framework was subsequently proposed to create testable hypothesis about TCR recognition.

Docking

Computer-simulated ligand binding or docking is a powerful technique for investigating intermolecular interactions. In general, the purpose of docking simulation is 2-fold—(i) to find the most probable translational, rotational and conformational juxtaposition of a given ligand-receptor pair, and (ii) to evaluate the relative goodness-of-fit or how well a ligand can bind to the receptor. Several docking techniques have been developed to address the

peptide/MHC combinatorial problem. Caflisch *et al.* [81] developed a combinatorial buildup algorithm to dock the influenza matrix peptide 58–68 to HLA-A*0201. Rosenfeld *et al.* [82, 83] utilized a multiple copy algorithm to identify probable termini peptide conformations and constructed the intervening sequence using a loop closure algorithm. Lim *et al.* [84] and Antes *et al.* [85] employed molecular dynamics (MD) simulation to examine the structures of class I peptide/MHC complexes. Bordner and Abagyan [86] applied the use of Monte Carlo simulations to predict the binding geometry of peptides to MHC class I alleles, while hybrid approaches that integrated the strength of Monte Carlo simulations and homology modeling were also applied to dock peptides to an array of class I and class II alleles [87–91].

PREDICTION OF MHC SUPERTYPES

The grouping of HLA allelic variants into superfamilies or supertypes on the basis of their structural features and/or binding specificities is important for development of epitope-based vaccines [92, 93]. Two groups of clustering techniques can be recognized in the literature reviewed—methods based on peptide specificities, and those that classify MHC alleles using 3D structural features.

Clustering using peptide specificities

A strategy for the development of epitope-based vaccines with wide population coverage is to identify HLA alleles that are present in most individuals from all major ethnic groups and ensuring that these alleles bind to at least one of the peptides in the vaccine. Accordingly, promiscuous peptides that bind more than one HLA allele are ideal for such purpose. By clustering MHC alleles on the basis of their peptide binding specificities, promiscuous T-cell epitopes that are representative of large proportion of human population can be identified. Sturniolo *et al.* [94] demonstrated the use of multiple quantitative matrices for predicting promiscuous peptides binding to HLA-DR alleles. Brusic *et al.* [95] combined peptide and MHC interaction sequences with a HMM to predict peptide binding to the HLA-A2 supertype. Guan *et al.* [39] employed the use of 2D-QSAR to investigate peptide specificities to four HLA-A3 alleles and formulated a refined HLA-A3 supertype motif. Lund *et al.* [96] constructed weight matrices representing the specificities of several HLA-DR alleles as well as all HLA-A and -B alleles in the SYFPEITHI database using a Gibbs sampling procedure. The distance matrices were clustered using the neighbor-joining method of Saitou and Nei [97]. This approach characterized HLA-A, -B and -DR alleles into five, seven and nine clusters, respectively according to their peptide-binding specificities.

Clustering using MHC structural features

An alternative approach for HLA supertype definition is to identify alleles with similar binding specificities from a structural interaction viewpoint. HLA alleles with similar binding specificities share common structural features within the peptide binding cleft. The binding clefts contain cavities (or anchor 'pockets') that correspond to primary and secondary anchor positions on the binding peptide. Doytchinova *et al.* [98, 99] demonstrated that only one to three amino acids within these binding pockets are sufficient to classify an allele to a particular class I or class II supertype. HLA-A, -B, -C, -DR, -DQ and -DP alleles were subsequently grouped into three, three, two, five DRs, three DQs and four DPs clusters, respectively. Kangueane *et al.* [100, 101] defined critical polymorphic functional residue positions for HLA-A, -B and -C alleles and grouped 47% of 295 A alleles, 44% of 540 B alleles and 35% of 156 C alleles to 36, 71 and 18 groups, respectively.

PREDICTION OF T-CELL EPITOPE REPERTOIRE INSIDE ANTIGENS

T-cell epitope prediction tools can be configured to identify high-density clusters of potential MHC-binding sequences within antigens. Meister *et al.* [22] developed a matrix-based clustering technique to identify high-density clusters of MHC alleles. By defining immunological hot spots as antigenic regions of up to 30 amino acids, Srinivasan *et al.* [102] and Zhang *et al.* [103] showed that T-cell epitopes clustered in certain regions of protein antigens such as SARS-CoV, dengue virus proteins, myelin oligodendrocyte glycoprotein, bee venom protein and hepatitis C virus 1B protein in a HLA supertype-dependent manner. By targeting these high-density regions of promiscuous T-cell epitopes,

the process of epitope discovery for vaccine development may be accelerated.

MODELLING ISSUES

The accuracy of a prediction model is highly dependent on the quantity and quality of available experimental data. Care should also be taken to ensure that there is no biasness in the data set. This section discusses the issues related to peptide data which have implications on the selection and performance of prediction model.

Data quantity

The availability of known peptide binders to specific alleles has a direct impact on the choice and quality of prediction model. Simple sequence motif models critically depend on the availability of training data set and are not applicable where data is unavailable. Where there is limited data or biasness in experimentally determined binding motifs, these models suffer from poor accuracy. A decade after peptide binding motif for pemphigus vulgaris (PV)-associated DR4 molecules was described [104], it was discovered that these motifs were insufficient for identification of PV epitopes presence of register shifts and polymorphisms in the binding registers [88]. In such scenarios, structure-based techniques are the only alternative predictors of peptide binders. However, the development of computational tools under this category is severely impeded by inherent complexities in terms of model building, data fitting and computational speed. As the number of known peptide binders increases, simple sequence motifs become more useful predictors, especially where rapid large-scale screening is involved. SVM outperforms ANN and decision tree using small training data set of 36 binders and 167 nonbinders [55]. An investigation on the predictive performance of ANNs revealed that this approach gradually outperforms motifs, matrices and HMMs with increasing peptide data [30]. ANN and HMM are the predictive methods of choice for MHC alleles with more than 100 known binders [30].

Data Quality

Noise and errors in the data sets have an adverse effect on the construction of useful predictive models. This may be a result of noncritical selection of peptides for constructing training data sets using existing binding motifs. A number of existing predictors have been built using 'virtual' binding and nonbinding peptides constructed from experimental binding motifs. While useful in practice, the utilization of such 'virtual' data requires dealing with imperfect and fuzzy measurements. Such forms of imprecise measurements are also observed when there is a need to combine data from multiple experimental sources. In such scenario, statistical techniques capable of handling fuzzy nonlinear data are recommended. Brusic *et al*. [105] investigated the impact of noise in data sets for constructing simple matrix models. They demonstrated that 5% of errors in a data set will double the number of data points required to build a matrix-based model. On the contrary, the same magnitude of error does not significantly affect the performance of ANNs [106].

Data biasness

Over-fitting occurs when a predictive model adapts too well to the training data and includes random disturbances in the training set as being significant. As these disturbances do not reflect the underlying distribution, the performance of the machine learning techniques on the given data set is affected. This over-fitting problem is typically avoided by using a regularizer [107, 108] that replaces the observed amino acid distribution by its estimator. Various techniques have been developed to avoid the over-fitting problem. Brusic *et al*. [46] pre-processed the training data set using a weighting scheme to penalize highly similar peptides. Peters *et al*. [34] introduced an additional term to the minimization function of SMM. The equation was subsequently solved using a generalized-reduced-gradient method. In order to facilitate transparent evaluation of newly developed prediction methods without data biasness and noises, there is a need for a framework for standardized side-by-side comparison of prediction methods as well as standardized training and testing data sets. The new framework proposed by Peters *et al*. [109] is excellent for such purposes.

A ROADMAP FOR THE PREDICTION OF MHC-BINDING PEPTIDES

In order to make sense of the bewildering array of tools available for prediction of MHC-binding peptides, we present a simplified solution as a roadmap, with a small set of selected options for each step. Specific resources, as listed in Table 1,

Table I: Listed in this table are some prediction software described in this review

Name	Methods	Coverage	URL
MULTIPRED	ANN, HMM	MHC class I and II binding	http://research.i2r.a-star.edu.sg/multipred/
SYFPEITHI	Binding matrices	MHC class I and II binding	http://www.syfpeithi.de/
EpiMer	Binding matrices	MHC class I and II binding	http://epivax.com
TEPITOPE	Binding matrices	MHC class I and II binding	http://www.vaccinome.com
VAGAT	ANN, HMM	Antigenome analysis	http://sdmc.i2r.a-star.edu.sg/vagat/
EpiDock	Homology modeling	MHC class I binding	http://bioinfo-pharma.u-strasbg.fr/cheminformatics-tools.php
PAProC	Network-based model	Proteosome cleavage	http://www.paproc.de/
NetChop	ANN	Proteosome cleavage	http://www.cbs.dtu.dk/services/NetChop/
PREDTAP	ANN, HMM	TAP binding peptides	http://antigen.i2r.a-star.edu.sg/predTAP/
SMM	Binding matrices	MHC class I binding	http://zlab.bu.edu/SMM
RANKPEP	Binding matrices	MHC class I and II binding Proteosome cleavage	http://bio.dfci.harvard.edu/Tools/rankpep.html
IEDB	ANN, SMM, Average Relative Binding (ARP) matrices	MHC class I and II binding Proteosome cleavage, TAP transport and MHC class I binding	http://www.immuneepitope.org/tools.do
EpiJen	Quantitative matrices	Proteosome cleavage, TAP transport and MHC class I binding	http://www.jenner.ac.uk/EpiJen/

have been selected according to their usefulness and performance. Some of the most reliable tools, such as MULTIPRED [102, 103], have already been used as base algorithms to develop more advanced methods for the analysis of antigenic hotspots with antigens. These programs have been demonstrated to be consistent independently or efficient as a part of different analysis pipelines and we wish to recommend these as a general '*modus operandi*' for small or large scale T-cell epitope-based projects.

For MHC alleles that have been extensively studied, sequence-based computational tools such as SYFPEITHI [25] or EpiMer [110] can effectively identify potential MHC-binding peptides. MULTIPRED [103] or TEPITOPE [111] can be used to predict promiscuous MHC class I and class II peptide ligands for a broad range of HLA-binding specificity at supertype level. VAGAT (http://sdmc.i2r.a-star.edu.sg/vagat/) is a viral antigenome analysis tool built on top of MULTIPRED [103] for systematic analyses of antigenic diversity from a set of viral protein sequences. Where novel MHC class I alleles are concerned, structure-based predictive strategies such as EpiDock [112], which combines homology modeling with a free energy scoring function FRESNO, can be applied for prediction of candidate binding peptides.

The PAProC [113, 114] or NetChop [115, 116] server can be used to predict potential cleavage sites of the human proteosomes. PRED^{TAP} [117] can be applied to predict transporter associated with antigen processing (TAP) binding peptides. For prediction of noncontinuous peptides [118], predicted peptide fragments may be combined and input into any of the above recommended software for prediction. The IEDB [109] and EpiJen [119, 120] provide integrated tools for predictions of antigen processing through the MHC class I antigen processing pathway. In addition, IEDB [109] contains tools for computing the population coverage of epitopes in different ethnicities, as well as the degree of conservancy of an epitope within a given protein sequence set at different degrees of sequence identity are also provided.

FUTURE DIRECTIONS

Recent advancements in computational modeling techniques and computational infrastructures are enabling immunologists to better explore the highly complex nature of the human immune system. The authors' view is that the next decade will bring increased focus on the development of computational techniques for large-scale analysis of the immune system at the system level.

At the level of T-cell epitopes, the proteasome- and TAP-dependent pathways play important roles in influencing the final peptide composition. In recent years, important steps toward the integration of computational tools modeling the different sub-components of the antigen processing pathway have been made. Dynamic activities over the past year have seen at least six reports of algorithms that integrated MHC class I peptide predictions with TAP and proteasomal cleavage specificities [109, 119–124]. These techniques are still in their infancy and need to be further developed and thoroughly tested. The recent discovery that T-cell epitopes may be formed by the fusion of two short noncontinuous peptide fragments [118] suggests that a much larger repertoire of T-cell epitopes exist, and T-cell epitope prediction is no longer confined to a linear search space. Accordingly, there is a need to adjust existing computational strategies to take into account the presence of nonlinear T-cell epitopes. Including considerations of secondary determinants such as longer class I binding peptides [125–127], expression levels of MHC locus products and their corresponding life-span may also improve the extant tools.

At the haplotype level, there is likely to be increasing focus on the development of computational tools targeting peptides binding to a complete set of MHC molecules in a model organism. An example of such system is $PRED^{BALB/c}$ [128] which focuses on prediction of peptides binding to the $H2^{d}$ haplotype of BALB/c mouse. Rapid progress in the development of computational infrastructures provides the foundation for large-scale simulation of more complex mammalian immune system. An example of such initiative is the European Virtual Human Immune System Project (http://www.immunogrid.org) that connects eight institutions to establish an infrastructure for modeling the human immune system at molecular, cellular and organ levels. These technologies permit a more complete view of the immune responses of a target organism with direct implications in peptide-based vaccine design. Therapeutically, it facilitates the identification of immunogenic epitopes as targets for selectively diminishing or altering immune reactions with minimal side effects according to the patient's genetic profile. This form of personalized immunotherapy can help prevent, ameliorate or cure disease and is particularly useful where treatment options are unsuccessful, limited or nonexistent [129, 130].

Key Point

- Several bioinformatics tools have been developed to identify T-cell epitopes to specific MHC alleles to facilitate vaccine development. This article examines existing strategies that utilize computational approaches for the study of peptide/MHC interactions. The most important bioinformatics tools and methods with relevance to the study of peptide/MHC interactions have been reviewed. We have also provided guidelines for predicting antigenic peptides based on the availability of existing experimental data.

CONCLUSIONS

The use of peptides that bind to MHC to induce specific T-cell mediated immune responses is a very attractive and well-studied field. Bioinformatics tools for prediction of T-cell epitopes are now a standard methodology [30, 131]. *In silico* T-cell epitope mapping, combined with *in vitro* and *in vivo* verification, accelerates the discovery process by approximately 10–20-fold [29]. Development of sophisticated bioinformatics tools will provide a platform for more in-depth analysis of immunological data and facilitate the construction of new hypotheses to explain the complex immune system function.

Acknowledgments

This project has been funded in part by the National Institute of Allergy and Infectious Diseases, National Institute of Health, Department of Health and Human Services, USA (Grant #5 U19 AI56541 & Contract #HHSN266200400085C).

References

1. Ferrari G, Kostyu DD, Cox J, *et al*. Identification of highly conserved and broadly cross-reactive HIV type 1 cytotoxic T lymphocyte epitopes as candidate immunogens for inclusion in Mycobacterium bovis BCG-vectored HIV vaccines. *AIDS Res Hum Retroviruses* 2000;**16**:1433–43.
2. Haselden BM, Kay AB, Larche M. Peptide-mediated immune responses in specific immunotherapy. *Int Arch Allergy Immunol* 2000;**122**:229–37.
3. Singh RR. The potential use of peptides and vaccination to treat systemic lupus erythematosus. *Curr Opin Rheumatol* 2000;**12**:399–406.
4. Wang E, Phan GQ, Marincola FM. T-cell-directed cancer vaccines: the melanoma model. *Expert Opin Biol Ther* 2001;**1**:277–90.

5. Williams TM. Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J Mol Diagn* 2001;**3**:98–104.
6. Roberts JD, Niedzwiecki D, Carson WE, *et al*. Phase 2 study of the g209-2M melanoma peptide vaccine and low-dose interleukin-2 in advanced melanoma: Cancer and Leukemia Group B 509901. *J Immunother* 2006;**29**: 95–101.
7. Bourdette DN, Edmonds E, Smith C, *et al*. A highly immunogenic trivalent T cell receptor peptide vaccine for multiple sclerosis. *Mult Scler* 2005;**11**:552–61.
8. Lopez JA, Weilenman C, Audran R, *et al*. A synthetic malaria vaccine elicits a potent CD8(+) and CD4(+) T lymphocyte immune response in humans. Implications for vaccination strategies. *Eur J Immunol* 2001;**31**:1989–98.
9. Knutson KL, Schiffman K, Disis ML. Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients. *J Clin Invest* 2001; **107**:477–84.
10. Falk K, Rötzschke O, Deres K, *et al*. Identification of naturally processed viral nonapeptides allows their quantification in infected cells and suggests an allele-specific T cell epitope forecast. *J Exp Med* 1991;**174**:425–34.
11. Falk K, Rötzschke O, Stevanovic S, *et al*. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 1991;**351**:290–6.
12. Jardetzky TS, Lane WS, Robinson RA, *et al*. Identification of self peptides bound to purified HLA-B27. *Nature* 1991; **353**:326–9.
13. Hunt DF, Henderson RA, Shabanowitz J, *et al*. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 1992; **255**:1261–3.
14. Roetzschke O, Falk K, Stefanovic S, *et al*. Exact prediction of a natural T cell epitope. *Eur J Immunol* 1991;**21**:2891–4.
15. Zhang QJ, Gavioli R, Klein G, Masucci SG. An HLA-A11-specific motif in nonamer peptides derived from viral and cellular proteins. *Proc Natl Acad Sci USA* 1993;**90**: 2217–21.
16. Lipford GB, Hoffman M, Wagner H, Heeg K. Primary in vivo responses to ovalbumin. Probing the predictive value of the Kb binding motif. *J Immunol* 1993;**4**:1212–22.
17. Sette A, Sidney J, Oseroff C, *et al*. HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. *J Immunol* 1993;**151**: 3163–70.
18. Sidney J, Oseroff C, del Guercio MF, *et al*. Definition of a DQ3.1-specific binding motif. *J Immunol* 1994;**152**: 4516–25.
19. Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* 1994; **152**:163–75.
20. Hammer J, Bono E, Gallazzi F, *et al*. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med* 1994; **180**:2353–8.
21. Rammensee HG, Friede T, Stevanović S. MHC ligands and peptide motifs: first listing. *Immunogenetics* 1995;**41**: 178–228.
22. Meister GE, Roberts CGP, Berzofsky JA, De Groot AS. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine* 1995;**13**:581–91.
23. D'Amaro J, Houbiers JGA, Drijfhout JW, *et al*. A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum Immunol* 1995;**43**:13–8.
24. Rajapakse M, Schmidt B, Brusic V. Multi-objective evolutionary algorithm for discovering peptide binding motifs. *Lecture Notes in Computer Science* 2006; 149–58.
25. Rammensee H, Bachmann J, Emmerich NP, *et al*. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999;**50**:213–9.
26. Chen W, Khilko S, Fecondo J, *et al*. Determinant selection of major histocompatibility complex class I-restricted antigenic peptides is explained by class I-peptide affinity and is strongly influenced by nondominant anchor residues. *J Exp Med* 1994;**180**:1471–83.
27. Jameson SC, Bevan MJ. Dissection of major histocompatibility complex (MHC) and T cell receptor contact residues in a Kb-restricted ovalbumin peptide and an assessment of the predictive power of MHC-binding motifs. *Eur J Immunol* 1992;**22**:2663–7.
28. Ruppert J, Sidney J, Celis E, *et al*. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* 1993;**74**:929–37.
29. Martin W, Sbai H, De Groot AS. Bioinformatics tools for identifying class I-restricted epitopes. *Methods* 2003;**29**: 289–98.
30. Yu K, Petrovsky N, Schonbach C, *et al*. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med* 2002;**8**:137–48.
31. Davenport MP, Ho Shon IAP, Hill AVS. An empirical method for the prediction of T-cell epitopes. *Immunogenetics* 1995;**42**:392–7.
32. Gulukota K, Sidney J, Sette A, DeLisi C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol* 1997;**267**:1258–67.
33. Schafer JR, Jesdale BM, George JA, *et al*. Prediction of well-conserved HIV-1 ligands using a matrix-based algorithm, EpiMatrix. *Vaccine* 1998;**16**:1880–4.
34. Reche PA, Glutting JP, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* 2002;**63**:701–9.
35. Peters B, Tong W, Sidney J, *et al*. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 2003;**19**: 1765–72.
36. Nielsen M, Lundegaard C, Worning P, *et al*. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 2004;**20**: 1388–97.
37. Rajapakse M, Wyse L, Schmidt B, Brusic V. Deriving matrix of peptide-MHC interactions in diabetic mouse by genetic algorithm. *Lecture Notes in Computer Science* 2005; **3578**:440–7.
38. Guan P, Doytchinova IA, Zygouri C, Flower DR. MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. *Appl Bioinformatics* 2003;**2**:63–6.

39. Guan P, Doytchinova IA, Flower DR. HLA-A3 supermotif defined by quantitative structure–activity relationship analysis. *Protein Eng* 2003;**16**:11–8.
40. Doytchinova IA, Blythe MJ, Flower DR. Additive method for the prediction of protein-peptide binding affinity. Application to the MHC Class 1 molecule HLA-A*0201. *J Proteome Res* 2002;**1**: 263–72.
41. Duda RO, Hart PE, Stork DG. *Pattern classification*. New York: Wiley-Interscience, 2001.
42. Savoie CJ, Kamikawaji N, Sasazuki T, Kuhara S. Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs. *Pac Symp Biocomput* 1999;182–9.
43. Segal MR, Cummings MP, Hubbard AE. Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics* 2001;**57**:632–42.
44. Zurada JM. *Introduction to Artificial Neural Systems*. St. Paul, MN, USA: PWS Publishing Co, 1999.
45. Brusic V, Rudy G, Harrison LC. Prediction of MHC binding peptides using artificial neural networks. In: Stonier RJ, Yu XS (eds). *Complex Systems: Mechanism of Adaptation*. Amsterdam: IOS Press, 1994; 253–60.
46. Brusic V, Rudy G, Honeyman M, *et al*. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 1998; **14**:121–30.
47. Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. *J Immunol Methods* 1995;**185**:181–90.
48. Milik M, Sauer D, Brunmark AP, *et al*. Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat Biotechnol* 1998;**16**: 753–6.
49. Buus S. Description and prediction of peptide-MHC binding: the human MHC project. *Curr Opin Immunol* 1999;**11**:209–13.
50. Nielsen M, Lundegaard C, Worning P, *et al*. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 2003;**12**: 1007–17.
51. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 1989; **77**:257–86.
52. Mamitsuka H. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 1989;**33**:460–74.
53. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press, 1998;51–68.
54. Han LY, Cai CZ, Ji ZL, *et al*. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res* 2004;**32**: 6437–44.
55. Zhao Y, Pinilla C, Valmori D, *et al*. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics* 2003;**19**:1978–84.
56. Dönnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC. Bioinformatics* 2002;**3**:25.
57. Bhasin M, Raghava GPS. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* 2004;**20**:421–3.
58. Bozic I, Zhang G, Brusic V. Predictive vaccinology: optimisation of predictions using support vector machine classifiers. *Lecture Notes in Computer Science* 2005;**3578**: 375–81.
59. Bhasin M, Raghava GPS. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* 2004;**22**: 3195–204.
60. Akutsu T, Sim KL. Protein threading based on multiple protein structure alignment. *Genome Inform* 1999; **10**:23–9.
61. Sezerman U, Vajda S, DeLisi C. Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Protein Sci* 1996;**5**:1272–81.
62. Altuvia Y, Schueler O, Margalit H. Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol* 1995;**249**:244–50.
63. Miyazawa S, Jernigan RL. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 1985;**18**: 534–52.
64. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 1996;**256**:623–44.
65. Altuvia Y, Sette A, Sidney J, *et al*. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol* 1997;**58**:1–11.
66. Schueler-Furman O, Elber R, Margalit H. Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des* 1998;**3**:549–64.
67. Kangueane P, Sakharkar MK, Lim KS, *et al*. Knowledge-based grouping of modeled HLA peptide complexes. *Hum Immunol* 2000;**61**:460–6.
68. Schueler-Furman O, Altuvia Y, Sette A, Margalit H. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci* 2000;**9**:1838–46.
69. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* 1999;**8**:361–9.
70. Jojic N, Reyes-Gomez M, Heckerman D, *et al*. Learning MHC I-peptide binding. *Bioinformatics* 2006;**22**:e227–35.
71. Bui HH, Schiewe AJ, von Grafenstein H, Haworth IS. Structural prediction of peptides binding to MHC class I molecules. *Proteins* 2006;**63**:43–52.
72. Doytchinova IA, Flower DR. Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J Med Chem* 2001;**44**:3572–3581.
73. Swindells MB, Thornton JM. Structure prediction and modelling. *Curr Opin Biotechnol* 1991;**2**:512–9.
74. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* 1993;**234**: 774–815.
75. Hammer J, Gallazzi F, Bono E, *et al*. Peptide binding specificity of HLA-DR4 molecules: correlation with

rheumatoid arthritis association. *J Exp Med* 1995;**181**: 1847–55.

76. Rognan D, Laumoeller SL, Holm A, *et al*. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* 1999;**42**:4650–8.
77. Logean A, Sette A, Rognan D. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg Med Chem Lett* 2001;**11**:675–9.
78. Michielin O, Luescher I, Karplus M. Modeling of the TCR-MHC-peptide complex. *J Mol Biol* 2000;**300**: 1205–35.
79. Michielin O, Karplus M. Binding free energy differences in a TCR-peptide-MHC complex induced by a peptide mutation: a stimulation analysis. *J Mol Biol* 2002;**324**: 547–69.
80. Almagro JC, Vargas-Madrazo E, Lara-Ochoa F, Horjales E. Molecular modeling of a T-cell receptor bound to a major histocompatibility complex molecule: Implications for T-cell recognition. *Protein Sci* 1995;**4**: 1708–11.
81. Caflisch A, Niederer P, Anliker M. Monte Carlo docking of oligopeptides to proteins. *Proteins* 1992;**13**:223–30.
82. Rosenfeld R, Zheng Q, Vajda S, DeLisi C. Computing the structure of bound peptides: Application to antigen recognition by class I major histocompatibility complex receptors. *J Mol Biol* 1993;**234**:515–21.
83. Rosenfeld R, Zheng Q, Vajda S, DeLisi C. Flexible docking of peptides to class I major-histocompatibility-complex receptors. *Genet Anal* 1995;**12**:1–21.
84. Lim JS, Kim S, Lee HG, *et al*. Selection of peptides that bind to the HLA-A2.1 molecule by molecular modelling. *Mol Immunol* 1996;**33**:221–30.
85. Antes I, Siu SW, Lengauer T. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics* 2006;**22**:e16–24.
86. Bordner AJ, Abagyan R. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins* 2006;**63**:512–26.
87. Tong JC, Tan TW, Ranganathan S. Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci* 2004;**13**:2523–32.
88. Tong JC, Bramson J, Kanduc D, *et al*. Modeling the bound conformation of pemphigus vulgaris-associated peptides to MHC class II DR and DQ Alleles. *Immunome Res* 2006;**2**:1.
89. Tong JC, Zhang GL, Tan TW, *et al*. Prediction of HLA-DQ3.2β ligands: evidence of multiple registers in class II binding peptides. *Bioinformatics* 2006;**22**:1232–38.
90. Ranganathan S, Tong JC. A practical guide to structure-based prediction of MHC binding peptides. In: Flower DR (ed.). *Methods in Molecular Biology: Immunoinformatics: predicting Immunogenicity in silico*. Humana Press, Totawa NJ (in press).
91. Ranganathan S, Tong JC, Tan TW. Structural immunoinformatics. In: Schoenbach C, Brusic V, Konagaya A (eds). *Immunomics Reviews: An official P. of the International Immunomics Society, vol 1: Immunoinformatics*. Springer (in press).
92. Sette A, Livingstone B, McKinney D, *et al*. The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. *Biologicals* 2001;**29**:271–6.
93. Sette A, Newman M, Livingston B, *et al*. Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. *Tissue Antigens* 2002;**59**:443–51.
94. Sturniolo T, Bono E, Ding J, *et al*. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999;**17**:555–61.
95. Brusic V, Petrovsky N, Zhang G, Bajic VB. Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol Cell Biol* 2002;**80**:280–5.
96. Lund O, Nielsen M, Kesmir C, *et al*. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 2004;**12**:797–810.
97. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
98. Doytchinova IA, Guan P, Flower DR. Identifying human MHC supertypes using bioinformatic methods. *J Immunol* 2004;**172**:4314–23.
99. Doytchinova IA, Flower DR. In silico identification of supertypes for class II MHCs. *J Immunol* 2005;**174**: 7085–95.
100. Kangueane P, Sakharkar MK, Rajaseger G, *et al*. A framework to sub-type HLA supertypes. *Frontiers in Bioscience* 2005;**10**:879–86.
101. Zhao B, Png AEH, Ren EC, *et al*. Compression of functional space in HLA-A sequence diversity. *Hum Immunol* 2003;**64**:718–28.
102. Srinivasan KN, Zhang GL, Khan AM, *et al*. Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens. *Bioinformatics* 2004; **20**(Suppl 1):i297–302.
103. Zhang GL, Khan AM, Srinivasan KN, *et al*. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res* 2005; **33**:W172–9.
104. Wucherpfennig KW, Yu B, Bhol K, *et al*. Structural basis for major histocompatibility complex (MHC)-linked susceptibility to autoimmunity: charged residues of a single MHC binding pocket confer selective presentation of self-peptides in pemphigus vulgaris. *Proc Natl Acad Sci USA* 1995;**92**:11935–9.
105. Brusic V, Schonbach C, Takiguchi M, *et al*. Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. *Proc Int Conf Intell Syst Mol Biol* 1997;**5**:75–83.
106. Hammerstrom D. Neural networks at work. *IEEE Spectrum* 1993;**30**:26–32.
107. Karplus K. Evaluating regularizers for estimating distributions of amino acids. *Proc Int Conf Intell Syst Mol Biol* 1995;**3**: 188–96.
108. Choo KH, Tong JC, Zhang L. Recent applications of hidden Markov models in computational biology. *Genom Proteom Bioinform* 2004;**2**:84–96.
109. Peters B, Sidney J, Bourne P, *et al*. The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* 2005;**3**:e91.

110. Meister GE, Roberts CGP, Berzofsky JA, De Groot AS. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. *Vaccine* 1995;**13**:581–91.

111. Sturniolo T, Bono E, Ding J, *et al*. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999;**17**:555–61.

112. Logean A, Rognan D. Recovery of known T-cell epitopes by computational scanning of a viral genome. *J Comput Aided Mol Des* 2002;**16**:229–43.

113. Kuttler C, Nussbaum AK, Dick TP, *et al*. An algorithm for the prediction of proteasomal cleavages. *J Mol Biol* 2000; **298**:417–29.

114. Nussbaum AK, Kuttler C, Hadeler KP, *et al*. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics* 2001;**53**:87–94.

115. Kesmir C, Nussbaum AK, Schild H, *et al*. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng* 2002;**15**:287–96.

116. Nielsen M, Lundegaard C, Lund O, Kesmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 2005;**57**:33–41.

117. Zhang GL, Petrovsky N, Kwok CK, *et al*. PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res* 2006;**2**:3.

118. Hanada K, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* 2004;**427**:252–6.

119. Doytchinova IA, Flower DR. Class I T-cell epitope prediction: improvements using a combination of proteasome cleavage, TAP affinity, and MHC binding. *Mol Immunol* 2006;**43**:2037–44.

120. Doytchinova IA, Guan P, Flower DR. EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinformatics* 2006;**7**:131.

121. Levitsky V, Zhang Q-J, Levitskaya J, Masucci MG. The lifespan of MHC-peptide complexes influences the efficiency of presentation and immunogenicity of two class I restricted cytotoxic T lymphocyte epitopes in the Epstein-Barr virus nuclear antigen-4. *J Exp Med* 1996;**183**: 915–26.

122. Larsen MV, Lundegaard C, Lamberth K, *et al*. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol* 2005;**35**:2295–303.

123. Tenzer S, Peters B, Bulik S, *et al*. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci* 2005;**62**:1025–37.

124. Donnes P, Kohlbacher O. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci* 2005;**14**:2132–40.

125. Tong JC, Kong L, Tan TW, Ranganathan S. MPID-T: database for sequence-structure-function information on TCR/peptide/MHC interactions. *Appl Bioinform* 2006;**5**: 111–4.

126. Probst-Kepper M, Hecht HJ, Herrmann H, *et al*. 'Conformational restraints and flexibility of 14-meric peptides in complex with HLA-B3501'. *J Immunol* 2004; **173**:5610–6.

127. Burrows SR, Rossjohn J, McCluskey J. Have we cut ourselves too short in mapping CTL epitopes? *Trends in Immunol* 2006;**27**:11–6.

128. Zhang GL, Srinivasan KN, Veeramani A, *et al*. PREDBALB/c: a system for the prediction of peptide binding to H2d molecules, a haplotype of the BALB/c mouse. *Nucleic Acids Res* 2005;**33**:W180–3.

129. Evavold BD, Sloan-Lancaster J, Allen PM. Tickling the TCR: selective T cell functions stimulated by altered peptide ligands. *Immunol Today* 1993;**14**:602–9.

130. Bielekova B, Martin R. Antigen-specific immunomodulation via altered peptide ligands. *J Mol Med* 2001;**79**: 552–65.

131. Schirle M, Weinschenk T, Stevanović S. Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T-cell epitopes from defined antigens. *J Immunol Methods* 2001;**257**:1–16.