Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools

Lianming Zhang, Keiko Udaka, Hiroshi Mamitsuka and Shanfeng Zhu

Submitted: 4th July 2011; Received (in revised form): 23rd August 2011

Abstract

Binding of short antigenic peptides to major histocompatibility complex (MHC) molecules is a core step in adaptive immune response. Precise identification of MHC-restricted peptides is of great significance for understanding the mechanism of immune response and promoting the discovery of immunogenic epitopes. However, due to the extremely high MHC polymorphism and huge cost of biochemical experiments, there is no experimentally measured binding data for most MHC molecules. To address the problem of predicting peptides binding to these MHC molecules, recently computational approaches, called pan-specific methods, have received keen interest. Pan-specific methods make use of experimentally obtained binding data of multiple alleles, by which binding peptides (binders) of not only these alleles but also those alleles with no known binders can be predicted. To investigate the possibility of further improvement in performance and usability of pan-specific methods, this article extensively reviews existing pan-specific methods and their web servers. We first present a general framework of pan-specific methods. Then, the strategies and performance as well as utilities of web servers are compared. Finally, we discuss the future direction to improve pan-specific methods for MHC-peptide binding prediction.

Keywords: MHC-peptide binding prediction; pan-specific; immunoinformatics

INTRODUCTION

Binding of short peptide fragments derived from antigenic pathogens to major histocompatibility complex (MHC) molecules is a core and the most selective step to initiate an adaptive immune response to eliminate the source pathogens. MHC molecules bind peptide fragments and present them on the cell surface to T-cell receptors, by which T-cell recognition of non-self peptide and subsequent immune response start [1, 2]. Precise identification of which peptides can bind to specific MHC molecules is of great significance for understanding the mechanism of immune response, as well as enhancing the discovery of immunogenic epitopes. In particular, promiscuous MHC epitopes that can bind to multiple MHC alleles are essential candidates for vaccine development. Epitope vaccines are considered as a safe and effective way against infectious diseases [3], such as hepatitis and EB virus, from which a large proportion of human population constantly suffer. Moreover, the peptides that bind to certain MHC molecules offer a promising immunotherapy to treat more serious human diseases, such as cancer [4], autoimmunity [5] and allergy [6].

Corresponding author. Shanfeng Zhu, School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China. Tel: +8621-65643786; Fax: +8621-6564253; E-mail: zhusf@fudan.edu.cn

Lianming Zhang is a postgraduate student at the Shanghai Key Lab of Intelligent Information Processing and School of Computer Science in Fudan University. His research interests include immunological bioinformatics and data mining.

Keiko Udaka is a professor at the Department of Immunology in Kochi Medical School. Her research interests include immunology, immunological bioinformatics and peptide-based immunotherapy.

Hiroshi Mamitsuka is a professor at the Institute for Chemical Research in Kyoto University, jointly appointed as a professor of the School of Pharmaceutical Sciences of the same university. His research interests include machine learning, data mining and their applications in bioinformatics and chemoinformatics.

Shanfeng Zhu is an associate professor at Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University, Shanghai, China. His research interests include machine learning, data mining and their applications in bioinformatics and information retrieval.

© The Author 2011. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com
MHC [human leukocyte antigen (HLA) for human] molecules have two general classes: class I (MHC-I or HLA-I for human) and class II (MHC-II or HLA-II for human). Furthermore, there are subclasses in each of these two classes. HLA-I mainly has A, B, and C subclasses, while HLA-II mainly has DP, DQ, and DR subclasses [7]. Further note that different from MHC-I molecule consisting of one chain, each MHC-II molecule has two chains, \( \alpha \) (A) and \( \beta \) (B). For example, HLA-DRB indicates \( \beta \)-chain of the subclass DR of human MHC.

MHC-I and MHC-II molecules play different roles in T-cell-mediated adaptive immunity [1, 2]. In the MHC-I pathway of antigen processing and presenting to T lymphocytes, endogenous antigens are first cleaved into peptide fragments by the proteasome, which are then generally translocated by the transporter associated with antigen processing (TAP) into the endoplasmic reticulum (ER). Finally, MHC-I molecules bind certain peptides and present them to cytotoxic T lymphocytes (CTL)-stimulating cellular immunity. On the other hand, in the MHC-II pathway, exogenous antigens are first taken into the cell through endocytosis, and then degraded to peptides within endosomes and lysosomes mainly by aspartic and cysteine proteases (e.g. cathepsin). MHC-II molecules are synthesized in the ER, form complexes with invariant chain (Ii), which blocks the peptide binding cleft of MHC-II, and facilitates MHC-II entering into golgi from ER. Later MHC-II complex fuses with endosome containing exogenous peptides. Finally, with the help of another MHC-like molecule, such as HLA-DM in humans, MHC-II can bind exogenous peptides and present them to T help cells. Both antigen processing and presentation are important in the process of T-cell-mediated adaptive immunity, but peptide binding to MHC molecules is the most selective step, which is the focus of this article. In a recent study, Assarsson et al. [8] state that <3% of possible peptides could bind MHC molecules strong enough to generate an immune response.

MHC molecules are highly diverse because of an extensive range of MHC polymorphism. In fact, the IMGT/HLA database of June 2011 [9] contains over 6000 HLA alleles, which include 4946 HLA-I and 1457 HLA-II alleles. Each encoded MHC molecule binds to a distinct set of peptides, but binding preferences of most alleles have not yet been experimentally characterized, mainly because of two reasons. First, biological experiments require immense amount of time and financial cost. Second, the number of possible peptides derived from pathogens is huge, while binding peptides (binders) will be merely a tiny fraction of all those possible peptides [8, 10]. Thus, a variety of computational methods, which are more efficient than doing experiments, have been developed for capturing peptides binding to specific MHC molecules [11].

Figure 1 shows a schematic overview of traditional allele-specific computational methods [12–20], in which only binding data to an allele can be used to train a model, and the model is only applied to predict peptides binding to the allele. Methods, e.g. NetMHC [12, 13], SMM [14] for MHC-I and SMM-align [17], NN-align [18], RTA [19] for MHC-II, are all outstanding allele-specific ones with decent accuracy [20]. They typically require at least 50–200 quantitative peptide binding measurements for each target allele [50–100 for MHC-I (14, 26) and 100–200 for MHC-II (17), respectively] to build a model with reasonable accuracy. However, again MHC polymorphism is extremely broad and biological experiments are both time- and cost-consuming, resulting in only a limited number of experimentally obtained binders. In fact, as of June 2011, only less than 30 alleles in a total of 1457 HLA-II alleles have more than 200 quantitative peptide binding measurements in the largest database of MHC binding peptides, IEDB [21]. This makes allele-specific methods difficult to learn and predict binding specificities of most MHC molecules. Thus, in order to address the problem of predicting peptides binding to MHC molecules with only a small number of experimentally obtained binders, a series of computational approaches, called pan-specific methods [22–36], have been developed and received intense interest recently.

Pan-specific methods use experimental data of multiple MHC alleles (source alleles) as input and attempt to predict binders of not only the input alleles but also the alleles with very few or even no known binders (target alleles). The idea behind pan-specific methods is to establish connections among different MHC alleles in terms of binding specificities, by which the binding specificities of source and target alleles can be captured. It is noted that in principle pan-specific methods can predict binders of any allele with known protein sequences [24, 32]. Compared to allele-specific methods, breakthrough in allele coverage and reasonable prediction accuracy enable pan-specific methods to be the most promising
solutions for MHC peptide binding prediction. It is worth mentioning that MULTIPRED [35, 36] is a pioneer work in this direction. Sette and Sidney [37] suggested that MHC molecules can be grouped into supertypes, where a group of MHC molecules in the same supertype share similar binding specificities. MULTIPRED incorporated binding data within same HLA supertype and trained supertype-specific models to predict promiscuous HLA binding peptides [35, 36].

This article extensively reviews existing pan-specific methods for MHC-peptide binding, focusing on both MHC-I and MHC-II. As illustrated in Figure 2, the peptide binding groove of MHC-I is different from that of MHC-II. The groove of MHC-I molecules has closed ends, which accommodate the whole binding peptide. Thus, MHC-I binding peptides are 8–11 consecutive residues, typically nonamers [38]. On the other hand, the groove of MHC-II molecules has open ends, which generally bind to longer peptides, normally 14–18 residues [39]. Only a small part (usually nine amino acid residues, called binding core) is fitted into the groove, with remaining peptide termini on both ends extending outside. The binding between a peptide and an MHC-II molecule depends upon the binding core, and the optimal binding core can be estimated as the one with the highest binding affinity. Given a peptide, the optimal binding core cannot be easily decided, and therefore prediction of peptides binding to MHC-II is much more challenging than that of MHC-I.

Currently, pan-specific methods for MHC-I include ADT [22], KISS [23], NetMHCPan [24, 25] and PickPocket [26], while TEPITOPE [27, 28], MHCIIMulti [29], SIADT [30], MultiRTA [31] and NetMHCIIpan (1.0 and 2.0) [32, 33] are pan-specific methods for MHC-II. MULTIPRED2 [34], build upon previous tools MULTIPRED [36] and PEPVAC [43], is also a pan-specific tool that makes use of NetMHCpan and NetMHCIIpan as prediction engines for both HLA-I and HLA-II supertypes and alleles. Some studies have also been carried out on the prediction of other pathway events in antigen processing and presenting, such as proteasome cleavage and TAP transport for MHC-I. Generally speaking, binding prediction itself can identify possible epitopes quite accurately, but higher accuracy should be achieved by incorporating predictions that model these two pathway events. One example is NetCTLpan [44], improved from the original NetCTL [45] and another state-of-art
pathway prediction method *MHC-pathway* [46]. It is a fine MHC-I pathway epitope discovery tool that integrated predictions of NetChop cleavage sites [47], TAP transport efficiency [48] and NetMHCpan MHC binding [25]. Such methods that incorporate the high-accuracy binding prediction method and one or two of the other event prediction models have received much interest. All these methods, including solo peptide binding prediction methods, normally can be applied to pre-screen target peptides within pathogenic strains cheaply, followed by experimental validation of smaller number of well-selected candidates, for example, in epitope-based vaccine design [34].

In the literature of MHC-peptide binding, [11] is a brief review on computational strategies in this field and [49] is a review on prediction tools, while [50–54] are not necessarily reviews but comparative studies on the performance of different prediction methods using several independent data sets. However, all these studies concentrate on allele-specific methods and do not contain the latest pan-specific methods. On the other hand, [55] is a benchmark study of pan-specific methods, which are restricted to MHC-I with little discussion on methodological strategies. In addition, [56–58] are more general recent reviews, which trace the history of MHC-I and MHC-II binding prediction methods, with highlighting the most useful and historically important approaches without putting focus on pan-specific methods. In this light, this review focuses on existing pan-specific methods and tools to examine the development in utility and performance. We first introduce a general framework of pan-specific methods. The strategies and performance of different pan-specific methods as well as utilities of their web servers are then compared. Finally, we discuss the future direction to further improve the performance of pan-specific MHC-peptide binding prediction.

**GENERAL FRAMEWORK AND EVALUATION OF PAN-SPECIFIC METHODS**

**Framework**

State-of-the-art pan-specific methods are generally data-driven, which are trained using a large amount of quantitative peptide binding measurements [25, 32, 33]. Figure 3 gives a general framework with two stages: (i) model training and (ii) prediction by the trained model. In the training stage, both peptides and MHC contact environment information are incorporated to establish the connection among different MHC molecules in binding specificities. An underlying assumption of pan-specific methods is that MHC molecules with similar contact environment have similar binding specificities. MHC contact environment is typically represented by MHC pseudo sequence, which consists of polymorphic residues that form the binding
### Table 1: Availability and utilities of pan-specific methods

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
<th>Method</th>
<th>Date</th>
<th>HLA</th>
<th>!HLA</th>
<th>Extend</th>
<th>Super</th>
<th>MultiSel</th>
<th>Input</th>
<th>PepLen</th>
<th>Package</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MHC-I pan-specific prediction tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KISS</td>
<td><a href="http://cbio.ensmp.fr/kiss/">http://cbio.ensmp.fr/kiss/</a></td>
<td>Kernel</td>
<td>12/2007</td>
<td>ABC</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>Pep</td>
<td>9</td>
<td>YES</td>
<td>[23]</td>
</tr>
<tr>
<td>PickPocket</td>
<td>No public server available</td>
<td>PSSM</td>
<td>3/2009</td>
<td>YES</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>NO</td>
<td>[26]</td>
</tr>
<tr>
<td>MULTIPRED2-I</td>
<td><a href="http://cvc.dfci.harvard.edu/multipred2/HTML/prediction1.php">http://cvc.dfci.harvard.edu/multipred2/HTML/prediction1.php</a></td>
<td>Integration/NetMHCpan</td>
<td>12/2010</td>
<td>ABC</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>FasSeq</td>
<td>8–11</td>
<td>NO</td>
<td>[34]</td>
</tr>
<tr>
<td><strong>MHC-II pan-specific prediction tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEPITOPE</td>
<td><a href="http://www.imtech.res.in/raghava/propred/">http://www.imtech.res.in/raghava/propred/</a></td>
<td>PSSM</td>
<td>7/1999</td>
<td>50 DRs</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>Seq, FasSeq</td>
<td>9</td>
<td>YES</td>
<td>[27,28]</td>
</tr>
<tr>
<td>MHCII Multi</td>
<td><a href="http://www.epitoolkit.org/mhcimulti/">http://www.epitoolkit.org/mhcimulti/</a></td>
<td>Multi-Instance learning</td>
<td>2008</td>
<td>DR</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>Seq, FasSeq</td>
<td>≥ 9</td>
<td>NO</td>
<td>[29]</td>
</tr>
<tr>
<td>SIADT</td>
<td>No public server available</td>
<td>Shifts-Invariant ADT</td>
<td>9/2008</td>
<td>DR, DP, DQ</td>
<td>YES</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>NO</td>
<td>[30]</td>
</tr>
<tr>
<td>MultiRTA</td>
<td><a href="http://bordnerlab.org/MultiRTA">http://bordnerlab.org/MultiRTA</a></td>
<td>RTA [19]</td>
<td>9/2010</td>
<td>DR, DP</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>Pep</td>
<td>≥ 9</td>
<td>NO</td>
<td>[31]</td>
</tr>
<tr>
<td>NetMHCIIpan-1.0</td>
<td><a href="http://www.cbs.dtu.dk/services/NetMHCIIpan-1.0/">http://www.cbs.dtu.dk/services/NetMHCIIpan-1.0/</a></td>
<td>ANN/SMM-align [17]</td>
<td>7/2008</td>
<td>DR</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>Pep, FasSeq</td>
<td>≥ 9</td>
<td>YES</td>
<td>[32]</td>
</tr>
<tr>
<td>NetMHCIIpan-2.0</td>
<td><a href="http://www.cbs.dtu.dk/services/NetMHCIIpan-2.0/">http://www.cbs.dtu.dk/services/NetMHCIIpan-2.0/</a></td>
<td>ANN/NN-align [18]</td>
<td>11/2010</td>
<td>DR</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>Pep, FasSeq</td>
<td>≥ 9</td>
<td>YES</td>
<td>[33]</td>
</tr>
<tr>
<td>MULTIPRED2-II</td>
<td><a href="http://cvc.dfci.harvard.edu/multipred2/HTML/prediction2.php">http://cvc.dfci.harvard.edu/multipred2/HTML/prediction2.php</a></td>
<td>Integration/NetMHCIIpan</td>
<td>12/2010</td>
<td>DR</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>FasSeq</td>
<td>9</td>
<td>NO</td>
<td>[34]</td>
</tr>
</tbody>
</table>

(i) Date gives the date of that the corresponding paper is published. (ii) MHC indicates alleles to be covered (HLA denotes HLA loci covered; !HLA shows whether prediction is available for non-human species or not; Extend shows whether the web server can be extended to a new allele (which is not in the online list) by providing the aligned or full amino acid sequence of the new allele. In contrast to NetMHCIIpan-2.0, NetMHCIIpan version 2.1 (updated in June 2011) allows users to input any full length MHC class II beta chain to predict peptides from any protein. Super indicates whether the server can predict peptides binding to the supertype. (iii) MultiSel shows whether the tool allows us to select multiple alleles when inputted. (iv) Input displays possible input format, where Pep (FasPep) means peptides in separate lines (in FASTA format) and Seq (FasSeq) means antigen protein sequence (in FASTA format). (v) PepLen gives the permitted length to segment peptide. (vi) Package shows "YES" if a stand-alone software package or scoring matrices are publicly available; otherwise NO. Here note that the latest NetMHCpan can also predict peptides binding to hundreds of MHC alleles of six non-human species: chimpanzee, rhesus macaque, gorilla, pig, mouse and BoLA. No server is available for PickPocket and SIADT, by which their attribute values are shown following the papers of these methods or simply using "--".
groove for the binding core of a given peptide [26, 27]. It is noted that the predictive performance of pan-specific methods, for novel MHC alleles in particular, have a strong dependence on the distance to training data, where distance is measured in terms of amino acid similarity between MHC pocket residues. This is because they learn the binding specificity by leveraging information from neighboring MHC molecules. The performance normally decreases as the distance increases, which has been demonstrated by several benchmarks [24–26, 32, 33]. The differences among pan-specific methods mainly lie in the following three points: (i) representation of peptides, (ii) representation of MHC contact environment and (iii) machine learning models employed. In this section, a variety of aspects in relation to pan-specific methods, including the above three points, are presented. Further description of each method is presented in Supplementary Section 1 in Supplementary Data.

**Representation of peptides**

MHC-I binding peptides are typically nonamers and represented in a position-wise manner, being categorized into two types: (i) 20 amino acids at each position, which is a simple approach, being used in PickPocket [26] and ADT [22]; (ii) considering the similarity between amino acids by some knowledge such as the BLOSUM matrix, which is used in NetMHCpan [24, 25] and KISS [23]. On the other hand, the representation of MHC-II binding peptides is usually more complicated since the peptide length can be variable. In this case, for example, not only nine residues in the binding core but also peptide flanking residues (which are flanking the binding core), as well as the peptide length, are considered as input. This representation is reflected in NetMHCIpan-1.0 [32] and NetMHCIpan-2.0 [33]. Another unique input can be found in MHCIIMulti [29], where a peptide is transformed into multiple nonamers (allowing overlaps), each

---

**Figure 3:** Schematic illustration of pan-specific methods [55]. (A and B) are two types of major data incorporated in training. (C) Shows the model trained using binding data of A1, making prediction for a new allele A2 and other alleles.
being further transformed into five-dimensional feature vectors by using physicochemical properties of amino acids.

**Representation of MHC contact environment**

According to the granularity that measures the MHC–peptide interaction, MHC contact environment can be represented in the following three ways: (i) allele-based, (ii) pocket-based and (iii) residue-based methods. (i) The allele-based method represents an MHC molecule by a vector of amino acid residues which are in contact with the binding peptide. This method is used in NetMHCpan [24, 25], KISS [23], NetMHCIIpan-1.0 [32] and NetMHCIIpan-2.0 [33]. (ii) Pockets are polymorphic cavities in the MHC binding groove, which is assumed to consist of nine pockets, each corresponding to one residue of the binding core. The pocket-based method represents an MHC molecule with nine pocket profiles, each being a set of MHC amino acid residues contacting the corresponding amino acid in the binding core. This manner can be found in PickPocket [26], TEPITOPE [27], MHCIIMulti [29], MultiRTA [31] and MULTIPRED [35]. (iii) In the residue-based method, all contacting pairs of peptide amino acid residues and MHC polymorphic amino acids are considered for the MHC-peptide binding. This method can be seen in ADT [22], SIADT [30].

**Machine learning methods**

Various typical machine learning methods have been adopted, which include artificial neural networks (ANN), kernel methods with support vector machines (SVM), logistic regression with regularization and k-nearest neighborhood (k-NN). One relatively minor machine learning technique, which fits the problem setting of pan-specific prediction well, is multiple instance learning (MIL). In MIL, a set of bags are given, where one bag has at least one instances. Labels are not assigned to instances but bags. If a bag is positive, one or more instance of the bag will be positive; otherwise, all instances in the bag are negative. Learning and predicting are done by using labeled bags. This setting is true for the problem of predicting MHCII-peptide binding, where a bag with instances corresponds to a peptide with nonamers, and a binder (i.e. a positive bag) must contain at least one positive nonamer, i.e. a binding core. Another popular approach, which is not necessarily a machine learning method, is position-specific scoring matrix (PSSM), being used in some methods such as TEPITOPE and PickPocket, since the resultant PSSM is visually comprehensible.

**Connections between MHC-I and MHC-II pan-specific methods**

Although peptides binding to MHC-I are very different from those to MHC-II, MHC-I and MHC-II pan-specific methods adopt the same techniques. For example, SIADT is an extension of ADT by considering all possibilities of the binding core, i.e., both methods are based on the idea of threading. Another typical example is that the same machine learning method is used by both MHC-I and MHC-II pan-specific methods. Specifically, NetMHCpan and NetMHCIIpan use ANN, KISS and MHCII Multi use SVM, while PickPocket and TEPITOPE are PSSM-based methods. Another point is that all methods incorporate MHC contact environment, which is the essence of pan-specific methods.

**Evaluation criteria**

Three types of experimental data are used—MHC binding peptides, endogenously presented ligands and T-cell epitopes [33], where typically MHC binding peptides are quantitative binding data and ligands/epitopes data are binary. Selection of evaluations measure depends upon types of data. However, area under the receiving operator characteristic (ROC) curve (AUC) [59] is a common evaluation criterion on both numerical and binary data, i.e. binding, ligand and epitope data. On the other hand, for evaluation based on quantitative binding data, Pearson correlation coefficient (PCC) [60] is used to assess linear correlation between prediction scores and (log-transformed) binding affinities. Spearman rank-order correlation (SRC) is also used when measuring nonlinear correlation [60]. When predicting peptides binding to MHC-II, the length of a given peptide can exceed nine, so another important evaluation measure is to identify the binding core that mediates MHC-peptide binding.

**Evaluation procedures**

There are three major validation experiments: (i) cross-validation, (ii) Leave-One-allele-Out (LOO) and (iii) independent test. (i) Cross-validation of pan-specific methods, for example, 5-fold cross-validation has the following manner: (a) the
pool of unique peptides is first randomly divided into 5-folds of equal size to ensure that all MHC binding data points related to a given peptide are put in the same fold and the same peptide cannot be found in more than 1-fold; (b) then choose 1-fold to test the model that was trained by the remaining 4-folds; (c) the step (b) is repeated over all 5-folds and the average performance of the 5 runs is computed. (ii) LOO means that prediction is made for one allele using the model trained by data of all other alleles [32, 33]. In other words, an uncharacterized allele is predicted by using already-known alleles, which is consistent with the practical setting where pan-specific methods are applied. (iii) The independent test is used to further evaluate trained models, publicly available servers in particular. This case uses independent data coming from another data source and without overlap with training data [55].

Databases and benchmark data sets
Predictive performance of data-driven computational models greatly relies on quantity and quality of experimental data available. Major databases concerning MHC-peptide binding prediction are listed in Table 2. To evaluate the performance of different computational prediction methods, several benchmark data sets have been compiled in different studies, which are given in Table 3. In this review, two more data sets Pan-I and Pan-II are prepared for evaluating MHC-I and MHC-II pan-specific methods, respectively. More details on these two data sets are available in the Supplementary Section 2 in Supplementary Data.

PERFORMANCE COMPARISON
For fair comparison, we attempt to use common training data (and test data independent from training data) for all pan-specific methods. However, we note that some methods are provided by web servers only, where training data might not be the same as those of other methods. We use binomial tests to evaluate the significance of the observed performance difference between two methods. The statistical significance is judged by setting the cut-off for the $P < 0.05$, where Method A > Method B means that Method A significantly outperforms Method B by the binomial test, while (Method A and Method B) means no significant difference is observed between Method A and Method B by the binomial test.

Comparison of MHC-I pan-specific methods: ADT, KISS, NetMHCpan and PickPocket
We compared the performance of the methods using three types of data: MHC binding peptides, ligands and T-cell epitopes. It is noted that the experimental results of PickPocket on HLA ligand and epitope data are unavailable since PickPocket does not have public available implementations.

**Predicting peptides binding to MHC**
Three large non-overlapped data sets are involved here, i.e. Peters [54], human data [55] and non-human data [26] of Zhang, which were all derived from IEDB [21]. Experimental results are taken from the related literature [26, 55]. On the Peters data set, results of 5-fold cross-validation and LOO indicated that NetMHCpan (NetMHCpan-2.0) significantly outperformed others; PickPocket and KISS performed comparable with each other, both significantly better than ADT. Table 4 shows the independent test results obtained by using the Peters and human data of Zhang for training and testing, respectively. Results showed that NetMHCpan and PickPocket outperformed ADT and KISS, whereas KISS and ADT did not show significant difference [26]. In summary, NetMHCpan > PickPocket > (KISS, ADT).

An additional result on the non-human data of Zhang indicated that PickPocket on average was
better than NetMHCpan [26], while ADT and KISS are unavailable for non-human peptide binding prediction.

### Identifying HLA ligands

As well as predicting MHC binding peptides, experimental results are taken from the related literature [23, 26, 55]. Table 5 shows the result for identifying endogenously presented ligands. NetMHCpan outperformed KISS, being significant for the performance average per-ligand but not significant for the per-allele. Both methods outperformed ADT significantly [55]. In summary, in per-ligand case, NetMHCpan > KISS > ADT.

### Predicting T-cell epitopes

We further evaluated the web servers of ADT, KISS and NetMHCpan-2.0 by identifying T-cell epitopes, using our original data set Pan-I, obtained from IEDB on March, 2011. Here, we discarded the epitopes which were used for training ADT and KISS in order to avoid overlaps between training and test data. Pan-I contains 208 T-cell epitopes covering

---

**Table 3: Benchmark data sets for evaluating pan-specific methods**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Brief</th>
<th>MHC</th>
<th>URL</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pan-I data sets</td>
<td>T-cell epitopes from IEDB, source protein from Uniprot</td>
<td>I</td>
<td><a href="http://www.biokdd.fudan.edu.cn/Service/panview.htm">http://www.biokdd.fudan.edu.cn/Service/panview.htm</a></td>
<td></td>
</tr>
<tr>
<td>Wang-1.0 data sets</td>
<td>MHC-II binding peptides, PDB structure, Th cell epitopes (2008)</td>
<td>II</td>
<td><a href="http://mhcbindingpredictions.immuneepitope.org/">http://mhcbindingpredictions.immuneepitope.org/</a></td>
<td>[53]</td>
</tr>
<tr>
<td>DFRMLI data sets</td>
<td>Preprocessed and scaled immunological binding and ligand data sets for machine learning (2008)</td>
<td>I and II</td>
<td><a href="http://bio.dfci.harvard.edu/DFRMLI/">http://bio.dfci.harvard.edu/DFRMLI/</a></td>
<td>[50, 52]</td>
</tr>
<tr>
<td>Pan-II data sets</td>
<td>Binary binding data from IEDB for evaluation</td>
<td>II</td>
<td><a href="http://www.biokdd.fudan.edu.cn/Service/panview.htm">http://www.biokdd.fudan.edu.cn/Service/panview.htm</a></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4: Comparison of MHC-I pan-specific methods on MHC binding data**

<table>
<thead>
<tr>
<th>Measure</th>
<th>KISS</th>
<th>ADT</th>
<th>PickPocket</th>
<th>NetMHCpan</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCC</td>
<td>0.455</td>
<td>0.488</td>
<td>0.55</td>
<td>0.62</td>
</tr>
<tr>
<td>SRC</td>
<td>0.44</td>
<td>0.522</td>
<td>–</td>
<td>0.60</td>
</tr>
<tr>
<td>AUC</td>
<td>0.734</td>
<td>0.756</td>
<td>–</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Performance values were excerpted from [26]. Online predictors trained using the Peters data set were evaluated on the Zhang data set in terms of three measures, i.e. PCC, SRC and AUC. The Peters data set has 29 336 HLA-peptide binding measurements as IC50 values for 35 human alleles. The Zhang data set has 6533 measurements for 33 human alleles.

**Table 5: Comparative performance results on identifying HLA-I ligands**

<table>
<thead>
<tr>
<th>Allele</th>
<th>No. of ligand</th>
<th>ADT</th>
<th>KISS</th>
<th>NetMHCpan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave per ligand</td>
<td>596</td>
<td>0.886</td>
<td>0.931</td>
<td>0.976</td>
</tr>
<tr>
<td>Ave per allele</td>
<td>34</td>
<td>0.919</td>
<td>0.945</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Performance values were excerpted from [54]. Online predictors were compared on the Zhang ligand data set that has 596 ligands covering 13 HLA-A alleles and 21 HLA-B alleles. Ave per ligand gives the average AUC over a total of 596 ligands, and Ave per allele gives the average over the Ave per ligand of all alleles.
29 HLA-I alleles, whose source proteins can be found in the UniProt database [65]. Table 6 shows the summary results obtained by applying the competing methods to Pan-I (see details in Supplementary Table S1 in Supplementary Data).

Results show that NetMHCpan significantly outperformed ADT and KISS. KISS performed significantly better than ADT on the per-epitope basis, but not significantly on the per-allele. Simply, NetMHCpan > (KISS, ADT).

**Summary**
NetMHCpan and PickPocket are the two best pan-specific methods. Concretely, NetMHCpan outperformed others significantly in most cases, whereas PickPocket outperformed NetMHCpan in one case. ADT worked better on quantitative MHC binding data while KISS performed favorably against ligand and epitope data, which might be due to the reason that KISS was trained by binary binding data [55].

**Comparison of MHC-II pan-specific methods: TEPITOPE, MultiRTA and NetMHCIIpan**
We conducted the performance comparison upon four types of data: HLA binding peptides, HLA ligands, T-cell epitopes and the binding core of HLA-peptide complexes. The web server of MHCIIMulti is not suitable for automatically submitting a large-scale data set of peptides, so we did not test MHCIIMulti [A previous study showed that MHCII Multi achieved a comparable performance with TEPITOPE on some data set (29)]. We did not test SIADT either since SIADT does not have public available implementation.

**Predicting MHC binding peptides**
We retrieved experimental results performed on an HLA binding data set, which called Nielsen-1.0 [32]. Table 7 shows the results of LOO on the Nielsen-1.0 data set taken from the related literature [31–33], which indicate that NetMHCIIpan-2.0 outperformed both NetMHCIIpan-1.0 and MultiRTA significantly, and MultiRTA and NetMHCIIpan-1.0 were comparable with each other. With respect to 11 alleles covered by TEPITOPE, all other methods significantly outperformed TEPITOPE. Overall, NetMHCIIpan-2.0 > (MultiRTA, NetMHCIIpan-1.0) > TEPITOPE.

Moreover, we have built our original HLA binding data set, named as Pan-II, which covers 17 HLA-DRB alleles and were retrieved from IEDB for independent test, meaning that there are no overlaps between their training data sets and Pan-II. Table 8 shows the results over Pan-II, indicating that out of all 17 alleles, NetMHCIIpan-2.0, NetMHCIIpan-1.0, MultiRTA and TEPITOPE achieved the highest AUC in eight, five, two and two alleles, respectively. In the average AUC over all 17 alleles, NetMHCIIpan-1.0 achieved the highest AUC of 0.722, which was closely followed by NetMHCIIpan-2.0 with AUC of 0.721, and then MultiRTA with AUC of 0.692. In a pair-wise comparison, NetMHCIIpan-2.0 performed significantly better than MultiRTA, whereas not against NetMHCIIpan-1.0 and TEPITOPE.

**Identifying HLA ligands**
We retrieved Nielsen-2.0 [33] data sets for HLA ligands, and the results over this data set were

### Table 6: Comparative performance results on identifying HLA-I epitopes

<table>
<thead>
<tr>
<th>Allele</th>
<th>No. of epitope</th>
<th>ADT</th>
<th>KISS</th>
<th>NetMHCpan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave per epitope</td>
<td>208</td>
<td>0.897</td>
<td>0.921</td>
<td>0.944</td>
</tr>
<tr>
<td>Ave per allele</td>
<td>29</td>
<td>0.893</td>
<td>0.886</td>
<td>0.950</td>
</tr>
</tbody>
</table>

208 epitopes were obtained from 13 HLA-A alleles, 14 HLA-B alleles and 2 HLA-C alleles. Ave per epitope gives the average AUC over a total of 208 epitopes, and Ave per allele gives the average over the Ave per epitope of all alleles.

### Table 7: LOO comparison of MHC-II pan-specific methods on the Nielsen-1.0 data set

<table>
<thead>
<tr>
<th>Measure</th>
<th>TEPITOPE</th>
<th>NetMHCIIpan-1.0</th>
<th>MultiRTA</th>
<th>NetMHCIIpan-2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave AUC</td>
<td>–</td>
<td>0.768</td>
<td>0.773</td>
<td>0.799</td>
</tr>
<tr>
<td>In Tepitope(II)</td>
<td>0.718</td>
<td>0.786</td>
<td>0.781</td>
<td>0.819</td>
</tr>
</tbody>
</table>

Performance values were excerpted from [31–33]. The Nielsen-1.0 data set has 14 607 numerical binding data, which cover 14 HLA-DRB alleles, II of which is characterized by TEPITOPE. Elements in the table are values of the AUC.
recomputed. Table 9 shows the results (see details in Supplementary Table S2 in Supplementary Data). NetMHCIIPan-2.0 performed the best, with the average per-ligand AUC of 0.829 and the per-allele of 0.797. NetMHCIIPan-2.0 significantly outperformed MultiRTA in both per-ligand and per-allele. We further checked the performance over the 17 alleles characterized by TEPITOPE. For these alleles, TEPITOPE achieved the highest average AUC of 0.811, but only significantly outperformed MultiRTA. The performance of MultiRTA, NetMHCIIPan-1.0 and NetMHCIIPan-2.0 were all slightly improved for the other 11 alleles.

Predicting T-cell epitopes
We retrieved Nielsen-2.0 [33] data sets for T-cell epitopes, originated from the IEDB database, being independent from training data. Table 10 shows the results (see details in Supplementary Table S3 in Supplementary Data). NetMHCIIPan-2.0 performed the best overall, achieving the average AUC of 0.751 for per-epitope and that of 0.779 for per-allele. For per-epitope, NetMHCIIPan-2.0 outperformed NetMHCIIPan-1.0 significantly, and both of them significantly outperformed MultiRTA. We further checked the performance over the 20 alleles used by TEPITOPE. For these 20 alleles, TEPITOPE achieved the highest average AUC of 0.755 for the per-allele, only significantly outperforming MultiRTA.

Identifying binding core of the HLA–peptide complex
Finally, we tested the performance on identifying the peptide binding core by using 20 unique

Table 8: Comparison of MHC-II pan-specific methods on independent binding data

<table>
<thead>
<tr>
<th>Allele</th>
<th>No. of bind</th>
<th>TEPITOPE</th>
<th>MultiRTA</th>
<th>NetMHCIIPan-1.0</th>
<th>NetMHCIIPan-2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1*01:01</td>
<td>732</td>
<td>270</td>
<td>0.763</td>
<td>0.767</td>
<td>0.778</td>
</tr>
<tr>
<td>DRB1*03:01</td>
<td>557</td>
<td>247</td>
<td>0.608</td>
<td>0.623</td>
<td>0.652</td>
</tr>
<tr>
<td>DRB1*04:01</td>
<td>779</td>
<td>386</td>
<td>0.737</td>
<td>0.744</td>
<td>0.763</td>
</tr>
<tr>
<td>DRB1*04:02</td>
<td>113</td>
<td>31</td>
<td>0.782</td>
<td>0.753</td>
<td>0.788</td>
</tr>
<tr>
<td>DRB1*04:04</td>
<td>204</td>
<td>65</td>
<td>0.784</td>
<td>0.764</td>
<td>0.802</td>
</tr>
<tr>
<td>DRB1*04:05</td>
<td>95</td>
<td>36</td>
<td>0.724</td>
<td>0.728</td>
<td>0.719</td>
</tr>
<tr>
<td>DRB1*07:01</td>
<td>507</td>
<td>222</td>
<td>0.65</td>
<td>0.646</td>
<td>0.639</td>
</tr>
<tr>
<td>DRB1*08:02</td>
<td>68</td>
<td>32</td>
<td>0.694</td>
<td>0.747</td>
<td>0.739</td>
</tr>
<tr>
<td>DRB1*11:01</td>
<td>575</td>
<td>219</td>
<td>0.743</td>
<td>0.734</td>
<td>0.729</td>
</tr>
<tr>
<td>DRB1*12:01</td>
<td>693</td>
<td>444</td>
<td>–</td>
<td>0.7</td>
<td>0.753</td>
</tr>
<tr>
<td>DRB1*13:01</td>
<td>168</td>
<td>40</td>
<td>0.722</td>
<td>0.71</td>
<td>0.693</td>
</tr>
<tr>
<td>DRB1*13:02</td>
<td>88</td>
<td>40</td>
<td>0.556</td>
<td>0.591</td>
<td>0.645</td>
</tr>
<tr>
<td>DRB1*15:01</td>
<td>251</td>
<td>149</td>
<td>0.717</td>
<td>0.639</td>
<td>0.713</td>
</tr>
<tr>
<td>DRB3*01:01</td>
<td>146</td>
<td>23</td>
<td>–</td>
<td>0.548</td>
<td>0.627</td>
</tr>
<tr>
<td>DRB3*02:02</td>
<td>648</td>
<td>315</td>
<td>–</td>
<td>0.685</td>
<td>0.735</td>
</tr>
<tr>
<td>DRB4*01:01</td>
<td>156</td>
<td>67</td>
<td>–</td>
<td>0.542</td>
<td>0.628</td>
</tr>
<tr>
<td>DRB5*01:01</td>
<td>217</td>
<td>151</td>
<td>0.888</td>
<td>0.851</td>
<td>0.874</td>
</tr>
<tr>
<td>Avg (17)</td>
<td>5997</td>
<td>2737</td>
<td></td>
<td>0.692</td>
<td>0.722</td>
</tr>
<tr>
<td>!In T epitope(3)</td>
<td>4354</td>
<td>1888</td>
<td>0.721</td>
<td>0.715</td>
<td>0.733</td>
</tr>
<tr>
<td>!In T epitope(4)</td>
<td>1643</td>
<td>849</td>
<td></td>
<td>0.619</td>
<td>0.686</td>
</tr>
</tbody>
</table>

The results were obtained from web servers or stand-alone packages. The 17 alleles include 13 alleles covered by TEPITOPE (InT epitope) and 4 alleles not covered by TEPITOPE (!InT epitope). The largest AUC value for each allele is highlighted in bold.

Table 9: Evaluation on ligand data

<table>
<thead>
<tr>
<th>SYFPEITHI</th>
<th>No. of ligand</th>
<th>TEPITOPE</th>
<th>MultiRTA</th>
<th>NetMHCIIPan-1.0</th>
<th>NetMHCIIPan-2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave per ligand</td>
<td>1164</td>
<td>0.760</td>
<td>0.799</td>
<td>0.829</td>
<td>0.829</td>
</tr>
<tr>
<td>Ave per allele</td>
<td>28</td>
<td>0.756</td>
<td>0.787</td>
<td>0.797</td>
<td>0.797</td>
</tr>
<tr>
<td>In Tepitope</td>
<td>17</td>
<td>0.811</td>
<td>0.733</td>
<td>0.767</td>
<td>0.785</td>
</tr>
<tr>
<td>!In Tepitope</td>
<td>11</td>
<td>0.791</td>
<td>0.818</td>
<td>0.814</td>
<td>0.814</td>
</tr>
</tbody>
</table>

MultiRTA is newly incorporated in this benchmark comparison. Number of ligand is the total number of ligands (for Ave per ligand) or alleles (for Ave per allele). The maximum AUC in each row is highlighted in bold.
Abstract

HLA-peptide complexes (with known binding cores) of seven HLA-DR alleles, which we retrieved from PDB (Protein Data Bank) [41]. It is noted that 15 out of 20 can be obtained from [32]. Results in Table 11 show that TEPITOPE achieved the best performance—only 1 out of 20 complexes were incorrectly predicted (see details in Supplementary Table S4 in Supplementary Data).

Summary

The best pan-specific method for MHC-II varies with experimental settings. For predicting HLA binding peptides, NetMHCIIpan-2.0 performed the best overall, being followed by NetMHCIIpan-1.0 and MultiRTA; for identifying the peptide binding core, TEPITOPE achieved the best performance, being followed by NetMHCIIpan-1.0 and MultiRTA. And for identifying HLA class II ligands and T-cell epitopes, if the target alleles are covered by TEPITOPE, TEPITOPE was the best choice, being followed by NetMHCIIpan-2.0; if not, NetMHCIIpan-2.0 was the best choice, being followed by NetMHCIIpan-1.0. Please also note that only MultiRTA can offer predictions for HLA DP alleles, while all other pan-specific methods focus only on HLA DR alleles.

CONCLUSION AND DISCUSSION

Predicting peptides binding to MHC molecules in silico has witnessed significant progress in the past decade. Pan-specific methods, in particular, have brought an important breakthrough concerning both prediction accuracy and allele coverage. At the moment, predictions with decent accuracy are available for HLA-I (A, B, C loci) and HLA-II (DR, DP loci) alleles if amino acid sequences are known. An important reason for this achievement is that the amount of high-quality experimental data currently available has increased and a larger number of MHC molecule sequences have been determined. They offer a solid basis to realize the high-performance of pan-specific methods, which make the epitope discovery effectively advance in practice right now and will further improve in the days to come. Thus, we have focused on pan-specific methods in this article.

Table 1 summarizes of the utilities of different pan-specific tools/methods. It clearly shows features of each tool. For example, NetMHCpan can predict binders, which are variable in length (8–11 amino acids), while only nonamers can be predicted by KISS. In fact, a variety of subclasses are covered by methods shown in the table, but still quite a few alleles are not covered by these methods yet. Therefore, the development of pan-specific methods remains to cover wider alleles and species, particularly, for non-human MHC-II alleles. In terms of usability of tools, e.g. web servers shown in Table 1, public availability and convenient interfaces are expected for further study and analysis. It would be better to offer stand-alone package, like
NetMHCpan and NetMHCIIpan. A good example for further analysis is MULTIPRED2, which incorporates two state-of-the-art methods, NetMHCpan and NetMHCIIpan, performing large-scale prediction over various HLA-I and -II supertypes. Therefore, it will be a very practical tool in vaccine design by offering visual analysis of antigenic regions and population coverage calculation of allele. Further improvement in interfaces as well as performance will make the tools closer to an ideal, practical application.

Beyond this table, we would like to emphasize that the predictive performance for MHC-II remains significantly lower than that for MHC-I, and performance for non-human alleles apparently lower than that for human alleles. Therefore, the development of highly accurate pan-specific methods for wider types of alleles and species would be the future direction of research.

In this article, we have compared the predictive performance of four pan-specific methods for MHC-I and six for MHC-II, respectively. There is no single method that always outperformed the others, and the best method varied with different data and alleles. For example, for MHC-I, NetMHCpan achieved the best performance in predicting peptides binding to human MHC molecules, while PickPocket outperformed NetMHCpan in predicting those to non-human MHC molecules. For MHC-II, NetMHCIIpan-2.0 achieved the best performance in predicting HLA-DR binding peptides, while TEPITOPE was better than NetMHCIIpan-2.0 for the alleles covered by TEPITOPE over a variety of problems, such as identifying ligands, epitopes and binding cores. These results indicate that we are unable to choose one method that can be universally applicable in all situations. A possible approach to deal with this situation and improve the overall predictive performance is to employ ensemble learning techniques, by which possible methods for predicting peptides binding to MHC can be optimally combined together, depending on the data and alleles. The ensemble learning strategy has already been widely presented and used in machine learning field [66], in particular, recently applied to the problem of this review in an allele-specific manner [53, 67–69]. Here, we point out two existent problems when combining different pan-specific methods. First, implementations of pan-specific methods are not necessarily open-accessed. In fact, publicly available stand-alone packages or source codes are most useful to make an ensemble, but some methods such as PickPocket and SIADT even do not have web servers. Second, web servers of some pan-specific methods, such as KISS and ADT, have not been updated for a long time. These methods could achieve higher predictive performances if inside engines of their web servers could be trained again by more up-to-date experimental data.

Overall, this review has provided rich and latest information on pan-specific methods, for predicting MHC-peptide binding in a comprehensive manner. We believe that this review would be helpful for applying and improving a variety of currently available pan-specific prediction tools and methods.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org.

Key Points
- The pursuit to predict the binding specificities of MHC molecules with very few or even no binding data has resulted in pan-specific MHC peptide binding prediction methods, which has drawn intensive interest.
- Different strategies have been adopted in different pan-specific methods and the best MHC-I and MHC-II pan-specific methods vary with different data and alleles.
- Overall, NetMHCpan and NetMHCIIpan-2.0 are best pan-specific methods for MHC-I and MHC-II, respectively.
- Pan-specific methods need to be further improved in terms of allele coverage, prediction accuracy and availability.

FUNDING
The National Nature Science Foundation of China (No. 61170097 and 60903076), the Japan Society for the Promotion of Science (JSPS) Invitation Fellowship and the Collaborative Research Program of Institute for Chemical Research, Kyoto University (grant #2011-19).

References
helper HER-2/neu peptide linked to the Ii-Key moiety of the invariant chain. *Int J Cancer* 2007;121(9):2031–41.


40. The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.