Machine Learning Methods for Predicting HLA–Peptide Binding Activity

Heng Luo[1,2], Hao Ye[1], Hui Wen Ng[1], Leming Shi[3], Weida Tong[1], Donna L. Mendrick[1] and Huixiao Hong[1]

[1]National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, USA. [2]University of Arkansas at Little Rock/University of Arkansas for Medical Sciences Bioinformatics Graduate Program, Little Rock, AR, USA. [3]Center for Pharmacogenomics, School of Pharmacy, Fudan University, Shanghai, China.

Supplementary Issue: Current Developments in Machine Learning Techniques in Biological Data Mining

ABSTRACT: As major histocompatibility complexes in humans, the human leukocyte antigens (HLAs) have important functions to present antigen peptides onto T-cell receptors for immunological recognition and responses. Interpreting and predicting HLA–peptide binding are important to study T-cell epitopes, immune reactions, and the mechanisms of adverse drug reactions. We review different types of machine learning methods and tools that have been used for HLA–peptide binding prediction. We also summarize the descriptors based on which the HLA–peptide binding prediction models have been constructed and discuss the limitation and challenges of the current methods. Lastly, we give a future perspective on the HLA–peptide binding prediction method based on network analysis.

KEYWORDS: HLA, peptide, binding, prediction, machine learning, MHC

SUPPLEMENT: Current Developments in Machine Learning Techniques in Biological Data Mining

CITATION: Luo et al. Machine Learning Methods for Predicting HLA–Peptide Binding Activity. *Bioinformatics and Biology Insights* 2015:9(S3) 21–29 doi: 10.4137/BBI.S29466.

TYPE: Review

RECEIVED: June 23, 2015. **RESUBMITTED:** July 30, 2015. **ACCEPTED FOR PUBLICATION:** August 02, 2015.

ACADEMIC EDITOR: J.T. Efird, Associate Editor

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 303 words, excluding any confidential comments to the academic editor.

FUNDING: The research was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (Heng Luo, Hao Ye, and Hui Wen Ng) administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. The research was also partially supported by grant funding from the National Institutes of Health (NIH) and National Institute of General Medical Sciences (NIGMS) (P20 GM103429) (formerly P20RR016460). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Huixiao.Hong@fda.hhs.gov; Donna.Mendrick@fda.hhs.gov

Background

Human leukocyte antigens (HLAs) and peptides. The major histocompatibility complexes (MHCs), a major component of the vertebrate immune system, are expressed on cell surfaces for cellular recognition and antigen presentation. In humans, the MHCs are called HLAs. Located at the short arm of chromosome 6, HLAs are one of the most polymorphic genes in humans and are different among countries and ethnicities.[1–3] According to the statistics of the international ImMunoGeneTics database/HLA database,[4] >12,500 HLA alleles have been recorded by March 2015. HLA alleles are systematically named in such a way, for example, HLA-A*02:01, where A indicates the HLA-A gene locus and 02:01 specifies the protein sequence for this allele. Detailed information of HLA nomenclature can be found at http://hla.alleles.org/. Three human MHC categories have been identified as Classes I, II, and III due to their different genetic loci. The Class I HLAs, including HLA loci A, B, C, E, F, and G, are codominantly expressed on the surface of all nucleated cells. They present intracellular-processed antigen peptides to helper CD8+ T-cells for cytotoxicity responses such as natural-killer-cell-induced apoptosis.[5–7] Class II HLAs, including HLA D locus, are selectively expressed on the surface of dendritic cells, B-cells, and other antigen-presenting cells. They present the antigen peptides to helper CD4+ T-cells to trigger acquired immune responses such as B-cell activation.[8–10] The Class III MHCs function in the complement system for the clearance of pathogens.[11,12] For Class I and Class II HLAs, studying their binding to peptides is essential to understand the immune system.

The Class I and Class II HLAs are similar in structure. Both classes of HLAs have a long binding groove that can bind peptides degraded from antigens. Though HLAs contain two chains, the binding grooves of Class I HLAs are determined by only α chain, while those of Class II HLAs consist of both α and β chains.[13–15] In addition, these two classes of HLAs bind to peptides with different lengths. While Class I HLAs bind to shorter peptides around 9-mers, Class II HLAs can

bind to a large variety of peptides around 15-mers or longer due to their open-ended binding grooves.[16–18] Though the peptide binders of Class II HLAs are generally longer, the core-binding regions are still around nine residues.[15] Therefore, when predicting the binding between Class II HLAs and peptides, extra processes are sometimes needed to determine which part of the peptide binds within the HLA pockets.[18]

In addition to the structural differences, Class I and Class II HLAs present the antigen peptides to trigger immune responses in different pathways as shown in Figure 1.[19,20]

Most peptides presented by Class I HLAs are from endogenous cytosolic proteins (eg, defective products and even viral proteins if the cell is infected by virus) synthesized by the cell itself. These proteins are degraded by proteasomes into peptides that are transported into endoplasmic reticulum (ER) by transporters associated with antigen processing and loaded onto Class I HLAs. After glycosylated in Golgi apparatus, the Class I HLA–peptide complexes are fused into cell membrane and presented to the T-cell receptors (TCRs) on CD8+ T-cells for cellular immune responses. If the CD8+ cell, or cytotoxic T-cell, recognizes the specific antigen, the CD8+ cell can trigger the presenting cell to undergo apoptosis. The peptides presented by Class II HLAs are usually from extracellular antigens. The exogenous antigens are engulfed into the endocytic route compartments and digested by proteases. Class II HLAs synthesized in ER and glycosylated in Golgi apparatus acquire the peptides in the vesicular compartments and present them on the cell surface. The Class II HLA–peptide complexes are then recognized by TCRs of CD4+ T-cells to further the immune response such as antibody synthesis.[19,20] The two pathways are not separated, and antigens that are mainly processed by one class of HLAs can be presented by the other via a cross-presentation pathway. However, details of this mechanism still remain obscure.[21]

HLAs play an important role in the immune system to present peptides to TCRs for immune responses; however, this process may result in adverse outcomes under certain circumstances. Autoimmunity can occur when HLAs may present peptides that are structurally similar to self-peptides to TCRs.[22] Exogenous drugs may react with the antigen protein, insert into the binding groove of HLAs, or interfere with the HLA–peptide–TCR complex to cause adverse events (not shown in Fig. 1).[23–25] The variety of HLAs, peptides, and TCRs all affect the immune response and make it challenging to understand the underlying mechanisms that could help with the prevention of adverse events. However, our recent study showed that by considering the binding peptide inside the HLA-binding groove, the performance of molecular modeling and prediction was improved.[13] Thus, understanding HLA–peptide binding can help interpret the interaction mechanisms between the drugs and HLAs. To address the complexity of the immune system and to improve the ability to understand and even to predict, the HLA–peptide binding

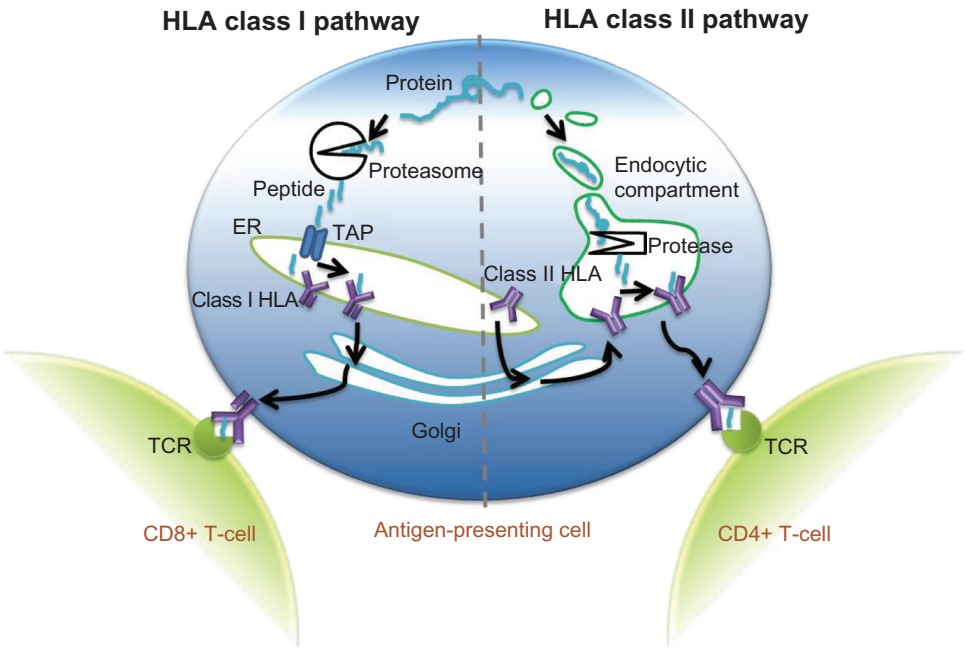

Figure 1. The typical pathways by which HLAs present antigen peptides to T-cells. In the HLA Class I pathway, endogenous antigen proteins are degraded by proteasomes into peptides that are transported via transporters associated with antigen processing (TAPs) into the ER. The peptides are loaded onto Class I HLAs and the complexes are sent to the Golgi apparatus for modification. Finally, the complexes are fused into the cell membrane where they can be recognized by TCRs on CD8+ T-cells. In the HLA Class II pathway, exogenous protein antigens are ingested by the cell into endocytic vesicular compartments and loaded onto Class II HLAs in the ER and processed by Golgi apparatus. The complexes are presented on the cell surface and recognized by TCR of CD4+ T-cells.

is definitely a crucial step. Various methods have already been developed to address such needs.

HLA–peptide binding prediction. The methods for HLA–peptide binding prediction can be divided into three categories: (1) position-specific scoring matrix (PSSM) based, (2) machine learning based, and (3) structure based.[26] The PSSM-based methods generate a matrix for each residue position inside a peptide given a specific HLA. When predicting the binding affinity for a new peptide, values for each residue at each position are attained and summarized for a score by a given formula. The PSSM methods were introduced when the available data were limited. They were gradually replaced by machine learning methods that showed larger data capability, fast prediction speed, and reliable accuracy.[27,28] Meanwhile, the structure-based methods, such as residue-based statistical energy function,[29] quantitative structure–activity relationship (QSAR) analysis,[30] and quantitative sequence–activity models,[31] are an alternative to the machine learning methods as more HLA–peptide binding structures are becoming available for analysis. The structure-based methods provide a better insight to understand the HLA–peptide binding at the structure level; however, the prediction accuracy, speed, and scope remain a challenge due to the limited number of available crystal structures.[26,32,33] Since the machine learning models are widely developed and used by major institutions, including the largest repository of HLA–peptide binding data, IEDB,[34,35] this review focused on the machine learning methods used for predicting HLA–peptide binding.

Current Status

Existing methods. Various machine learning approaches have been used for HLA–peptide binding prediction, including artificial neural network (ANN), decision tree, hidden Markov model (HMM), regression methods, support vector machine (SVM), and consensus methods; the latter combines with several of the former. Table 1 gives an overall summary of these tools including their descriptors, supported HLAs and peptides, and performance.

Artificial neural network. Since its first application to HLA-A*02:01 in 1995,[36] ANNs have been widely used to predict peptide binding for a large number of HLA alleles. To construct an ANN model, the peptide sequences are transformed to numeric descriptors that are then fed to several layers of artificial neurons. The value of each artificial neuron is deducted from the previous layer via mathematical formulae, and a final prediction value is calculated. The parameters within the formulae are determined during the training process by back propagation. Multiple papers and servers have implemented the ANN method, including ANNPred/nHLAPred,[37] IEDB,[34] MULTIPRED,[38,39] NetMHC/NetMHCII,[40,41] and NetMHCpan/NetMHCIIpan.[42,43]

ANNs can be utilized to make both qualitative and quantitative predictions for both classes of HLAs. Reliable performances have been achieved regarding this method. It is still under active development and improvement on quite a few servers including IEDB and NetMHCpan/NetMHCIIpan. However, ANNs require a fixed number of input neurons; therefore, peptides of various lengths need to be proceeded with extra processes to have a fix-length sequence.

Decision tree. The decision trees are a group of splitting tree structure constructed from the training samples.[44] The splitting rules are determined from the training process. When a new sample arrives, it undergoes a flowchart-like structure and finally reaches a classification prediction. It was first introduced to make predictions for HLA-A*02:01 in 1999.[45] Zhu et al implemented a C4.5 decision tree classifier to identify peptide binding for 16 HLA-A alleles.[46]

The decision trees are easier to interpret than ANNs due to their rule-based nature. Though this method is widely used in the machine learning field, it is less implemented in predicting HLA–peptide binding.[18,44]

Hidden Markov model. As a widely used method for pattern recognition, HMMs have been utilized for HLA–peptide binding predictions. In an HMM, a peptide is converted to different states for different positions. At each position or state, the probabilities of amino acids are calculated, and a final value is given by combining all the probabilities. There are different ways to construct HMMs, including fully connected HMMs,[47] HMMs optimized with successive state splitting algorithm,[48] and profile HMMs[49] that can merge overlapping patterns.

HMMs have the advantage of processing peptides with various lengths; however, different HMMs have to be developed separately for binders and nonbinders. Due to the model separation, HMMs have been developed only for a very limited number of HLAs.

Regression. Regression models have been developed for quantitative predictions of HLA–peptide binding. MHCPRED implemented a QSAR regression to predict HLA–peptide binding affinity.[50] In this method, individual amino acid contributions at each position are calculated using partial least squares. SVRMHC utilized support vector regression (SVR), a regression derivative of SVM, for quantitative predictions.[51] Multiple instance learning (MIL) and its regression derivation, multiple instance regression (MIR), were also used to predict HLA–peptide binding in MHCMIR/MHC2MIR.[52,53]

Regression models have the advantage of making quantitative predictions, which not only identify whether a peptide is a binder or nonbinder toward an HLA but also tell how strong the binding is. However, to make an accurate quantitative prediction can be more challenging than to make a qualitative one.

Support vector machine. The SVM creates a hyperplane in the high-dimensional space of training data to classify them into different groups. SVM has been used by a few servers for HLA–peptide binding prediction including MHC2PRED,[54] MULTIPRED,[55] SVMHC,[56,57]

Table 1. An overview of major machine learning tools for predicting HLA–peptide binding sorted by category and method. The tools were divided into two categories, qualitative or quantitative, depending on the outputs. The underlying method, descriptors, performance, and URL were harvested from the original papers. The supported number and class of HLAs and corresponding length of peptides were harvested from either the original papers or their websites. Some tools utilize extra process to deal with peptides with various lengths, which are listed in the "extra process" column.

CATEGORY	NAME	METHOD	DESCRIPTOR	PERFORMANCE	HLA (CLASS)	PEPTIDE LENGTH (HLA CLASS)	EXTRA PROCESS	URL
Qualitative	ANNPred[37]	ANN	Sparse encoding	Accuracy: 87.3%±5.9%	30(I)	9-mers(I)	N/A	http://www.imtech.res.in/raghava/nhlapred/neural.html
	MULTIPRED[38,39]	ANN/HMM/SVM	Sparse encoding	AUC >0.80	23(I), 6(II)	9-mers(I), 9-mer cores(II)	N/A	http://antigen.i2r.a-star.edu.sg/multipred/
	nHLAPred[37]	ANN/PSSM	Sparse encoding	Accuracy: 93.6%±2.92%	30(I)	9-mers(I)	N/A	http://www.imtech.res.in/raghava/nhlapred/comp.html
	Zhu et al.[46]	Decision Tree	N/A	Accuracy: ~0.8	16(I)	9-mers(I)	N/A	N/A
	S-HMM[48]	HMM	N/A	AUC: 0.85~0.89	1(II)	9~25-mers (II)	N/A	N/A
	ocHMM[49]	HMM	Physicochemical property grouping	Accuracy: 0.35~0.99	2(I)	Various(I)	N/A	N/A
	Salomon et al.[61]	Kernel	BLOSUM62[$]	AUC: 0.82~0.96	25(II)	9~33-mers (II)	N/A	N/A
	KISS[59]	SVM	Heckerman et al[81#]	AUC: 0.86~0.90	35(I)	9-mers(I)	N/A	http://cbio.ensmp.fr/kiss/
	MHC2PRED[54]	SVM	Sparse encoding	Accuracy: ~80%	42(II)	9-mers or longer(II)	Matrix optimization techniques (MOTs)	http://www.imtech.res.in/raghava/mhc2pred/
	POPI[58]	SVM	Physicochemical properties	Accuracy: ~60%	23(I), 21(II)	9-mers(I), 9-mer cores(II)	N/A	http://iclab.life.nctu.edu.tw/POPI/
	SVMHC[56,57]	SVM/PSSM	Sparse encoding	MCC: 0.85	32(I), 51(II)	9-mers(I), 9-mer cores(II)	N/A	http://abi.inf.uni-tuebingen.de/Services/SVMHC
Quantitative	NetMHC/ NetMHCII[40,41]	ANN	Sparse encoding/ BLOSUM50	AUC: 0.914(I), 0.787(II)	78(I), 14(II)	8~11-mers(I), various(II)	NN-align	http://www.cbs.dtu.dk/services/
	NetMHCpan/ NetMHCIIpan[42,43]	ANN	Sparse encoding/ BLOSUM50[#]	Pearson: 0.77(I), AUC: 0.847(II)	150(I), 35(II)*	8~14-mers(I), 9~19-mers(II)	Similar to NN-align	http://www.cbs.dtu.dk/services/
	IEDB[35,62,63]	ANN/Consensus	N/A	AUC: 0.96(I), 0.76(II)	50(I), 54(II)	Various(I/II)	N/A	http://tools.immuneepitope.org/main/tcell/
	NetMHCcons[64]	Consensus	N/A	Better than single methods	101(I)*	8~15-mers(I)	N/A	http://www.cbs.dtu.dk/services/NetMHCcons/
	MHCMIR/ MHC2MIR[52,53]	MIL/MIR	BLOSUM62[$]	AUC: 0.73~0.89	26(II)	9~25-mers(II)	N/A	http://datamining-iip.fudan.edu.cn/service/MHC2MIL/index.html
	MHCPRED[50]	QSAR regression	N/A	q^2: 0.3~0.8	11(I), 3(II)	9-mers(I), 9-mer cores(II)	N/A	http://www.ddg-pharmfac.net/mhcpred/MHCPred/
	SVRMHC[51]	SVR	Sparse encoding/ 11 physicochemical properties	q^2: −0.6~0.7	36(I), 6(II)	9-mer cores(I/II)	Iterative self-consistent (ISC)	http://svrmhc.biolead.org

Notes: *NetMHCpan/NetMHCIIpan/NetMHCcons can predict any HLA allele with a known sequence, thus the HLA number is unlimited. [#]The descriptors contain both the peptides and HLAs. [$]The BLUSUM62 matrix was used for distance calculation.
Abbreviation: N/A, not available/applicable.

Prediction Of Peptide Immunogenicity (POPI),[58] and Kernel-based Inter-allele peptide binding prediction SyStem (KISS).[59] The quantitative derivative of SVM was implemented by SVRMHC as we mentioned before.[51]

SVM usually achieves a high accuracy.[60] Similar to ANN, SVM requires a fixed dimension or length of input data, limiting its applicability. However, Salomon and Flower proposed a kernel method derived from SVM that can handle peptides with various lengths using similarity scores.[61]

Consensus method. A consensus method uses a combination of the predictions from its component models. Each of the component models first makes a prediction separately and the final prediction is then made considering all the predictions from the component models. IEDB recommended a consensus approach on its server.[35,62,63] In IEDB, the binding prediction is made from four PSSM-based models for Class I HLAs, while for Class II HLAs, the result comes from nine models including several PSSM-based models and machine learning models. Three machine learning models (QSAR regression-based MHCPRED,[50] SVM-based MHC2PRED[54] and SVR-based SVRMHC[51]) were utilized by IEDB for HLA–peptide binding prediction. When predicting a HLA–peptide binding, not every model is able to return a value. However, if three or more models provide predictions, the three top-performed models are selected and the median value is used as the final consensus score. A similar consensus approach was implemented by NetMHCcons[64] that combines ANN-based NetMHC, NetMHCpan, and PSSM-based PickPocket. Instead of median values, NetMHCcons uses the average log-transformed values as the final prediction scores.

The consensus method generally outperforms single models since it can preferably select the top-performing methods from benchmark tests. However, the applicability of the consensus method is limited by its individual components since it requires outputs simultaneously from those models.

Descriptors. *Overview.* Most machine learning methods take only peptide sequences as input features; therefore, individual models have to be developed for each HLA. One exception is KISS[59] that implements a kernel function of both peptide descriptors and HLA similarities. Another exception is NetMHCpan/NetMHCIIpan[42,43] that uses the sequences of both the peptides and the HLAs to construct a single pan-specific model for an entire class of HLAs. The developers of NetMHCpan/NetMHCIIpan studied different HLA structures and identified a series of residues on HLAs that closely interact with peptides to form a pseudo-sequence. Both the peptide sequences and HLA pseudo-sequences are input into the machine learning model at the same time. Such models make it possible to make predictions for HLA alleles with little or no experimental data as long as their sequences are known.

Some machine learning models such as HMMs,[48] kernel functions,[61] and MIL/MIR models[52,53] naturally process peptides with various lengths. Most models only deal with fixed lengths of peptides; therefore, separate models have to be developed for peptides with different lengths. This is problematic for Class II HLA binders since peptides partially interact with Class II HLAs. In order to solve this problem, extra processes were implemented to identify the interacting region of peptides. SVM-based method MHC2PRED[54] utilized matrix optimization techniques to identify the 9-mer binding cores from Class II HLA binders. SVR-based SVRMHC[51] used iterative self-consistent[65] to find the 9-mer cores with the information of HLA anchor positions. ANN-based NetMHC/NetMHCII[40,41] and NetMHCpan/NetMHCIIpan[42,43] took advantage of alignment-based NN-align or similar processes[41,66] to get both the 9-mer cores and the peptide flanking residues for Class II HLA binders. Both the 9-mer cores and the flanking residues were used as features to develop the models.

The machine learning models did not directly make use of sequences as features except the HMMs. Usually machine learning models accept either binary categories (such as 0 and 1) or continuous numeric inputs. Therefore, input sequences need to be transformed to descriptors. The commonly used descriptors are sparse encoding, blocks substitution matrix (BLOSUM), and physicochemical properties.

Sparse encoding. Sparse encoding is simple but widely used by servers such as ANNPred/nHLAPred,[37] MHC2PRED,[54] NetMHC/NetMHCII,[40,41] NetMHCpan/NetMHCIIpan,[42,43] SVMHC,[56,57] and SVRMHC.[51] The concept of sparse encoding is similar to dummy variables in the machine learning field. Since each position within a sequence can be any of 20 different amino acids, a position is presented by a 20-number vector of 19 zeros and a single one such as 10000 …, 01000 …, and 00100 … depending on what is the actual amino acid. Each type of amino acids at a specific position is represented by a single and unique variable. Thus, a 9-mer peptide is converted to $20 \times 9 = 180$ binary variables.

Blocks substitution matrix. The BLOSUMs are widely used in protein sequence alignment. The scores within BLOSUMs represent pairwise evolutionary distances between amino acids. Some machine learning methods utilize BLOSUMs to calculate sequence distances. For example, Salomon and Flower used the BLOSUM62 matrix to calculate peptide similarity scores in their kernel method.[61] They also compared 83 different matrices including physicochemical and structural distance matrices and found BLOSUM62 matrix was among the top three. MHCMIR/MHC2MIR[52,53] implemented the BLOSUM62 matrix to compute subsequence similarities as well. For some other machine learning models, the BLOSUMs are just used as descriptors. In addition to sparse encoding, NetMHC/NetMHCII[40,41] and NetMHCpan/NetMHCIIpan[42,43] implemented the BLOSUM50 matrix as descriptors. However, we did not find any benchmark data that compare sparse encoding versus BLOSUM.

Physicochemical properties. Besides the substitution matrices, the physicochemical properties of amino acids, such as hydrogen bond number, polarity, and hydrophobicity were used as descriptors. Zhang et al used physicochemical properties to group amino acids in their HMM.[49] POPI used 20 physicochemical properties to represent each amino acid according to AAindex database 9.0.[58,67] SVRMHC utilized both sparse encoding and their 11-factor physicochemical property descriptors in their method.[51,68] They compared both types of descriptors and found that the two types of descriptors showed different performance on different HLA alleles. No conclusion was drawn to indicate which one is absolutely better than the other.

Performance. Recently published machine learning models achieved an area under receiver operating characteristic curve (AUC) around 0.85–0.95 for Class I HLAs and 0.75–0.85 for Class II HLAs.[69] For some HLA alleles such as HLA-A*02:04, the AUC reached 0.98.[28] The existing machine learning models were developed using the data sets harvested at different times. Some HLA–peptide binding data are qualitative and some are quantitative, and the data set sizes are different, making direct performance comparison between the models difficult. Therefore, some benchmark tests were conducted using unified data sets.[35,70–73]

For prediction of peptides binding Class I HLAs, Peters et al tested three methods using a data set of 48 Class I MHCs.[70] From the prospective of AUC, their ANN model (AUC = 0.957) outperformed their PSSM-based models (AUC = 0.934–0.952). They also benchmarked 16 publicly available tools such as MHCPRED,[50] MULTIPRED,[38,39] NetMHC,[40] and SVMHC.[56,57] Five Class I HLAs were evaluated but only two of them had been predicted by all these methods. For HLA-A*02:01, the performance was NetMHC (ANN) > MULTIPRED (ANN) > MHCPRED (QSAR regression) = SVMHC (SVM) > MULTIPRED (HMM), while for HLA-A*24:02, the result was NetMHC (ANN) > MULTIPRED (ANN) > MULTIPRED (HMM) > MHCPRED (QSAR regression) > SVMHC (SVM). Likewise, Trost et al tested 16 tools using Peters data[70] and other literature data and found similar results.[71] In 2008, Lin et al evaluated 30 servers on the binding data of tumor antigens toward seven Class I HLAs.[72] The classification result indicated a rank of NetMHC (ANN) > IEDB (ANN) > MHCPRED (SVM), while NetMHC (ANN) and IEDB (ANN) showed the best performance in quantitative predictions.

For prediction of peptides binding Class II HLAs, Wang et al evaluated several methods using their experimental data set that contains 16 Class II MHCs.[35] Their result indicated a performance order of Consensus method > SVRMHC (SVR) > MHC2PRED (SVM) > MHCPRED (QSAR regression). Lin et al evaluated 21 methods using 103 test peptides from four protein antigens and seven Class II HLAs.[73] Their classification result indicated that NetMHCIIpan (ANN) was the best followed by two PSSM-based methods and MULTIPRED (SVM). They also showed MHCPRED (QSAR regression) and MULTIPRED (HMM) had an AUC >0.775 when predicting the binding of promiscuous peptides.

Some methods were developed several years ago and rarely updated. Due to the limited availability of different models on the supported HLAs, it is hard to benchmark the existing models on a large number of HLAs.[70] However, the current benchmarks indicated that the ANN-based and consensus models such as IEDB and NetMHC have a good overall performance.

Challenge

Though various methods and tools have been developed to predict HLA–peptide binding, challenges in this field remain for researchers to address.

Limited support for HLAs. Most machine learning tools develop models for individual HLAs separately. Therefore, the peptides binding models have been developed for a very limited number of HLAs. In order to train a reliable model, 15–50 minimal binding peptides for a specific HLA[37,51,57] are a precondition. Though more binding data are becoming available, the experimental data for different HLAs are still disparately distributed. Incorporating HLA sequences as input features to machine learning models is one of the solutions. For example, NetMHCpan/NetMHCIIpan[42,43] utilized both the pseudo-sequence from the HLAs and the peptide sequence as features. Such a method can theoretically predict binding between any HLAs and peptides given their sequences; however, the prediction accuracy for HLAs with little or no experimental data is lower.

Peptide length. While some methods such as HMMs can naturally accept peptides with different lengths, others can only accept fixed lengths of input peptide sequences. Therefore, separate models have to be developed to fit different lengths of peptides. This practice has challenges when it is applied to peptides that bind to Class II HLAs since the peptides are very diverse in length and only partially interact with Class II HLAs. Extra processes such as NN-align[41,66] were implemented to identify a 9-mer core-binding region of these peptides so that they can be proceeded by the general machine learning models. However, Nielsen et al pointed out the prediction models for Class II HLAs generally have an AUC 0.10 less than those for Class I HLAs.[69] The problem of peptide length variety may be one of the causes.

Performance. Though the models can get a good performance during cross-validation, they may have problems when used on new data sets or applications. In a benchmark test of four real protein antigens by Lin et al.[73], no predictor showed good performance in predicting promiscuous peptides. The researchers mentioned that future improvement of the HLA–peptide binding prediction includes minimizing false positives.

Prospective

Network approach. With more data becoming available, it is possible to analyze and predict the HLA–peptide binding from a network viewpoint. HLA–peptide binding data can be transformed into a network where the HLAs and peptides are presented as the nodes and the binding data (categories or affinities) as the edges. Network-based prediction algorithms, such as collaborative filtering algorithm,[74] network-based inference,[75,76] and neighbor-edges based and unbiased leverage algorithm (Nebula), the latter developed in our laboratory,[77] have been proposed for predictions of HLA–peptide binding.

The network approaches are able to process different HLA alleles (or even different HLA classes) and peptides of various lengths without extra processes. Our submitted manuscript showed that Nebula outperformed existing methods. Since Nebula is not necessarily a machine learning method and does not require a training process, predictions even on a large network of 120,000 HLA–peptide binding pairs can be made within a second. In addition, the HLA–peptide binding network can be analyzed and clustered into modules that may reveal specific binding properties and patterns to better understand HLA–peptide interactions.

Combining multiple approaches. The existing consensus methods combining several PSSM-based and machine learning-based methods showed generally improved performance than a single method.[35,62–64] Such methods take advantage of the best performing models to make better predictions with reduced outliers; however, the consensus methods are restrained by the prediction abilities of their component methods such as a limited number of supported HLAs.

As more crystal structures of HLA–peptide complexes become available, studying such structures can aid the interpretation and prediction of HLA–peptide binding. Based on the existing crystal structures, the structures of HLAs with known sequences can be modeled via homology modeling with high identities.[13] Various structure-based methods, including molecular docking and dynamics, can be utilized for HLA–peptide binding predictions.[78–80] Some modeling process, such as molecular dynamics, may take a large amount of calculation time. However, with the development of cloud computing technologies, parallel computing can accelerate this process hundredfold or more. These structure-based approaches can be used for HLA alleles or peptides with little or no experimental data, which may complement the data-dependent machine learning methods. Combining such methods can be promising for the future development in this field.

Conclusion

Understanding and predicting HLA–peptide binding is an essential step for studies of the immune system, T-cell epitopes, and adverse drug reactions. The methods that are popular in the field of computer sciences are widely used for HLA–peptide binding predictions. Different types of descriptors were utilized to transfer the peptide and HLA sequences into model-acceptable numbers. Some extra processes were implemented to deal with peptides with various lengths. Among the machine learning-based methods, the ANNs and consensus methods were found among the top-performed methods.

However, the existing methods have different kinds of limitations such as lack of supported HLAs, a problem dealing with peptides of various lengths and high false positives in experimental validations. We have proposed the network method and a combination approach that uses multiple types of methods to address some of these challenges.

The findings and conclusions in this article have not been formally disseminated by the US Food and Drug Administration (FDA) and should not be construed to represent the FDA determination or policy.

Author Contributions

Wrote the first draft of the manuscript: HL, DLM, HH. Contributed to the writing of the manuscript: HY, HWN, LS, WT. All authors reviewed and approved the final manuscript.

REFERENCES

1. Jin P, Wang E. Polymorphism in clinical immunology – from HLA typing to immunogenetic profiling. *J Transl Med*. 2003;1(1):8.
2. Trowsdale J. The MHC, disease and selection. *Immunol Lett*. 2011;137(1–2):1–8.
3. Illing PT, Vivian JP, Purcell AW, Rossjohn J, McCluskey J. Human leukocyte antigen-associated drug hypersensitivity. *Curr Opin Immunol*. 2013;25(1):81–9.
4. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG. The IMGT/HLA database. *Nucleic Acids Res*. 2013;41(Database issue):D1222–7.
5. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature*. 1999;401(6756):921–3.
6. Bushkin Y, Demaria S, Le JM, Schwab R. Physical association between the CD8 and HLA class I molecules on the surface of activated human T lymphocytes. *Proc Natl Acad Sci U S A*. 1988;85(11):3985–9.
7. Spaggiari GM, Contini P, Carosio R, et al. Soluble HLA class I molecules induce natural killer cell apoptosis through the engagement of CD8: evidence for a negative regulation exerted by members of the inhibitory receptor superfamily. *Blood*. 2002;99(5):1706–14.
8. Mangalam A, Rodriguez M, David C. Role of MHC class II expressing CD4+ T cells in proteolipid protein(91–110)-induced EAE in HLA-DR3 transgenic mice. *Eur J Immunol*. 2006;36(12):3356–70.
9. Poncet P, Arock M, David B. MHC class II-dependent activation of CD4+ T cell hybridomas by human mast cells through superantigen presentation. *J Leukoc Biol*. 1999;66(1):105–12.
10. Lang P, Stolpa JC, Freiberg BA, et al. TCR-induced transmembrane signaling by peptide/MHC class II via associated Ig-alpha/beta dimers. *Science*. 2001;291(5508):1537–40.
11. Sim E, Cross SJ. Phenotyping of human complement component C4, a class-III HLA antigen. *Biochem J*. 1986;239(3):763–7.
12. Holland MC, Lambris JD. The complement system in teleosts. *Fish Shellfish Immunol*. 2002;12(5):399–420.
13. Luo H, Du T, Zhou P, et al. Molecular docking to identify associations between drugs and class I human leukocyte antigens for predicting idiosyncratic drug reactions. *Comb Chem High Throughput Screen*. 2015;18(3):296–304.
14. Saper MA, Bjorkman PJ, Wiley DC. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution. *J Mol Biol*. 1991;219(2):277–319.
15. Stern LJ, Brown JH, Jardetzky TS, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*. 1994;368(6468):215–21.
16. Rudensky A, Janeway CA Jr. Studies on naturally processed peptides associated with MHC class II molecules. *Chem Immunol*. 1993;57:134–51.
17. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol*. 1999;17:51–88.

18. Lafuente EM, Reche PA. Prediction of MHC-peptide binding: a systematic and comprehensive overview. *Curr Pharm Des*. 2009;15(28):3209–20.
19. Villadangos JA, Schnorrer P. Intrinsic and cooperative antigen-presenting functions of dendritic-cell subsets in vivo. *Nat Rev Immunol*. 2007;7(7):543–55.
20. Felix NJ, Allen PM. Specificity of T-cell alloreactivity. *Nat Rev Immunol*. 2007;7(12):942–53.
21. Platzer B, Stout M, Fiebiger E. Antigen cross-presentation of immune complexes. *Front Immunol*. 2014;5:140.
22. Wucherpfennig KW, Strominger JL. Molecular mimicry in T cell-mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. *Cell*. 1995;80(5):695–705.
23. Illing PT, Vivian JP, Dudek NL, et al. Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature*. 2012;486(7404):554–8.
24. Bharadwaj M, Illing P, Theodossis A, Purcell AW, Rossjohn J, McCluskey J. Drug hypersensitivity and human leukocyte antigens of the major histocompatibility complex. *Annu Rev Pharmacol Toxicol*. 2012;52:401–31.
25. Wei CY, Chung WH, Huang HW, Chen YT, Hung SI. Direct interaction between HLA-B and carbamazepine activates T cells in patients with Stevens-Johnson syndrome. *J Allergy Clin Immunol*. 2012;129(6):1562.e–69.
26. Liao WW, Arthur JW. Predicting peptide binding to major histocompatibility complex molecules. *Autoimmun Rev*. 2011;10(8):469–73.
27. Paul S, Kolla RV, Sidney J, et al. Evaluating the immunogenicity of protein drugs by applying in vitro MHC binding data and the immune epitope database and analysis resource. *Clin Dev Immunol*. 2013;2013:467852.
28. 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(5):1007–17.
29. Kumar N, Mohanty D. MODPROPEP: a program for knowledge-based modeling of protein-peptide complexes. *Nucleic Acids Res*. 2007;35(Web Server issue):W549–55.
30. Hattotuwagama CK, Doytchinova IA, Flower DR. Toward the prediction of class I and II mouse major histocompatibility complex-peptide-binding affinity: in silico bioinformatic step-by-step guide using quantitative structure–activity relationships. *Methods Mol Biol*. 2007;409:227–45.
31. Li Z, Wu S, Chen Z, et al. Structural parameterization and functional prediction of antigenic polypeptome sequences with biological activity through quantitative sequence-activity models (QSAM) by molecular electronegativity edge-distance vector (VMED). *Sci China C Life Sci*. 2007;50(5):706–16.
32. Jojic N, Reyes-Gomez M, Heckerman D, Kadie C, Schueler-Furman O. Learning MHC I–peptide binding. *Bioinformatics*. 2006;22(14):e227–35.
33. Zhang H, Wang P, Papangelopoulos N, et al. Limitations of Ab initio predictions of peptide binding to MHC class II molecules. *PLoS One*. 2010;5(2):e9272.
34. Vita R, Zarebski L, Greenbaum JA, et al. The immune epitope database 2.0. *Nucleic Acids Res*. 2010;38(Database issue):D854–62.
35. Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol*. 2008;4(4):e1000048.
36. Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. *J Immunol Methods*. 1995;185(2):181–90.
37. Bhasin M, Raghava GP. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci*. 2007;32(1):31–42.
38. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res*. 2005;33(Web Server issue):W172–9.
39. Zhang GL, DeLuca DS, Keskin DB, et al. MULTIPRED2: a computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles. *J Immunol Methods*. 2011;374(1–2):53–61.
40. Lundegaard C, Lund O, Nielsen M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9 mers. *Bioinformatics*. 2008;24(11):1397–8.
41. Andreatta M, Schafer-Nielsen C, Lund O, Buus S, Nielsen M. NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PLoS One*. 2011;6(11):e26781.
42. Hoof I, Peters B, Sidney J, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*. 2009;61(1):1–13.
43. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*. 2013;65(10):711–24.
44. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. 2008;26(9):1011–3.
45. 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;4:182–9.
46. Zhu S, Udaka K, Sidney J, Sette A, Aoki-Kinoshita KF, Mamitsuka H. Improving MHC binding peptide prediction by incorporating binding data of auxiliary MHC molecules. *Bioinformatics*. 2006;22(13):1648–55.
47. Mamitsuka H. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins*. 1998;33(4):460–74.
48. Noguchi H, Kato R, Hanai T, et al. Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J Biosci Bioeng*. 2002;94(3):264–70.
49. Zhang C, Bickis MG, Wu FX, Kusalik AJ. Optimally-connected hidden Markov models for predicting MHC-binding peptides. *J Bioinform Comput Biol*. 2006;4(5):959–80.
50. Guan P, Doytchinova IA, Zygouri C, Flower DR. MHCPred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res*. 2003;31(13):3621–4.
51. Wan J, Liu W, Xu Q, Ren Y, Flower DR, Li T. SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics*. 2006;7:463.
52. EL-Manzalawy Y, Dobbs D, Honavar V. Predicting MHC-II binding affinity using multiple instance regression. *IEEE/ACM Trans Comput Biol Bioinform*. 2011;8(4):1067–79.
53. Xu Y, Luo C, Qian M, Huang X, Zhu S. MHC2MIL: a novel multiple instance learning based method for MHC-II peptide binding prediction by considering peptide flanking region and residue positions. *BMC Genomics*. 2014;15(suppl 9):S9.
54. Bhasin M, Raghava GP. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*. 2004;20(3):421–3.
55. Zhang GL, Bozic I, Kwoh CK, August JT, Brusic V. Prediction of supertype-specific HLA class I binding peptides using support vector machines. *J Immunol Methods*. 2007;320(1–2):143–54.
56. Donnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*. 2002;3:25.
57. Donnes P, Kohlbacher O. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res*. 2006;34(Web Server issue):W194–7.
58. Tung CW, Ho SY. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*. 2007;23(8):942–9.
59. Jacob L, Vert JP. Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics*. 2008;24(3):358–66.
60. Han J, Kamber M. *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, San Francisco, CA; 2006.
61. Salomon J, Flower DR. Predicting class II MHC-Peptide binding: a kernel based approach using similarity scores. *BMC Bioinformatics*. 2006;7:501.
62. Kim Y, Ponomarenko J, Zhu Z, et al. Immune epitope database analysis resource. *Nucleic Acids Res*. 2012;40(Web Server issue):W525–30.
63. Moutaftsi M, Peters B, Pasquetto V, et al. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat Biotechnol*. 2006;24(7):817–9.
64. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*. 2012;64(3):177–86.
65. Doytchinova IA, Flower DR. Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics*. 2003;19(17):2263–70.
66. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*. 2009;10:296.
67. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res*. 2000;28(1):374.
68. Liu W, Meng X, Xu Q, Flower DR, Li T. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*. 2006;7:182.
69. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. *Immunology*. 2010;130(3):319–28.
70. Peters B, Bui HH, Frankild S, et al. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol*. 2006;2(6):e65.
71. Trost B, Bickis M, Kusalik A. Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. *Immunome Res*. 2007;3:5.
72. Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol*. 2008;9:8.
73. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics*. 2008;9(suppl 12):S22.
74. Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web. ACM, New York, NY; 2001:285–95.
75. Cheng F, Zhou Y, Li W, Liu G, Tang Y. Prediction of chemical-protein interactions network with weighted network-based inference method. *PLoS One*. 2012;7(7):e41064.

76. Cheng F, Liu C, Jiang J, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*. 2012; 8(5):e1002503.
77. Luo H, Ye H, Ng HW, et al. Understanding and predicting binding between human leukocyte antigens and peptides by network analysis. *BMC Bioinformatics*. 2015;16(suppl 14):S9.
78. Rognan D, Scapozza L, Folkers G, Daser A. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry*. 1994;33(38):11476–85.
79. 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(5):675–9.
80. Bui HH, Schiewe AJ, von Grafenstein H, Haworth IS. Structural prediction of peptides binding to MHC class I molecules. *Proteins*. 2006;63(1):43–52.
81. Heckerman D, Kadie C, Listgarten J. Leveraging information across HLA alleles/supertypes improves epitope prediction. *J Comput Biol*. 2007;14(6):736–46.