The Utility and Limitations of Current Web-Available Algorithms To Predict Peptides Recognized by CD4 T Cells in Response to Pathogen Infection

Francisco A. Chaves, Alvin H. Lee, Jennifer L. Nayak, Katherine A. Richards and Andrea J. Sant

*J Immunol* 2012; 188:4235-4248; Prepublished online 30 March 2012;
doi: 10.4049/jimmunol.1103640
http://www.jimmunol.org/content/188/9/4235

Supplementary Material  http://www.jimmunol.org/content/suppl/2012/03/30/jimmunol.1103640.DC1

References  This article cites 109 articles, 31 of which you can access for free at:
http://www.jimmunol.org/content/188/9/4235.full#ref-list-1

Why *The JI*? Submit online.

- **Rapid Reviews! 30 days** from submission to initial decision
- **No Triage!** Every submission reviewed by practicing scientists
- **Fast Publication!** 4 weeks from acceptance to publication

*average

Subscription  Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscription

Permissions  Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts  Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts

*The Journal of Immunology* is published twice each month by
The American Association of Immunologists, Inc.,
1451 Rockville Pike, Suite 650, Rockville, MD 20852
Copyright © 2012 by The American Association of Immunologists, Inc. All rights reserved.
Print ISSN: 0022-1767 Online ISSN: 1550-6606.
The Utility and Limitations of Current Web-Available Algorithms To Predict Peptides Recognized by CD4 T Cells in Response to Pathogen Infection

Francisco A. Chaves,* Alvin H. Lee,* Jennifer L. Nayak,† Katherine A. Richards,* and Andrea J. Sant*

The ability to track CD4 T cells elicited in response to pathogen infection or vaccination is critical because of the role these cells play in protective immunity. Coupled with advances in genome sequencing of pathogenic organisms, there is considerable appeal for implementation of computer-based algorithms to predict peptides that bind to the class II molecules, forming the complex recognized by CD4 T cells. Despite recent progress in this area, there is a paucity of data regarding the success of these algorithms in identifying actual pathogen-derived epitopes. In this study, we sought to rigorously evaluate the performance of multiple Web-available algorithms by comparing their predictions with our results—obtained by purely empirical methods for epitope discovery in influenza that used overlapping peptides and cytokine ELISPOTS—for three independent class II molecules. We analyzed the data in different ways, trying to anticipate how an investigator might use these computational tools for epitope discovery. We come to the conclusion that currently available algorithms can indeed facilitate epitope discovery, but all shared a high degree of false-positive and false-negative predictions. Therefore, efficiencies were low. We also found dramatic disparities among algorithms and between predicted IC₅₀ values and true dissociation rates of peptide–MHC class II complexes. We suggest that improved success of predictive algorithms will depend less on changes in computational methods or increased data sets and more on changes in parameters used to “train” the algorithms that factor in elements of T cell repertoire and peptide acquisition by class II molecules. The Journal of Immunology, 2012, 188: 4235–4248.

CD4 T cells are known to play a key role in protective immunity to infectious organisms, and much current research uses epitope-specific probes to study the role that CD4 T cells play in immunity to complex pathogens. Further success in identification of the peptides that are the focus of an adaptive CD4 T cell response is essential for understanding the mechanisms of protective immunity and the factors that influence the dynamics and specificity of host–pathogen interactions. CD4 T cell epitope identification is also needed for vaccine evaluation, tetramer-based studies of T cell phenotype, and for development of peptide-based vaccines. With increasing success in genome sequencing of complex bacterial and viral pathogens (reviewed in Refs. 1–5), candidate proteins for vaccines are increasing, but identification of epitopes that are the focus of immune responses remains a bottleneck in this research.

A number of empirical approaches have historically been used for epitope discovery—including biochemical isolation and proteolytic fragmentation of antigenic proteins (6, 7), derivation of genetic constructs that encode all or selected segments of candidate pathogen-derived proteins (8–11), elution and sequencing of peptides from pathogen-infected cells or tumor cells (12–16), and individual epitope mapping—using arrays of synthetic peptides (17–22). These approaches, typically coupled with T cell assays to identify the immunologically active peptide within the candidate Ag, are time consuming and involve significant expenditure of effort and resources to be successful. The labor-intensive nature of these approaches is a particularly large obstacle for complex pathogens that express hundreds of proteins, of which only a small fraction may be the target of T cells or B cells or which may serve a protective role as vaccine candidates.

The considerations of time and expense required for empirical approaches have led to the development and refinement of algorithms that use different logic bases and sources of data to predict epitopes that will be presented by particular MHC molecules (reviewed in Refs. 23–28). Because the major selective force in peptide binding to MHC involves side chains of amino acids (“anchors”) in the peptide with depressions (“pockets”) in the MHC molecule, the algorithms focus on scoring these interactions as a means to predict CD4 epitopes. Some methods such as matrix-based algorithms operate with the general model that each amino acid adds or detracts from the binding of the peptide to the MHC protein in a largely predictable, independent, and quantifiable manner (29, 30). Large data sets, or “training data,” are used to construct and refine the algorithms that ultimately search for the highest 9-mer core in a peptide and output the predicted binding affinity of every candidate peptide. Other less rigid algorithms that operate using such methods as neural networks (31, 32) and particle swarm optimization (33) have also been developed and used.
Finally, Sette and coworkers (34) describe a “Consensus” approach that essentially averages the predicted ranking hierarchy of a given set of peptides scored by what their studies suggest to be the best-performing three to four Web-available algorithms.

In general, the predictive algorithms developed for MHC class I peptides that activate CD8 T cells significantly outperform those that predict MHC class II-presented peptides that elicit CD4 T cells, in large part because of the nature of their respective peptide binding pockets. MHC class I molecules are closed at their periphery, thus limiting the size of the peptide that binds to 8–10 aa (35–37). Therefore, the amino acids that contribute the key anchor positions for pocket interactions are easily identified. In contrast, the binding pocket of MHC class II is open at its periphery. Elution and sequencing studies indicate that peptides bound by class II molecules typically range from 9 to 25 aa in length (38, 39), and long peptides are well presented to CD4 T cells (40, 41). Often, the “register” of these peptides, the amino acids that comprise the 9-aa core within the MHC-binding groove and that dictate MHC binding and T cell recognition, is not known. For several peptides analyzed in detail, multiple registers are presented by class II (42–45). Therefore, simple knowledge that a peptide is presented by a class II molecule does not provide insight into what amino acids control its presentation. Because much of the data that train the algorithms are long peptides, it has been challenging to know how identified antigenic peptides bind to the class II molecule and thus to predict what peptides within an uncharacterized set will be presented by the same or related class II molecules. Databases containing epitopes presented by class II molecules have steadily accumulated (46–48) and been used to refine existing algorithms, but the issue of uncertain binding cores persists.

Because CD4 T cells are known to be critical regulators of both B cell and CD8 T cell responses to pathogens, to provide direct function for responses against intracellular pathogens, and to be critical for vaccine success, it is clear that despite the theoretical challenges, there is considerable appeal in using computation-based algorithms to identify candidate CD4 T cell epitopes, particularly for complex pathogenic organisms that express hundreds of proteins. “Benchmark” studies that assessed the accuracy of the predictions typically measured the ability of the algorithms to predict binding to MHC class II molecules using existing or new data sets (25, 29, 32–34, 49–51). Evaluation of peptide performance using functional tests of the actual immunogenicity of the predicted peptides have been much more limited. The studies performed are typically restricted to a handful of peptides, are sometimes tested by immunization with free peptides rather than a complex Ag, and, often, the presenting class II molecule for CD4 T cell recognition is not unequivocally identified (20, 34, 52–55). In recent years, our laboratory has empirically and comprehensively investigated the peptide specificity of CD4 T cells elicited in response to primary influenza infection using a completely unbiased approach involving overlapping peptide libraries and cytokine ELISPOT assays. Multiple strains of mice have been studied, including HLA-DR transgenic mice and common inbred strains of mice (21, 22, 56). We have identified and quantified the predictors of the role that CD4 T cells play in protective immunity to pathogenic organisms. The third goal was to gain the insight needed to improve performance of the algorithms for future efforts to facilitate epitope discovery. To our knowledge, the analyses presented in this study are the first to comprehensively evaluate the performance of algorithms in comparison with the results of empirical and unbiased epitope discovery of multiple pathogen-encoded Ags and unrelated class II molecules.

Materials and Methods

Mice

The HLA-DR1 and the HLA-DR4 transgenic mice were obtained from D. Zaller (Merck) through Taconic Laboratories and were maintained in the specific pathogen-free facility at the University of Rochester according to institutional guidelines. C57BL/10 mice were purchased from The Jackson Laboratory (Bar Harbor, ME). Mice were used at 2–5 mo of age.

Ethics statement

All animal protocols used in this study adhere to the Association for Assessment and Accreditation of Laboratory Animal Care International, the Animal Welfare Act, and the Policy on Humane Care and Use of Laboratory Animals Guide. All protocols have been approved by the University of Rochester Committee on Animal Resources (animal welfare assurance no. A329-01). The protocols under which this study was conducted were approved on March 4, 2006 (protocol no. 2006-030), and April 10, 2008 (protocol no. 2008-023).

Influenza infection of mice

A/New Caledonia/2099 was produced in allantoic fluid of embryonated eggs. Briefly, eggs were purchased from SPAFAS (North Franklin, CT) and incubated at 70˚F and 100% humidity for 9 d, followed by the infection of the allantoic cavity with 100 µl human (H1N1) influenza virion A/N Caledonia/2099 (kindly provided by Dr. John Tureen, University of Rochester). Virus was collected from the allantoic fluid, and the EID<sub>50</sub> for the produced virus was determined. HLA-DR1, HLA-DR4, transgenic, and C57BL/10 (I-A<sup>b</sup>) mice were infected intranasally with A/New Caledonia/2099 at 50,000 EID<sub>50</sub> in 30 µl PBS. Groups of mice were 2–5 mo old at the time of infection and were anesthetized by intraperitoneal injection with tribromoethanol prior to infection. Ten to fourteen days postinfection, the mice were euthanized, and spleens and mediastinal lymph nodes were isolated and used as sources of CD4 T cells for ELISPOT analyses. Lymphocytes were pooled from four to eight mice and depleted of B cells, CD8 cells, and macrophages by negative selection using MACS depletion (Miltenyi Biotech, Bergisch Gladbach, Germany) according to the manufacturer’s instructions.

Dissociation kinetics

The dissociation experiments were performed using the procedure as previously described (57–60). Briefly, 50 µl 100 nM purified soluble DR1 molecules expressed in S2 cells were incubated with 1 µl 250 µM fluorescein-labeled binding peptides at a concentration of 100 nM and pH 5.3 in McIlvaine’s buffer containing 0.025% NaN<sub>3</sub> and 200/99 at 50,000 EID<sub>50</sub> in 30 µl PBS. Groups of mice were 2–5 mo old at the time of infection and were anesthetized by intraperitoneal injection with tribromoethanol prior to infection. Ten to fourteen days postinfection, the mice were euthanized, and spleens and mediastinal lymph nodes were isolated and used as sources of CD4 T cells for ELISPOT analyses. Lymphocytes were pooled from four to eight mice and depleted of B cells, CD8 cells, and macrophages by negative selection using MACS depletion (Miltenyi Biotech, Bergisch Gladbach, Germany) according to the manufacturer’s instructions.
tion assays were similarly performed by incubating the purified MHC class II–peptide complexes at 37 °C in McElvaine’s buffer plus 0.025% NaN3 at a final pH of 5.3 in the presence of either unlabeled influenza HA [306–318] (PKYYVKNTKLAT) for DR1 or Eo [52–68] (ASFEAQGALANLV-DKK) for I-Ad to prevent possible rebinding of the dissociating peptide. The dissociating peptide was separated from the remaining complex with a BioSep SEC-S 3000 column (Phenomenex, Torrance, CA) using a Shimadzu chromatograph (Shimadzu Corporation, Columbia, MD). The intensity of the remaining class II–peptide complex was determined using a fluorescence detector (at wavelengths of 495-nm excitation and 525-nm emission; Shimadzu). The intensity of the peak belonging to the complex decreases over time, allowing quantification of the dissociation $t_{1/2}$. The graph of the intensity of the MHC class II–peptide complex versus the dissociation time is used to generate the dissociation curve. Baselines values were measured before each dissociation experiment to ascertain that there were no traces of previous samples, and the program for elution of the complexes by HPLC–SEC includes a wash time that ensures complete removal of all previous materials. Dissociation curves fit to single exponential decay curves with $r^2 > 0.9$. The half-life is the time at which the initial fluorescence intensity of the complexes decays to half, which is stated as the $t_{1/2}$ dissociation, proportional to the stability of that complex.

**Peptides**

17-mer peptides overlapping by 11 aa to cover the entire sequences of the HA and neuraminidase (NA) proteins from the A/New Caledonia/2009 influenza virus (H1N1), the nonstructural protein 1 (NS1) sequence from the A/New York/444/2001 influenza virus (H1N1), and the nucleocapsid protein (NP) and matrix protein 1 (M1) sequences from A/New York/348/2001 influenza virus (H1N1) were used. Peptide arrays were obtained from the National Institutes of Health Biodefense and Emerging Infections Research Repository, National Institutes of Allergy and Infectious Diseases. The A/New York/348/2003 aa sequences for NP and M1 are 100% conserved, and the A/New York/444/2001 NS1 aa sequence is >99% conserved compared with A/New Caledonia/2009 (H1N1) influenza virus. The peptides were reconstituted to 10 mM in PBS, with or without added DMSO for hydrophobic peptides, and 1 mM DTT for cysteine-containing peptides. Working solutions were prepared in DMEM (Invitrogen, Carlsbad, CA), filter sterilized, and stored at −20 °C. The final concentration of the individual peptides used in the ELISPOT assays was 10 μM.

**Use of Web-based MHC class II–peptide binding prediction servers**

The primary sequences for HA, NA, NP, NS1, and M1 from A/New Caledonia/2009 (H1N1) were obtained from the Protein Data Bank. The sequences were run using the Web-available predictors NetMHCII 2.2, NetMHCIIpan-2.1, and SMM-align at http://www.cbs.dtu.dk/services/ and Consensus, ARB, and Tepitope at http://tools.immuneepitope.org/main/. If only 15-mer peptides were scored by algorithms (Consensus, ARB, and Tepitope), the empirically tested peptides were matched to the functionally assayed 17-mer sequences by alignment with their N termini. The raw scores, the $IC_{50}$, or the percentile rank values were used to determine the peptides that were selected by the algorithms. For methods that report $IC_{50}$ values such as NetMHCII 2.2 (referred to as NetMHCII hereafter), NetMHCIIpan-2.1 (referred to as NetMHCIIpan hereafter), SMM-align (referred to as SMM hereafter), and ARB, the predictors qualify a peptide as follows: 1) nonbinders for peptides with $IC_{50} > 500$ nM; 2) weak binders for peptides with $IC_{50}$ between 50 and 500 nM; and 3) strong binders for peptides with $IC_{50} < 50$ nM. These specific cutoffs were used to analyze predictive performance in some experiments based on their use by the developers or users of predictive algorithms (49, 61–64). For Consensus, the percentile rank was used to rank the peptides, whereas for Tepitope, the raw score was used for ranking. When a given functionally validated peptide was penalized because its predicted binding core was at either the N or C terminus of the scored peptide epitope, we evaluated each 9-aa core within the adjacent scored peptides having the same core as the epitope and assigned the score that had the best binding value. When scoring the selected viral “proteome,” the input sequences were the entire amino acid sequences of the five proteins joined together, in the order HA, NA, NP, NS1, and M1, with nonnative peptides at the junctions of the proteins eliminated from consideration. Only the experimentally tested peptides were scored for analysis.

**Derivation of receiver operating characteristic curves and area under the curve scores**

Algorithm prediction accuracy was measured via receiver operating characteristic (ROC) curve, and the area under the curve (AUC) values were calculated as described by others (23–25, 32, 34). Each algorithm’s binding predictions were compared with the empirically defined set of “epitopes” and categorized under a binary classification system. At each discrimination threshold, peptides were grouped into one of four categories: true positive (TP), predicted binders as defined by the threshold that elicited response in T cell assays; false positive (FP), predicted binders as defined by current threshold that did not elicit response in T cell assays; true negative (TN), peptides predicted relative to threshold that did not elicit response in T cell assays and false negative (FN), when peptides predicted as nonbinding relative to threshold elicited response in T cell assays. The true-positive rate (TPR) [TPR = TP/(TP + FN)] was plotted against the false-positive rate (FPR) [FPR = FP/(FP + TN)] for each threshold to form the ROC curve. The area under each algorithm’s curve was calculated to give the AUC score.

**CD4 T cell epitope discovery**

Peptide-specific cytokine ELISPOT assays were performed as previously described (65) and as represented schematically in Fig. 1. Briefly, 96-well filter plates (Millipore, Billerica, MA) were coated with 2 μg/ml purified rat anti-mouse IL-2 (clone JES6-1A12; BD Biosciences, San Jose, CA) in PBS, washed, and incubated with media to block nonspecific binding. CD4 T cells isolated and purified from previously infected mice (at 100,000 to 300,000 cells per well) were cocultured with fibroblasts expressing the T2 transfectants and were isolated from infected mice (at 10 μM). After overnight coculture, plates were processed to visualize IL-2–producing cells as described (21, 22, 56), and cytokine ELISPOTs were enumerated with an ImmunoSpot reader series 2A using ImmunoSpot software version 2.

**Results**

Identification of CD4 T cell epitopes through an entirely empirical method

Recently, our laboratory has identified the full repertoire of influenza-specific CD4 epitopes from five major influenza proteins in several inbred strains of mice that express distinct class II molecules [SJL, C57BL/10, HLA-DR1 (“DR1”) transgenic mice] by using a completely empirical approach (21, 22, 65). HLA-DR4 (“DR4”) epitopes were additionally identified for this study. Mice were infected by intranasal inhalation, which represents the natural route of infection for this pathogen in humans. At 10–14 d postinfection, CD4 T cells were isolated from lymphoid tissue, purified, and assessed directly for their peptide specificity using synthetic peptides and cytokine ELISPOT assays, allowing direct ex vivo quantification of Ag or peptide-reactive lymphocytes. CD4 T cell specificity was identified from a sequential, iterative method of epitope identification using overlapping synthetic 17-mer peptides representing the entire translated sequence of these five viral proteins, as schematically illustrated in Fig. 1. These selected viral proteins that were examined (HA, NA, NP, M1, and NS1) differ with regard to their expression within the influenza virion and in infected cells, allowing us to evaluate whether these variables influenced epitope selection or algorithm performance. HA and NA are expressed as transmembrane proteins in the plasma membrane of infected cells and in the virion envelope. NP is localized within the cytosol and nucleus, and M1 is associated with the inner leaflet of the plasma membrane and is abundant in the virion. NS1 is notable because it is excluded from the virion particle.

For epitope discovery, the approach depended on the number of candidate peptides. For small viral proteins represented by <50 different peptides (M1 and NS1), each peptide was tested individually in ELISPOT assays. For larger proteins (HA, NA, and NP), we used the peptide pooling matrix strategy outlined on the right in Fig. 1, adapted by us (21, 22, 65) from an approach originally described by Tobery and coworkers (19, 66). Three different alleles (HLA-DR1, HLA-DR4, and I-A<sup>δ</sup>) were the focus of the current study because they have been studied extensively and thus we imagined that the performance of the algorithms...
would be highest for these. The individual peptide epitopes recognized by the CD4 T cells from infected mice are listed in Supplemental Table I, with the responses adjusted and presented to represent cytokine ELISPOTs elicited by each peptide per million CD4 T cells tested. Peptides were considered positive if they reproducibly recruited >30 cytokine-producing cells per million CD4 T cells, and data presented represents the average value from at least three and typically more than five independent assays.

Comparison of algorithm predictions versus the empirically defined epitopes

To analyze the relationship between predicted MHC class II binding peptide and experimentally identified epitopes, we scored the 17-mer peptides used for empirical identification of epitopes using the Web-available algorithms. We chose algorithms that integrate much of the available information regarding peptide binding to class II molecules and limited our studies to Web-accessible algorithms. In the analyses described in this study, we evaluated ARB, SMM, Tepitope, NetMHCIIpan, NetMHCII 2.2, and the Consensus method. The “score” of each peptide is provided in different formats for the different algorithms. Some of the algorithms, such as ARB, NN-align, and SMM (29, 30, 32), calculate a “predicted IC50” value for each peptide scanned and recommend a cutoff for strong, weak, and nonbinding peptide, typically in the range 10–500 nM. Other algorithms, such as Tepitope (67) and Consensus (34), rank the tested peptides from which an arbitrary cutoff can be chosen, without any estimate of absolute affinity. The Consensus method ranks the peptides based on assimilation of the prediction scores from what their studies have revealed to be the most accurate algorithms. The choice of how to analyze the performance of the algorithms in the current study was driven both by our desire to objectively analyze their accuracy and also by anticipation of the way these tools might be used by other investigators.

We began the analyses using the Consensus and the NetMHCII methods, both of which have unique strengths and have been widely used or evaluated by others (25, 29, 49, 50, 62, 68–71). The Consensus method calculates the median percentile rank of each peptide based on the predictions of three independent, high-performing algorithms. Because the scoring output is a percentile rank, there is no absolute value of affinity implied, and candidate epitopes are identified on inclusion within the top scorers. The recently developed NetMHCII (32) corrects for biases due to replicate binding cores within the training data and is one of the few algorithms that factors in effects of the amino acids that flank the 9-aa core of the peptides. It also has added flexibility in that it allows evaluation of peptides of different lengths. This method provides output in the form of predicted IC50 values. Both algorithms are able to predict epitopes for the three alleles of MHC class II molecules studied here, and both were found by their originators to outperform other individual state-of-the-art algorithms.

To globally visualize the predicted binding versus the actual hierarchy of CD4 T cell responses to influenza, the number of cytokine ELISPOTs elicited by each peptide tested within the five viral proteins (the “proteome”) were plotted versus the predicted affinity (NetMHCII) or percentile rank (Consensus). Predicted values are represented as their reciprocal, so that the height of the epitope on the y-axis represents a peptide’s actual or predicted “strength.” The CD4 cytokine ELISPOT data are represented as spots per million for every peptide analyzed, with an arbitrary cutoff of 100 spots per million CD4 T cells indicated, as this might be a minimal frequency (0.01%) useful for such methods at tetramer-based flow cytometry or intracytoplasmic cytokine staining. These graphs, shown in Fig. 2A, are thus divided into four quadrants, where the top-right quadrant represents the “double-positive” peptides—those peptides both predicted by the algorithm and that were also true CD4 T cell epitopes, as detected by cytokine ELISPOT assays. The bottom-left quadrant represents the double-negative peptides, those neither predicted nor observed. The top-left quadrant represents the “false-positive” peptides, those predicted to be above the threshold used, and the bottom-right quadrant represents the false negatives, those peptides found to be epitopes in the influenza-specific CD4 T cell response but not predicted to be high-affinity binders. The number of peptides that fell into each quadrant for each allele is tabulated in Fig. 2C, which also indicates the fraction of peptides in each category. Also quantified in Fig. 2C are minor epitopes, those that recruit low but detectable numbers of cells >30 but <100 cells per million. All regions scored are illustrated schematically in Fig. 2B.

These analyses indicated that both the number and percentage of each candidate peptide that is in the “double-positive” quadrant varies significantly with the allele and algorithm. For example, DR1 has the highest fraction of double-positive peptides chosen by NetMHCII (16 of 30 total positives, or 53%), most notably the four highly dominant peptides, recruiting >400 cells per million CD4 T cells and estimated to be of very high affinity. These true-
positive peptides were offset by a total of 61 false-positive peptides. DR4 had only one peptide that was both predicted and found empirically, whereas I-Ab has none. The range for Consensus ranged from 25–41% true positives, depending on the allele, and between 19 and 27 peptides predicted but not found, with DR1 having three peptides that were highly ranked and found to be immunodominant. In all cases, although the algorithms selected some epitopes, there were many epitopes that were not predicted to be MHC binding peptides by either algorithm and many peptides that were predicted to be binders that were not CD4 epitopes. Because some cellular proteins have preferential access to the MHC class II presentation pathway (39, 72–74), we analyzed the individual peptide epitopes for each viral protein separately to determine if algorithm performance varied among the proteins examined. Supplemental Figs. 1 and 2 show the results of these analyses for the three alleles analyzed and the predictions made by NetMHCII and Consensus. Although there were not striking disparities between predicted versus true epitopes, two rather consistent trends were noted. First, NetMHCII predictions for NS1, M1, and NP appeared to have a somewhat higher “hit rate” across alleles and fewer false positives. For example, for DR1, 11 NS1-derived peptides were predicted to be high-affinity binders, and six of these were true epitopes. Second, we noted that epitopes within HA seemed to have an opposite pattern, where 30 peptides were predicted to be high-affinity binders to this allele, but only eight of these were true epitopes. Supplemental Fig. 2 shows a similar type of analysis performed using the Consensus method. By the nature of the scoring for this algorithm, all protein/allele combinations will have similar numbers of peptides that fall above and below any cutoff, independently of whether they are truly high-affinity binders. Using a 10% cutoff as a “prescreen” would have indeed allowed identification of some of the major epitopes that were recognized by CD4 T cells after influenza infection. In the case of DR1 of the 25 major epitopes, only six were identified to be in the top 10% rank, two in NS1 and two in M1. For DR4, of the eight major epitopes, only three were identified as strong binders, and they were all in NP. For I-A^b, there were 11 immunodominant epitopes, and of these, three epitopes were within the top 10% predicted and two were in NP.

The relationship between cutoff choices for selection of peptides compared with the yield of actual epitopes for both algorithms is most readily seen in Fig. 3, which represents each peptide contained in the tested proteome plotted in order of decreasing predicted binding affinity or rank by NetMHCII and Consensus, respectively, with two potential cutoffs shown, corresponding to IC_{50} values of 50 or 500 nM for NetMHCII and 10 or 20% predicted top binders for Consensus. From this analysis, several conclusions were made. First, the epitopes identified empirically through functional studies are mostly contained within the top 50% of the 332 peptides scored. This indicates that the actual epitopes recognized by CD4 T cells are enriched in the high-scoring group of peptides. Second, although there is clustering of epitopes in this top half of the peptides, the most robust CD4 T cell specificities are not clustered within the highest of the predicted affinities. This is particularly apparent for the human DR4 molecule. Thus, although this type of display allows use of algorithms to select peptides, the “cutoff” for selection of peptides for further confirmation needs to be generous if it is hoped to capture the most dominant of epitopes with the candidate peptides. Finally, using NetMHCII, the ranking of affinity versus response is
different for each allele, so selection of a consistently appropriate cutoff may be difficult. For example, for DR1, the use of a 50 nM predicted affinity cutoff would select ∼80 of 332 peptides and would have successfully identified 30% of the true epitopes, including the three most immunodominant. The same predicted affinity for DR4 would select for only 15 peptides and only four epitopes, none of which are the most immunodominant. For the DR4 allele, the 500 nM cutoff would be more effective. For I-Ab, the 50 nM cutoff would select only two peptides, of which none of the 11 major epitopes would be found, and here the lower threshold of 500 nM would be more effective. Thus, the algorithms do not perform equivalently for all alleles, in terms of epitopes "captured" at any given cutoff. The bottom panel of Fig. 3 shows Consensus prediction, where the peptides are simply ranked. Consensus was the most successful for I-Ab where most of the peptide epitopes would have been identified with the 20% cutoff, and many identified even with the 10% cutoff. Finally, although the lowest ranked peptides contain some epitopes. In conclusion, both methods facilitated identification of epitopes, and if only a few epitopes are needed, a very stringent cutoff will select a subset of peptides that contains at least a few epitopes, but many epitopes, including major immunodominant epitopes, will be missed.

Evaluation of the relationship between predicted affinity and epitope dominance

We next extended our analyses to additional algorithms. We had two goals: first to assess the agreement among algorithms with each other and second to examine the relationship between immunodominance and the affinity estimates generated by the algorithms. Shown in Fig. 4 is this analysis performed with several algorithms that present estimated affinity in the form of IC₅₀, each of which is represented by a different symbol. Only SMM, NetMHCII, and ARB incorporate I-Ab, so the analysis of this MHC class II molecule was restricted to these three algorithms. For simplicity, we only included peptide epitopes that recruited at least 50 spots per million CD4 T cells. For several epitopes, there is good agreement among all of the algorithms for predicted binding, and these estimates could allow selection of epitopes. For example, for DR1, among the 32 dominant peptide epitopes, 20 were predicted by at least three algorithms to possess high-affinity binding, and three highly dominant M1 peptides and one NS1 peptide were predicted to be high affinity by all four algorithms, suggesting that some peptides have features that are well recognized to promote binding to this allelic form of human class II.

For DR4 and I-Ab, the relationship between predicted IC₅₀ among algorithms versus epitope dominance was much weaker. Among the eight major epitopes restricted by DR4 that recruited at least 100 CD4 T cells per million, two were predicted to bind with high affinity by more than two algorithms, whereas for I-Ab, of the 12 major epitopes, only two were selected by at least two algorithms to have high-affinity binding. This analysis also revealed that a number of epitopes that were identified empirically were not predicted to be MHC binding by any of the algorithms. For DR1, 9 of 32 were judged to be nonbinders by at least three algorithms, for DR4, 4 of 8 were judged to be nonbinders, and for
The vast majority of epitopes were predicted to be poor binders or nonbinders. There was also a significant range in the IC50 values predicted for individual peptides, as evidenced by the high degree of spread among the different symbols in Fig. 4.

Analysis of the efficiency of predictive algorithms for response “capture”

Because there are many false-positive peptides predicted by each algorithm, we concluded that secondary screens are important and next sought to evaluate different thresholds that might be used for selection of candidate peptides. Of particular interest was the cost (how many peptides would be tested in secondary screens) versus benefit (how many epitopes found) of different thresholds. In this analysis, we selected the subset of algorithms that provides output in the form of predicted affinity (SMM, NetMHCIIpan, NetMHCII, and ARB). Fig. 5A shows the fraction of the response predicted or what would have been “captured” by preselection of epitopes by the indicated algorithm, at either an IC50 value of 50 or 500 nM. The cutoffs were chosen based on common use in the literature by those who derive or use the predictive algorithms (49, 61–64). Although these two values are somewhat arbitrary and other cutoffs, such as 5 nM for very high affinity or 1000 nM for very low affinity, could be used, the two cutoffs selected provide a framework that can be used to compare the predictive efficacy of the individual algorithms. The number of cytokine-producing cells elicited by all of the peptides identified by an algorithm were summed and then divided by the total number of cytokine-producing CD4 T cells response identified empirically. One can consider this to be the “benefit” of using the algorithms to select peptides for further study. Fig. 5B presents the number of peptides that were selected by each algorithm at the two cutoff values and thus the number that would need to be made if all of the peptides predicted to be in that group of binders were tested in functional studies.

Use of a 50 nM IC50 value of predicted affinity did capture at least 30% of the response for DR1, with increasing performance using SMM < NetMHCIIpan < NetMHCII < ARB. However, the apparent greater performance came at the “cost” of increasing numbers of peptides to be tested (Fig. 5B, open bars) for these algorithms, based on additional peptides estimated to be better than the 50 nM IC50 value. For the less well studied DR4 molecule, ARB outperformed the other algorithms and would have allowed identification of 40% of the response using the 50 nM IC50 value.

FIGURE 4. Affinity predictions from different algorithms for influenza-specific CD4 T cell epitopes. Shown by bars for each allele is the magnitude of the response recalled by the indicated peptide, represented as cytokine spots per million CD4 T cells on the right-hand y-axis. Shown by the symbols are the affinity predictions from each algorithm for the peptides relevant to the indicated allele, with the axis shown in the left-hand y-axis. The black line indicates the threshold of 50 nM, where all symbols above that line are peptides predicted to bind with high affinity. The solid red line indicates the threshold of 100 spots/million for dominant peptide epitopes.

FIGURE 5. Algorithm performance in predicting CD4 T cell response specificity using different affinity cutoffs. The cytokine ELISPOTs for all epitopes recognized by CD4 T cells in each strain were summed to represent the total response. Predictions of MHC class II binding peptides were calculated by each algorithm, and the epitopes that were successfully predicted at either 50 nM (open bars) or 500 nM (filled bars) IC50 values were identified. (A) Responses to these two subsets of peptides were summed and divided by the total response elicited by all of the peptides to yield “% predicted,” the fraction of the response that would have been identified by using only the algorithm-selected peptides at the indicated threshold. (B) The total number of peptides predicted by each algorithm for each allele at either the 50 or 500 nM threshold is shown. The total number of peptides that would have been made using no selection is 332.
have selected a response predicted at 500 nM predicted affinity cutoff, which would have required testing of only 60 peptides. At the 500 nM cutoff, NetMHCIIpan appeared to outperform the other algorithms in terms of cost versus benefit, with >60% of the response captured and only requiring testing of 90 peptides. For I-Ab, ARB seemed to be the best performing, with >30% of the response predicted at 500 nM predicted affinity cutoff, which would have selected ~80 peptides to be tested in secondary screens. This comparative retrospective analysis shows that it is difficult to identify algorithms or parameters that will efficiently identify epitopes for untested class II molecules.

Analysis of algorithm performance by ROC curves

The large discrepancies within the absolute affinities predicted by the different algorithms introduced the rather large caveat of having to pick a threshold adequately to compare predictive ability between algorithms. As observed in the previous section, a good predictive algorithm in a relative sense can hypothetically be too stringent or lenient with its predicted affinities and be eclipsed by an algorithm with less predictive ability but more moderate affinity estimate. To circumvent this caveat, the binary classifier system—sorting predictions into true positive, false positive, true negative, and false negative—was imposed at all possible thresholds, and for each threshold, a true-positive rate and false-positive rate were calculated and plotted as ROC curves for each algorithm (Fig. 6), a method that is commonly used when multiple algorithms are compared for predictive performance (23–25, 32, 34) and that is generally used to measure the distinguishing capability of a classification or diagnostic system (75). The AUC values were calculated to score the overall predictive ability of each algorithm without the need for a binding cutoff and compared with what would have been predicted by chance alone. From the given analysis, NetMHCIIpan and NetMHCII seem to score consistently higher than the mean AUC value. However, aside from Tepitope’s relatively poor performance in DR1, there is no algorithm that significantly outperforms or underperforms relative to the other algorithms. Furthermore, the differences between algorithms are increased for more well studied MHC class II alleles. The range and SD between AUC values is largest in DR1 and near negligible in I-Ab.

Potential advantage of percentile ranking versus predicted affinity for selection of peptides as candidates for further studies

The overall inconsistency in using affinity estimates to predict epitopes, as well as the tendency of some algorithms to estimate high affinity to many peptides (ARB) and other to estimate low affinity to many of the peptides (SMM), prompted us to consider the possibility that the affinity estimates can be inaccurate in an absolute sense but might be useful in a relative sense. We therefore asked if ranking of peptides in a hierarchy of affinity rather than absolute affinity estimates would lead to more consistent results and more efficient use of the algorithms while allowing us still to estimate the actual fractional response successfully predicted. To evaluate this, we simply ranked all the peptides by their estimated IC_{50} values and then used a simple percentile cutoff rather than affinity threshold to select peptides (Fig. 7). Using this method of ranking, we were also able to extend our studies to Consensus and Tepitope, which simply rank peptides. When this fixed percentile ranking rather than estimated affinity was used to select peptides, a much better “performance” was apparent for all alleles and algorithms, with as much as 40–60% of the response predicted at even the 10% cutoff, depending on the allele. There was a particularly striking agreement when comparing the different algorithms for DR1 using this method. This result indicates that algorithm outputs of IC_{50} values for peptide–class II interactions are the most useful for simple ranking of candidate peptide epitopes, rather than for use as absolute values for selection. This conclusion from our studies is consistent with recently published approaches that assimilate the rankings of peptides, rather than their estimated affinities, to select peptides for further empirical study (34, 71, 76).

An additional and quite striking observation from this analysis is that although irregular in shape, the lines that describe the prediction effectiveness of each algorithm in capturing the response essentially overlap. They all have a fairly sharp slope to the 10% cutoff (successfully predicting ~20–30% of the response), then a diminished slope from ~10–60% cutoff, and then approach a plateau after the 60–70% cutoff, representing the low yield of additional epitopes with increasing peptides tested. The conclusion from these analyses is that ranking of peptides by different algorithms normalizes their prediction efficiency, and when compared in this way, they all perform similarly to each other. The conclusion is in agreement with the ROC analyses discussed earlier. The individual peptides “captured” at each cutoff varies somewhat for each algorithm (data not shown), but any given algorithm would apparently allow identification of ~25% of the response by selection of 32 peptides of 332, at least for these three alleles of MHC class II molecules.

![FIGURE 6. ROC analysis and AUC values of predictive algorithms. The predictive ability of SMM, NetMHCIIpan, NetMHCII, ARB, Tepitope, and Consensus are shown in terms of ROC curves. The “standard” (y = x) line represents the curve formed had peptides been chosen at random. Predictive ability is gauged by how far the points on the curve fall to the upper left of the standard line. Because better performing algorithms encompass more area under their respective curves, AUC values are a numerical representation of algorithm efficiency, with 1 representing a perfect predictor and 0.5 representing a nonpredictive or random algorithm.](http://www.jimmunol.org/doi/abs/10.4049/jimmunol.1801110)
The cytokine ELISPOTs for all of the epitopes recognized by CD4 T cells in each strain of mice were summed to represent the total response. Prediction of MHC class II binding peptides were estimated by each algorithm, and epitopes were ranked by each algorithm in the order of diminishing predicted affinity. Cytokine ELISPOT responses elicited at each incremental rank of the peptides selected by each algorithm were summed and divided by the total response elicited by all of the peptides to yield the number of peptide reactive CD4 T cells successfully predicted at the indicated threshold. Shown by the dotted line is the point at which 50% of the response would have been predicted or “captured” through the use of the indicated algorithm.

**Kinetic measure of class II–peptide affinity as a predictor of immunodominance**

One issue that is raised by the data in Figs. 2 and 4 is why some peptides that are not predicted to bind to the relevant MHC class II at all do in fact successfully elicit CD4 T cell responses to influenza infection. One possibility to explain these false negatives is that some peptides can bind very weakly to the “presenting” MHC class II molecule but nonetheless have properties that allow elicitation of a CD4 T cell response; for example, TCR contacts that recruit a tremendously high number of reactive CD4 T cells, a parameter that is not currently evaluated by available algorithms. The other possibility is that the IC50 values estimated by the algorithms do not accurately assess the true affinity of some MHC class II–peptide complexes. Our laboratory has previously shown that dissociation kinetics of MHC class II–peptide complexes rather than competition assays are the best predictor of immunodominicity, both in responses to complex proteins and in multipeptide-based vaccines (57, 77–80). In analyses of class II–peptide complexes with dissociation rates that differ by seven orders of magnitude, McConnell’s group (81) showed that the association rates were essentially the same, indicating that dissociation kinetics provide accurate measurements of the affinities of peptide–class II complexes. To evaluate whether this parameter is a better indicator of immunodominity for influenza-specific responses, we tested the dissociation rates of several DR1 peptide complexes using purified class II and purified fluorescent-labeled peptides. Two peptides that elicited greater than 100 CD4 T cells per million but that were not predicted by any of the algorithms to possess high-affinity binding to DR1 were chosen for this biochemical study. One peptide was derived from HA (375-SGYAADQKSTQNAING), and one was from NP (73-ERRNKYLEEHPSAGK). Two control peptides analyzed in earlier work (56) were studied in parallel. Dissociation half-times of the peptide–class II complexes (t1/2) cytokine ELISPOTs, and predicted IC50 values for these peptides are shown in Table I. It is clear from this study that all of the immunogenic peptides tested have very high affinity and stable interactions with the presenting class II molecule. The HA [375–389] and NP [73–88] peptides that were not predicted to bind with high affinity to DR1 by any of the algorithms had dissociation half-times of 600 and 500 h, respectively, indicating that they persist on DR1, even at acidic pH, for >20 d. Only the one HA peptide previously examined by our group (ELLVLLENERTLFHD, t1/2 350 h) was predicted to be a high-affinity binder (low IC50) by most of the algorithms. Strikingly, among the four peptides, there was no correlation between the dissociation rates and predicted IC50 values. The two peptides that had the most rapid dissociation rates (HA-162 and HA-440 with t1/2 of 85 and 350 h, respectively) had the highest estimated affinity (mean of 106 and 148 nM IC50, respectively), whereas the peptides that had the most stable interactions with DR1 (HA-375 and NP-73, t1/2 of 600 h and 500 h, respectively) had the lowest mean estimated affinity (3354 and 1300 nM, respectively). Table II shows a similar type of analyses performed with the murine I-A^d molecule using the algorithms that can be used to score antigenic peptides presented by I-A^d and predict IC50 values (NetMHCII, SMM, and ARB). In these studies, known immunodominant or cryptic epitopes from model or pathogenic origin, whose binding register and dissociation kinetics were previously determined by us (57, 59, 79, 80, 82), were studied. We also compared peptide variants of these peptides whose binding stability to the class II molecule could be modulated by changes in anchor residues, as we have described (57, 59, 79, 82). Again, as was seen with some of the epitopes presented by HLA-DR1, immunodominant peptides such as LACK [161–173] and an HA

### Table I. Relationship between the predicted IC50 values, kinetic stability, and the immunodominance of influenza-derived peptides presented by HLA-DR1 molecules

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Sequence</th>
<th>Cytokine Spots t1/2 (h) pH 5.3</th>
<th>SMM IC50</th>
<th>NetMHCIIpan IC50</th>
<th>NetMHCII IC50</th>
<th>ARB IC50</th>
<th>Mean IC50</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP-73</td>
<td>ERRNKYLEEHPSAGKDP</td>
<td>167 500 3684 860 151 506 1300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA-375</td>
<td>SGYAADQKSTQNAING</td>
<td>118 600 8264 1930 3058 166 3354</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA-162a</td>
<td>KRNLWLTGKGL</td>
<td>163 85 302 54 30 40 106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA-440a</td>
<td>ELLVLLENERTLFHD</td>
<td>146 350 523 42 13 15 148</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shown are the number of CD4 T cells recruited per million, the t1/2 dissociation values for the DR1–peptide complexes, measured as described in Materials and Methods, and affinity (IC50 value) of the peptide for DR1 predicted by the indicated algorithm.

*Peptide with a sequence extension to improve peptide solubility but which did not change the affinity value predicted by the algorithm.
peptide [126–138, T128 > V] designed to improve its anchor interaction with the P1 pocket (57, 59, 80, 82) were predicted by the algorithms to have variable but generally weak IC$_{50}$ values that were not substantially different from peptides that very poorly recruit CD4 T cells in vivo such as hen egg lysozyme (HEL) [11–25] or HA [126–138] (57, 82). In all cases, however, the measured off-rates of the peptides from the class II molecule correlated with their immunodominance. Table III additionally shows that in many cases, incorrect or differing binding registers were predicted by the different algorithms from what our studies have shown to be the correct binding register (57, 59, 82). We conclude from these studies that the ability to bind strongly to the presenting class II molecule is indeed a strong predictor of immunodominance and that peptide dissociation assays accurately measure this parameter, but that algorithms sometimes fail to recognize the features of high-affinity binding.

**Discussion**

In this study, we have combined the results of an unbiased approach for influenza CD4 T cell epitope discovery with implementation of predictive algorithms to select immunodominant peptide epitopes. Advances in CD4 epitope discovery are becoming increasingly important for the tracking and functional characterization of CD4 T cells during or after infection or vaccination and are also essential for derivation of peptide-based vaccines. Purely empirical strategies for epitope discovery, using either the method we have used or other approaches such as tetramer-based epitope mapping (83–86), are daunting because of their cost and labor-intensive nature. Also, for pathogens with large genomes, purely empirical approaches are impractical. Therefore, the potential benefits of computer-based algorithms to facilitate identification of CD4 T cell epitopes are significant. The increasing depth of available databases for peptides presented or bound by class II molecules (46, 87–89), coupled with enhanced sophistication of computer-based algorithms, has led to increasing optimism toward use of predictive algorithms to facilitate identification of CD4 T cell epitopes. From a practical standpoint, the high false-positive rate category varied, depending on the specific allele and algorithm and the cutoff used for selection of peptides. For example, with lower thresholds may need to be used, with the increasing “yield” balanced by increasing “cost” of greater numbers of synthetic peptides to test. Agreement of algorithms and prediction efficiencies were enhanced through use of a ranking system rather than predicted IC$_{50}$ value to select peptides.

Second and quite striking for each of the algorithms tested, there is a high number of “false-positive” peptides identified: peptides that were predicted to be high-affinity binders to the host class II molecule but that were not recognized by CD4 T cells elicited in response to influenza infection. The number of peptides in this category varied, depending on the specific allele and algorithm and the cutoff used for selection of peptides. For example, with high stringency (either 50 nM for algorithms that predict IC$_{50}$ values or 10% for algorithms that rank peptides), the range in false positives for DR1 ranged from 20 to 65 peptides for the 332-peptide “proteome,” consisting of five viral proteins tested. When the threshold included all peptides that are predicted to bind to the relevant MHC class II molecule, there were >100 false-positive peptides. From a practical standpoint, the high false-positive rate for many class II–protein combinations indicates in general that the algorithms should be used only as a first step in epitope discovery. We conclude that secondary screens are essential to

Severe several important conclusions were possible from the analyses we have performed. First, use of predictive algorithms can facilitate epitope discovery. With all of the three allelic forms of class II molecules tested in this study, the identified epitopes were highly represented in the top third to half of the peptides predicted by the different algorithms. Although the degree of enrichment varied with the allele and the algorithm, this result does suggest that use of predictive algorithms can be a good first step toward epitope identification, at least for the alleles analyzed in this study and for pathogens with limited genome sizes or for protein vaccines. Specific strategies to use this enrichment step for these types of applications will depend on the goals of the investigator. For example, if one seeks to identify only a few epitopes for tracking a response to a particular pathogen, fairly stringent selection (i.e., selection of the top 5–10% of peptides) will likely be successful. If more criteria need to be fulfilled (highly immunodominant peptides or CD4 epitopes within a particular protein), lower thresholds may need to be used, with the increasing “yield” balanced by increasing “cost” of greater numbers of synthetic peptides to test. Agreement of algorithms and prediction efficiencies were enhanced through use of a ranking system rather than predicted IC$_{50}$ value to select peptides.
identify the true epitopes, after which focused studies can be pur-

sued. For large-genome organisms, however, secondary screens of

large numbers of candidate peptides may be unfeasible using the
currently available computational strategies.

There are several possibilities to explain the high number of false-

positive peptides. The first and simplest explanation is that these

peptides have been “chosen” correctly by the algorithm to be high-

affinity binders for the MHC molecules expressed in the host but

lack some property needed to be a CD4 T cell epitope. One ele-

ment that could be lacking in MHC-binding peptides that do not
elicit CD4 T cells could be their ability to be processed and pre-

sented by the class II molecule, due to the source protein abun-
dance or the location of the peptide within the viral protein. A

second feature that might diminish a peptide’s antigenicity despite

high-affinity binding to class II molecules is that it possesses few

amino acid residues in solvent-exposed positions that are potent in

recruitment of CD4 T cells. The predicted affinity of some peptides

may be an overestimate, even for those peptides that have “at-

tractive” TCR contacts and are abundantly released and available

for peptide binding during Ag presentation. We do know that the

reproducibility in responses among individual animals is reason-

ably good (Supplemental Fig. 3), suggesting that we have not

missed epitopes due to pooling of lymphocytes from multiple

animals, each of which might display greatly varied responses.

Pooling of CD4 T cells from multiple mice is a necessary exper-

imental strategy because of the number of different peptide epi-
topes tracked in parallel in our studies and the low yields of CD4

T cells from individual mice. Finally, from a purely technical

standpoint, it is possible that poor peptide quality or solubility

limits our T cell detection for occasional peptides, thus accounting

for false positives. The peptides used in our studies were HPLC

purified, but it is possible that peptide quantity or purity was

overestimated by the supplier or alternatively that peptide solubility

was limited, thus diminishing the signal in the ELISPOT assays.

Our analyses also revealed a high false-negative rate; peptides

that were recognized in the CD4 T cell response but not predicted to

be MHC binders by most of the algorithms. The analyses of this

study indicated even with well-characterized class II molecules that

at least 25% of the peptides that elicit strong CD4 T cell responses

were not predicted by most of the algorithms to bind to the MHC

molecule in question. Although in the animal studies reported here,

we have studied the immunodominance pattern at the peak of the

immune response (day 10–14), we have found that the hierarchies

in peptide-specific CD4 T cells remain in the memory phase of the

immune response (56). These results suggest that epitopes dis-

covered in our studies are representative of those persisting in the

memory phase and in human circulating CD4 T cells. In fact,

some of the HLA-DR epitopes discovered in our transgenic

models predicted to be of very low affinity by all of the algorithms
tested, including NP-73 and NP-402 for DR1 and M1-97 for DR4,

have been reported in the literature from human subjects (84, 90).

Finally, we are confident that the false-negative rate is not due to

incorrect assignment of the class II restriction element in the mice.

Only the HLA-DR1 mice express an additional class II molecule, and

for these peptide epitopes, MHC restriction was unambigu-

ously assigned through the use of transfected APC in ELISPOT

assays, as we have described (21).

The false-negative score for many epitopes suggests that the

algorithms in general fail to recognize certain types of peptides that

are presented by MHC class II molecules, even those that have been

well studied, such as those in this study. In favor of this possibility

are results of our studies with the I-A\textsuperscript{d} class II molecule, which has

been extensively analyzed for its peptide binding and cocystal-
lized with antigenic peptide (91). I-A\textsuperscript{d} was widely thought to have

a “promiscuous” binding pocket with specificity primarily dictated

by the P4 and P6 pockets (39, 91, 92), but we recently found the

P1 pocket to be a strong determinant of peptide binding particu-

larly with glutamic acid at this position (59, 82). Our studies

suggest that this charged residue binds via a novel salt bridge within

this pocket. Because of this alternative type of interaction in the

P1 pocket, the “value” of glutamic acid in antigenic peptides was

previously missed.

One unexpected finding revealed by our analyses is that the

efficiency in prediction of epitopes seems to similarly plateau for all

the algorithms tested, even if ranking is used to order peptides. It is

remarkable that algorithms developed recently do not outperform

others developed more than a decade ago when the efficiency of

prediction was analyzed in this way. This conclusion prompts us to

question whether the approaches used for development and re-

finement of computation tools have fundamental limitations that

will not improve even with more sophisticated computational

methods. We suggest that rather than focus on modifications of

computational strategies, increased performance of computational

aids to epitope prediction may require a serious re-evaluation of

the data that are used to “train” the algorithms. One improvement

that might significantly enhance performance would be dispro-

portionately to weigh data that provide the most information on

the peptide core and stability of binding to the host MHC molecule.

Most of the newly developed methods are trained on data derived

primarily from competitive binding assays. Because of their rel-

atively low cost and high-throughput capacity, these assays are

very useful for identification of peptides that do or do not bind to

given class II allele. However, the quantitative ability of peptides

to compete for binding may reflect properties that are distinct from

those that promote accumulation onto class II molecules, clearly

a key event for immunogenicity (57, 77, 80, 82). From the sim-

plest point of view, IC\textsubscript{50} values are measures of the ability of

a given peptide to diminish accumulation of a labeled “indicator

peptide.” Such competitive assays are based on the assumptions

that data arise from a simple equilibrium between single, ho-

mogenous monovalent receptor and ligand and follow the laws of

mass action (93–96). It is now clear from the work of many groups

that the reactions between peptide and MHC II proteins involve

long-lived intermediates, heterogeneous initial states, and a pre-
equilibrium between active and inactive forms of the class II

protein and that binding reactions compete with inactivation of the

MHC protein, an event that is potentiated by formation of unstable

MHC–peptide complexes (97–101). The complexities in class II–

peptide interactions, coupled with the possibilities of multiple

binding cores of antigenic peptides, indicate that predictions based

solely on competitive binding assays will be limited in their ability

quantitatively to assess the ability of the test peptides to bind

stably to the class II molecule. Dissociation assays and elution and

sequencing or crystallographic studies together provide key in-

sight into the amino acids that promote accumulation of peptides

onto the class II molecules. Although these assays are cost and

labor prohibitive as routine methods, more investment in genera-

tion of these data and more weight given to these data could

enhance the predictive capacity of computational strategies.

Incorporation of data from these assays might also improve ac-

curacy of the 9-aa core predicted to bind in the MHC class II

pocket. We hypothesize that limitations in our understanding of

peptide–class II interactions have led to incorrect core assign-

ments, a hypothesis recently supported by discrepancies between

predicted cores and cocRYstallization of peptide–class II com-

plexes (25, 63) and by our own data shown in this study. Faulty

core designation may “contaminate” the prediction accuracy of

preferences in binding going forward.
True breakthroughs in algorithm performance for prediction of CD4 T cell epitopes is also likely to be facilitated by considerations of peptide sequences that contribute to recruitment of diverse TCRs. Although strength of MHC binding is the key parameter that determines a peptide’s immunogenicity, solvent-exposed residues within the peptide likely determine the upper range of the magnitude of a CD4 T cell response that can be recruited by a peptide. Recent theoretical and functional studies and detailed consideration of the molecular events in thymic positive and negative selection events support the view that particular solvent-exposed amino acids within MHC-bound peptides dramatically influence TCR repertoire development and elicitation (102–108). Notably, recent syntheses of computation and functional studies suggest that the best antigenic epitopes will have TCR contact amino acids that both survive negative selection and have the ability to recruit diverse TCR, and these studies quantify the amino acids that tend to favor this “equation” (109, 110). Incorporation of these characteristics of TCR recruitment into computational methods of epitope prediction will likely increase the efficiency of computer-based peptide epitope selection. We suggest that more explicit experimental evaluation of these issues, coupled with the computational power of many groups of investigators, will dramatically enhance the performance of computation-based predictors, a key step needed for efficient epitope discovery.

Acknowledgments
We thank Michael (Rusty) Elliot for comments on the manuscript and faculty at the David H. Smith Center for Vaccine Biology and Immunology for helpful discussions.

Disclosures
The authors have no financial conflicts of interest.

References


