Structural bioinformatics

Peptide length-based prediction of peptide–MHC class II binding

Stewart T. Chang1, Debashis Ghosh2, Denise E. Kirschner3,5 and Jennifer J. Linderman4,5,*

1Program in Bioinformatics, 2Department of Biostatistics, 3Department of Microbiology and Immunology,
4Department of Chemical Engineering and 5Department of Biomedical Engineering, University of Michigan,
Ann Arbor, MI, USA

Received on April 14, 2006; revised on September 7, 2006; accepted on September 8, 2006
Advance Access publication September 25, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Algorithms for predicting peptide–MHC class II binding are typically similar, if not identical, to methods for predicting peptide–MHC class I binding despite known differences between the two scenarios. We investigate whether representing one of these differences, the greater range of peptide lengths binding MHC class II, improves the performance of these algorithms.

Results: A non-linear relationship between peptide length and peptide–MHC class II binding affinity was identified in the data available for several MHC class II alleles. Peptide length was incorporated into existing prediction algorithms using one of several modifications: using regression to pre-process the data, using peptide length as an additional variable within the algorithm, or representing register shifting in longer peptides. For several datasets and at least two algorithms these modifications consistently improved prediction accuracy.

Availability: http://malthus.micro.med.umich.edu/Bioinformatics

Contact: linderma@umich.edu

1 INTRODUCTION

Major histocompatibility complex (MHC) molecules, also known as human leukocyte antigens (HLAs), are a vital component to the development of the immune response to pathogens (Kaufmann, 2005). These molecules act as receptors for peptides derived from foreign antigens as well as self peptides and enable the long-term display of antigens on the cell surface. T cells recognize antigenic peptides in the context of MHC, and depending on the class of MHC involved, recognition can lead to the death of the presenting cell or its activation. In either case peptide–MHC binding is an important prerequisite event and has far-reaching consequences to the ensuing response.

Prediction of peptide–MHC binding therefore represents an important goal in bioinformatics, particularly as applied to immunology, and a number of computational approaches have been developed [reviewed in Buus (1999); see also Robinson et al. (2003) for other MHC-specific bioinformatics tools]. The simplest are based on motifs, i.e. requirements for particular amino acids at positions within the peptide as determined from pool sequencing of eluted peptides (Falk et al., 1991; Rammenese, 1995 and references therein). Such approaches have largely been superseded by algorithms using matrices to score the relative contribution of amino acids at each position within the peptide (Parker et al., 1994; Davenport et al., 1995; Marshall et al., 1995). Machine learning methods including hidden Markov models and artificial neural networks have also been applied, with peptide sequence serving as input and binding/non-binding as output (Brusic and Harrison, 1994; Honeyman et al., 1998; Mamitsuka, 1998). More recently, attempts have been made to predict the structure of the peptide–MHC complex and free energy changes associated with binding [Altuvia et al. (1997), Rognan et al. (1999), Schueler-Furman et al. (2000), Davies et al. (2003) and Schaferoth and Floudas (2004); for a review of current structural information and nomenclature see Kaas and Lefranc (2005)]. It is also possible to combine some of these approaches, as Sturniolo et al. (1999) did using matrices to represent each pocket lining the peptide-binding groove.

Continued progress in the development of these algorithms faces a number of challenges including how to handle differences between the two classes of MHC. Most prediction algorithms were first developed in the context of peptide–MHC class I binding which involves peptides of a narrow range of lengths, usually 8–10 amino acids. These algorithms were then applied to peptide–MHC class II binding, typically with little or no modification.

Despite the fact that both classes of MHC share superficial similarities and bind a core of nine amino acids within peptides (Jones, 1997), important differences exist. In particular the open-ended nature of MHC class II peptide-binding groove allows for a wide range of peptide lengths (Brown et al., 1993). Peptides binding MHC class II usually vary between 13 and 17 amino acids in length, though shorter or longer lengths are not uncommon (Chicz et al., 1992; Serkarz and Maverakis, 2003). As a result peptides are hypothesized to shift within the MHC class II peptide-binding groove, changing which 9mer window (register) sits directly within the groove at any given time. In contrast the capped nature of the MHC class I peptide-binding groove does not allow variation in length or such register shifting.

Variation in peptide length may have important consequences for the binding and function of antigenic peptides (Malcherek et al., 1994; Vogt et al., 1994). For instance, Srinivasan et al. (1993) found that a 23mer peptide derived from cytomegalovirus c was 32 times more immunogenic than a 10mer peptide containing the same putative binding core. A direct relationship between peptide length and binding affinity has been observed for some MHC class II alleles, but whether this holds true for most alleles remains unknown, as does an explanation for why this relationship exists (Bartnes et al., 1999; Fleckenstein et al., 1999; Arnold et al., 2002;
were not used in this study. Our datasets comprised the sequences and IC\textsubscript{50} et al. in beta version at http://www.immuneepitope.org, Peters et al. are also available, including the Immune Epitope Database (currently
the perl
tration of peptide required to inhibit 50\% of reporter peptide–MHC binding.

2.2 Regression of binding affinity versus peptide length

Both parametric and non-parametric fits were made to plots of affinity versus length in the data. Parametric fits were made with one, two and three fitted parameters (linear, quadratic, and cubic, respectively) using the open-source statistical program R (http://www.R-project.org). R Development Core Team, 2005) and the function lm. Non-parametric local regression
fits were made using the R function loess with default settings (Cleveland and Devlin, 1988). To evaluate fit quality, analysis of variance
was performed using the R function anova. An F statistic was generated which we used to compare linear with non-linear parametric fits (Motulsky and Christopoulos, 2004).

Non-parametric local regression fits were evaluated using a permutation test. In this test each pIC\textsubscript{50} value was reassigned to a different peptide sequence at random, and a loess fit was re-derived for the shuffled values. This was repeated 1000 times, and the smallest 25 (2.5\%) and largest 25 (2.5\%) fitted values at each length were excluded. The local regression fit to the original, non-shuffled dataset was then compared with the remaining 95\% of permuted values at each length and was determined to be significant if it fell outside of this interval.

2.3 Simulations of register shifting

To simulate the effects of register shifting on peptide–MHC class II binding affinity over a range of peptide lengths, we derived a formula for the expected value of the affinity of a single hypothetical peptide with multiple registers:

\[ E[K(X)] = \sum K(x_i)p(x_i), \]

where \( K(X) \) is the equilibrium association constant, or affinity, of a peptide \( X \), \( K(x_i) \) is the affinity of a complex with a single register \( x_i \), and \( p(x_i) \) is the probability of register \( x_i \) occurring. We assume that \( p(x_i) \) can be approximated by the proportion of complexes having register \( x_i \):

\[ p(x_i) = \frac{N(x_i)}{\sum N(x_i)}, \]

where \( N(x_i) \) denotes the number of complexes having register \( x_i \) and the sum is taken over all possible registers. Belmares and McConnell (2001) found that the kinetics of shifting between two registers could be accurately represented as \( x_1 \rightarrow P + M \rightarrow x_2 \) where \( P \) and \( M \) are peptide and MHC, respectively. Based on this result, at equilibrium \( [x_1] = K(x_1)[P][M] \) and \( [x_2] = K(x_2)[P][M] \). Because both \( x_1 \) and \( x_2 \) exist in the same solution, it follows that:

\[ \frac{N(x_1)}{N(x_1) + N(x_2)} = \frac{K(x_1)}{K(x_1) + K(x_2)}. \]

More generally,

\[ \frac{N(x_i)}{\sum N(x_i)} = \frac{K(x_i)}{\sum K(x_i)}. \]
Combining Equations (1), (2) and (4), we obtain the following result for the expected affinity of a given complex when multiple registers are available:

$$E[K(X)] = \frac{\sum K(x_i)}{\sum K(x_i)}.$$  \hspace{1cm} (5)

This result can also be applied to log-transformed measures of affinity such as log K(X). Henceforth we refer to Equation (5) (or its log-transformed counterpart) as the equilibrium-based formula for reconciling multiple registers.

We assume that every overlapping 9mer window within a peptide can result in binding to MHC and therefore set the lower and upper limits of summation at 1 and \( l - 8 \), respectively, where \( l \) represents peptide length and is varied between 9 and 25, the shortest and longest lengths typically observed in our datasets. \( K(x_i) \) was generated from a lognormal distribution with mean \( 10^{-3} \) and SD \( 10^{0.3} \), based on the observation that most values for the equilibrium dissociation constant \( K_D \) of peptide–MHC binding fall in the range of \( 10^{-7} \) to \( 10^{-3} \) M (McFarland and Beeson, 2002). Moreover, a lognormal distribution was chosen based on the equation for standard free energy change, \( \Delta G = -RT \ln (1/K_D) \) where \( R \) and \( T \) are the gas constant and temperature, respectively (Eisenberg and Crothers, 1979), and the assumption that free energy change for peptide–MHC binding is normally distributed. For each value of \( l \) between 9 and 25, a set number of values were generated (in our case, either 10 or 100), resulting in a scatter plot of simulated pIC\(_{50}\) values versus length. A curve was then fit to this plot using local regression (the loess function in R) with default settings.

### 2.4 Peptide–MHC binding affinity prediction

Two algorithms were selected to generate baseline predictions against which the effects of modifications based on length could be compared. One of these algorithms was the iterative self-consistent (ISC) partial-least-squares (PLS) algorithm of Doytchinova and Flower (2003). We implemented this matrix-based algorithm for predicting peptide–MHC binding affinity in perl and R. Briefly, this algorithm uses partial-least-squares regression to identify underlying factors (also known as latent variables) relating multiple predictor variables to an outcome variable. In the case of peptide–MHC binding, 180 predictor variables were used to denote the presence or absence of the 20 possible amino acids within each 9mer window, and the outcome variable was binding affinity as pIC\(_{50}\).

The initial step of the algorithm were performed using perl scripts: splitting each dataset into training and test sets; generating all possible 9mers for each training set peptide; selecting only those 9mers having position 1 anchor residues (F, I, L, M, V, W and Y); and converting 9mers thus selected into bit strings. PLS regression was then performed in R using the bit-encoded 9mers and their corresponding pIC\(_{50}\) values. PLS is available for R as the pls.pcr library (available at http://cran.r-project.org) and was called from within a perl script using the IPC::Open2 module. Default settings were used for PLS; however, some options in the commercial software used by Doytchinova and Flower (2003) were not available in R, namely scaling method and column filtering. Subsequent steps in the algorithm were performed using additional perl scripts: selecting those 9mers in the training set yielding predicted pIC\(_{50}\) values closest to experimental pIC\(_{50}\) values during cross-validation and repeating the algorithm until the selected set of 9mers matched the previously selected set, i.e. when self-consistency was achieved. For computational expediency we limited the number of PLS iterations for any given peptide to 10. At that point the final PLS model was extracted and used to generate predictions on the test set.

For test set peptides having more than one 9mer with an anchor residue in position 1, multiple predictions were generated and a rule was needed to make a final prediction. One option is to assume only one register predominates and to take the highest score from among the predictions. More complicated rules are also possible such as the combination rule of Doytchinova and Flower (2003) whereby the mean of the pIC\(_{50}\) predictions is chosen if they fall within a one log range; otherwise, the highest is chosen.

To measure the performance of the algorithm we used 5-fold cross-validation (5x-CV), setting aside one-fifth of each dataset to use as a test set and using the other four-fifths as the training set. This process was repeated on the same dataset four additional times until a prediction was made for each peptide in the dataset and complete coverage was achieved. (This instance of cross-validation was independent of the leave-one-out cross-validation used in the ISC–PLS algorithm.) The accuracy of each set of predictions was scored by calculating the area under receiver operating characteristic curve (A\(_{ROC}\)). This calculation can be done in R using the prediction and performance functions of the ROCR library. By repeating each 5x-CV multiple times, we were able to calculate the standard error of the A\(_{ROC}\) scores which could then be used to determine whether two mean A\(_{ROC}\) scores significantly differed by Student’s t-test. Pearson correlation coefficients between predicted and experimentally determined pIC\(_{50}\) values were also used to score performance and are provided in the online Supplementary Data (Lund et al., 2005).

A second algorithm that was selected was the TEPITOPE algorithm of Stunziolo et al. (1999). In this algorithm amino acid-binding profiles are generated for each pocket within the peptide-binding groove, and these profiles are combined according to MHC sequence. We did not regenerate these matrices but rather used the matrices available on the ProPred website (http://www.imtech.res.in/raghava/propred, Singh and Raghava, 2001). Using the appropriate matrix a sum was calculated for each peptide in a selected Antigen dataset. To this value we added an approximation of the binding affinity of an all-alanine 9mer (pIC\(_{50} = 6.169\), Doytchinova and Flower, 2003) generating a final prediction. Performance was scored by calculating the A\(_{ROC}\).

### 2.5 Incorporating length into existing prediction algorithm

Peptide length was incorporated into the ISC–PLS algorithm using one of three modifications. In Modification 1 (Mod. 1) a local regression fit was first made to the peptide lengths and pIC\(_{50}\) measurements in each training set. (In the event that the pIC\(_{50}\) value for either the shortest or the longest peptide was excluded from the training set but included in the test set, a local regression fit at that length could not be generated; instead, we assigned the average fitted values at the remaining lengths.)

The value of the fit was then subtracted from the original pIC\(_{50}\) value for each peptide, and the resulting difference, i.e. the residual, was then used in place of the original pIC\(_{50}\) value. The ISC–PLS algorithm was performed as described earlier providing initial predictions on the test set. To these predictions the value of the regression fit was added yielding final predictions. Alternatively, in Alternative Mod.1 (Alt. 1), peptide length was appended as the 181st predictor variable to the bit-encoded training set and test set of 9mers. The remainder of the algorithm was then performed as described earlier. Finally, in Modification 2 (Mod. 2) the formula derived to represent register shifting [Equation (5)] was used to reconcile predictions made on multiple candidate 9mers, i.e. registers, within a test set peptide. This modification occurred at the last stage of the ISC–PLS algorithm and was used in place of the combination rule described above.

Only Mod. 2 was used to incorporate length into the TEPITOPE/ProPred algorithm. When TEPITOPE/ProPred is applied to peptides with multiple registers, the highest score among the different registers is typically taken to be the score of the entire peptide (Brusic et al., 1998; Nielsen et al., 2004; Murugan and Dai, 2005). We reconciled individual register scores using the equilibrium-based formula (Equation 5) but did not regenerate the pocket profiles and therefore did not apply Mod. 1 or Alt. 1 in this case.

### 3 IMPLEMENTATION

#### 3.1 Peptide length significantly affects binding affinity to MHC class II

To determine the nature of the relationship between peptide length and peptide–MHC class II binding affinity, we derived a number of
regression fits to binding data for several MHC class II alleles from the AntiJen database. In all cases homologous sequences were first removed from the datasets using a pre-filtering algorithm, UniqueProt (Mika and Rost, 2003). Parametric fits were then made based on polynomials with one, two, or three fitted parameters (linear, quadratic, and cubic, respectively). Analysis of variance from these fits showed that for these MHC class II alleles the nature of the relationship was most likely non-linear (Table 1). A quadratic or cubic fit resulted in a significant reduction in sum of squares in all three cases at the 0.05 level.

To better characterize the apparent non-linearities in the length-affinity data we then made non-parametric fits to the data and analyzed the fits. Local regression was used to make non-parametric fits, and analysis was done using a permutation test. In this test binding affinities were reshuffled among peptide lengths to create 1000 new datasets, and a local regression fit was re-derived for each dataset. If the fit to the original data fell outside of the middle 95% of permutation fits at any particular length, the non-linearity at that length was determined to be significant. In each dataset we found that the non-linearity between length and affinity was significant at one or more lengths (Fig. 2). Lengths associated with strongest affinity could be identified, as could lengths associated with weakest affinity. For example, for DRB1*0401 affinity appeared strongest for peptides of 12 amino acids and weakest for peptides of 20 amino acids. When the datasets were combined and the local regression fits were regenerated, the same trends were seen (Fig. 2D): shorter peptide lengths, of ~12 amino acids, were associated with higher affinity, while longer peptide lengths, of ~20 amino acids, were associated with lower affinity.

Non-linearities may have been present in the length-affinity data for several reasons, including the ability of peptides to shift registers within the MHC class II peptide-binding groove. To simulate the effect of register shifting on the mean affinity observed for peptides of different lengths, we used a simple statistical model based on two assumptions: first, that longer peptides are likely to contain more registers than shorter peptides, and secondly, that the measured affinity of a given peptide–MHC complex approximates the weighted average of the affinities of all the registers in a peptide [Equation (5)]. For a simulated peptide of a given length \( l \), the affinities of \( l - 8 \) registers were generated and averaged. This process was repeated until the average affinities of either 10 or 100 peptides at each length (i.e. each value of \( l \)) were obtained, resulting in datasets of two sizes (one of the same magnitude as those typically obtained from databases, the other an order of magnitude larger). At this point a regression curve was derived (Fig. 3). For the larger sized dataset the fitted curve was non-linear and monotonically increasing (Fig. 3A). The same trend was seen in the smaller dataset; in this case, however, deviations were also possible, resulting in maxima at mid-length peptides (Fig. 3B).

Together these results suggest that register shifting may be one mechanism behind the non-linearities in the length-affinity relationship from experimental datasets.

We also estimated the lengths of the N- and C-terminal portions of each peptide extending outside of the MHC class II peptide-binding groove to determine if particular lengths at either end of the peptide were favorable or unfavorable for binding. 9mer cores were identified by position 1 anchor residues (F, I, L, M, V, W and Y), and the lengths remaining at each end were calculated. Local regression fitting and permutation testing were done as with overall peptide length. In most cases fits to N- and C-terminal peptide extensions were determined to be significant at one or more lengths (Fig. 4 and additional data not shown). In comparing fits we found that extensions of 2–4 amino acids at the N-terminus and extensions of 1–2 at the C-terminus generally appeared favorable for binding (Fig. 4 and additional data not shown). Likewise, longer extensions (8 and 10 amino acids at the N- and C-termini, respectively) generally appeared unfavorable for binding (Fig. 4 and additional data not shown). We also found that in at least some cases fits to overall peptide length could be decomposed into N- and C-terminal contributions. For example binding to DRB1*0401 was strongest when N- and C-termini were 2 and 1 amino acids, respectively (Fig. 4). Together with the 9mer core, these lengths sum to

### Table 1. Evidence of non-linear relationships in length-affinity data for several MHC class II alleles

<table>
<thead>
<tr>
<th></th>
<th>DRB1*0101</th>
<th>DRB1*0401</th>
<th>DRB1*1501</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic, ( F )</td>
<td>11.745 (0.001)</td>
<td>8.575 (0.004)</td>
<td>3.670 (0.057)</td>
</tr>
<tr>
<td>Cubic, ( F )</td>
<td>5.849 (0.016)</td>
<td>0.708 (0.401)</td>
<td>4.871 (0.028)</td>
</tr>
</tbody>
</table>

\( F \)-statistics are shown for analysis of variance results with \( P \)-values in parentheses.
3.2 Modifying algorithms to account for peptide length consistently improves performance

We incorporated peptide length into two peptide–MHC class II binding prediction algorithms in one of three ways. First, as a pre-processing event (Mod. 1 in Fig. 1) a local regression fit was made for affinity versus length in the training/fitting data and the value of the fit was subtracted from each affinity measurement. The resulting residuals were used in place of the original pIC50 values in the training set. After the algorithm was used to make initial predictions for the target set peptides, the value of the regression fit for each target set peptide length was added to yield final predictions. Alternatively (Alt. 1 in Fig. 1) length was also incorporated directly into the existing algorithm as an additional variable (in the case of ISC–PLS, as the 181st variable). Training/fitting was then performed as published, and predictions were made on test set peptide sequences and peptide lengths. Lastly we used a formula derived from the equilibrium-based statistical model to reconcile predictions made by existing algorithms on multiple registers within the peptide (Mod. 2 in Fig. 1). We point out that Mod. 1 and Alt. 1 are similar modifications that both consider peptide length directly (by fitting length as a discrete variable); in contrast Mod. 2 considers binding registers (i.e. 9mers with a valid position 1 anchor) and the relationship among them. Therefore, Mod. 1 and Alt. 1 are not used together, although either can be used with Mod. 2.

Incorporating peptide length by one or more modifications into the ISC–PLS algorithm improved the performance of the algorithm for all alleles examined (Table 2). Performance was measured by area under receiver operating characteristic curves (AUC) at the N-terminus and (B) at the C-terminus. 95% boundaries of permutation distributions are shown (dotted) with fits to the original, non-shuffled data (solid).

Fig. 4. Local regression fits of peptide–MHC class II binding affinity versus lengths of portions of the peptide extending outside of the peptide-binding groove for the HLA-DRB1*0401 dataset: (A) at the N-terminus and (B) at the C-terminus. 95% boundaries of permutation distributions are shown (dotted) with fits to the original, non-shuffled data (solid).

reconcile register predictions. In the case of DRB1*0401, all three modifications resulted in the same magnitude of increase in performance. Finally in the case of DRB1*1501 only an application of both the regression fit (Mod. 1) and the equilibrium-based formula (Mod. 2) resulted in the greatest increase in performance. Differences in which modifications resulted in the greatest increase in performance may be suggestive of allele- or dataset-specific mechanisms behind the length-affinity relationships.

We also incorporated peptide length into the TEPITOPE/ProPred algorithm (Sturniolo et al., 1999) and without re-deriving the pocket-specific matrices that define that algorithm found that increases in performance could be obtained by use of the equilibrium-based formula alone (Table 3). Typically in applications of TEPITOPE/ProPred to MHC class II, predictions on multiple registers are reconciled by taking the highest scoring register to be representative of the whole peptide (Brusic et al., 1998; Nielsen et al., 2004; Murugan and Dai, 2005). We therefore used this rule to generate baseline predictions against which we could compare the performance of the equilibrium-based formula. Applying the formula for register shifting increased algorithm performance for all three datasets examined.

We also investigated whether our modifications might be applied to alleles for which fewer data exist. In analyzing the data for two other alleles, DRB1*0404 and DRB1*0405, we found no significant non-linearities in regression fits of length versus affinity (Supplementary Data). Consistent with the results of these fittings, we observed no increase in performance after applying either
Table 3. Binding prediction accuracy of ProPred algorithm for different MHC class II alleles when peptide length was incorporated

<table>
<thead>
<tr>
<th>Combination rule</th>
<th>DRB1*0101</th>
<th>DRB1*0401</th>
<th>DRB1*1501</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProPred:</td>
<td>0.685</td>
<td>0.741</td>
<td>0.669</td>
</tr>
<tr>
<td>Highest scoring register</td>
<td>0.667*</td>
<td>0.754*</td>
<td>0.635*</td>
</tr>
<tr>
<td>Mod. 2: equilibrium formula</td>
<td>0.702</td>
<td>0.764</td>
<td>0.680</td>
</tr>
</tbody>
</table>

Matrices were obtained from the ProPred website and used to calculate a score for each register within a peptide. To each score the approximate affinity of an all-alanine 9mer to MHC was added (ΔIC50 = 6.169; Doychinova and Flower, 2001). ΔIC50 scores between predicted and experimentally determined IC50 are shown, using a threshold of 500 nM (Sette et al., 1994) to distinguish binding from non-binding peptides.

*Highest ProPred-predicted scores from all eligible registers were used as baseline predictions following recent precedents (Brusic et al., 1998; Nielsen et al., 2004; Murugan and Dai, 2005).

Mod. 1 or Alt. 1 to the ISC-PLS algorithm when training sets were derived from these datasets (Supplementary Data). An increase in performance was observed, however, for the larger of the two datasets using Mod. 2 (Supplementary Data). These results suggest that our proposed modifications, like matrix-based prediction algorithms, are subject to limitations based on the size of the training set.

4 DISCUSSION AND CONCLUSION

Information is typically lost during the prediction of peptide–MHC class II binding because most algorithms focus exclusively on 9mers within the peptide. An underlying assumption is that properties of the parent peptides that cannot be captured in their 9mers are irrelevant. This assumption may be true for MHC class I binding which involves peptides of nine amino acids almost exclusively but may not be true for MHC class II binding. Peptides that bind MHC class II are variable in length and may contain segments that extend past the ends of the peptide-binding groove, also known as peptide-flanking residues or PFR (Brown et al., 1993). PFR–MHC interactions may in turn affect peptide–MHC binding in a manner that is consistent and useful to prediction. Longer peptides also allow for register shifting, i.e. the ability of peptides to bind MHC using different core 9mers. PFR–MHC interactions and register shifting represent two possible mechanisms by which variability in peptide length affects affinity to MHC class II.

In this study we found that non-linear relationships exist between peptide length and peptide–MHC class II binding affinity in a number of aggregate datasets available online. When these nonlinearities are examined in more detail, they were found to be significant at several lengths, suggesting some lengths were more favorable for binding than others. This is consistent with the data from a number of experimental studies (Malcherek et al., 1994; Vogt et al., 1994; Bartness et al., 1999; Fleckenstein et al., 1999). In these studies affinity was generally found to increase with length up to the longest lengths examined, typically between 15 and 17 amino acids.

In our simulations register shifting was found to be one mechanism that could account for the direct relationship between length and binding affinity. However, our analysis of aggregate datasets suggests that additional mechanisms also contribute to the effect of length on affinity. For example, register shifting alone cannot explain why certain lengths at the N- and C-termini are advantageous or disadvantageous for binding DRB1*0401. In this case other mechanisms such as hypothesized PFR–MHC interactions that are either attractive or repulsive may also be playing a role (Sercarz and Maveraakis, 2003).

Incorporating peptide length into existing binding prediction algorithms by one or more of our modifications consistently improved performance for multiple MHC class II alleles. Three modifications were used—one at the level of the training set data (Mod. 1), another within the algorithm itself (Alt. 1), and the last after 9mer predictions were generated (Mod. 2)—and all resulted in performance gains over reference algorithms ISC–PLS and TEPITOPE/ProPred. Baseline AROC scores for different algorithms varied between 0.57 and 0.73. By comparison AROC scores for modified algorithms varied between 0.68 and 0.77, consistent with the range of scores listed in MHCBench (http://www.imtech.res.in/raghava/mhcbench/). The modification resulting in the largest performance increase differed by allele, and this may in part reflect differences in the mechanisms by which length affects affinity. For DRB1*0401, for example, using the formula for register shifting resulted in performance gains that were statistically indistinguishable from those obtained using other modifications. For DRB1*0101, however, modifications based on regression modeling resulted in significantly greater performance increases. These data therefore support roles for both register shifting and other mechanisms.

Previous studies have provided indirect evidence that accounting for variability in peptide length could improve prediction. Godkin et al. (1998), for example, found that matrices based on 15mers generally outperformed matrices based on shorter lengths, showing the usefulness of considering information outside of the 9mer core. Likewise, Bui et al. (2005) have proposed deriving a separate matrix for each length of peptide (Bui et al., 2005). Despite the suggestion that explicit consideration of peptide length could improve binding prediction (McFarland and Beeson, 2002), to our knowledge no previous study has implemented this idea. Our results affirm the use of peptide length in binding prediction. In addition our modifications are sufficiently general that they could be incorporated into other current algorithms based on scoring 9mers.

Thus far experimental evidence of either register shifting or PFR–MHC interactions has involved only a small sampling of MHC class II alleles and been of indeterminate generality. For example, register shifting has been demonstrated to occur with alleles I-A^d and I-A^b in mice and DR2 in humans (McFarland et al., 1999; Li et al., 2000; Seamons et al., 2003; Bankovich et al., 2004). Solved structures exist for a somewhat wider array of alleles, including I-A^d and I-A^b in mice and DR1, DR3 and DR4 in humans (see McFarland and Beeson, 2002 for a review). Although these structures show the presence of PFRs in peptide–MHC class II complexes, they fail to capture the dynamics of either register shifting or PFR–MHC interactions.

Our analysis of regression fits to different aggregate binding datasets suggests that longer PFRs (i.e. in peptides longer than ~16 amino acids) may generally be deleterious to binding. At the same time, however, PFRs of a certain minimum length increase the probability of a peptide having multiple binding registers which, our simulations show, increases overall binding affinity. An optimal peptide length for binding each MHC class II variant may therefore exist. Further computational analysis of aggregate datasets may provide a complement to more direct, observation-based studies.
in continuing to elucidate the role of peptide length in MHC class II binding. In addition these findings may be of use to the design of peptide vaccines which often comprise only short segments of disease-relevant protein antigens (Larche and Wraith, 2005). Including PFRs of optimal lengths may help to ensure efficacious binding to MHC.

ACKNOWLEDGEMENTS

We thank Irini Doytchinova, Darren Flower, Pandjassaram Kanguane and Chao-Yie Yang for helpful discussions. This work was supported by NIH grants HL68526 and LM009027 and a University of Michigan Rackham Predoctoral Fellowship to STC.

Conflict of Interest: none declared.

REFERENCES