BIOINFORMATICS

Vol. 22 no. 14 2006, pages e227–e235
doi:10.1093/bioinformatics/btl255

Learning MHC I—peptide binding

Nebojsa Jojic[1,†,*], Manuel Reyes-Gomez[1,†], David Heckerman[1], Carl Kadie[1] and Ora Schueler-Furman[2]

[1]Microsoft Research, Redmond WA 98052, USA and [2]Dept. of Molecular Genetics and Biotechnology Hadassah Medical School, The Hebrew University of Jerusalem, Israel

ABSTRACT

Motivation and results: Motivated by the ability of a simple threading approach to predict MHC I—peptide binding, we developed a new and improved structure-based model for which parameters can be estimated from additional sources of data about MHC-peptide binding. In addition to the known 3D structures of a small number of MHC-peptide complexes that were used in the original threading approach, we included three other sources of information on peptide-MHC binding: (1) MHC class I sequences; (2) known binding energies for a large number of MHC-peptide complexes; and (3) an even larger binary dataset that contains information about strong binders (epitopes) and non-binders (peptides that have a low affinity for a particular MHC molecule). Our model significantly outperforms the standard threading approach in binding energy prediction. In our approach, which we call adaptive double threading, the parameters of the threading model are learnable, and both MHC and peptide sequences can be threaded onto structures of other alleles. These two properties make our model appropriate for predicting binding for alleles for which very little data (if any) is available beyond just their sequence, including prediction for alleles for which 3D structures are not available. The ability of our model to generalize beyond the MHC types for which training data is available also separates our approach from epitope prediction methods which treat MHC alleles as symbolic types, rather than biological sequences. We used the trained binding energy predictor to study viral infections in 246 HIV patients from the West Australian cohort, and over 1000 sequences in HIV clade B from Los Alamos National Laboratory database, capturing the course of HIV evolution over the last 20 years. Finally, we illustrate short-, medium-, and long-term adaptation of HIV to the human immune system.

Availability: http://www.research.microsoft.com/~jojic/hlaBinding.html

Contact: jojic@microsoft.com

1 BACKGROUND AND DATASETS

The development of computational methods that predict protein folding and binding is of considerable interest to the scientific community. In addition to furthering our understanding of basic chemical-physical principles that govern the complexity of protein structure, results in this area may also lead to important medical applications. Current research in this area focuses on complex physics-based models using a large number of particles to describe not only the proteins, but also the solvent molecules that surround them.

This paper is motivated by the following observation: Protein binding is at heart of many biological processes which have been heavily studied at a higher level, and so a number of studies have provided indirect sources of information that could be mined to infer unknown parameters of a physics-based binding model. For example, many of the binding configurations bear significant similarities, and therefore the known structures of representative protein complexes can be useful in inferring geometry of binding or binding strength for new proteins. In addition, in some cases, there are publicly available datasets of experimental binding energies (or affinities) for mutated proteins and certain molecules. Other biological experiments are concerned only with the result of a binding process within a more complex system, and so their results may provide only binary information (whether or not the proteins of interest bound or not in a specific context). If interpreted jointly, these diverse sources of data could significantly contribute to our understanding of a system, improve our ability to predict binding partners, and may eventually allow us to manipulate interactions of interest.

Here we focus on one example of such joint data interpretation grounded in a simple physics-based binding model whose purpose is the prediction of the binding energy of peptides to Major Histocompatibility Complex (MHC) class I molecules. MHC class I molecules participate in the detection of foreign proteins expressed within cells. Proteins in the cell are processed to peptides of 8-11 residues length, and some of them are loaded onto MHC molecules which travel to the cell surface and present them to other components of the immune system. In particular, presented peptides may be recognized by cytotoxic T cells, which can destroy the cells deemed to be operating improperly because they present unexpected MHC-peptide complexes. The ''unusual'' complex can be formed as a consequence of a variety of events, such as cell damage, mutation (e.g., cancer), or viral infection, and more recently, organ transplantation.

Due to the importance of this process, it has been experimentally studied in a variety of ways. We describe how we used these studies to train a novel adaptive double threading model of MHC-peptide binding which does not only point out peptides with very low binding energies (good binders, or potential epitopes), but also ranks the peptides with intermediate levels of binding. Adaptivity and double threading make our model appropriate for predicting binding for alleles for which very little data (if any) is available beyond just their sequence, including prediction for alleles for which 3D structures are not available. Armed with this tool, we

*To whom correspondence should be addressed.
†The first two authors contributed equally.

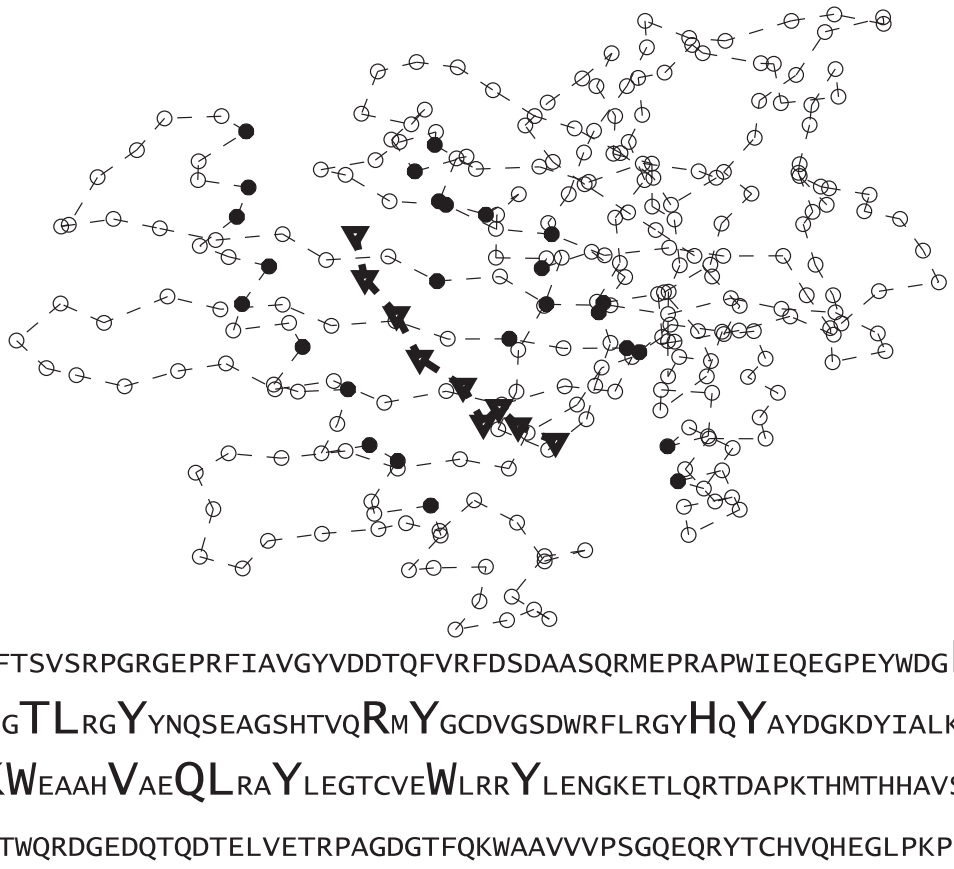

Fig. 1. 3D structure ofMHCA0201 bound to peptide GILGFVFTL (PDB code 1hhi; Madden *et al*., 1993). The centers of the peptide residues are marked in 3D space by triangles and the centers of MHC's residues are marked by circles. Residues in the peptide binding groove of the MHC (i.e. within 4A of the peptide) are marked by filled circles. The lower panel shows the MHC A0201 sequence, with groove sites indicated by increased font size.

are able to demonstrate the effects of immune pressure on HIV sequence evolution within a host and on a population level.

To train the model we can use the following types of data:

MHC I sequence data. We focus here on human MHC class I molecules: they are encoded in three regions of the human genome, labeled A, B, and C. These regions are among the most variable in the human genome, with dozens to hundreds of different MHC variants in each region. Since each individual inherits genes from two parents, each of us has at least three and up to six different MHC molecules operating in our cells. As different MHC molecules typically bind to different peptides, it has been very important to immunologists to classify MHC types. For example, organ transplant recipients may reject organs of donors with non-matching MHC types, as the cells in these organs will present MHC-peptide complexes that are new to the immune system of the recipient. Modern MHC typing is performed by sequencing, and the sequence data for all known MHC variants is available.

MHC-peptide complex structure data. The importance of peptide-MHC interactions to the immune response has motivated crystallographers to solve the structures of a range of different MHC-peptide complexes. In several cases, the structural variability of a specific MHC allele could be assessed by solving the structure of this allele when bound to a range of different peptides. An example of such a MHC-peptide complex structure and sequence is given in Fig. 1.

The present study is based on a dataset of 37 different MHC-peptide complex structures that was also used by (Furman *et al*., 2000). The structures were downloaded from the RCSB protein data bank (Berman *et al*., 2000: http://www.rcsb.org/pdb/).

MHC-peptide binding affinities. The relative binding ability of different peptides to a specific MHC molecule can be directly assessed by competition experiments. The peptide concentration that leads to 50% inhibition of a standard peptide, IC50, is measured, and the relative binding energy can be described as the ratio between the IC50 of the standard peptide and that of a test peptide (Sette *et al*., 1994). The result of such experiments is a set of *relative* binding energies (negative logarithms of the relative concentrations), for different MHC-peptide combinations. This study used a dataset of 870 different combinations from Furman *et al*., 2000, which capture a large range of different binding energies, as discussed in Section 3.

Known good binders (epitopes) and nonbinders. Viral or cancer epitopes, and other excellent binders are often discovered by EliSPOT essays that capture the reaction between T-cells of exposed patients with peptides containing suspected epitopes. Other peptides are known to evoke only very low reactivity in binding essays. Large databases of known epitopes, as well as nonbinders, for various MHC molecules are publicly available. We have used the SYFPEITHI database (Rammensee *et al*., 1999: http://www.syfpeithi.de/), Los Alamos National Laboratory

HIV Database (http://www.hiv.lanl.gov/) and the MHCBN Database (Bhasin *et al*., 2003: www.imtech.res.in/raghava/mhcbn/third.html). These databases provided us with ''binary'' energy data for many peptides (by simply indicating if a peptide is a strong binder, or a non-binder with respect to a particular MHC type).

2 THE BINDING ENERGY MODEL

Our binding energy model is based on the geometry of MHC-peptide complexes, and is motivated by the *threading* approach (Jones *et al*., 1992). Its implementation in (Furman *et al*., 2000) is here augmented by including learnable parameters. We demonstrate that these parameters can be estimated by using all of the described data jointly.

In general, threading aims at evaluating the compatibility of a certain protein sequence with a certain protein structure: The sequence is threaded onto the structure, and a list of contacting amino acid pairs is extracted, based on contacting residue positions (defined as residues in close proximity, e.g. that have at least one pair of atoms less than 4.5A apart). In order to allow estimation of the binding energy of any peptide with an MHC molecule whose structure in complex with some other peptide is known, we assume that the proximity pattern to the peptide in the groove does not change dramatically with the peptide's sequence.

Assuming that energy is additive, and that the pairwise potentials depend only on the amino acids themselves—and not on their context in the molecule—the energy becomes a sum of pairwise potentials taken from a symmetric 20×20 matrix of pairwise potentials between amino acids. These parameters are computed based on the aminoacid binding physics, or from statistical analyses of amino acid pair contact preferences in large sets of available protein structures. Several sets of pairwise potentials have been described in the literature, each derived in a different way (for review see Melo *et al*., 2002). Obviously, the choice of pairwise potential matrix can dramatically alter performance of the energy predictor (Furman *et al*., 2000).

The advantage of the original threading-based approach lies in its independence on binding data. In this approach, as long as a structure of the MHC-peptide complex is available, an allele can in principle be characterized without the need of multiple tedious binding experiments. However, the very same data used for verification of the original threading approach could be used to refine it in a data-driven way. Furthermore, over the last few years a large amount of additional data about binding peptides has been produced for a range of different alleles. Combining the threading approach with a machine learning philosophy of fitting to data, we show that it is possible to estimate a pairwise potential matrix and also learn additional parameters that make the results less sensitive to approximations made in the original threading model.

In order to motivate the parameterization of our model, we start with a slightly more general mathematical definition of the basic threading model, which predicts the binding energy E as a function of the structural template m, the MHC sequence $\mathbf{s}$, and the peptide sequence $\mathbf{e}$, as

$$E(m,\mathbf{s},\mathbf{e}) \approx \sum_i \sum_j \phi_{\mathbf{s}_i,\mathbf{e}_j} h(d^m_{i,j}), \tag{1}$$

where i and j are sequence positions in the MHC molecule and the peptide respectively, ϕ are the pairwise potentials discussed[1], and $d^m_{i,j}$ is the distance between the i-th MHC residue and the j-the peptide residue in the m-th 3D structure (as we have different structures for different molecules)[2]. Finally, in the threading approach, function h is simply the step function

$$h(d) = \begin{cases} 1, & d \le d_{thr} \\ 0, & d > d_{thr} \end{cases}. \tag{2}$$

The threading model is based on the rational approach, which uses physical models to predict the binding energy for a new MHC-peptide complex when a crystal structure (indexed by m) and the sequence of both the MHC molecule and the peptide ($\mathbf{s}$ and $\mathbf{e}$, respectively) are given.

In order to use the abundant direct or indirect information about binding to improve the threading model, and to allow reliable predictions even in the absence of the known structural templates, we make a few adjustments to this model. First, we consider parameters ϕ as hidden variables, with the previously published pairwise potential matrix serving as a basis of the prior on ϕ to avoid over training. Second, instead of the step function, we use a soft step (sigmoid),

$$h(d) = \frac{1}{1+e^{-a(d-d_{thr})}} \tag{3}$$

increasing the robustness of the predictor to slight variations in the geometry of the structural model (residue pairs with a close-to-threshold distance might suddenly be turned off if the distance is only slightly above the threshold). The parameters of h can be learned, setting the threshold (d_{thr}) and the softness a of the step. Finally, we add weights w^m_j to allow our model to adapt to the errors introduced by the strong assumption that all close residue pairs (as defined by h) will contribute to the energy independently. A stringent threshold parameter will produce a very sparse set of pairs i, j that contribute to the energy, and in this case each pair can be assumed to contribute independently. However, many important interactions might be missed by applying a stringent threshold. A loose threshold on the other hand will result in the inclusion of non-relevant residue pairs (amino acid pairs that in fact do not significantly interact in the structure). Including these additional contributions into the energy function might blur the signal. In addition, residues will likely interact simultaneously with several neighbors, which could question the additive model.

In order to address this problem, we add MHC-specific weights $w^m_{i,j}$ to the threading equation with altered function h:

$$E(m,\mathbf{s},\mathbf{e}) \approx \sum_i \sum_j w^m_{i,j} \phi_{\mathbf{s}_i,\mathbf{e}_j} h(d^m_{i,j}), \tag{4}$$

For these weights we use a Gaussian prior favoring $w^m_{i,j} = 1$. The model is designed so that it reduces to standard threading when priors are strong enough to ignore the dataset of energies E for various peptide and MHC combinations. However, in our experiments the priors are left weak enough so that the data can dominate the learning process, and the priors simply serve as a measure against over-fitting.

Note that several variants of the model can be derived from this basic form, depending on how many parameters we want to

[1] ϕ is a 20×20 matrix of potentials for different pairs of amino acids.

[2] In fact, the 3D structure of a MHC-peptide complex may vary slightly for different peptides, in which case a consensus distance is used. See Furman *et al*., 2000 for details

train and how much data we have. For instance, the weights w can be shared across all MHC types, leaving only the sequence s of a molecule to define its behaviour in the model. Furthermore, a single consensus geometry for all types can be used, removing index m from the model completely. Another way of reducing $E(m, \mathbf{s}, \mathbf{e})$ to simply $E(\mathbf{s}, \mathbf{e})$, is to treat structure index m as a hidden variable and infer it with help of proper priors, sequence similarity, or in cross-validation during training. The simpler variants, more capable of generalization, are especially interesting when the goal is prediction of binding energies for new alleles for which no binding data is available. In fact, all forms of this model are based on a physics-based approach which primarily uses the protein sequences into account when evaluating the binding affinity. The MHC type is not primarily captured by its symbolic name (e.g., A0201), but by its sequence as shown in Fig. 1. Thus, applications beyond epitope or energy prediction for each molecule in isolation are possible, e.g., studying the effect of MHC mutations on the efficacy of the immune system in different infections.

We assume Gaussian noise in the energy data (perhaps there are better models motivated by the physics of the process), and we fit the model by standard variational learning, which is needed because of the bilinear dependence of E on ϕ and w. As the optimization criterion becomes quadratic (ignoring parameters of h for a moment), the variational inference essentially iterates between a linear regression to find ϕ variables (penalized appropriately by the prior) and a regression that estimates weights w, again taking into account the Gaussian prior favoring $w_{i,j} = 1$. Refinement of step function parameters (d_{thr} and a) is interleaved with these two steps. For MHC molecules for which we do not have the 3D structure on which to define $d_{i,j}$, we use the available structure of a related MHC molecule with the highest sequence similarity. This is motivated by the fact that across all MHC molecules, the geometry of the groove (i.e. the residues that are in proximity of the peptide) does not change significantly, even when the amino acid content is significantly different.[3] If we view this model as generative, then m can be considered as a hidden variable influencing the sequence $\mathbf{s}$, thus allowing inference of m from $\mathbf{s}$. In principle, in inference of m, both $\mathbf{s}$ and d should be taken into account, but we avoided that in our initial experiments for simplicity. The prior parameters can be tuned through cross validation on the training set.

The dataset of binding energies can be directly used in training our model, but the dataset of known good binders and non-binders requires a treatment of missing energy values. We simply used the lowest binding energy in the binding energy dataset for good binders (epitopes), and similarly, the highest binding energy for the non-binders. Alternatively, the spread between the binding energies of the binders and non-binders can be maximized, or a cost function different than quadratic can be used which punishes bad but not good binding energies for good binders, and does the opposite for non-binders.

It is important to note that we fit all MHC-peptide complexes together, as ϕ parameters are shared across all data. The w_{ij}^m parameters, on the other hand, are specific to a particular MHC geometry (obtained by crystallography). Joint training helps energy prediction for individual MHC types (training only on a limited number of MHC molecules degrades the performance of the predictor on the test data even for the MHC molecules *included* in training). Also note that the model is set up so that it would provide an energy prediction after training even for MHC molecules for which no data other than their sequence is given. The ϕ parameters estimated from the existing data would then be used together with uniform weights $w_{ij}^m = 1$, as dictated by the prior.

[3]In fact, different MHC molecules align well and only 10% of the residues show sequence variability. The ''groove'' residues, however, are the most variable with about 30% of them showing sequence variability, even between two molecules coded in the same region of the genome (A, B or C).

Table 1. Summary of the IC50 dataset used in Sect. 3

	Good binders	Intermediate	Non binders
A0201, peptide length 9	62	254	202
A0201, peptide length 10	27	138	100
A6801, peptide length 9	21	74	35
B2709, peptide length 9	11	11	44

3 MODEL PERFORMANCE ON DIFFERENT TYPES OF DATA

In this section, we empirically illustrate how the model behaves in different situations, such as the usage of binary and/or continuous energy data, with different training set sizes and MHC compositions.

The experimental binding energies (or equivalently IC50 ratios, whose negative log corresponds to energy) for peptides in the set used in this section covered a large range, with only some of the peptides having very low energies (epitopes). To illustrate, we divide peptides into three categories: good binders (IC50 ratio >0.1), non-binders (IC50 ratio <0.0001), and intermediate binders with values in between, as suggested by Furman *et al*., 2000. Table 1 summarizes the data in terms of the MHC molecules, peptide lengths and the binding strength.

In order to compare our method to standard threading, we report the performance of our predictor in terms of peptide ranking measured by Spearman correlation factor, as proposed by Furman *et al*., 2000. This measure varies between -1 and 1, with values close to one indicating that sorting the peptides by their predicted energies produces a similar ranking as sorting by the experimentally measured energies. In a first step, we verified that the numbers obtained by the original threading approach (Furman *et al*., 2000) could be reproduced. In contrast to the threading approach, the method presented here requires training, and for this purpose, the data was divided 100 times into random training/testing partitions (70% for used for training, with the data distribution for both sets kept similar to the above table), and we report the average performance, as well as the variance across the experiments.[4]

Table 2 indicates that our model outperforms the threading model when the direct and indirect information about MHC binding is used to train the model.

Note that for our model the potentially most influential type of data are binding energy measurements (i.e. IC50 values), but this

[4]Threading approach, on the other hand, is rational, not data-driven and so it uses no training data and provides a single number as an output.

Table 2. Comparison of the standard threading and the trained bilinear model

	Threading	Bilinear model	Standard deviation
A0201, 9mers	0.57	0.78	0.03
A0201, 10mers	0.61	0.82	0.03
A6801, 9mers	0.20	0.67	0.13
B2705, 9mers	0.39	0.71	0.09

Table 3. The ability to predict binding for one type by training on other two (transfer)

	Full transfer	Partial transfer
A0201	NA	0.6067 (196 + 869)
A6801	0.23 (584)	0.2974 (584 + 869)
B2705	0.33 (648)	0.5958 (648 + 869)

Full transfer refers to the use all the available training data (continuous and binary) for two MHC types and predicting binding on the third based on its sequence. Partial transfer refers to using all available data for two types as well as the binary energies (but not continuous) of the third type to predict binding energies in the test set for the third type. The results are quantified in terms of Spearman correlation factor between predicted and true binding energies. The numbers in parenthesis are the numbers of training samples (continuous+binary) in different experiments. Full transfer for A0201 could not be performed as removing all A0201 data did not leave enough data for training. See Sect. 5 for results on larger datasets.

kind of data is scarce and is not available for many protein binding problems (but see Sect. 5 about the recent availability of this data for some MHC types). It is therefore of interest to investigate whether the present approach could also be applied to MHC types not experimentally tested in this way, by using information from related, experimentally scrutinized alleles. For this purpose, we evaluate the ability of our model to predict binding energies when some types of training data are not available for MHC types of interest. For each of three MHC types (A0201, A6801 and B2705), two models of nonamer binding were trained: the first using only the experimental binding energies for the remaining two MHC molecules (simulating the situation where the peptide binding to a new MHC allele is modeled), and the second using both the experimental binding energies for the remaining two MHC molecules and 869 binary energies for all three MHC types (simulating the situation where binary data is available for the allele of interest, e.g., through related research, such as epitope discovery, or tracking evolution of a pathogen; but the direct IC50 experiments are not available). In all cases, of course, the test set of known binding energies, was unrelated to the training data. These experiments are summarized in Table 3 and they illustrate how much the peptide preference of a particular allele can be characterized by including binding data for other MHC alleles.

As can be seen, without the information about the specific allele in the training set (column 1), the performance is reduced to values similar to the original threading approach, highlighting the significant contribution of this source of information (compare to Table 2). Note that this experiment could not be performed for A0201 due to insufficient data (around 100 examples, whereas just the number of parameters in the potential matrix is over 200). On the other hand, addition of binary energies from the alleles significantly improved the prediction (see column 2), indicating good generalization capabilities of the model. It is important to note that this experiment was performed on a small dataset in order to study the effects of prior knowledge (3D structure, MHC sequence, and threading model) as well as the value of binary data. In the next section, we revisit the issue of predicting binding for an allele based only on its sequence and the IC50 data for other alleles, but this time using much more data that recently became available.

In order to further evaluate the performance of our method on the data for which only binary energies are known, we used the whole set of binding energies in Table 1, all available 3D structures (for inference of m, when the structure of an MHC molecule is not known), and some of the binary data for training, leaving the rest of the binary data for testing. Again, the training and testing sets are chosen randomly 10 times, and both average performance and the standard deviation are reported. The training set spanned 9 MHC types (A0201, A6801, B2705, A1101, B3501, B5301, A0301, B4402, and B0702), with peptides of lengths 9–10. Since both threading and our method output binding energy, and not a binary decision, we compared the two in terms of ROC curves obtained by varying the good-binder (or epitope) threshold and measuring the number of false positives and false negatives. Our method again significantly outperformed threading (some examples are in Fig. 2), and produced results almost as good as the recently published state of the art in (binary) epitope prediction[5] (Heckerman *et al*., 2006) (more figures available at *www.research.microsoft.com/∼jojic/hlaBinding.html*). Note that for A0301 and B0702 we did not have crystal structures, and yet, our adaptive double threading approach was able to adequately predict peptide binding based on the known sequence of the allele, and a structure of a related allele. Additional examples of predictions based on structures of related alleles, compared to predictions based on the actual crystal structure are available at the above web site.

While the results in this section indicate that the use of binary data is justified, we should point out the important caveat. The epitope data in literature comes form different sources, and some ways of experimentally discovering epitopes do not capture only MHC binding but also other processes that lead to immune reaction (e.g., cleavage and T-cell binding). This means that any tunable model, including ours, when trained on lots of binary data, may capture some of these other effects, becoming better at predicting known epitopes, but worse in predicting strictly MHC binding. At the same time, the constrains in the model structure make our model more suited to modeling IC50-derived energies, then to general purpose classification, and may thus limit its performance in binary epitope classification, when this classification includes factors other than MHC-peptide binding.

For example, when we trained a recently published epitope predictor (Heckerman *et al*., 2006) on binary data only, we find that this method produces good binary classification results, but without significant correlation of the epitope probabilities with true binding energies for intermediate binders in the test set. On the other hand, the model presented here when trained on the same binary data, still recovers peptide ranking for intermediate binders with statistical significance, but with much less accuracy than is the

[5]Epitope prediction algorithms specialize on binary classification and usually do not predict well the quality of binding for intermediate binders.

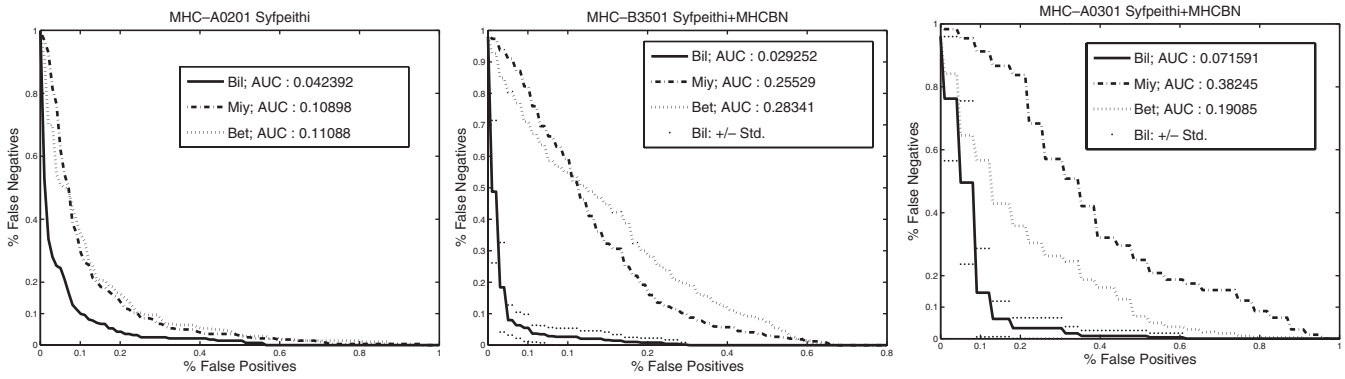

Fig. 2. ROC curves produced by varying the epitope binding energy threshold and computing the number of false positives and false negatives on the SYFPEITHI database. Note that for B3501 only binary energy data was available, while for A0301 no crystal structure was available (The known structure of the MHC molecule with the highest sequence similarity was used, as described in the text.) For the standard threading approach, we used two previously proposed matrices, labeled 'bet' and 'miy' (Furman *et al*., 2000, Betancourt *et al*., 1999, Miyazawa *et al*., 1985), while for our trained bilinear model ('bil'), we also provide standard deviation curves computed over different data splits into training and testing. Over all MHC types, the area under the ROC curve was between 2.5 to 15 times lower for the bilinear model than for either of the threading models.

case when the model is trained on IC50 data. We found that on the binary classification task on epitope data, in comparison to Heckerman *et al*., 2006, the bilinear model suffers a 1% increase in false positive and false positive rates due to its bias towards capturing only the MHC-peptide binding component of being an epitope. This indicates that the tunability of the model makes it possible to tradeoff its energy predictions with its epitope classification capabilities, but it the model may not necessarily extract the single effect (MHC-peptide binding) fully. We are now investigating ways to separate the effects of MHC binding from other effects in binary data and train a combined predictor.

Next, in Sect. 4 we investigate if the predictive power of our model can be used to study the global changes the immune system causes in pathogens, once the model is trained on all available binary and continuous data. Then, in section 5 we evaluate our model on the recently published dataset of IC50 energies.

4 VIRUSES EVOLVE TO MODULATE THEIR BINDING TO MHC MOLECULES

MHC A0201 is one of the most frequent MHC types, especially in the Western world. Using our model, we computed the average binding energy of all HIV 9mers (taking each overlapping peptide from all proteins of the current consensus sequence for clade B) and found it to be equal to 9.74 (the units are of no importance, only the ranking of energies matters). On the other hand, the average binding energy in a randomized HIV is 9.3. The randomized HIV contained the same set of proteins with same lengths but with random aminoacid sequences. The difference in average binding energies has a very strong statistical significance ($p < 10^{-5}$ based on 50 different randomizations), and can be explained by viral evolution—higher average binding energy translates into a smaller total number of presented peptides which trigger immune reaction. Similar patterns should be expected from other viruses, variable enough to use mutation as an escape mechanism. (It is possible that less variable viruses, evolving over a very long time, may still have the same property, and we are planning on investigating this next.)

It has been shown previously that some HIV mutations correlate (weakly) with the MHC types of the host (Moore *et al*., 2002). The binding energy estimators that we developed allow us now to begin to explain these correlations. In Fig. 3, we demonstrate significant correlation ($p < 0.05$) between the average A0201 binding energy and the viral load in the A0201 positive patients from the WA cohort obtained by Moore *et al*., 2002 (as would be expected, in A0201 negative patients we do *not* find any correlation).

For each chronically infected and untreated A0201 positive patient in the cohort, we plot the patient's viral load v.s the sum of 9mer and 10mer average binding energies for A0201 (each patient's HIV was sequenced providing a source of 9mers and 10mers for this computation). The virus whose peptides bind well to a particular MHC molecule is typically under strong immune pressure in patients with this MHC type, and is forced to mutate away from its fittest form towards a form that binds less well to MHC. But, as HIV damages the immune system, the high viral load in the figure indicates a removal of the pressure to escape A0201 binding. Therefore, the negative trend in the figure could be explained by reversion of the viral sequence towards the wild type with higher replicative fitness and lower adaptation to A0201, in patients whose immune system is starting to fail, but other alternative explanations are possible (such as that the intermediate binders in the sequence, become better binders as that serves some purpose to the virus, which after all, infects the immune system). We are investigating these trends further experimentally.

Finally, in Fig. 3, we also track the average binding energy of MHC A0201 to HIV peptides over the last 23 years. The sequences of various proteins from over 1000 patients were obtained from the Los Alamos National Laboratory database. To smooth out the sampling density over time, all sequences were grouped into 3 year time intervals: 1982–1984, 1985–1987, . . . , 2003–2005. The apparent upward trend is statistically weak, but may still indicate that HIV as a population is adapting to the immune systems of the host population. Recently, a trend of HIV fitness attenuation has also been indicated (Arien *et al*., 2005) which would be consistent with this. In order to find out if the trend of modulation

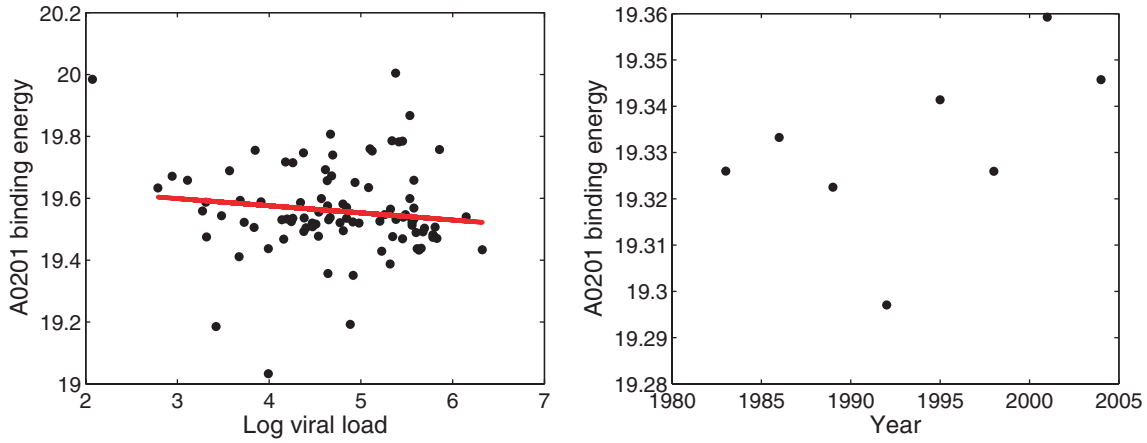

Fig. 3. HIV-MHC A0201 binding energy trends as a function of viral load in individual patients, and the time of sampling.

of MHC binding is significant, we plan to take all MHC alleles into account.

5 PERFORMANCE OF THE MHC-PEPTIDE BINDING MODEL ON A RECENTLY PUBLISHED LARGE DATASET OF IC50 VALUES

Recently, Peters *et al*., 2006 have assembled a large database of experimentally acquired binding energies for a range of MHC molecules. The experiments were performed at the Sette and Buus labs, and Peters *et al*., 2006 go on to also test a variety of published algorithms that can predict MHC-peptide binding. Their published dataset may prove to be the most useful community resource for studying MHC-peptide binding so far. In particular, a very useful feature of their dataset is that the data has been acquired relatively uniformly, with some potential variability due to the fact that the experiments were performed in two different labs.[6] The data consists entirely of IC50 values for 49 different MHC class I alleles, both human and animal. Here, we focus on human alleles from this dataset (the total of 35 A and B alleles), for which the total of 29,371 IC50 values for different MHC-nonamer combinations were tested. Peters *et al*., 2006 show that among the tools available in their labs and on the web, the best performing tool is a neural network proposed by Nielsen *et al*., 2003.

Such a rich dataset and comprehensive comparisons provide several opportunities for additional evaluations of our approach. In particular, as discussed in Sect. 3, the use of heterogeneous data improves overall results, but may skew our model away from solely predicting MHC-binding energy and towards partially capturing additional effects present in the binary experimental data. Nielsen *et al*., 2003 consists solely of the continuous binding energy measurements, and can thus be used to better asses some of the interesting properties of our model. In addition, the amount of data reduces the effects of regularization priors.

[6]In fact, the authors provide a brief analysis of this potential source of error in the paper.

5.1 Predicting binding for new alleles: adaptive double threading

First, we asses the ability of our model to predict binding for a new MHC allele, for which no other data is available but its sequence. This is possible as our trained model performs double threading: not only does it thread a peptide onto the known structure and content of the particular MHC molecule, but it can also use another MHC molecule's structure to thread the new allele's sequence and the peptide on it. When the 3D structure for the allele of interest is available, it is used, but otherwise the best structure from the available database of structures is used (inference of variable m in the model).

To illustrate this empirically, we first focused on the well studied allele A0201, and trained our model on three data subsets and tested the trained models on A0201 test sets in five-fold cross-validation. The first model was trained on IC50 training data for all 35 available molecules in the dataset. The second model was trained on all the data for 34 molecules, but no data whatsoever for A0201 allele, including its 3D structure. The third model is trained on an even more limited dataset which further excluded all A02 types (in this data, A0202, A0203 and A0206), leaving 31 alleles for training. The first model, which was exposed to around 2400 binding energies for A0201 in each fold, achieved the Spearman correlation factor of 0.82, which is comparable to the best result (0.83) reported in Peters *et al*., 2006, and better than all other techniques tested there. However, a more interesting observation is that the second model, which had *no* exposure to A0201 data in training, still predicted A0201 with the Spearman correlation at 0.8, which is only slightly lower than that of the first model. It is important to note that the model did not have the A0201 structure available, and so it could not reduce to standard threading. The model chose to use the structure of the most similar available allele by sequence similarity (A1101). Furthermore, the third model, which had no exposure to any of the A02 types in training, nor the A02 3D structures, still predicted binding with Spearman correlation factor of 0.42. All results are strongly significant with p values virtually zero. Similarly, the Spearman correlation factor for A1101 binding prediction goes from 0.79 to 0.61 when all A1101 data is excluded from training, and only

A1101 sequence is fed to the trained model, but not its structure. This illustrates that our model degrades gracefully as the data related to a particular allele is removed from training, as long as the data for other alleles is available. Therefore, the model is leveraging data for multiple alleles in each of its predictions, and can potentially be used to predict binding for new alleles, given only by their sequences.

5.2 Geometry estimation

The known 3D structure of several alleles is the basis of our bilinear model and it affects the predictions through terms $h(d_{ij}^m)$ in (4). For large pairwise distances d_{ij}, these terms will be virtually zero, thus making the appropriate amino acid pairs irrelevant in prediction. Another way of thinking about the effect of these distances is as a way of regularizing the combined set of parameters $u_{ij} = w_{ij}h(d_{ij}^m)$ in the model $E(m, \mathbf{s}, \mathbf{e}) \approx \sum_i \sum_j u_{ij}^m \phi_{\mathbf{s}_i, \mathbf{e}_j}$, in which distance function is merged with the weights for the pairs. But, if enough data is available, this regularization should not have to be so strongly informed by the structure, and could instead be based on usual norm-regularization. Since our model is grounded in physics, so estimated weights u_{ij} may in fact capture the relevant structure: wherever the distances are large the importance of the pair should be low, and thus the inferred weight should be close to zero.

To test this hypothesis, we selected 8 diverse A alleles and trained the model without the step function h, and with regularization of the norm of u. We limited the pairs i, j only to those that involved variable sites on the MHC molecule (as the conserved sites will have no discriminating effect in training). Then, we compared the learned pairwise weights u_{ij} with the appropriate Euclidean distances d_{ij} between allele and peptide residues in the consensus A0201 structure. Indeed, the Spearman correlation factor between the absolute value of the estimated weights u_{ij} and distances d_{ij} in the 3D structure was negative (-0.16), as expected, and the result is statistically significant ($p < 0.05$). Therefore, by training our model, it is possible, at least to a certain extent, to recover relevant parts of the 3D structure of the binding configuration.

We also note that we have experimented with a simple linear version of the model for binary prediction, which learns directly the products $v_{i,j,\mathbf{s}_i,\mathbf{e}_j} = u_{ij}^m \phi_{\mathbf{s}_i,\mathbf{e}_j}$, without constraining the weights to satisfy a bilinear form. Such a model is forced to learn a weight for any combination of amino acids at any pair of positions in the MHC molecule and the peptide, and is thus vastly over-parameterized. Therefore most of the weights should be equal to zero to avoid over-training. However, we have found that, when nonzero weights are selected using a wrapper method (Kohavi *et al*., 1997), the linear model makes binary predictions as well as the bilinear model, and it also tends to choose i, j pairs with small distances for its nonzero weights, thus performing some structure estimation, as well. We are extending these experiments to the non-binary case.

5.3 Comparison to other techniques

We have also trained our model have trained our model on the nonamers for 35 human alleles on the same folds as Peters *et al*., 2006 and compared with the techniques they analyzed in five fold cross validation. These techniques treat each different MHC allele in isolation from other, which means that they tend to get punished for not using all available data when the allele is not supported by a large amount of training data. On the other hand, when a lot of data for an allele is available, these techniques may have an advantage as they do not have to sacrifice performance on one allele in order to better capture the others and generalize.

Our model achieved an overall test Spearman correlation factor of 0.75, in line with the best performer of Peters *et al*., 2006, which was a neural network proposed by Nielsen *et al*., 2003, and whose Spearman correlation factor on this data was 0.76. In terms of binary classification, the Nielsen *et al*., 2003 beats our model in 18 out of 35 alleles in this data, with our method typically outperforming when the available training data for an allele is small, as would be expected given the ability of our model to generalize over different alleles. Both our model and Nielsen *et al*., 2003 seem to outperform all other techniques compared in Peters *et al*., 2006 by a significant margin. It should be noted again, however, that this data consists of a consistently measured IC50 values for different peptides, and for binary classification tests, only the test data is binarized by thresholding.

The full set of comparisons is available at:
http://www.research.microsoft.com/~jojic/hlaBinding.html.

6 CONCLUSIONS

We have introduced a new model of MHC-peptide binding, which rather than focusing on binary classification of epitopes, can be used to estimate a high range of binding energies for high resolution MHC types (four digits, based on MHC sequencing). Both in terms of peptide ranking and binary classification performance, our model significantly outperforms the threading model which was the basis of our bilinear model with hidden variables. In individual allele predictions, our model is comparable to the best among the models in the recent comprehensive study (Peters *et al*., 2006). Furthermore, as the model is physics-based there is a potential for its use in settings where the existing models cannot be used. For example, we demonstrated that we can predict binding for new alleles and infer (to a certain extent) the geometry of the binding configuration from binding energy data. The predictive power of our model enabled us to capture HIV evolution patterns in response to the immune pressure of the human hosts (the threading model alone did not show statistically significant trends). We are now investigating medium- and long-term evolutionary response of other pathogens to the pressure created by the cellular arm of the human immune system. The model can also be used to provide binding energies for epitome learning (Jojic *et al*., 2005).

ACKNOWLEDGEMENTS

We thank Corey Moore, Mina John and Simon Mallal, for providing the data from the WA cohort.

REFERENCES

O. Schueler-Furman, Y. Altuvia, A. Sette and H. Margalit, ''Structure-based prediction of binding peptides to MHC class I molecules:Application to a broad range of MHC alleles,'' Protein Science(2000)9:1838–1846.

A. Sette, J. Sidney, MF. del Guercio, S. Southwood, J. Ruppert, C. Dahlberg, H.M. Grey, R. T. Kubo, ''Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays,'' Mol. Immunol.(1994)31:813–822.

M. Bhasin, H. Singh, and G. Raghava, ''MHCBN: A comprehensive database of MHC binding and non binding peptides,'' Bioinformatics(2003)19:665–666.

H. Rammensee, J. Bachmann, N. Emmerich, O.A. Bachor, and S. Stevanovic ''SYFPEITHI: database for MHC ligands and peptide motifs,'' Immmunogenetics (1999)50:213–219.

C. Moore, M. John, I.R. James, F.T. Christiansen, C.S. Witt, and S.A. Mallal ''Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level,'' Science(2002)296:1439–1443.

C. Yanover, and T. Hertz ''Predicting protein-peptide binding affinity by learning peptide-peptide distance functions,'' Recomb(2005).

K. Arien, R. M. Troyer, Y. Gali, R.L. Colebunders, E. J. Arts, and G. Vanham ''Replicative fitness of historical and recent HIV-1 isolates suggest HIV-1-attenuation over time,'' AIDS(2005)19:1555–1564.

N. Jojic, V. Jojic, B. Frey, C. Meek, and D. Heckerman, ''Modeling genetic diversity with epitomes: Rational design of HIV vaccine cocktails,'' NIPS 2005.

D. R. Madden, D. N. Garboczi, and D. C. Wiley, '' The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2,'' Cell(1993)75:693–708.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, ''The Protein Data Bank,'' Nucleic Acids Research(2000)28:235–242.

D. T. Jones, W. R. Taylor, and J. M. Thornton, ''A new approach to protein fold recognition,'' Nature(1992)358:86–89.

M.R. Betancourt, and D. Thirumalai, ''Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes,'' Protein Sci(1999)8:361–369.

S. Miyazawa, and R. L. Jernigan R. L., ''Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation,'' Macromolecules(1985)18:534–552.

D. Heckerman, C. Kadie, and J. Listgarten, Leveraging Information Across HLA Alleles/Supertypes Improves Epitope Prediction,'' Recomb 2006.

F. Melo, R. Sanches, and A. Sali, ''Statistical potentials for fold assessment,'' Protein Science(2002)11:430–448.

B. Peters, H. H. Bui, S. Frankild, M. Nielsen, C. Lundegaard, *et al*., ''A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules,'' PLoS Computational Biology(2006) In press. DOI: 10.1371/journal.pcbi.0020065.eor.

M. Nielsen, C. Lundegaard, P. Worning, S. L, Lauemoller, K. Lamberth, *et al*., ''Reliable prediction of T-cell epitopes using neural networks with novel sequence representations,'' Protein Science(2003)12: 1007–1017.

R. Kohavi, D. Sommerfield, and J. Dougherty, ''Data Mining using MLC++, a Machine Learning Library in C++,'' International Journal of Artificial Intelligence Tools,(1997)6:537–566.