Title

Prediction of Genomewide Conserved Epitope Profiles of HIV-1: Classifier Choice and Peptide Representation

Authors

Xiao, Yuanyuan
Segal, Mark R

Publication Date

2005-09-16

1 Introduction

The recognition of antigenic peptides by T cells plays an essential role in triggering a successful immune response that degrades detrimental foreign or self proteins. While the sufficient conditions for a peptide to be recognized by T cells are unknown, convincing evidence from experimental work suggests that only peptides that bind to MHC (major histocompatibility complex) molecules (the term Human Leukocyte Antigen, HLA, is used for human) can induce T cell responses. These peptides are called T cell epitopes (Buus et al. (1987)). Identification of such epitopes, in particular those that are conserved across diverse pathogen strains, is fundamental for accelerating vaccine development and improving immunotherapy. Accordingly, considerable effort has been invested in predicting peptide binding to MHC. Much of this effort has been computational/statistical since purely experimental approaches are practically and combinatorially prohibitive (Sung and Simon (2004)).

There are two classes of MHC molecules. MHC class I molecules mostly bind peptides originating from the cytoplasm (endogenous protein or intracellular pathogens) whereas MHC class II molecules bind peptides derived from exogenous antigens. X-ray crystallographic analysis has shown that both classes of MHC molecule have a large peptide binding groove formed by two α-helices overlaying a β sheet. Binding between antigenic peptides and MHC molecules is stabilized by hydrogen bonds formed between MHC protein side-chains and the peptide main chain carbonyl and amide groups, in addition to binding interactions between peptide side-chains and pockets within the MHC peptide-binding groove. For peptides that bind to HLA-A2, an allele of MHC class I in human, the sides of the binding groove typically restrict the size of the bound peptides to be nonamers (9mers). These frequently possess two hydrophobic amino acid anchor residues: Lysine at position 2 and Valine at position 9. There are six pockets (A-F) in the binding groove that can accomodate specific side chains of the binding peptide. Pocket B and pocket F are important for accomodating the primary anchor residues at postion 2 and position 9 (Ruppert et al. (1993); Kubo et al. (1994); Parker et al. (1992); Parker et al. (1994); Falk et al. (1991)). However, these two anchors are necessary, but not sufficient, for high affinity binding with several other position/residue combinations playing important roles.

Experimentally determined structures of class II MHC-peptide complexes are similar to class I complexes. One basic difference, however, is that the binding groove is open at both ends which allows for the protrusion of bound peptides. Consequentially, class II molecules are compatible with longer peptides

(10-mers to 20-mers) than class I, and a specific peptide can bind in different registers (Rudensky et al. (1991); Chicz et al. (1993)). There is one deep pocket in the binding groove that binds the side chain at the primary anchor position, while several shallow pockets accommodate sides chains at secondary anchor positions. The greater flexibility in binding of class II molecules has resulted in lower prediction accuracy compared to class I molecules (Brusic et al. (1997)).

Previously, T-cell epitopes have been identified by binding assays that examine T-cell responses to synthesized peptides generated from target antigens. However, these approaches are too expensive and labour intensive for genome scale study of viruses, bacteria or parasites. The development of prediction rules for peptide binding to MHC molecules can appreciably reduce and refine subsequent experimental work. A wide variety of classification/prediction methods have been used in this context. Initially, simple motif-based rules were proposed that specified the presence of particular amino acids in certain positions – peptides with such motifs being designated as binders. So-called position weight matrices provided an extension of these motif-based methods, whereby a matrix of weights for each amino acid at each peptide position was specified. Both these methods assume independent contributions of individual peptide positions and consequently had limited success in predicting MHC binding. This was especially true for class II molecules, where motifs are less well defined due to the complex nature of its binding. To obtain improved predictive performance a multitude of data mining / classification procedures have been applied to this problem. These include artificial neural networks (ANN; Brusic et al. (1997); Milik et al. (1998)), hidden Markov models (HMM; Mamitsuka (1998)), classification trees (Segal et al. (2004)), discriminant analysis (Mallios (2001)), multivariate regression (Lin et al. (2004)) and support vector machines (SVM; Dönnes and Elofsson (2002); Bhasin and Raghava (2004)).

While some comparative studies of the aforementioned methods have been reported, these have been limited to demonstrating the superiority of one or two particular methods over motif-based approaches. A more comprehensive comparison of techniques in the context of MHC - peptide binding prediction has not been conducted. Part of the purpose of this paper is to report on such a study. We investigate six methods: classification trees, ANNs and SVMs, as well as the more recently devised aggregate/ensemble methods bagging, random forest and boosting. To our knowledge, ensemble classifiers have not been applied in this setting despite demonstrated exceptional performance in benchmark studies (Breiman (1996); Breiman (2001a); Hastie et al. (2001)).

Ensemble approaches synthesize results from weak and/or unstable base classifiers to develop an improved classifier. The mechanism whereby such improvement is realized is transparent for bagging and random forests: averaging over the constituent base classifiers reduces prediction variance. Differing explanations have been proffered for the success of boosting; see Hastie et al. (2001), Breiman (2001a) and Bühlmann and Yu (2003). It is notable that the base classifiers most frequently employed by ensemble methods are classification trees. We have previously contended (Segal et al. (2004); Segal et al. (2001)) that tree-based methods are natural for handling amino acid sequence predictors since they are adept at dealing with multi-level (here 20), unordered categorical covariates. Arguably, it is the inability of select classifiers to readily handle such covariate types that has led, at least in part, to the use of other representations of peptide sequence as described next.

The binding affinity of a given peptide to MHC is dictated by biophysical properties of the amino acids composing the peptide. Accordingly, Milik et al. (1998) employed hand-picked property variables, such as hydrophobicity, polarity, charge and volume as inputs to their ANN models. Lin et al. (2004) chose two QSAR (quantitative structure-activity relationship) molecular structure descriptors, isotropic surface area (ISA; measures the side chain surface area) and electronic charge index (ECI; sum of absolute electronic charge of the side chain), on the basis of their presumed importance to binding. Sung and Simon (2004), adopted a more comprehensive strategy to representing amino acid sequence data via properties. They extracted ten orthogonal factors obtained from 188 physical properties of the 20 amino acids via principal components analysis (Kidera et al. (1985)). These ten factors account for 86% of the total variance. The first four factors (in order of explained variance) load heavily on individual properties: alpha-helix preference, bulk, beta-structure preference, and hydrophobicity. The first factor shows moderate association with ISA, whereas the fourth factor is strongly associated with both ISA and ECI. So, in addition to investigating impact of classifier choice, we also contrast the differing peptide representation schemes by comparing four specifications: amino acid sequence (hereafter denoted AA), the 10 orthogonal biophysical properties (denoted 10-P), the two QSAR descriptors (denoted QSAR) and the combined set of 12 property variables (denoted (10+QSAR)-P).

We undertake these comparisons of predictive performance for both an MHC class I molecule (HLA-A2) and an MHC class II molecule (HLA-DR4), classification being relatively harder for the latter (Brusic et al. (1997); Sung and Simon (2004)). However, we go beyond just summarizing error rates for the respec-

tive model × representation × allele fits. Consideration is given to eliciting variable importance in order to enhance interpretation of results. Further, using classifier predictions for HLA-DR4, we pursued genomewide epitope profiling of HIV-1. We obtained and aligned sequence from 30 diverse HIV-1 strains and assessed the conservation of predicted epitopes across strains. These predictions were further validated against known T-cell epitopes as identified in the literature and the JenPep database (Blythe et al. (2002); Doytchinova and Flower (2003)).

The paper is organized as follows. Section 2 presents details on obtaining and preprocessing the pertinent HLA-A2 and HLA-DR4 binding and non-binding datasets. Descriptions of the six competing classifers and their implementations are also provided. Results obtained from contrasting the six methods and four representation schemes, along with the epitope profile and validation study findings, are presented in Section 3. In broad terms we find that the amino acid representation is as good as any and that ensemble and SVM classifiers perform best. Finally, Section 4 provides some concluding discussion.

2 Materials and Methods

2.1 Data Acquisition

HLA-A2

The public database MHCPEP (`http://wehih.wehi.edu.au/mhcpep/`) was used to build a working pool of binders. MHCPEP is a curated database comprising over 13,000 peptide sequences known to bind MHC molecules (Brusic et al. (1996)). Entries in MHCPEP are compiled from published reports as well as from direct submissions of experimental data. Duplicate entries of HLA-A2 binders were deleted as were non-9-mer peptides, resulting in a total of 485 binders. Since experimentally confirmed non-binders are relatively sparse, we generated 9-mers utilizing amino acid frequencies as in SwissProt (Emmert et al. (1994)) and make the assumption that such randomly generated 9mers are highly unlikely to bind to the MHC. This was reinforced by deleting any generated peptide (from the non-binder pool) that was present in the binding pools, giving rise to a total of 500 non-binders. Figure 1 gives barplots that illustrate specific amino acid frequencies at each 9mer position for our binding and non-binding peptides for HLA-A2.

HLA-DR4

Peptides binding HLA-DR4 in the MHCPEP database have variable lengths. Brusic et al. (1997) used position weight matrices to locate the 9-mer core for each binding sequence. Sung and Simon (2004) employed an iterative algorithm to identify the 9-mer core for each binding sequence by excluding, in each cycle, the subsequence that is farthest from the centroid of the binding pool. In order to facilitate comparisons of the predictive performance of our six classification methods and Sung and Simon's "peptide property model", we used the same 621 9-mer core binding sequences as obtained from the last cycle of their algorithm. Our non-binding pool comprises 600 synthesized 9-mer peptides, generated analogously to the HLA-A2 non-binders described above. Amino acid frequencies at each position for both HLA-DR4 binders and non-binders are contrasted in Figure 2.

2.2 Classification Techniques

We are now confronted with a two class classification problem – discriminate between binding and non-binding peptides on the basis of (one of the representations of) their sequence. We next give a brief description of the classification approaches employed, further detail being available in the respective citations, with Hastie et al. (2001) providing a good overview.

Classification Trees

The classification tree paradigm is described in Breiman et al. (1984). Tree construction involves four components. These are: (1) A set of binary (yes/no) questions, or *splits*, phrased in terms of the covariates that serve to partition the covariate space. A tree structure derives from splitting recursively. The subsamples created by assigning cases according to these splits are termed *nodes*; (2) A *split function*, that can be evaluated for any split of any node, which is used to evaluate competing splits; (3) A means for determining appropriate tree size; and (4) Statistical summaries for the nodes of the tree. The first item deals with handling covariates and so is germane to which sequence representation is adopted. In most implementations (e.g., Therneau and Anderson (1997)) allowable splits are defined as follows: (a) each split depends upon the value of only a *single* covariate; (b) for ordered (continuous or categorical) covariates – *cf properties* – only order preserving splits are permitted; (c) for unordered categorical covariates – *cf amino acids* – all possible splits into disjoint category subsets are theoretically allowed. A computational shortcut (see Breiman et al. (1984), Theorems 4.5 and 9.4) reduce the number of splits actually examined from an impractical $2^{L-1} - 1$ to $L - 1$ for a covariate with

L levels. It is this *exhaustive* handling of *groups* of amino acids that makes classification trees attractive in this setting. Additionally, by the very recursive nature of tree construction, trees are geared to detecting interactions and so can capture between (peptide) position dependencies.

We employed the common strategy of growing an initial tree to maximal depth (by appropriately specifying tuning parameters), so that each terminal node only contains instances from a single class (e.g. binders). This strategy avoids the need to prespecify stopping rules and can uncover unanticipated structure. The likely over-fitting is then remedied by pruning back to an appropriate size as determined either by cross validation or use of an independent test dataset. We implemented classification trees via the `R` package `rpart`.

Despite the abovementioned utility of classification trees with regard to handling sequence-based predictors, these techniques have some general deficiencies. Foremost amongst these is modest prediction performance when compared with more flexible methods, such as ANNs or SVMs. Bagging and random forests, described next, were devised to address these shortcomings.

Bagging and Random Forests

In a series of recent papers, Breiman has demonstrated that consequential gains in classification or prediction accuracy can be achieved by using (large) ensembles of trees, where each tree in the ensemble is grown corresponding to some introduced randomness. Final classifications are obtained by aggregating (plurality voting) over the ensemble, typically using equal weights. Bagging (Breiman (1996)) represents an early example in which each tree is constructed from a bootstrap (Efron and Tibshirani (1993)) sample drawn with replacement from the (training) data. The simple mechanism whereby bagging reduces prediction error (for squared error loss) for unstable predictors, such as trees, is well understood in terms of variance reduction resulting from averaging (Hastie et al. (2001)). Such variance gains can be enhanced by reducing the correlation between the quantities being averaged. It is this principle that motivates random forests, which effect such correlation reduction by a further injection of randomness. Instead of determining the optimal split of a given node of a (constituent) tree by evaluating all allowable splits on all covariates, as is done with single tree methods or bagging (item 2 above), a subset of the covariates drawn at random is used. The size of this subset, m_{try}, constitutes the primary tuning parameter of the random forest procedure. Breiman (2001b) argues that random forests enjoy exceptional prediction accuracy for a wide range of settings of m_{try}. Here, we used ensembles of size 100 and the recommended value of m_{try}, which is the square root of the number of variables. Results were

largely insensitive to varying these quantities. We implemented bagging using the `R` package `ipred` and random forests using standalone FORTRAN software available from `http://www.stat.berkeley.edu/users/breiman/rf.html`.

Boosting

Boosting has enjoyed considerable recent success as an effective "off-the-shelf" classifier. While boosting was originally presented as a procedure that combines outputs from many so-called weak classifiers (learners) to produce an ensemble, it is fundamentally distinct from bagging and random forests (Hastie et al. (2001)). Insights into the basis for boosting's success, including its tendency to avoid overfitting, have been provided by viewing the method as additive modeling (Friedman et al. (2000)) and stagewise functional gradient descent (Friedman (2001), Bühlmann and Yu (2003)). We adopt the AdaBoost formulation (exponential loss) which was one of the earlier proposed boosting algorithms (Freund and Schapire (1997)). After tuning, we used classification trees that have a maximum depth of 4 (root node is counted as depth 0) as the weak learner and 100 iterations. We implemented AdaBoost using custom software and the `R` package `gbm`.

Support Vector Machines

Extensive descriptions of SVMs can be found in Christianini and Shaw-Taylor (2000) and Hastie et al. (2001). A key component of SVM methodology is basis expansion, effected by transforming the input vector (here a sequence representation) into a high dimensional feature space via use of a prescribed kernel. There are some standard choices for the kernel including polynomial, radial basis function (RBF) and sigmoid. Since prior work in the peptide – MHC binding context found that RBF kernels gave optimal performance (Bhasin and Raghava (2004)) and these have been recommended as a good default (Hsu et al. (2003)) we adopt this choice. We determined remaining tuning parameters of the SVM (penalty/cost and width of the RBF kernel) by cross-validation. For the amino acid sequence (AA) representation we employed (19) indicator variables - we obtained these using treatment contrasts (`R` default in package `e1071`) with A (arginine) as the baseline group (also the default). This gave an input vector of length $171 (= 19 \times 9)$ for the 9-mer peptides, while 10-P and (10+QSAR)-P had $90 (= 10 \times 9)$ and $108 (= 12 \times 9)$ inputs respectively. Fitting made recourse to the `R` package `e1071`.

Artificial Neural Networks

Informative overviews of ANNs are provided by several monographs including Bishop (1995) and Ripley (1996). In most, if not all, ANN applications to

MHC - peptide binding classification problems a fully-connected, feed-forward architecture with one hidden layer is chosen for the network. As noted by Hastie et al. (2001), there are many delicacies surrounding training neural networks as the associated model is generally overparameterized and the optimization problem is nonconvex and potentially unstable. We employed three-fold cross validation to effect such training, focusing on optimizing the number of hidden nodes and weight decay (momentum) parameters. The inputs were the same as used for SVMs above. Implementation was effected using the `R` package `nnet`.

2.3 Model Training, Validation and Performance

Due to the absence of designated test datasets, we used multiple levels of cross-validation to tune the above classifiers and evaluate their predictive performance. The following scheme was used: a) randomly split the data into 10 parts and reserve (set aside) one part as a (pseudo) test dataset with the remaining parts constituting the learning dataset; b) develop predictive models using the respective classifiers using the learning dataset; c) obtain test error rates by applying the models to the test dataset; d) cycle through all 10 possible withheld test datasets, repeating steps a)-c); e) repeat the entire procedure, steps a) through d), 10 times corresponding to differing random partitions of the data. For the SVM and ANN classifiers, tuning parameters were, determined using three-fold cross validation of each learning set as constructed in step a). For the other classifiers (limited) sensitivity analyses indicated that default settings were adequate.

The predictive performance of each classifier and sequence representation scheme was assessed using receiver operating characteristic (ROC) curves, generated by thresholding classifier class predictions. An ROC curve was derived for each test dataset and the corresponding area under the ROC curve (aROC) was computed. To summarize the performance of each method we calculated the mean aROC and its associated standard error over the differing cross-validation folds. Values of $aROC = 50\%$ indicate random choice; $aROC > 80\%$ indicate moderate accuracy; and $aROC > 90\%$ indicates high prediction accuracy. In addition, for comparability with prior summaries (Sung and Simon (2004)), we also compared sensitivities that correspond to 80% specificity (sensitivity_{80}).

2.4 Prediction of HIV-1 Epitopes

Our epitope profiling of HIV-1 is focused on predictions obtained for HLA-DR4 binding. This emphasis derives from the fact that binding to HLA-DR4 is more complex than for HLA-A2 and accordingly we see more variation in predictive performance across classifiers and representations. We obtained the accession numbers of 32 full genome-length reference HIV-1 strains from Los Alamos National Laboratory HIV Sequence Database (`http://www.hiv.lanl.gov/content/index`). This diverse collection of strains span subtypes A, B, C, D, E, F, H, J and O. The corresponding gag, pol and env amino acid sequences were retrieved from GenBank (`http://www.ncbi.nlm.nih.gov/Genbank/`). Two pol sequences were incomplete and were therefore excluded from further analysis. To detect conserved epitopes, all sequences were aligned using Clustal W (Thompson et al. (1994)) with default settings. We showcase results for a single classifier (bagging) but examine each of the four sequence codifications. For each of these models we construct image plots displaying (thresholded) binding affinity at each (including overlaps) 9-mer of the aligned HIV-1 sequences. Hits that straddle the different strains represent highly conserved epitopes that constitute putative vaccine targets. Informal validation of the epitope predictions is assessed by recourse to the HIV Molecular Immunology Database (`http://www.hiv.lanl.gov/content/immunology/index`), which contains a comprehensive list of experimentally confirmed HIV-1 T cell epitopes.

2.5 Validation Using Known Epitopes

To further evaluate the impact of differing peptide representation schemes we obtained a benchmark dataset that has previously been used for model validation in this context (Doytchinova and Flower (2003)). Peptide sequences, of lengths ranging from 12 to 20 amino acids, for 25 known T-cell epitopes binding to HLA-DR4, were downloaded from the JenPep database (Blythe et al. (2002)). Bagging classifiers, based on each of the four representations, were trained using HLA-DR4 binders collected from MHCPEP and synthesized non-binders (see Section 2.1). These were then tested on each 9-mer subsequnce of the 25 known epitopes. We score each peptide for each representation using the highest scoring 9-mer subsequence based on the corresponding classifier. To determine a binding threshold for these scores we generated a set of 2,500 non-binding 9-mer peptides according to amino acid frequencies from SwissProt. These, in turn, were queried against the four classifiers and the maximum score within the synthetic dataset for each model was identified and used as

the threshold: peptides with scores exceeding the corresponding threshold are predicted as "binders".

3 Results

Figure 1 and 2 illustrate frequencies (by binding status) of the 20 amino acids at positions 1-9 for HLA-A2 and HLA-DR4 respectively. For HLA-A2, amino acids A, I, L, M, T and V are over-represented in binders at (known) anchor postions 2 and 9. However, such frequency disparities are not apparent for HLA-DR4, underlying the greater complexity of class II binding.

We evaluated the impact of classifier and sequence representation choices by examining all $24 = 6 \times 4$ combinations of classifiers (single tree, bagging, random forest, boosting, SVM and ANN) and sequence representation schemes (AA, QSAR, 10-P and (QSAR+10)-P). Performance evaluation made recourse to cross-validation and ROC curves, as described in Section 2.3. Tables 1 and 2 present areas under the ROC curves (aROC) and the sensitivity rates at 80% specificity for HLA-A2, with the corresponding ROC curves displayed in Figure 3. The tree ensemble methods – bagging, random forests and boosting – exhibit excellent performance irrespective of data representation. SVM is comparably accurate, except when solely QSAR properties were used for peptide representation. ANNs and single tree were poorer, with the latter being the worst. Peters et al. (2003) show similar results for HLA-A2, with single tree, ANN and SVM providing increasingly better performance in that order. The same ordering with respect to sensitivity (as opposed to aROC) are obtained by Zhao et al. (2003). Variable importance measures for boosting (extracted from the `R` package `gbm` coincide with those obtained from random forest. Postions 2 and 9 are found to be the most important for binding, corresponding to known anchor positions. However, using only anchor positions for classification degrades predictive performance significantly for all classifiers except the single tree.

Analogous results fro HLA-DR4 are given in Tables 3 and 4 and Figure 4. Here, there were slightly more marked differences in predictive performance according to which sequence representation scheme was used. Overall, the amino acid coding gave consistently (across classifiers) best results. As for HLA-A2, use of solely QSAR properties fared worst indicating that, for the broad spectrum of classifiers examined, this selection is inadequate for capturing the complexities of binding for HLA-DR4 molecules. Again, the tree ensemble methods and SVM provided superior performance to a single tree

and ANN. Bhasin and Raghava (2004) also show better performance of SVM over ANN in their comparison studies of HLA-DR4 binding. All classifiers, except for the single tree, performed significantly better than the classifier developed by Sung and Simon (2004) for which the corresponding aROC is 0.84 and $\text{Sensitivity}_{80\%}$ is 0.72. Variable importance measures for boosting (extracted from the `R` package `gbm`) coincide with those obtained from random forest, and positions 9, 8 and 6 are found to be the most important for binding, whereas anchor positions reported previously are positions 1,4,6 and 9 (Muntasell et al. (2002)).

3.1 Illustration of Representation Impact

To further showcase differences deriving from using different sequence representations we focus on single trees as applied to HLA-DR4 peptide binding. Single trees were selected despite their inferior predictive performance (see Table 3) since (i) results were sensitive to representation scheme; and (ii) output readily facilitates interpretation.

The trees depicted in Figures 5 were obtained by (cost-complexity) pruning of initial trees grown to maximal depth. For comparison purposes we contrast trees with 4 or 5 terminal nodes. In the tree diagrams ovals designate internal nodes and rectangles designate terminal nodes. Within each node both the predicted class (0: non-binding or 1: binding) and sample size ratio (#{non-binding}/#{binding}) is given. For the AA coding (see Figure 5a), the root node 1 is partitioned on the basis of position 8 – peptides having amino acids A or L at this position are assigned to the right daughter node which is enriched with binders (87 non-binders and 354 binders). For comparison, both of the property trees use two consecutive splits to achieve a similar enrichment. For the 10-P tree (Figure 5b), the first split partitions on "α-helix preference" at position 8, which corresponds to assigning amino acids A, E, F, L and M to the right daughter node. The subsequent split of this node uses bulkiness to assign amino acids A and L to recover above enrichment. The QSAR tree (Figure 5c) first splits on postion 8 using ECI and assigns amino acids G, A, V, L, I to the right daughter. This node is further split based on ISA of position 9 to produce a terminal node with 91 non-binders and 328 binders. Thus, the AA tree requires fewer splits to achieve either the same or better separation.

As mentioned in Section 2.2, classification trees are theoretically exhaustive in determining the optimal split for an unordered categorical covariate. For the AA representation, there are $2^{19} - 1 = 524,287$ such splits per position

whereas for 10-P and QSAR-property encoding there are $19 \times 10 = 190$ and $19 \times 2 = 38$ possible splits per position. Further, any property based split can be captured via an AA split but not vice versa.

3.2 HIV-1 Epitope Profiling

To appraise predictive models based on differing sequence representations and assess effects on T-cell epitope identification on a genomewide scale, we analyzed a set of HIV-1 reference sequences. Extensive research has shown that HIV-1 specific CD4+ T cells play an important role in the control of HIV-1 replication. To identify T-cell epitopes in HIV-1 that might serve as vaccine targets, it is purposeful to determine conservation of HIV-1 derived peptides that display affinity to multiple MHC molecules. Sung and Simon (2004) adopted such an approach, using their "peptide property model" to scan reference HIV-1 genomes for MHC binding potential. We mimic their analysis, albeit focused on contrasting results from the four different sequence representations. Accordingly, we do not attempt to investigate multi-allele antigenicity, but direct attention solely to HLA-DR4 in view of its complex binding patterns. Initially, HLA-DR4 training data were used to fit the ensemble classifier bagging, which was chosen among the six classifiers (see Table 3) for (i) its overall good predictive performance and (ii) its relative sensitivity toward the four different sequence representations. Subsequently, overlapping 9-mers from the reference HIV-1 strains were queried via the binding model, giving rise to a series of binding potentials spanning the entire sequence of the HIV-1 strains. Unlike Sung and Simon (2004), we used ClustalW (Thompson et al. (1994)) to align the set of HIV-1 strains, which enhances the ability to gauge epitope conservation.

Figure 6 (a) gives image plots of the affinity of gag proteins of a diverse set of (aligned) HIV-1 strains to HLA-DR4. For the bagging classifier the binding probability was estimated by the percentage of trees (in the ensemble) that classify the target sequence as a "binder", where target sequence at postition n (in the image plot) corresponds to the 9-mer peptide spanning positions n to $n + 8$ along the aligned gag sequence. These probabilities were thresholded so as to display a similar number of binders (in red) to Sung and Simon (2004) and across the four different sequence representations. Results for proteins pol and env are shown in Supplementary Data. Correlations between the binding probabilities predicted based on the different sequence representations are schematically illustrated in Figure 6(b). Predictions based on QSAR properties showed the worst association with the others whereas, as expected,

correlation between 10-P and (10+QSAR)-P is the highest (0.9). AA is more strongly correlated with 10-P and (10+QSAR)-P (0.75-0.77) than with QSAR (0.51-0.59).

A red line spanning the different strains of HIV-1 in Figure 6(a) suggests a highly conserved epitope that could be a vaccine target against HIV-1. We listed such peptides derived from gag, pol and env in Table 5, which are conserved in at least 90% of the strains and are classified as binders by any of the four representation schemes. Note that the use of alignment generates far more hits of conserved binders compared to the unaligned approach of Sung and Simon (2004) (see their corresponding Table 2), even though the number of binders per strain were conditioned to be the same. The identities of the representations against which the peptides exhibited significant binding are given in parentheses next to the target peptides. To gauge the accuracy of these epitope predictions we used the HIV Molecular Immunology Database (`http://hiv-web.lanl.gov/content/immunology/index.html`), which provides a comprehensive collection of annotated and searchable HIV-1 T-cell epitopes, to identify known and experimentally confirmed MHC binding peptides in gag, pol and env from within our list of binders. Such peptides are marked with the corresponding references in Table 5. Overall, the classification model based on AA had the highest hit rate with 41 experimentally confirmed binders, whereas for QSAR, 10-P and (10+QSAR)-P, the numbers are 28, 33 and 31 respectively.

3.3 Validation Using Known T-cell Epitopes

We performed further validation using 25 known T-cell epitopes binding to DRB1*0401 collected from the JenPep database (Blythe et al. (2002)) as described in Section 2.5. Results are presented in Table 6 which lists binding scores from the four different sequence representations. Those peptides that are predicted as non-binders (false negatives) are indicated in italics. Using the AA representation identified 21 (of the 25) epitopes, whereas using QSAR, 10-P and (10+QSAR)-P identified 19, 20 and 20 respectively.

4 Discussion

The objectives of this paper have been two-fold: (i) to compare classification techniques in the context of peptide binding to MHC molecules; and (ii) to

illustrate the impact of differing peptide respresention schemes on classification accuracy. We based our evaluation on both an MHC class I molecule (HLA-A2) and an MHC class II molecule (HLA-DR4).

The MHC - peptide binding classification problem is characterized by the following features: (1) short (9-mer) peptides providing multilevel (20) unordered categorical covariates (amino acids); (2) peptide anchor positions and complex between-position interactions influencing binding affinity; and (3) a premium on interpretable classification rules. Many "strong learners" (Bühlmann and Yu (2003)), such as ANNs and SVMs, that are suitable for handling the challenges posed by item (2), do not efficiently handle covariate types as in (1). Accordingly, this has resulted in many analyses making recourse to select biophysical properties of amino acids, rather than using the amino acids themselves. For example, Lin et al. (2004) use two hand-picked QSAR molecular structure descriptors, Milik et al. (1998) use six hand-picked properties, while Sung and Simon (2004) use ten factors derived from principal components analysis (PCA) of 188 biophysical properties. Not only do these contrasting approaches immediately beg questions of which, and how many, properties to employ but, more importantly, concerns surrounding information loss arise. Such concerns extend to the situation where pre-selection of properties is based on PCA and a large proportion of total variation is accounted for: variation not captured may be important for classification. Our results on representation scheme are unequivocal. The use of amino acids (AA) themselves does at least as well (and in several cases significantly) better than the other representation schemes, irrespective of classifier employed and/or epitope validation approach used, and avoids altogether the above selection issues. Therefore, it is our recommended representation.

Some clear-cut conclusions can also be drawn with regard to classifier choice. That single classification trees fare uniformly worst is not surprising, since such relatively poor performance has been well documented (Friedman et al. (2000); Breiman (2001a); Hastie et al. (2001); Bühlmann and Yu (2003)). However, the fact that ANNs are dominated by the remaining methods is notable, as they have been the most widely used method in the MHC - peptide binding setting. While ANNs, in turn, dominate simple motif-based classifiers (Brusic et al. (1997); Yu et al. (2002)), our results indicate that they are not competitive with more credible approaches such as SVMs and ensemble methods. This poor performance is compounded by the tuning sensitivity and black-box nature of ANNs. SVMs have seen more recent application to peptide binding prediction (Dönnes and Elofsson (2002); Zhao et al. (2003); Bhasin and Raghava (2004)) where, as was the case for our investigation, good prediction perfor-

mance results were obtained. Differing recommendations have been advanced regarding choice of kernel, with Zhao et al. (2003) advocating a linear kernel whereas Bhasin and Raghava (2004) found the radial basis kernel performed best. This disparity might reflect target application as the former is based on simpler MHC class I binding, while the latter pertains to more complex MHC class II molecules. We used radial basis kernels for both classes and achieved competitive performance.

We have introduced the use of random forests and boosting to MHC - peptide binding classification problems and our results demonstrate that these techniques are consistently more accurate than heretofore used alternatives. Furthermore, they are relatively robust with respect to tuning. Random forests readily provide accurate estimates of test set error and measures of covariate importance, making them all the more appealing. For these reasons we believe they constitute the classifier of choice for problems with sequence-based predictors.

While our comparisons of the differing peptide representation schemes with regard to epitope prediction indicated that the AA representation was at least as effective as properties, this constitutes only partial validation. Far more compelling would be demonstration that predicted epitopes not (yet) identified in a relevant pathogen – here we have focused on HIV-1 – database are true epitopes as opposed to false positives. Of course, this requires experimental verification. However, such confirmation can at least be directed by considering degrees of conservation, akin to the depiction in Figure 6.

References

Adams, S. L., Biti, R. A., and Stewart, G. J. (1997). T-cell response to HIV in natural infection: optimized culture conditions for detecting responses to gag peptides. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, 15:257–263.

Bedford, P. A., Clarke, L. B., Hastings, G. Z., and Knight, S. C. (1997). Primary proliferative responses to peptides of HIV Gag p24. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, 14:301–306.

Bhasin, M. and Raghava, G. P. S. (2004). SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*, 20(3):421–423.

Bishop, C. (1995). *Neural networks for Patter Recognition*. Oxford University Press, Oxford.

Blythe, M. J., Doytchinova, I. A., and Flower, D. R. (2002). JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics*, 18:434–439.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical modeling: the two cultures. *Statistical Science*, 16:199–215.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth, Belmont, California.

Brusic, V., Rudy, G., Honeyman, M., Hammer, J., and Harrison, L. (1997). Prediction of MHC class II-binding peptides using an evolution algorithm and artificial neural network. *Bioinformatics*, 14(2):121–130.

Brusic, V., Rudy, G., Kyne, A. P., and Harrison, L. C. (1996). MHCPEP – a database of MHC-binding peptides: update 1995. *Nucleic Acids Research*, 24(1):242–244.

Bühlmann, P. and Yu, B. (2003). Boosting with the l_2 loss: Regression and classification. *JASA*, 98:324–339.

Buus, S., Sette, A., Colon, S. M., Miles, C., and Grey, H. M. (1987). The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenic peptides. *Science*, 235:1353–1358.

Calvo-Calle, J. M., Hammer, J., Sinigaglia, F., Clavijo, P., Moya-Castro, Z. R., and Nardin, H. (1997). Binding of malaria T cell epitopes to DR and DQ molecules *i*n vitro correlates with immunogenicity. *J. Immunol.*, 159:1362–1373.

Carmichael, A., Jin, X., and Sissons, P. (1996). Analysis of the human env-specific cytotoxic T-lymphocyte (CTL) response in natural human immunodeficiency virus type 1 infection: low prevalence of broadly cross-reactive env-specific CTL. *J. Virol.*, 70:8468–8476.

Chicz, R. M., Urban, R. G., Gorga, J. C., Vignali, D. A. A., Lane, W. S., and Strominger, J. L. (1993). Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J. Exp. Med*, 178:27–47.

Christianini, N. and Shaw-Taylor, J. (2000). *Support Vector Machines*. Cambridge University Press, Cambridge.

Congia, M., Patel, S., Cope, A. P., Virgiliis, S. D., and Sonderstrup, G. (1998). T cell epitopes of insulin defined in HLA-DR4 transgenic mice are derived from preproinsulin and proinsulin. *Proc. Natl. Acad. Sci. USA*, 95:3833–3838.

de Lalla, C., Sturniolo, T., Abbruzzese, T., Abbruzzese, L., Hammer, J., Sidoli, A., Sinigaglia, F., and Panina-Bordignon, P. (1999). Identification of novel T cell epitopes in *Lol p5α* by computational prediction. *J. Immunol.*, 163:1725–1729.

Diepolder, H. M., Gerlach, J.-T., Zachoval, R., HOffmann, R. M., Jung, M.-C., Wierenga, E. A., Scholz, S., Santantonio, T., Houghton, M., and Southwood, S. (1997). Immunodominant CD4+ T-cell epitope within nonstructural protein 3 in acute hepatitis C virus infection. *J. Virol.*, 71:6011–6019.

Dong, X., An, B., Kierstead, L. S., Storkus, W. J., Amoscato, A. A., and Salter, D. (2000). Modification of the amino terminus of a class II epitope confers resistance to degradation by CD13 on dendritic cells and enhances presentation to T cells. *J. Immunol.*, 164:129–135.

Dönnes, P. and Elofsson, A. (2002). Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3:25.

Doytchinova, I. A. and Flower, D. R. (2003). Towards the in silico indentification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics*, 19(17):2263–2270.

Efron, B. and Tibshirani, R. (1993). *An introduction to the Bootstrap*. Chapman and Hall, London.

Emmert, D. B., Stoehr, P. J., Stoesser, G., and Cameron, G. N. (1994). The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Research*, 22(17):3445–3449.

Endl, J., Otto, H., Jung, G., Dreisbusch, B., Donie, F., Stahl, P., Elbracht, R., Schmitz, G., Meinl, E., and Hummer, M. (1997). Identification of naturally processed T cell epitopes from glutamic acid decarboxylase presented in the context of HLA-DR allels by T lymphocytes of recent onset IDDM patients. *J. Clin. Invest.*, 99:2405–2415.

Falk, K., Rotzschke, O., Stefanovic, S., Jung, G., and Rammensee, H.-G. (1991). Allele specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*, 351:290–296.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407.

Fugger, L., Rothbard, J. B., and Sonderstrup-McDevitt, G. (1996). Specificity of an HLA-DRB1*0401-restricted T cell response to type II collagen. *Eur. J. Immunol.*, 26:928–933.

Gahery-Segard, H., Pialoux, G., Charmeteau, B., Sermet, S., Poncelet, H., Raux, M., Tartar, A., Levy, J. P., Gras-Masse, H., and Guillet, J. G. (2000). Multiepitopic B- and T cell responses induced in humans by a human immunodeficiency virus type 1 lipopeptide vaccine. *J. Virol.*, 74:1694–1703.

Gaston, J. S., Deane, K. H., Jecock, R. M., and Pierce, J. H. (1996). Identification of 2 Chlamydia trachomatis antigens recognized by synovial fluid T cells from patients with Chlamydia induced reactive arthritis. *J. Rheumatol.*, 23:130–136.

Geretti, A. M., Baalen, C. A. V., Borleffs, J. C., Els, C. A. V., and Osterhaus, A. D. (1994). Kinetics and specificities of the T help-cell response to gp120 in the asymptomatic stage of HIV-1 infection. *Scand. J. Immunol.*, 39(4):355–362.

Goudebout, P., Zeliszewski, D., Golvano, J. J., Pignal, C., Gac, S. L., Borras-Cuesta, F., and Sterkers, G. (1997). Binding analysis of 95 HIV gp120

peptides to HLA-DR1101 and -DR0401 evidenced many HLA-class II binding regions on gp120 and suggested several promiscuous regions. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, 14(2):91–101.

Gross, D. M., Forsthuber, T., Tary-Lehmann, M., Etling, C., Ito, K., Nagy, Z. A., Field, J. A., Steere, A. C., and Huber, B. T. (1998). Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. *Science*, 281:703–706.

Hammer, J., Gallazzi, F., Bono, E., Karr, R. W., Guenot, J., Valsasnini, P., Nagy, Z. A., and Sinigaglia, F. (1995). Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association. *J. Exp. Med.*, 181:1847–1855.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer, New York.

Honeyman, M. C., Stone, N. L., and Harrison, L. C. (1998). T-cell epitopes in type 1 diabetes autoantigen tyrosine phosphatase IA-2: potential for mimicry with rotavirus and other environmental agents. *Mol. Med.*, 4:231–239.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.*, 4(1):23–55.

Kovats, S., Whiteley, P. E., Concannon, P., Rudensky, A. Y., and Blum, J. S. (1997). Presentation of abundant endogenous class ii DR-restricted antigens by DM-negative B cell lines. *Eur. J. Immunol.*, 27:1014–1021.

Kubo, R. T., Sette, A., Grey, H. M., Appella, E., Sakaguchi, K., Zhu, K., Arnott, N.-Z., Sherman, D., Shabanowitz, N., Michel, H., Bodnar, W. M., Davis, T. A., and Hunt, D. F. (1994). Definition of specific peptide motifs for four major HLA-A alleles. *J. Immunol.*, 152:3913–3924.

Li, K., Adibzadeh, M., Halder, T., Kalbacher, T., Heinzel, S., Muller, C., Zeuthen, J., and Pawelec, G. (1998). Tumour-specific MHC-class-II restricted response after *in vitro* sensitization to synthetic peptides corresponding to gp100 and Annexin II eluted from melanoma cells. *Cancer Immunol. Immunother.*, 47:32–38.

Lin, Z., Wu, Y., Zhu, B., and Wang, L. (2004). Toward the quantitative prediction of t-cell epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201. *Journal of Computational Biology*, 11(4):683–694.

Livingston, B., Crimi, C., Newman, M., Higashimoto, Y., Appella, E., Sidney, J., and Sette, A. (2002). A rational strategy to design multiepitope immunogens based on multiple th lymphocyte epitopes. *J. Immunol.*, 168:5499–5506.

Mallios, R. R. (2001). Predicting class II MHC/peptide multi-level binding wit an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics*, 17(10):942–948.

Mamitsuka, H. (1998). Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Protein*, 33:460–474.

Manca, F., Fenoglio, D., Valle, M. T., Pira, G. L., Kunkl, A., Ferraris, A., Saverino, D., Lancia, F., and Mortara, L. (1995). Human CD4+ t cells can discriminate the molecular and structural context of T epitopes of HIV and HIV p66. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, 9:227–237.

McNicholl, J. M., Whitworth, W. C., Oftung, F., Fu, X., Shinnick, T., Jensen, P. E., Simon, M., Wohlhueter, R. M., and Karr, R. W. (1995). Structural requirements of peptide and MHC for DR(alpha, beta1*0401)-restricted T cell antigen recognition. *J. Immunol.*, 155:1951–1963.

Milik, M., Sauer, D., Brunmark, A. P., Yuan, L., Vitiello, A., Jackson, R., Peterson, P. A., Skolnick, J., and Glass, C. A. (1998). Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nature Biotechnology*, 16:753–756.

Muntasell, A., Carrascal, M., Serradell, L., van Veelen, P., Verreck, F., Koning, F., Raposo, G., Abiän, J., and Jaraquemada, D. (2002). HLA-DR4 molecules in neuroendocrine epithelial cells associate to a heterogeneous repertoire of cytoplasmic and surface self peptides. *J. Immunol.*, 169:5052–5060.

Muraro, P. A., Vergelli, M., Kalbus, M., Banks, D. E., Nagle, J. W., Tranquill, L. R., Nepom, G. T., Biddison, W. E., McFarland, H. F., and Martin, R. (1997). Immunodominance of low-affinity major histocompatibility complex-binding myelin basic protein epitope (residues 111-129) in HLA-DR4 (B1*0401) subjects is associated with a restricted T cell receptor repertoire. *J. Clin. Invest.*, 100:339–349.

Nehete, P. N., Schapiro, S. J., Johnson, P. C., Murthy, K. K., Satterfield, W. C., and Sastry, K. J. (1998). A synthetic peptide from the first conserved region in the envelop protein gp160 is a strong T-cell epitope in HIV-infected chimpanzees and humans. *Viral Immunol.*, 11(3):147–158.

Parker, K. C., Bednarek, M. A., and Coligan, J. E. (1994). Scheme for ranking potential HLA-A2 binidng peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, 153:163–175.

Parker, K. C., Bednarek, M. A., Hull, L. K., Utz, U., Cunningham, B., Zweerink, H. J., Biddison, W. E., and Coligan, J. E. (1992). Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. *J. Immunol.*, 149:3580–3587.

Peters, B., Tong, W., Sidney, J., Sette, A., and Weng, Z. (2003). Examining the independent binding assumption for bindng of peptide epitopes to MHC-I molecules. *Bioinformatics*, 19(14):1765–1772.

Reece, J. C., Geysen, H. M., and Rodda, S. J. (1993). Mapping the major human T helper epitopes of tetanus toxin. *J. Immunol.*, 151:6175–6184.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, New York.

Rosenberg, E. S., Billingsley, J. M., Caliendo, A. M., Boswell, S. L., Sax, P. E., Kalams, S. A., and Walker, B. D. (1989). Vigorous HIV-1-specific CD4+ T cell response associated with control of viremia. *Science*, 278:1447–1450.

Rudensky, A. Y., Preston-Hurlburt, P., Hong, S.-C., Barlow, A., and Janeway, C. A. (1991). Sequence analysis of peptides bound to MHC class II molecules. *Nature*, 353:622–627.

Ruppert, J., Sidney, J., Celis, E., Kubo, R. T., Grey, H. M., and Sette, A. (1993). Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. *Cell*, 74:929–937.

Schrier, R. D., Gnann, J. W., Landes, R., Lockshin, C., Richman, D., McCutchan, A., Kennedy, C., Oldstone, M. B. A., and Nelson, J. A. (1989). T-cell recognition of HIV synthetic peptides in a natural infection. *J. Immunol*, 142:1166–1176.

Segal, M. R., Barbour, J. D., and Grant, R. M. (2004). Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 2.

Segal, M. R., Cummings, M. P., and Hubbard, A. (2001). Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics*, 57:632–642.

Sitz, K. V., Ratto-Kim, S., Hodgkins, A. S., Robb, M. L., and Birx, D. L. (1999). Proliferative response to human immunodeficiency virus type 1 (HIV-1) gp120 peptides in HIV-1-infected individuals immunized with HIV-1 rgp120 or rgp160 compared with nonimmunized and uninfected controls. *J. Infect. Dis.*, 179(4):817–824.

Sung, M.-H. and Simon, R. (2004). Genomewide conserved epitope profiles of HIV-1 predicted by biophysical properties of MHC binding peptides. *Journal of Computational Biology*, 11(1):125–145.

Therneau, T. and Anderson, E. (1997). An introduction to recursive partitioning using the rpart routines. Technical report, Mayo Foundation.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680.

Topalian, S. L., Gonzales, M. I., Parhurst, M., Li, Y. F., Southwood, S., Sette, A., Rosenberg, S. A., and Robbins, P. F. (1996). Melanoma-specific CD4+ T cells recognize nonmutated HLA-DR-restricted tyrosinase epitopes. *J. Exp. Med.*, 183:1965–1971.

van der Burg, S. H., Kwapperberg, K. M., Geluk, A., van der Kruk, M., Pontesilli, O., Hovenkamp, E., Franken, K. L., van Meijgaarden, K. E., Drijfhout, J. W., Ottenhoff, T. H., Melief, C. J., and Offringa, R. (1999). Identification of a conserved universal Th epitope in HIV-1 reverse transcriptase that is processed and presented to HIV-specific CD4+ T cells by at least four unrelated HLA-DR molecules. *J. Immunol.*, 162:152–160.

Wahren, B., Rosen, J., Sandstrom, E., Mathiesen, T., Modrow, S., and Wigzell, H. (1989). HIV-1 peptides induce a proliferative response in lymphocytes from infected persons. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, 4:448–456.

Wilson, C. C., Palmer, B., Southwood, S., Sidney, J., Higashimoto, Y., Appella, E., Chesnut, R., Sette, A., and Livingston, B. D. (2001). Identification

and antigenicity of broadly cross-reactive and conserved human immunodeficiency virus type-1 derived helper T-lymphocyte epitopes. *J. Virol.*, 75:4195–4207.

Yu, K., Petrovsky, N., Schönbach, C., Koh, J. L. Y., and Brusic, V. (2002). Methods for prediction of peptides binding to MHC molecules: a comparative study. *Molecular Medicine*, 8(3):137–148.

Zhao, B., Pinilla, C., Valmori, D., Martin, R., and Simon, R. (2003). Application of support vector machines of T-cell epitopes prediction. *Bioinformatics*, 19(15):1978–1984.

Table 1: Comparisons of various models and coding methods on HLA-A2 classification: aROC

Models	aROC on test set			
	AA	QSAR	10 properties	(QSAR+10) properties
single tree	0.898 (0.008)	0.904 (0.010)	0.904 (0.005)	0.895 (0.008)
bagging	0.965 (0.002)	0.954 (0.002)	0.963 (0.001)	0.963 (0.002)
random forest	0.970 (0.002)	0.971 (0.001)	0.972 (0.002)	0.974 (0.001)
boosting	0.969 (0.002)	0.963 (0.001)	0.969 (0.002)	0.968 (0.002)
svm	0.968 (0.001)	0.923(0.01)	0.972 (0.001)	0.970 (0.001)
neural net	0.953 (0.003)	0.906 (0.005)	0.937 (0.004)	0.938 (0.003)

Table 2: Comparisons of various models and coding methods on A2 classification: $\text{Sensitivity}_{80\%}$

Models	$\text{Sensitivity}_{80\%}$			
	AA	QSAR	10 properties	(QSAR+10) properties
single tree	0.882	0.882	0.882	0.872
bagging	0.950	0.927	0.945	0.939
random forest	0.958	0.957	0.961	0.967
boosting	0.952	0.939	0.945	0.948
svm	0.954	0.885	0.959	0.960
neural net	0.940	0.847	0.916	0.916

Table 3: Comparisons of various models and coding methods on HLA-DR4 classification: aROC

Models	aROC on test set			
	AA	QSAR	10 properties	(QSAR+10) properties
single tree	0.804 (0.011)	0.746 (0.008)	0.755 (0.011)	0.763 (0.010)
bagging	0.911 (0.003)	0.877 (0.003)	0.891 (0.002)	0.891 (0.003)
random forest	0.916 (0.003)	0.911 (0.002)	0.914 (0.004)	0.919 (0.004)
boosting	0.917 (0.002)	0.907 (0.004)	0.914 (0.004)	0.914 (0.003)
svm	0.917 (0.002)	0.846(0.003)	0.919 (0.002)	0.917 (0.002)
neural net	0.881 (0.006)	0.772 (0.009)	0.870 (0.004)	0.846 (0.009)

Table 4: Comparisons of various models and coding methods on HLA-DR4 classification: $\text{Sensitivity}_{80\%}$

Models	$\text{Sensitivity}_{80\%}$			
	AA	QSAR	10 properties	(QSAR+10) properties
single tree	0.724	0.605	0.670	0.682
bagging	0.850	0.796	0.809	0.818
random forest	0.857	0.843	0.855	0.861
boosting	0.862	0.842	0.850	0.854
svm	0.865	0.725	0.865	0.855
neural net	0.808	0.592	0.784	0.749

Table 5: Conserved HLA-DR4 epitopes in gag, pol and env of HIV-1 predicted by the bagging model. Amino acid sequence encoding is denoted as "1", and encodings based on QSAR properties, 10- and (10+QSAR)-P are labeled "2", "3" and "4" respectively.

#	Sequence(Model)	#	Sequence(Model)	#	Sequence(Model)
			GAG		
1	MGARASVLS(all)	34	IVWASRELE(234)	42	ERFAVNPGL(124)
57	CRQILGQLQ(4)	75	LRSLYNTVA(12)	78	LYNTVATLY(134)
81	TVATLYCVH(2)	94	IKDTKEALD(2)	140	GQMVHQAIS(2)^{A}
152	LNAWVKVVE(2)	168	VIPMFSALS(134)^{B}	171	MFSALSEGA(1)^{B}
172	FSALSEGAT(34)^{B}	184	LNTMLNTVG(2) ^{C}	194	HQAAMQMLK(13)^{D}
201	LKETINEEA(13) ^{D}	206	NEEAAEWDR(1)	213	DRVHPVHAG(2)^{E}
229	REPRGSDIA(2)^{E}	233	GSDIAGTTS(3)	234	SDIAGTTST(134)^{E}
266	IILGLNKIV(2) ^{AB}	276	MYSPTSILD(234)^{B}	297	VDRFYKTLR(2)^{C}
298	DRFYKTLRA(1)^{C}	300	FYKTLRAEQ(4)	301	YKTLRAEQA(1)
302	KTLRAEQAS(34)	310	SQEVKNWMT(34)	314	KNWMTETLL(1)^{F}
315	NWMTETLLV (134)^{F}	319	ETLLVQNAN (14)^{F}	330	CKTILKALG(34)
334	LKALGPAAT(134)	336	ALGPAATLE(2)	346	MMTACQGVG(2)
353	VGGPGHKAR(2)	357	GHKARVLAE(24)	359	KARVLAEAM(all)
360	ARVLAEAMS(all)				
			POL		
73	GQLKEALLD(234)	76	KEALLDTGA(1)	77	EALLDTGAD(134)
79	LLDTGADDT(1)	105	GIGGFIKVR(2)	122	ICGHKAIGT(2)
125	HKAIGTVLV(13)	128	IGTVLVGPT(4)	130	TVLVGPTPV(1)
131	VLVGPTPVN(2)	139	NIIGRNLLT(all)	142	GRNLLTQIG(2)
146	LTQIGCTLN(all)	153	LNFPISPIE(2)	154	NFPISPIET(1)
160	IETVPVKLK(3)	181	LTEEKIKAL(34)	185	KIKALVEIC(2)
215	VFAIKKKDS(134)	223	STKWRKLVD(2)	225	KWRKLVDFR(4)
226	WRKLVDFRE(234)	231	DFRELNKRT(1)	232	FRELNKRTQ(2)
257	KKKSVTVLD(all)	281	KYTAFTIPS(124)^{A}	284	AFTIPSINN(1)^{A}
286	TIPSINNET(1)	297	IRYQYNVLP(13)	316	QSSMTKILE(2)^{G}
341	DLYVGSDLE(134)	357	IEELRQHLL(13)^{H}	387	YELHPDKWT(4)
412	IQKLVGKLN(134)^{I}	415	LVGKLNWAS(all)^{H}	430	KVRQLCKLL(4)
431	VRQLCKLLR(2)^{H}	437	LLRGTKALT(1234)^{HI}	440	GTKALTEVI(2)^{H}
443	ALTEVIPLT(134)^{H}	447	VIPLTEEAE(134)^{H}	451	TEEAELELA(134)^{H}
452	EEAELELAE(34)	503	NLKTGKYAR(1)	523	LTEAVQKIT(2)
526	AVQKITTES(4)	555	TWWTEYWQA(34)	556	WWTEYWQAT(1)
560	YWQATWIPE(3)	565	WIPEWEFVN(2)	568	EWEFVNTPP(34)

continued...

569	WEFVNTPPL(1)^{A}	577	LVKLWYQLE(all)^{AI}	593	ETFYVDGAA(124)^{I}
627	TTNQKTELQ(13)	633	ELQAIYLAL(all)	636	AIYLALQDS(1)
644	SGLEVNIVT(24)	651	VTDSQYALG(134)	654	SQYALGIIQ(2)
681	IKKEKVYLA(134)	690	WVPAHKGIG(2)^{A}	716	FLDGIDKAQ(2)
731	HSNWRAMAS(14)^{A}	733	NWRAMASDF(3)	734	WRAMASDFN(2)
745	PVVAKEIVA(2)	776	WQLDCTHLE(2)	781	THLEGKVIL(2)
785	GKVILVAVH(2)	787	VILVAVHVA(2)	788	ILVAVHVAS(234)
792	VHVASGYIE(2)	799	IEAEVIPAE(34)	810	QETAYFLLK(134)
813	AYFLLKLAG(all)	836	FTGATVRAA(13)	854	FGIPYNPQS(4)
877	IGQVRDQAE(34)	883	QAEHLKTAV(4)	884	AEHLKTAVQ(12)
887	LKTAVQMAV(all)^{A}	912	GERIVDIIA(2)	937	NFRVYYRDS(1)^{A}
949	LWKGPAKLL(13)	956	LLWKGEGAV(2)	983	IRDYGKQMA(234)
			ENV		
34	LWVTVYYGV(13)^{J}	35	WVTVYYGVP(4)^{J}	44	VWKEATTTL(1)^{J}
45	WKEATTTLF(13)^{JK}	48	ATTTLFCAS(3)^{J}	51	TLFCASDAK(4)^{J}
52	LFCASDAKA(14)^{J}	86	LVNVTENFN(2)^{J}	201	ITQACPKVS(2)^{JKL}
251	IRPVVSTQL(134)^{HJ}	254	VVSTQLLLN(1)^{HJ}	256	STQLLLNGS(1)^{HJ}
257	TQLLLNGSL(34)^{HJ}	258	QLLLNGSLA(134)^{J}	342	LKQIASKLR(3)^{J}
447	SNITGLLLT(3)^{M}	494	LGVAPTKAK(1)^{J}	516	GALFLGFLG(3)^{K}
518	LFLGFLGAA(12)^{K}	519	FLGFLGAAG(34)^{K}	520	LGFLGAAGS(all)^{K}
523	LGAAGSTMG(4)^{K}	530	MGAASMTLT(3)^{K}	533	ASMTLTVQA(1)^{K}
538	TVQARQLLS(34)	541	ARQLLSGIV(2)	548	IVQQQNNLL(13)^{E}
549	VQQQNNLLR(134)^{E}	552	QNNLLRAIE(2)^{E}	555	LLRAIEAQQ(2)
558	AIEAQQHLL(1)	559	IEAQQHLLQ(3)	561	AQQHLLQLT(134)
574	KQLQARILA(34)	575	QLQARILAV(1)	605	TTAVPWNAS(1)^{N}
654	EKNEQELLE(14)	660	LLELDKWAS(1)^{N}	666	WASLWNWFN(2)
669	LWNWFNITN(2)^{E}	681	YIKLFIMIV(2)	685	FIMIVGGLV(2)^{E}
686	IMIVGGLVG(2)^{E}	688	IVGGLVGLR(2)^{E}	691	GLVGLRIVF(2)
695	LRIVFAVLS(all)	698	VFAVLSIVN(2)	708	VRQGYSPLS(all)
711	GYSPLSFQT(4)	752	SLALIWDDL(1)	753	LALIWDDLR(2)
764	CLFSYHRLR(3)	771	LRDLLLIVT(2)	793	LKYWWNLLQ(2)
804	SQELKNSAV(4)	811	AVSLLNATA(14)	814	LLNATAIAV(2)^{J}
816	NATAIAVAE(all)^{J}	818	TAIAVAEGT(34)^{J}	845	RRIRQGLER(34)^{E}
847	IRQGLERIL(2)^{E}				

^{A}Wilson et al. (2001) ^{B}Rosenberg et al. (1989) ^{C}Adams et al. (1997) ^{D}Gahery-Segard et al. (2000) ^{E}Wahren et al. (1989) ^{F}Bedford et al. (1997) ^{G}Livingston et al. (2002) ^{H}Manca et al. (1995) ^{I}van der Burg et al. (1999) ^{J}Geretti et al. (1994) ^{K}Nehete et al. (1998) ^{L}Goudebout et al. (1997) ^{M}Sitz et al. (1999) ^{N}Schrier et al. (1989)

Table 6: Comparisons of HLA-DR4 binding prediction using a set of 25 know T-cell epitopes

T-cell epitope	source	AA	QSAR	10-P	10+QSAR
QNLLKAEKGNKAAAQR	Histone H1-like protein HC1[A]	0.12	0.15	0.17	0.16
LLESIQQNLLKAEKGN	Histone H1-like protein HC1[A]	0.11	*-0.04*	0.08	0.02
EYLNKIQNSLSTEWSPCSVT	Circumsporozoite protein[B]	0.31	0.12	0.13	0.07
AGFKGEQGPKGEP	Collagen alpha1 (II) chain[C]	0.24	0.17	0.28	0.25
FFRMVISNPAATHQDIDFLI	Glutamate decarboxylase[D]	0.27	0.17	0.06	0.07
LPRLIAFTSEHSHF	Glutamate decarboxylase[D]	0.08	0.02	0.07	*-0.05*
MNILLQYVVKSFD	Glutamate decarboxylase[D]	*-0.15*	0.11	*-0.07*	*-0.07*
IAFTSEHSHFSLK	Glutamate decarboxylase[D]	0.10	0.08	0.07	0.02
PKYVKQNTLKLATGMRNVP	Hemagglutinin [Fragment][E]	0.41	0.40	0.44	0.45
GYKVLVLNPSVAAT	Genome polyprotein[F]	0.11	0.12	0.18	0.20
KHKVYACEVTHQGLSS	Ig kappa chain C region[G]	0.17	0.11	0.17	0.17
KVQWKVDNALQSGNS	Ig kappa chain C region[G]	0.28	0.03	0.17	0.24
KVDNALQSGNS	Ig kappa chain C region[H]	0.01	*-0.01*	*-0.08*	*-0.14*
QPLALEGSLQK	Insulin[I]	*-0.17*	*-0.04*	*-0.12*	*-0.11*
YVIEGTSKQ	Integrin alpha-L[J]	*-0.05*	*-0.32*	*-0.14*	*-0.07*
EFVVEFDLPGIKA	18kDa antigen[K]	0.04	0.08	0.21	0.20
LSRFSWGAEGQRPGFGYGG	Myelin basic protein[L]	0.17	0.18	0.22	0.20
WNRQLYPEWTEAQRLD	Melanocyte protein Pmel 17[M]	0.24	*-0.04*	0.23	0.19
AKYDAFVTALTE	Major pollen allergen Pha a 5.3[N]	0.29	0.17	0.25	0.28
AFNDEIKASTGG	Pollen allergen Phl p 5a[N]	*-0.13*	*-0.04*	*-0.02*	*0.04*
VIVMLTPLVEDGVKQC	Protein-tyrosine phosphatase-like N[O]	0.17	0.19	0.20	0.20
AKFYRDPTAFGSG	Proteoglycan link protein[P]	0.27	0.23	0.24	0.29
QYIKANSKFIGITEL	Tetanus toxin[Q]	0.13	0.11	0.13	0.10
QNILLSNAPLGPQFP	Tyrosine[R]	0.41	0.40	0.44	0.45
DYSYLQDSDPDSFQD	Tyrosine[R]	0.40	0.34	0.43	0.42

[A]Gaston et al. (1996) [B]Calvo-Calle et al. (1997) [C]Fugger et al. (1996)
[D]Endl et al. (1997) [E]Carmichael et al. (1996) [F]Diepolder et al. (1997) [G]Kovats et al. (1997)
[H]Dong et al. (2000) [I]Congia et al. (1998) [J]Gross et al. (1998) [K]McNicholl et al. (1995)
[L]Muraro et al. (1997)[M] Li et al. (1998) [N]de Lalla et al. (1999) [O]Honeyman et al. (1998)
[P]Hammer et al. (1995) [Q]Reece et al. (1993) [R]Topalian et al. (1996)

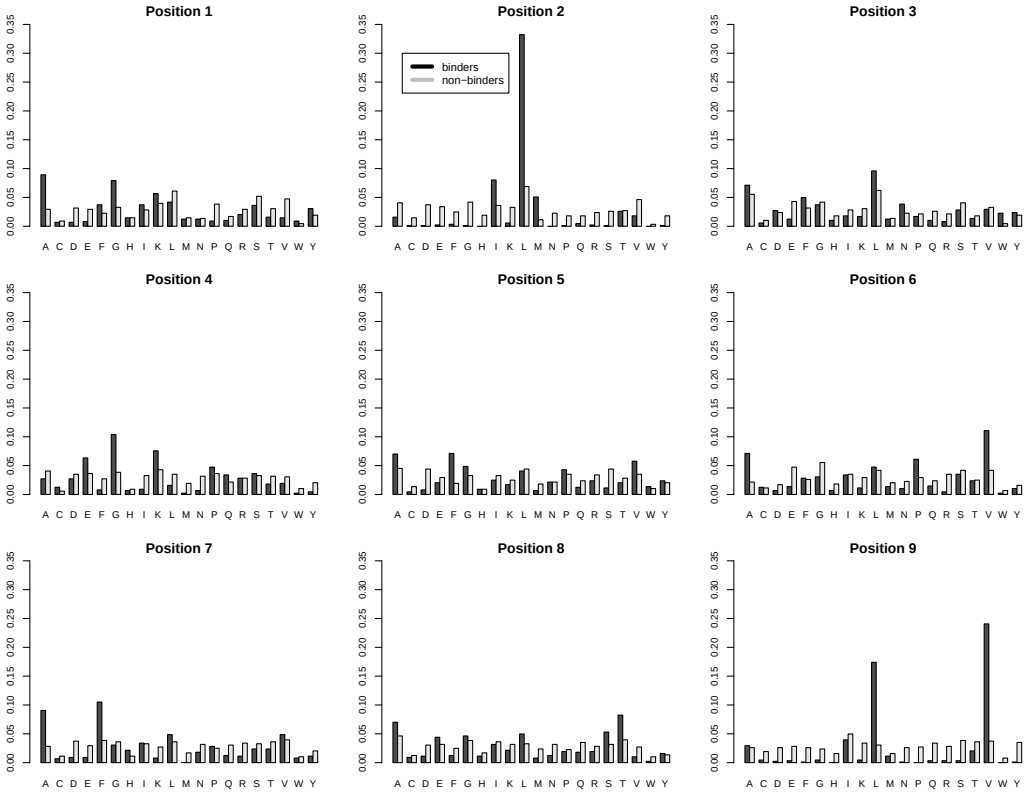


Figure 1: Amino acid frequencies for HLA-A2 at postions 1-9.

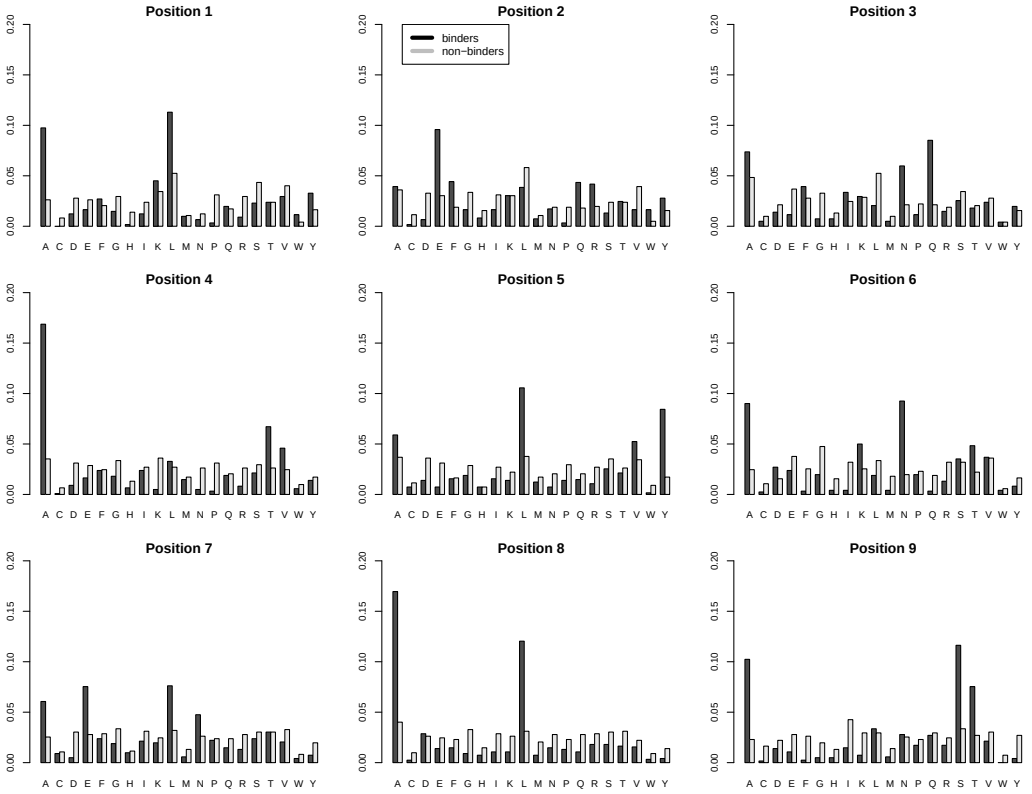


Figure 2: Amino acid frequencies for HLA-DR4 at postions 1-9.

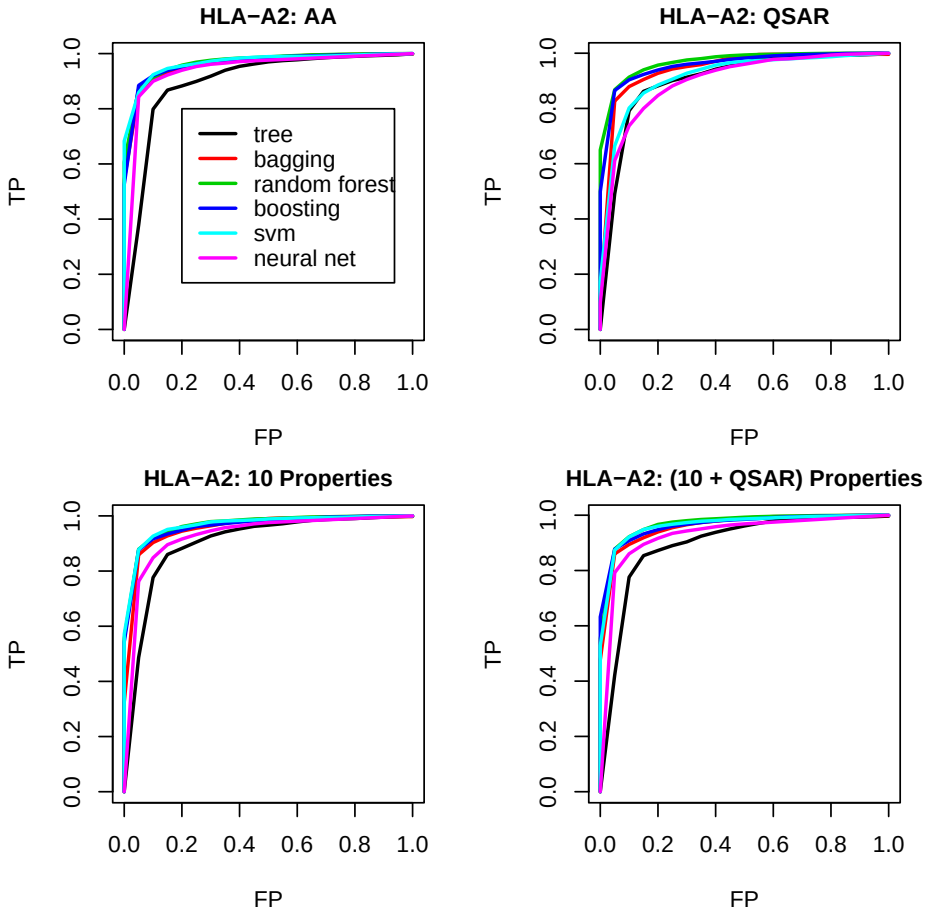


Figure 3: The ROC curves for prediction of HLA-A2 binding.

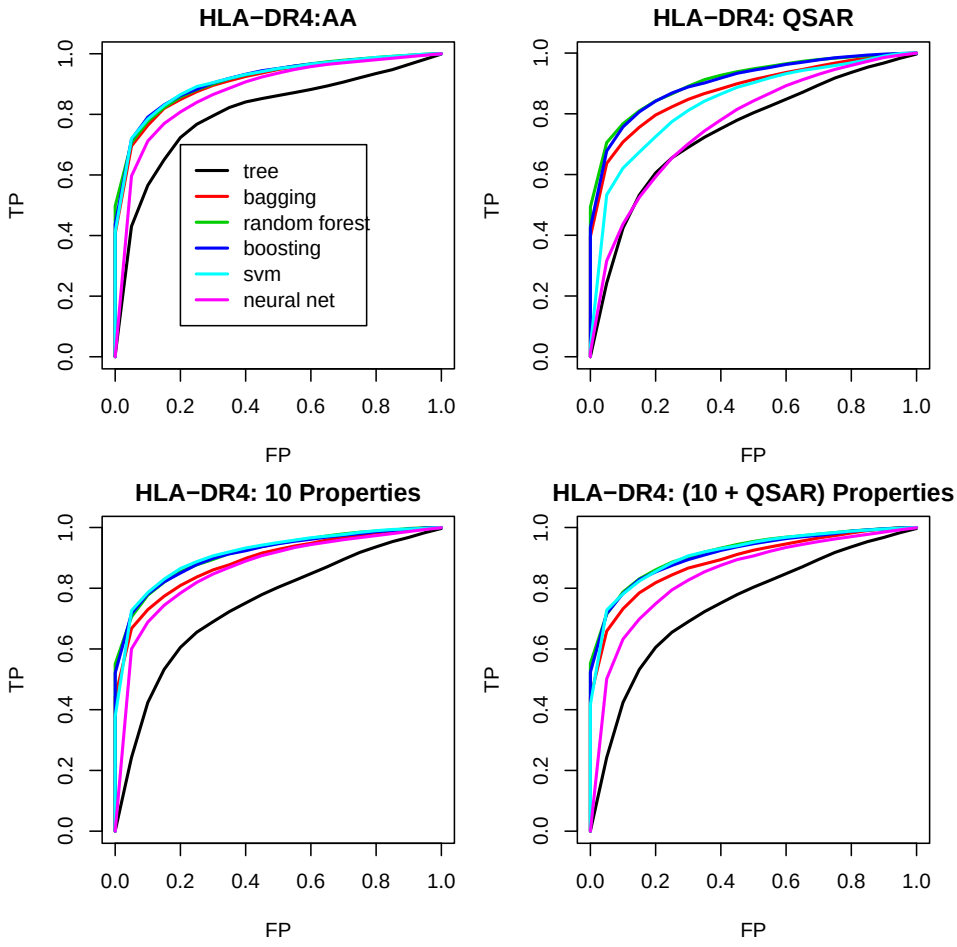


Figure 4: The ROC curves for prediction of HLA-DR4 binding.

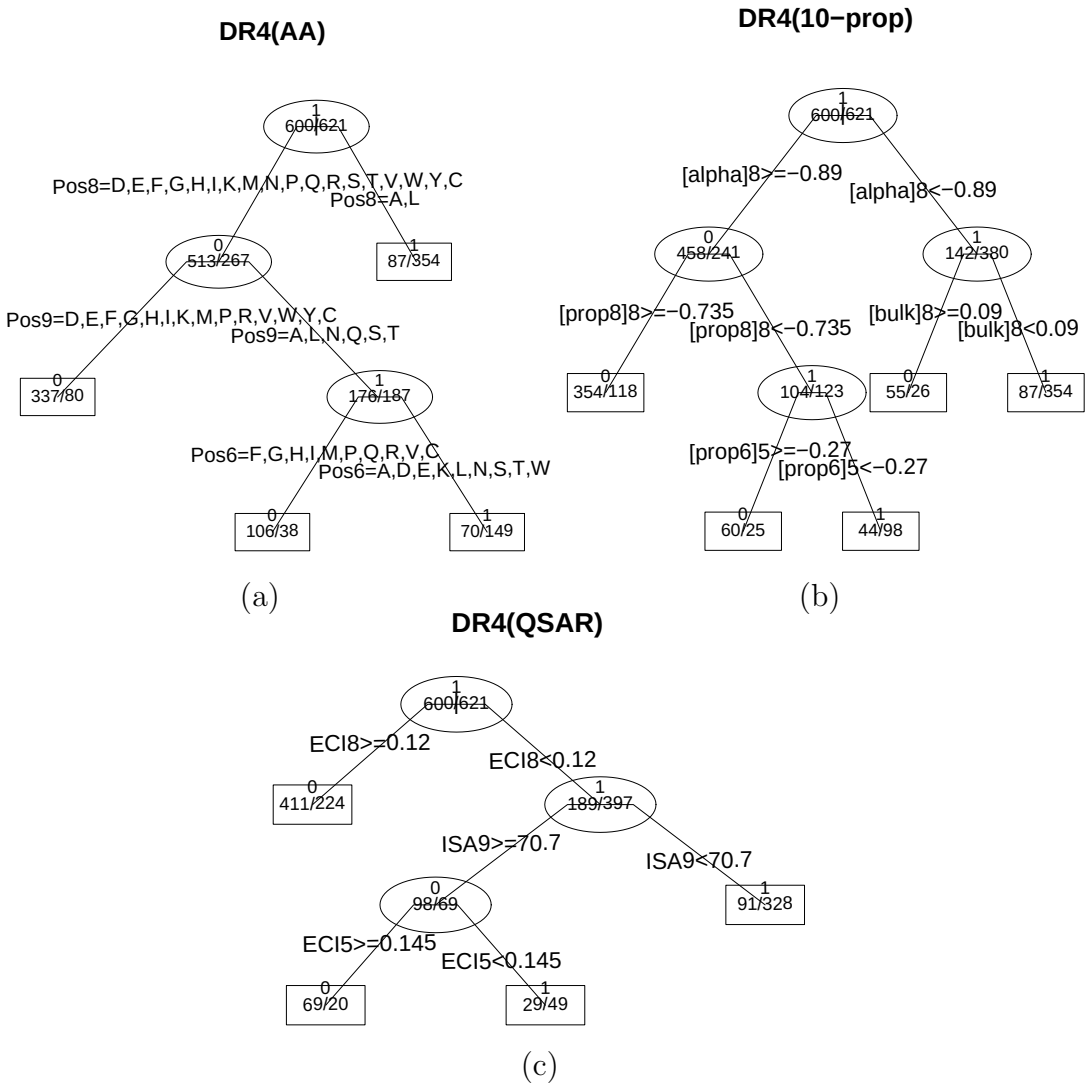
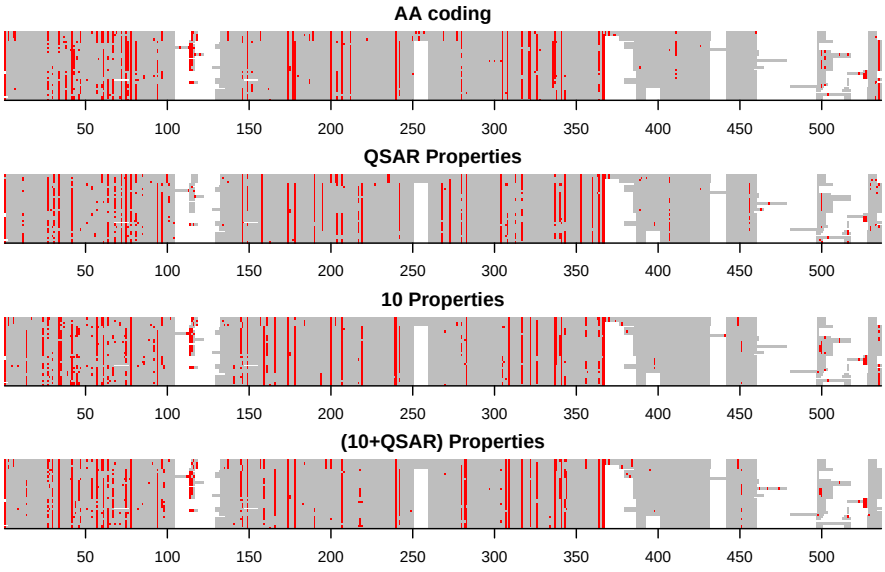
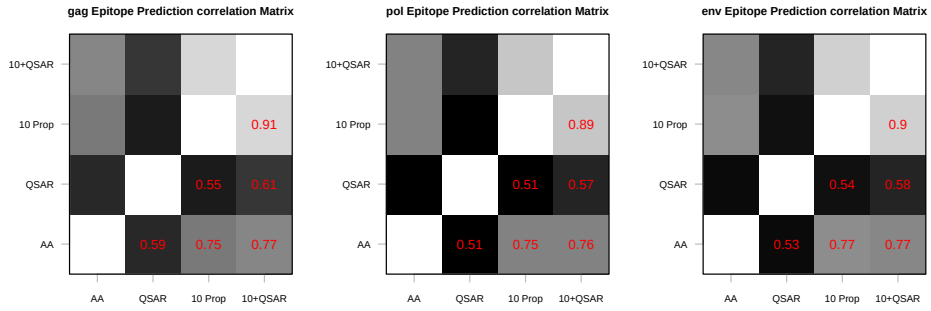


Figure 5: CART trees for prediction of HLA-DR4 peptide binding. Trees are built based on (a) amino acid sequence encoding; (b) 10 orthogonal biophysical property variables; (c) Two QSAR structural descriptors. Trees are pruned to 4-5 terminal nodes for easy comparisons.



(b)



(a)

Figure 6: HLA-DR4 epitope profiles of the HIV-1 strains. (a) HLA-DR4 profiles of the gag protein of 32 HIV-1 strains. The x axis represents the aligned amino acid positions and the y axis displays the HIV-1 strains. Predicted binders are illustrated in red, non-binders in grey and gaps in aligned sequences in white. (b) Correlation matrices of binding affinities of gag, pol and env to HLA-DR4 between the four different amino acid codings.