

## Special Feature

# Large-scale computational identification of HIV T-cell epitopes

CHRISTIAN SCHÖNBACH, YU KUN and VLADIMIR BRUSIC

*Biodiscovery Group, Kent Ridge Digital Labs, Singapore*

**Summary** Bioinformatics-driven T-cell epitope-identification methods can enhance vaccine target selection significantly. We evaluated three unrelated computational methods to screen Pol, Gag and Env sequences extracted from the Los Alamos HIV database for HLA-A\*0201 and HLA-B\*3501 T-cell epitope candidates. The hidden Markov model predicted 389 HLA-B\*3501-restricted candidates from 374 HIV-1 and 97 HIV-2 sequences. The artificial neural network (ANN) model, and Bioinformatics and Molecular Analysis Section (BIMAS) quantitative matrix predictions for A\*0201 yielded 1122 HIV-1 and 548 HIV-2 candidates. The overall sequence coverage of the predicted A\*0201 T-cell epitopes was 2.7% (HIV-1) and 3.0% (HIV-2). HLA-B\*3501-predicted epitopes covered 0.9% (HIV-1) and 1.4% (HIV-2) of the total sequence. Comparison of 890 ANN- and 397 BIMAS-derived HIV-1 A\*0201-restricted epitope candidates showed that only 13–19% of the predicted and 26% of the experimentally confirmed T-cell epitopes were captured by both methods. Extrapolating these results, we estimated that at least 247 predicted HIV-1 epitopes are yet to be discovered as active A\*0201-restricted T-cell epitopes. Adequate comparison and combined usage of various predictive bioinformatics methods, rather than uncritical use of any single prediction method, will enable cost-effective and efficient T-cell epitope screening.

**Key words:** artificial neural network, epitope coverage, hidden Markov model, HIV, HLA, peptide, T-cell epitope prediction.

## Introduction

More than 18 years after HIV-1 was isolated, research has yet to result in an effective preventive or therapeutic vaccine.<sup>1</sup> Multiple problems, including an incomplete understanding of HIV protective immunity, have contributed to the current situation. Research in the past decade has established clearly that the cellular immune response is crucial for mounting a host response against HIV.<sup>2</sup> Seronegative multiple HIV-1-exposed individuals typically develop CTL responses, although there is no proof that these individuals will be protected from developing AIDS.<sup>3</sup> In other studies, HIV-1 antigen-specific CD8<sup>+</sup> CTL were shown to reduce viraemia by controlling HIV replication *in vivo*; therefore, delaying the onset of the disease.<sup>4</sup> Low virus load and slow progression to AIDS in chronically infected individuals is correlated to strong CTL responses against HIV epitopes.<sup>5</sup> However, infected individuals, despite the presence of HIV-specific CTL responses, ultimately develop AIDS.<sup>6</sup> Amino acid substitutions in CTL epitopes caused by negative selection pressures on HIV can abolish MHC class I binding or T-cell receptor recognition, potentially resulting in immune escape of the virus.<sup>7,8</sup> Until recently, research into CTL responses to HIV focused on single MHC molecules, a small number of CTL T-cell epitopes, and systemic immunity of HIV-positive individuals. Large-scale screening of HIV T-cell epitopes has only started recently.<sup>9,10</sup>

The translation of these findings into the design of peptide and DNA vaccines requires screening of all expressed HIV antigens for promiscuous, cross-reactive T-cell epitopes in the context of HLA alleles common in populations at risk. Experimental screening of retroviral vector-expressed combinatorial peptide libraries inside the cell, with 10 000 or more peptides, can provide a high-throughput solution. Yet, this method is technically complicated and expensive; therefore, it is not an ideal solution for large-scale screening.<sup>11</sup>

Structural modelling is useful for a detailed analysis of single peptide–single MHC molecule interactions, but is not suitable for large-scale screening of antigens.<sup>12</sup> Computational predictive methods of MHC peptide binding are based on binding motifs,<sup>13</sup> quantitative matrices,<sup>14,15</sup> artificial neural networks (ANN)<sup>16,17</sup> or hidden Markov models (HMM).<sup>18</sup> A detailed comparison of predictive methods has shown that predictions resulting from binding motifs have the lowest accuracy, and typically generate large numbers of false-positive predictions of HLA binding peptides (K Yu & C Schönbach, unpubl. data). Quantitative matrices can predict binding MHC peptides with high specificity,<sup>14,15</sup> but they accurately predict only a subset of peptides, namely those containing canonical motifs.<sup>12</sup> A considerable number of potential T-cell epitopes and their variants will be missed if simple prediction methods are used.<sup>19,20</sup>

In the present study, we evaluated the results and problems of different predictive methods to help establish an effective, computational, large-scale screening strategy for HIV T-cell epitopes. We applied ANN and HMM models to screen more than 500 sequences of the structural proteins Env, Gag and Pol of HIV-1, HIV-2 and Simian Immunodeficiency Virus (SIV) for HLA-A\*0201 and HLA-B\*3501 candidate T-cell epitopes. Our analysis focused on T-cell epitope coverage and

Correspondence: V Brusic, Biodiscovery Group, Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore. Email: vladimir@krdl.org.sg

Received 7 February 2002; accepted 14 February 2002.

variation among viruses and clade types. This work is expected to improve vaccine design strategies by defining appropriate use of computational screening of vaccine candidates.

## Materials and Methods

### Selection of protein coding sequences

The protein sequences Env, Gag and Pol were retrieved from HIV and SIV FASTA sequence formatted alignments from the HIV sequence database (<http://hiv-web.lanl.gov/seq-db.html>; October 2000). This 'screening dataset' contained 533 non-redundant Gag, Env and Pol sequences of which 374 were derived from HIV-1, 97 from HIV-2 and 62 from SIV. Gag sequences were composed of 88 HIV-1, 27 HIV-2 and 15 SIV sequences. Env data contained 199 HIV-1, 46 HIV-2 and 32 SIV sequences. The Pol sequence set was derived from 87 HIV-1, 24 HIV-2 and 14 SIV sequences.

### Training set of HLA-A\*0201 and HLA-B\*3501 binding peptides

The training sets for each HLA-A\*0201 and HLA-B\*3501 comprised known binding and non-binding 9-mer peptide sequences. The positive data were retrieved from the FIMM HIV Immunology (<http://hiv-web.lanl.gov/immunology/>)<sup>21</sup> and MHCPEP<sup>22</sup> databases. Negative data were extracted from a collection of MHC non-binding peptides (V Brusic, unpubl. data). All retrieved peptide sequences that showed 100% identity to any of the HIV and SIV sequences of the screening dataset were removed from the peptide training set. The positive training set for the HLA-A\*0201 binding peptide contained 424 sequences of which 170 were reported as T-cell epitopes, 56 were naturally processed and the remainder were synthetic peptides. The positive training set for HLA-B\*3501 binders consisted of 20 T-cell epitope sequences and 74 peptides that were only tested in binding assays. The negative training set consisted of 787 HLA-A\*0201 and 128 HLA-B\*3501 non-binding peptides.

### Artificial neural network models

The ANN models were constructed as previously described.<sup>17</sup> We trained a fully connected three-layer feed-forward ANN using PLANET software.<sup>23</sup> The training set consisted of binding and non-binding 9-mer peptides. The ANN architecture comprised 180 input units, corresponding to the binary representation of 9-mer peptides, two hidden layer units and a single output unit. The learning algorithm took the form of error back-propagation. Training was performed for 300 cycles. The values for momentum and learning rate were 0.5 and 0.2, respectively. Each prediction result was calculated as the average of four independent prediction runs. Each amino acid was encoded as a binary string of length 20 with a unique position set to '1' and all other positions set to '0'. A 9-mer peptide was represented as a sparse binary string of length 180, sequentially combining representations of each amino acid. The output value was scaled 0–10, representing a range from no affinity to very high binding affinity. Binding scores used for ANN training were 1, 4, 6 and 8 for no-, low-, moderate- and high-affinity binders, respectively.

### Hidden Markov model

The aligned HLA-A\*0201 and HLA-B\*3501 binding peptides of the training set were used to train the first-order profile HMM models using the HMMER package.<sup>24</sup> The training models were built with the

*Fastmodelmaker*, rather than the default *Maxmodelmaker* option of HMMER, because the model architecture can be determined heuristically and without residue-position ambiguity. For example, we noticed that *Maxmodelmaker* wrongly assigns identical scores to two overlapping peptides SLYNTVATL and LYNTVATLE. *Hmmcalibrate* was used to refine the model, and *Hmmsearch* was used to score the test data. *Hmmcalibrate* was used with a fixed length of sequences of nine amino acids; the cut-off expectation value in *Hmmsearch* was set to  $E = 60$ . The observed range of output values was between  $-0.3$  and  $-23$ . Higher values corresponded to predicted binders and lower values corresponded to predicted non-binders.

### Quantitative matrices and motifs

Matrix- and motif-based predictions were used to compare the performance and coverage of our methods. The Bioinformatics and Molecular Analysis Section (BIMAS) ([http://bimas.dcrf.nih.gov/molbio/hla\\_bind/](http://bimas.dcrf.nih.gov/molbio/hla_bind/)) HLA-A\*0201 matrix was derived experimentally from the measurements of half-time dissociation rates of peptide-HLA complexes.<sup>14</sup> The B35CS matrix was constructed from ranking scores of amino acid position frequencies of HLA-B\*3501 experimentally determined binding and non-binding synthetic peptides.<sup>15</sup> Our modification of the originally published HLA-B\*3501 matrix included assigning zero values to non-anchor amino acids at positions P2 and P9, and the value of 1.0 to proline at position 2.

### Parameter selection

Predictive models should have high specificity (SP) and sensitivity (SE), and these were calculated as  $SE = TP/(TP + FN)$  and  $SP = TN/(TN + FP)$ . TP stands for true positives (peptides that are both predicted and experimentally measured binders); TN stands for true negatives (predicted and experimental non-binders); FP stands for false positives (predicted binders, experimental non-binders); and FN stands for false negatives (predicted-non-binders, experimental binders). All ANN predictions were carried out with a scoring threshold equal to 4.0 (on the scale of 0–10);  $SP = 0.97$  and  $SE = 0.39$ . The parameters of the BIMAS model were set to  $SP = 0.95$ ,  $SE = 0.38$  and the threshold as 100.00, which is comparable to the parameters of the ANN model. The parameters of the HMM model for the HLA-B\*3501 molecule were set as  $SP = 0.82$ ,  $SE = 0.55$  and the threshold as  $-5.00$ . These thresholds were determined using predictions of published, experimentally confirmed T-cell epitopes.

## Results

### Frequency and coverage of predicted T-cell epitope candidates

First, we analysed the accuracy of the predictive models used in the present study (data not shown). In summary, the ANN models for HLA-A\*0201 showed high accuracy, particularly for high levels of specificity. The ANN models for HLA-B\*3501 were not accurate for any level of specificity (probably because of the small number of training peptides). Hidden Markov models had better accuracy for HLA-B\*3501. Therefore we used ANN models to predict HLA-A\*0201 binding peptides and HMM for HLA-B\*3501. These predictions were compared to the results of matrix-based predictions.

We screened 533 Gag, Env and Pol sequences of HIV-1, HIV-2 and SIV clades for 9-mer binders as potential T-cell epitope candidates. In total, 1282 and 452 potential T-cell

**Table 1** Number and coverage of artificial neural network (HLA-A\*0201) and hidden Markov model (HLA-B\*3501) predicted epitopes

|              | Gag        |            | Env        |            | Pol        |            | All        |            |
|--------------|------------|------------|------------|------------|------------|------------|------------|------------|
|              | HLA-A*0201 | HLA-B*3501 | HLA-A*0201 | HLA-B*3501 | HLA-A*0201 | HLA-B*3501 | HLA-A*0201 | HLA-B*3501 |
| HIV-1†       | 64/88      | 68/88      | 762/199    | 120/199    | 64/87      | 95/87      | 890/374    | 283/374    |
| Frequency    | 0.73       | 0.77       | 3.83       | 0.60       | 0.73       | 1.10       | 2.38       | 0.64       |
| Coverage (%) | 1.31       | 1.40       | 4.05       | 0.63       | 0.65       | 0.98       | 2.66       | 0.85       |
| HIV-2†       | 33/27      | 10/27      | 161/46     | 59/46      | 38/24      | 37/24      | 232/97     | 106/97     |
| Frequency    | 1.22       | 0.37       | 3.50       | 1.31       | 1.58       | 1.54       | 2.39       | 1.10       |
| Coverage (%) | 2.14       | 0.65       | 4.69       | 1.72       | 1.34       | 1.11       | 3.00       | 1.40       |
| SIV†         | 32/15      | 8/15       | 83/32      | 28/32      | 45/14      | 27/14      | 160/61     | 63/61      |
| Frequency    | 2.13       | 0.53       | 2.56       | 0.88       | 3.21       | 0.14       | 2.62       | 1.03       |
| Coverage (%) | 3.56       | 0.89       | 4.94       | 1.67       | 4.95       | 2.97       | 4.59       | 1.81       |
| Total        | 129/131    | 86/131     | 1006/277   | 207/277    | 147/125    | 159/125    | 1282/532   | 452/532    |

†The first value indicates the number of predicted epitopes, the second value represents the number of sequences. Frequency, average number of predicted epitopes per sequence; coverage, an estimate of the per cent antigen sequence that contains potential T-cell epitopes  $[(\text{no. epitopes} \times 9) / (\text{length of all sequences}) \times 100]$ .

epitopes were predicted for HLA-A\*0201 and HLA-B\*3501, respectively (Table 1). The absolute number of predicted candidates was highest among Env sequences, followed by Pol and Gag in both HLA-A\*0201 and HLA-B\*3501. The number and length of screening sequences varied widely because of natural variation and the presence of partial sequences. Therefore, we computed the relative frequency and coverage of predicted epitopes for each antigen/virus combination. The minimum frequency of 0.37 epitope candidates per sequence was found for HIV-2 Gag HLA-B\*3501. HLA-A\*0201-restricted epitope candidates occurred at a maximum frequency of 3.83.

Then we estimated which of the structural antigens provided the highest and lowest coverage of T-cell epitope candidates. First, we compared the coverage of predicted T-cell epitopes in each antigen for the three viruses. The coverage of HLA-A\*0201-restricted epitope candidates in Gag, Env and Pol increased from 2.66 (HIV-1) to 3.00 (HIV-2) and to 4.59 (SIV). A similar pattern was observed in the coverage of predicted HLA-B\*3501 T-cell epitopes: 0.85 (HIV-1), 1.40 (HIV-2) and 1.81 (SIV). The high predicted human (HLA-A\*0201 and HLA-B\*3501) T-cell epitope coverage for SIV, relative to HIV, may be explained by the lack of positive selection pressure on SIV by the human immune system. We do not know the explanation for the difference between HIV-1 and HIV-2 coverage in our results. When comparing the coverage within one virus across the Gag, Env and Pol antigens, the overall coverage of HLA-A\*0201-restricted predicted T-cell epitopes was 2.5 times higher than for HLA-B\*3501. Because the sensitivities of our predictors are similar, this result indicates that HLA-A\*0201 T-cell epitopes are more frequent than HLA-B\*3501 T-cell epitopes.

#### Comparison of artificial neural network and Bioinformatics and Molecular Analysis Section predictions

The comparison of T-cell epitope predictions derived from different methods may be misleading if the sensitivity threshold of each method is different. We used the results from experimentally determined HIV-1 HLA-A\*0201-restricted T-cell epitopes (Table 2) to select comparable sensitivity

thresholds for ANN and BIMAS predictions. Therefore, we selected the thresholds for the ANN and BIMAS predictions that produced sensitivities of 0.39 and 0.38, respectively.

From 374 HIV-1 sequences, BIMAS predicted 397 potential T-cell epitopes. From 97 HIV-2 sequences, 151 HIV-2 T-cell epitope candidates were predicted. A larger number of potential T-cell epitopes were ANN predicted: 890 HIV-1 and 231 HIV-2 from the same sequences (Table 3). The overall predicted T-cell epitope coverage of the ANN predictions was approximately twice as high as that of BIMAS. Because the coverage of a predictive method does not allow conclusions on the efficacy of epitope prediction, we calculated the ratio of peptides that were predicted by both ANN and BIMAS. Surprisingly, only 14% of the HIV-1 and 19% of the HIV-2 T-cell epitope candidates were predicted by both ANN and BIMAS.

#### Validation of predictions with experimentally confirmed T-cell epitopes

The unexpected finding of largely disparate predictions prompted us to validate ANN, HMM, BIMAS and the B35CS matrix methods with 18 HLA-A\*0201 and eight HLA-B\*3501 published and experimentally confirmed HIV-1 T-cell epitopes. Eight (44%) ANN-predicted and seven (39%) BIMAS-predicted HLA-A\*0201 T-cell epitopes were true positives. However, only five (26%) epitopes were correctly predicted by both ANN and BIMAS. If we extrapolate these findings, at least 154 of the 397 BIMAS-predicted and 396 of the 890 ANN-predicted HIV-1 epitopes, of which 247 were predicted by both methods, are expected to be confirmed as T-cell epitopes.

The validation of our HLA-B\*3501 predictions showed that the HMM model underperformed compared to the matrix-based B35CS predictions (Table 2). Out of eight known HLA-B\*3501 T-cell epitopes, five were correctly predicted by HMM. Six epitopes were predicted correctly by B35CS, whereas BIMAS failed to predict any epitope with the default threshold of  $T_{1/2} > 100$ . BIMAS predictions with lower threshold settings (e.g.  $T_{1/2} = 5$ ) included some of the confirmed T-cell epitopes among a large number of false positives (data not shown).

**Table 2** Validation of predictions with experimentally confirmed T-cell epitopes of HIV-1 positive individuals

| HLA    | Epitopes  | HXB2       | Score | HLA-A*0201 |       | HLA-B*3501 |       |       | Source                |
|--------|-----------|------------|-------|------------|-------|------------|-------|-------|-----------------------|
|        |           |            |       | ANN        | BIMAS | HMM        | B35CS | BIMAS |                       |
| A*0201 | SLYNTVATL | Gag77–85   | 5     | TP         | TP    | –          | –     | –     | HIVdb                 |
| A*0201 | SLYNTVAVL | Gag77–85   | 6     | TP         | FN    | –          | –     | –     | FIMM <sup>19</sup>    |
| A*0201 | SLFNTVATL | Gag77–85   | 6     | TP         | TP    | –          | –     | –     | HIVdb                 |
| A*0201 | SLYNTIAVL | Gag77–85   | 6     | TP         | FN    | –          | –     | –     | FIMM                  |
| A*0201 | TLNAWVKVV | Gag151–159 | 6     | TP         | FN    | –          | –     | –     | HIVdB                 |
| B*3501 | PIPIVGDIY | Gag254–262 | –7.3  | –          | –     | FN         | FN    | FN    | HIVdb                 |
| A*0201 | VLAEAMSQV | Gag362–370 | 4     | FN         | TP    | –          | –     | –     | Altfeld <sup>31</sup> |
| B*3501 | DPNPQEVVL | Env78–86   | –6.8  | –          | –     | FN         | TP    | FN    | HIVdb                 |
| A*0201 | IISLWDQSL | Env108–116 | <3    | FN         | FN    | –          | –     | –     | HIVdB                 |
| A*0201 | KLTPLCVTL | Env121–129 | 4     | FN         | FN    | –          | –     | –     | HIVdB                 |
| B*3501 | RPIVSTQLL | Env252–260 | –4    | –          | –     | TP         | TP    | FN    | HIVdB                 |
| B*3501 | LPCRKIQII | Env416–424 | –4.1  | –          | –     | TP         | TP    | FN    | HIVdB                 |
| B*3501 | TAVPWNASW | Env606–614 | <–8.0 | –          | –     | FN         | FN    | FN    | HIVdb                 |
| A*0201 | WLWYIKIFI | Env678–686 | 6     | TP         | TP    | –          | –     | –     | HIVdb                 |
| A*0201 | LLQYWSQEL | Env799–807 | 3     | FN         | FN    | –          | –     | –     | HIVdB                 |
| A*0201 | LLNATAIAV | Env814–822 | 7     | TP         | TP    | –          | –     | –     | HIVdb                 |
| A*0201 | RVIEVLQRA | Env828–836 | <3    | FN         | FN    | –          | –     | –     | HIVdb                 |
| B*3501 | TVLVDVGDY | Pol262–270 | <–8.0 | –          | –     | FN         | TP    | FN    | HIVdb                 |
| B*3501 | SPAIFQSSM | Pol311–319 | –3    | –          | –     | TP         | TP    | FN    | HIVdb                 |
| B*3501 | NPDIVIQY  | Pol330–338 | –6.5  | –          | –     | FN         | TP    | FN    | HIVdb                 |
| B*0201 | HPDIVIQY  | Pol330–338 | –3.4  | –          | –     | TP         | TP    | FN    | HIVdb                 |
| A*0201 | VIQYMDDL  | Pol334–342 | <3    | FN         | FN    | –          | –     | –     | HIVdb                 |
| B*0201 | IPLTEEAEL | Pol448–456 | –3.8  | –          | –     | TP         | TP    | FN    | HIVdb                 |
| A*0201 | ILKEPVHGV | Pol464–472 | 3     | FN         | FN    | –          | –     | –     | HIVdb                 |
| A*0201 | IVGAETFYV | Pol589–597 | 3     | FN         | TP    | –          | –     | –     | HIVdb                 |
| A*0201 | LLWKGEHAV | Pol956–964 | 6     | TP         | TP    | –          | –     | –     | HIVdb                 |

ANN, artificial neural network; BIMAS, Bioinformatics and Molecular Analysis Section; HMM, hidden Markov model; score, ANN and HMM prediction; TP, true positive; FN, false negative.

**Table 3** Coverage, frequency and overlap of artificial neural network (ANN) and Bioinformatics and Molecular Analysis Section (BIMAS) predicted HLA-A\*0201-restricted epitopes

|              | Gag         |       | Env          |         | Pol         |       | All          |         |
|--------------|-------------|-------|--------------|---------|-------------|-------|--------------|---------|
|              | BIMAS       | ANN   | BIMAS        | ANN     | BIMAS       | ANN   | BIMAS        | ANN     |
| HIV-1†       | 60/88       | 64/88 | 263/199      | 762/199 | 74/87       | 64/87 | 397/374      | 890/374 |
| Overlap      | 26 (20.97%) |       | 137 (13.35%) |         | 17 (12.41%) |       | 180 (13.98%) |         |
| Frequency    | 0.68        | 0.73  | 1.32         | 3.83    | 0.87        | 1.06  | 1.06         | 2.38    |
| Coverage (%) | 1.23        | 1.31  | 1.39         | 4.05    | 0.76        | 0.65  | 1.19         | 2.66    |
| HIV-2†       | 21/27       | 33/27 | 107/46       | 161/46  | 23/24       | 37/24 | 151/97       | 232/97  |
| Overlap      | 12 (22.22%) |       | 50 (18.66%)  |         | 10 (16.67%) |       | 72 (18.80%)  |         |
| Frequency    | 0.78        | 1.22  | 2.33         | 3.50    | 0.96        | 1.54  | 1.56         | 2.39    |
| Coverage (%) | 1.36        | 2.14  | 3.11         | 4.69    | 0.83        | 1.34  | 1.95         | 3.00    |

†The first value indicates the number of predicted epitopes, the second value represents the number of sequences. Overlap, number of epitopes that were predicted by both BIMAS and ANN; frequency, average predicted epitope hit per sequence; coverage, an estimate of the per cent antigen sequence that contains potential A\*0201 restricted T-cell epitopes [(no. epitopes × 9)/[sequence length × no. sequences] × 100].

**Discussion**

*Conserved cross-clade epitopes and variants*

A challenge for HIV vaccine design is to identify epitopes that are promiscuous and to provide sufficient coverage of cross-clade, intraclade and circulating recombinant isoforms. HIV-1 clade B strains occur predominantly in the USA and Europe, while clade A and C strains prevail in Africa and Asia, where the majority of HIV-infected persons reside. However, 35% of newly emerging strains are circulating recombinant isoforms (CRF), which can cause significant antigenic shifts.<sup>25</sup> The estimate on emerging CRF argues for a

thorough comparative analysis of predictive methods and a judicious application of different computational methods. This is particularly important before launching large-scale experimental studies because of the need for identification of clade-specific, cross-clade conserved and CRF-shared T-cell epitopes.

The preliminary results of cross-clade, intraclade and CRF hits of predicted T-cell epitopes (Table 4), indicates differences in T-cell epitope distribution. The dataset contained 157 Env sequences of HIV-1 strains belonging to clade A-U, 21 to CRF and 21 to other recombinants (Table 4). The average non-redundant ‘intraclade hit rate’ (identical predicted T-cell

**Table 4** Number and hit rate of predicted HLA-A\*0201 Env T-cell epitopes across clades

| Clade        | Sequences | Intra-clade |         | Cross-clade |        | Circulating recombinant isoforms |        | Recombinants |        | Total non-intra |
|--------------|-----------|-------------|---------|-------------|--------|----------------------------------|--------|--------------|--------|-----------------|
|              |           | NRE         | Hits    | NRE         | Hits   | NRE                              | Hits   | NRE          | Hits   | Hits            |
| A            | 16        | 75.00       | 150.00  | 25.00       | 443.00 | 13.00                            | 90.00  | 26.00        | 104.00 | 637.00          |
|              |           | 4.67        | 9.38    | 0.18        | 3.14   | 0.62                             | 4.29   | 1.24         | 4.90   | 3.48            |
| B            | 82        | 289.00      | 920.00  | 43.00       | 145.00 | 22.00                            | 91.00  | 21.00        | 32.00  | 268.00          |
|              |           | 3.52        | 11.22   | 0.57        | 1.93   | 1.05                             | 4.30   | 1.00         | 1.52   | 2.29            |
| C            | 26        | 84.00       | 138.00  | 9.00        | 21.00  | 7.00                             | 18.00  | 7.00         | 10.00  | 49.00           |
|              |           | 3.23        | 5.31    | 0.07        | 0.16   | 0.33                             | 0.86   | 0.33         | 0.48   | 0.28            |
| D            | 8         | 20.00       | 26.00   | 1.00        | 2.00   | 0.00                             | 0.00   | 5.00         | 5.00   | 7.00            |
|              |           | 2.50        | 3.25    | 0.01        | 0.01   | –                                | –      | 0.24         | 0.24   | 0.04            |
| F            | 1         | 5.00        | 5.00    | 3.00        | 18.00  | 0.00                             | 0.00   | 4.00         | 6.00   | 24.00           |
|              |           | 5.00        | 5.00    | 0.02        | 0.11   | –                                | –      | 0.19         | 0.28   | 0.14            |
| F1           | 4         | 11.00       | 11.00   | 1.00        | 6.00   | 0.00                             | 0.00   | 3.00         | 3.00   | 9.00            |
|              |           | 2.75        | 2.75    | 0.01        | 0.04   | –                                | –      | 0.14         | 0.14   | 0.05            |
| F2           | 2         | 8.00        | 8.00    | 0.00        | 0.00   | 0.00                             | 0.00   | 0.00         | 0.00   | 0.00            |
|              |           | 4.00        | 4.00    | –           | –      | –                                | –      | –            | –      | –               |
| G            | 7         | 21.00       | 26.00   | 6.00        | 0.00   | 0.00                             | 1.00   | 12.00        | 7.00   | 8.00            |
|              |           | 3.00        | 3.71    | 0.04        | –      | –                                | 0.05   | 0.57         | 0.33   | 0.05            |
| H            | 3         | 21.00       | 23.00   | 0.00        | 0.00   | 1.00                             | 1.00   | 2.00         | 2.00   | 3.00            |
|              |           | 7.00        | 7.67    | –           | –      | 0.05                             | 0.05   | 0.10         | 0.10   | 0.02            |
| J            | –         | –           | –       | –           | –      | –                                | –      | –            | –      | –               |
| K            | 1         | 2.00        | 3.00    | 0.00        | 0.00   | 0.00                             | 0.00   | 1.00         | 1.00   | 1.00            |
|              |           | 2.00        | 3.00    | –           | –      | –                                | –      | 0.05         | 0.05   | 0.01            |
| N            | 1         | 11.00       | 11.00   | 1.00        | 3.00   | 0.00                             | 0.00   | 1.00         | 1.00   | 3.00            |
|              |           | 11.00       | 11.00   | 0.01        | 0.02   | –                                | –      | 0.05         | 0.05   | 0.02            |
| O            | 5         | 37.00       | 52.00   | 0.00        | 0.00   | 0.00                             | 0.00   | 0.00         | 0.00   | 0.00            |
|              |           | 7.40        | 10.40   | –           | –      | –                                | –      | –            | –      | –               |
| U            | 1         | 6.00        | 6.00    | 0.00        | 0.00   | 0.00                             | 0.00   | 0.00         | 0.00   | 0.00            |
|              |           | 6.00        | 6.00    | –           | –      | –                                | –      | –            | –      | –               |
| CRF          | 21        | 0.00        | 0.00    | 0.00        | 0.00   | 63.00                            | 90.00  | 0.00         | 0.00   | 0.00            |
|              | –         | –           | –       | –           | 3.00   | 4.28                             | –      | –            | –      | –               |
| Recombinants | 21        | 0.00        | 0.00    | 0.00        | 0.00   | 0.00                             | 0.00   | 110.00       | 110.00 | 0.00            |
|              | –         | –           | –       | –           | –      | –                                | –      | 5.20         | 5.20   | –               |
| Total        | 199       | 590.00      | 1379.00 | 89.00       | 638.00 | 106.00                           | 291.00 | 192.00       | 281.00 | 1009.00         |

NRE, number of non-redundant predicted epitopes in intraclade sequences; hits, number of redundant hits to intraclade sequences. Frequency: of epitopes per intraclade sequence; of epitopes that occur in other clades (e.g. 25 epitopes/[157 clade A-U sequences – 16 clade A sequences]); number and frequency of epitopes of one clade that occurred in other clades, circulating recombinant isoforms and other recombinants.

epitope in different sequences of the same clade) was 3.78. The estimated interclade hit rates (predicted T-cell epitope of one clade is identical to predicted epitope of other clade sequences) varied between 0.57 (clade B) to 0.01 (clades D, F1, N). Predicted clade T-cell epitopes with decreasing frequencies of 1.05 (clade B) to 0.05 (clade H) were found in CRF. T-cell epitope predictions using clade A and B sequences only are expected to produce high numbers of conserved interclade and CRF-specific epitopes. Therefore, the design of a broad interclade and CRF prophylactic vaccine should start out with sequences of clades A and B strains.

Recently, De Groot *et al.*<sup>9</sup> described a matrix-based computational selection of conservative HIV-1 sequences from 10 000 sequences across clades (world clade approach), followed by the prediction of 10-mer peptides binding to alleles that belong to the B7 supertype family. The highest scoring, most conserved predicted epitopes (24) were selected for experimental confirmation of B7 binding, with 15 being confirmed as T-cell epitopes. The preselection of the most conserved and highest binding-score predicted peptides may produce narrow cross-clade specific sets of T-cell epitopes that may limit the effectiveness of a prophylactic vaccine. However, the combination of the world clade approach with

cross-clade and intraclade analysis will improve the selection process of the best peptide candidates for vaccine formulation.

For the design of therapeutic vaccines, the balanced coverage of intraclade and CRF T-cell epitopes and their variants is preferred. Sensitive and specific prediction methods (or a combination thereof) will facilitate the identification of large numbers of active T-cell epitopes for inclusion in polytope vaccines. We looked at experimentally confirmed HLA-A\*0201-restricted T-cell variant epitopes SLYNTVATL, SLFNTVATL, SLYNTVAVL and SLYNTI AVL. All four variants were predicted correctly by the ANN model. BIMAS identified only SLYNTVATL and SLFNTVATL, which reflects the frequency dependence of amino acids at a given position in the coefficient matrix. Amino acid patterns of variants that had been observed rarely when the BIMAS matrix was constructed are not likely to be identified. Therefore, the number of ANN-predicted SLYNTVATL variants (23 peptides) is more than twice the number of BIMAS-predicted variants (11 peptides).

This finding has significant implications in the selection and validation process of vaccine targets regarding immunodominant, antagonistic and non-responsive T-cell epitopes. SLYNTVATL is often considered to be the immunodominant epitope of Gag p17 because studies of Goulder<sup>26</sup> and

Brander<sup>27</sup> showed that the majority of HIV-infected individuals (22) responded to this Gag T-cell epitope. In their experimental system, several variants did not diminish CTL recognition. However, Sewell *et al.* found that some natural variants, such as 3F, 3S, 3C, 3 L or 3F and 5 A, antagonized the CTL response.<sup>28,29</sup> The different level of immunogenicity (dominance *vs* antagonism) of variant peptides appears to be caused by diverse structural interactions of the T-cell receptor and coreceptor that are independent of HLA-peptide binding.<sup>29</sup>

Even if the predictions of the immunogenicity of T-cell epitopes and their variants were improved by incorporating additional data (TCR and coreceptor molecule interactions or empirical association rules derived from CTL assays), it would be difficult to predict that SLYNTVATL does not induce any anti-Gag CTL response in vaccinated uninfected individuals.<sup>30</sup> Since it appears that vaccine-induced T-cell epitopes differ from epitopes induced by natural infections, the current approach of predictive model building and virtual screening for vaccine candidates should be adjusted towards prophylactic and therapeutic vaccine requirements.

Considering the disparate coverage of predicted T-cell epitopes, computational data-driven methods for predicting vaccine targets require: (i) systematic evaluation of computational models; (ii) model refinement using additional experimental data<sup>19</sup> and combining of existing models; (iii) consideration of the allelic variants of HLA; and (iv) integration of T-cell epitope information from literature and database annotations into the prediction results in order to facilitate the interpretation and T-cell epitope selection process.

### Acknowledgement

We would like to thank Associate Professor N Petrovsky for critical appraisal of this manuscript.

### References

- Nathanson N, Mathieson BJ. Biological considerations in the development of a human immunodeficiency virus vaccine. *J. Infect. Dis.* 2000; **182**: 579–89.
- Walker BD, Chakrabarti S, Moss B *et al.* HIV-specific cytotoxic T lymphocytes in seropositive individuals. *Nature* 1987; **328**: 345–8.
- Rowland-Jones SL, Dong T, Fowke KR, Kimani J, Krausa P, Newell H *et al.* Cytotoxic T cell responses to multiple conserved HIV epitopes in HIV-resistant prostitutes in Nairobi. *J. Clin. Invest.* 1998; **102**: 1758–65.
- Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MB. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* 1994; **68**: 6103–10.
- Ogg GS, Jin X, Bonhoeffer S *et al.* Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. *Science* 1998; **279**: 2103–6.
- van der Burg SH, Klein MR, Pontesilli O *et al.* HIV-1 reverse transcriptase-specific CTL against conserved epitopes do not protect against progression to AIDS. *J. Immunol.* 1997; **159**: 3648–54.
- Kelleher AD, Long C, Holmes EC *et al.* Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. *J. Exp. Med.* 2001; **193**: 375–86.
- Barouch DH, Kunstman J, Kuroda MJ *et al.* Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes. *Nature* 2002; **415**: 335–9.
- De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G. An interactive web site providing major histocompatibility ligand predictions: application to HIV research. *AIDS Res. Hum. Retroviruses* 1997; **13**: 529–31.
- De Groot AS, Bosma A, Chinai N *et al.* From genome to vaccine: in silico predictions, ex vivo verification. *Vaccine* 2001; **19**: 4385–95.
- Tolstrup AB, Duch M, Dalum I, Pedersen FS, Mouritsen S. Functional screening of a retroviral peptide library for MHC class I presentation. *Gene* 2001; **263**: 77–84.
- Doytchinova IA, Flower DR. Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A\*0201. *J. Med. Chem.* 2001; **44**: 3572–81.
- Rammensee HG, Bachmann J, Emmerich NN, Bachor OA, Stevanovic S. SYFPEITHI. Database for MHC ligands and peptide motifs. *Immunogenetics*, **50**: 213–19.
- Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 1994; **152**: 163–75.
- Schönbach C, Ibe M, Shiga H *et al.* Fine tuning of peptide binding to HLA-B\*3501 molecules by nonanchor residues. *J. Immunol.* 1995; **154**: 5951–8.
- Brusic V, Rudy G, Harrison LC. Prediction of MHC binding peptides using artificial neural networks. In: Stonier R, Yu XH (eds). *Complex Systems: Mechanism of Adaptation*. Ohmsha: IOS Press, 1994; 253–60.
- Honeyman MC, Brusic V, Stone NL, Harrison LC. Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.* 1998; **16**: 966–9.
- Mamitsuka H. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 1998; **33**: 460–74.
- Brusic V, Bucci K, Schönbach C, Petrovsky N, Zeleznikov J, Kazura JW. Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding. *J. Mol. Graph. Model.* 2001; **19**: 405–11.
- Prilliman KR, Lindsey M, Jackson KW, Cole J, Bonner R, Hildebrand WH. Complexity among constituents of the HLA-B\*1501 peptide motif. *Immunogenetics* 1998; **48**: 89–97.
- Schönbach C, Koh JLY, Sheng X, Wong L, Brusic V. FIMM, a database of functional molecular immunology. *Nucleic Acids Res.* 2000; **28**: 222–4.
- Brusic V, Rudy G, Harrison LC. MHCPEP – a database of MHC-binding peptides. *Nucleic Acids Res.* 1997; **26**: 368–71.
- Miyata Y. *A user's guide to PlaNet, Version 5.6*. Colorado: Computer Science Department, University of Colorado, 1991.
- Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998; **14**: 755–63.
- Peeters M, Sharp PM. Genetic diversity of HIV-1: the moving target. *AIDS* 2000; **14**: S129–40.
- Goulder PJ, Sewell AK, Lalloo DG *et al.* Patterns of immunodominance in HIV-1-specific cytotoxic T lymphocyte responses in two human histocompatibility leukocyte antigens (HLA) – identical siblings with HLA-A\*0201 are influenced by epitope mutation. *J. Exp. Med.* 1997; **185**: 1423–33.
- Brander C, Yang OO, Jones NG *et al.* Efficient processing of the immunodominant, HLA-A\*0201-restricted human immunodeficiency virus type 1 cytotoxic T-lymphocyte epitope despite multiple variations in the epitope flanking sequences. *J. Virol.* 1999; **73**: 10191–8.

- 28 Sewell AK, Harcourt GC, Goulder PJ, Price DA, Phillips RE. Antagonism of cytotoxic T lymphocyte-mediated lysis by natural HIV-1 altered peptide ligands requires simultaneous presentation of agonist and antagonist peptides. *Eur. J. Immunol.* 1997; **27**: 2323–9.
- 29 Price DA, Meier U-C, Klenerman P, Purbhoo MA, Phillips RE, Sewell AK. The influence of antigenic variation on cytotoxic T lymphocyte responses in HIV-1 infection. *J. Mol. Med.* 1998; **76**: 699–708.
- 30 Ferrari G, Neal W, Jones A *et al.* CD8 CTL responses in vaccines. Emerging patterns of HLA restriction and epitope recognition. *Immunol. Lett.* 2001; **79**: 37–45.
- 31 Altfeld MA, Livingston B, Reshamwala N *et al.* Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif. *J. Virol.* 2000; **75**: 1301–11.