

Epitope Discovery with Phylogenetic Hidden Markov Models

Miguel Lacerda,¹ Konrad Scheffler,² and Cathal Seoighe^{*,1,3}

¹School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland

²Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa

³Centre for High-Performance Computing, CSIR Campus, Rosebank, Cape Town, South Africa

*Corresponding author: E-mail: cathal.seoighe@nuigalway.ie.

Associate editor: Daniel Falush

Abstract

Existing methods for the prediction of immunologically active T-cell epitopes are based on the amino acid sequence or structure of pathogen proteins. Additional information regarding the locations of epitopes may be acquired by considering the evolution of viruses in hosts with different immune backgrounds. In particular, immune-dependent evolutionary patterns at sites within or near T-cell epitopes can be used to enhance epitope identification. We have developed a mutation–selection model of T-cell epitope evolution that allows the human leukocyte antigen (HLA) genotype of the host to influence the evolutionary process. This is one of the first examples of the incorporation of environmental parameters into a phylogenetic model and has many other potential applications where the selection pressures exerted on an organism can be related directly to environmental factors. We combine this novel evolutionary model with a hidden Markov model to identify contiguous amino acid positions that appear to evolve under immune pressure in the presence of specific host immune alleles and that therefore represent potential epitopes. This phylogenetic hidden Markov model provides a rigorous probabilistic framework that can be combined with sequence or structural information to improve epitope prediction. As a demonstration, we apply the model to a data set of HIV-1 protein-coding sequences and host HLA genotypes.

Key words: epitopes, immune-mediated evolution, phylogenetic hidden Markov model.

Introduction

An effective cellular immune response depends on the presentation and recognition of foreign peptides called epitopes. Antigens from endogenous pathogens, such as a virus, are degraded by the proteasome within the cytoplasm and transported to the endoplasmic reticulum (ER) by the transporter associated with processing. In the lumen of the ER, the peptide fragments undergo N-terminal trimming before attaching to human leukocyte antigen (HLA) class I molecules for delivery to the cell surface and presentation to cytotoxic T lymphocytes (CTLs). The subsequent binding of the epitope/HLA class I complex to the T-cell receptor and CD8 coreceptor on a CTL induces clonal expansion and the release of cytotoxins that lyse the virus-infected cell (Vyas et al. 2008). Mutations of amino acid residues in or around epitopes may inhibit any of these stages in the antigen processing pathway, enabling viral escape from the immune response. This is particularly evident in rapidly evolving viruses, such as the HIV that undergoes error-prone reverse transcription before integration as a provirus in the host's DNA.

The genes encoding HLA molecules are highly polymorphic with distinct alleles having different peptide-binding domains. HLA class I molecules are capable of presenting peptide sequences of 8–11 amino acids with different HLA alleles having distinct peptide-binding preferences. The antigenic peptides that are ultimately presented to CTLs therefore differ between individuals with different sets of HLA alleles. Consequently, viral escape from CTL

recognition will depend on the HLA genotype of the host and is thus said to be “HLA restricted.” Individuals expressing a common HLA allele will have the potential to present a common set of epitopes and thus to harbor similar escape variants of the virus. Indeed, several studies have reported robust correlations between polymorphisms within or near HIV epitopes and the HLA genotype of the host (Brumme et al. 2008; Carlson and Brumme 2008; Navis et al. 2008; Peters et al. 2008; Kawashima et al. 2009; Rousseau et al. 2008). These studies highlight the important role of the cellular immune response in driving HIV evolution at sites in and around HLA class I epitopes, resulting in adaptation of the virus to HLA alleles both at the level of individual-infected hosts and at the population level.

Accurate prediction of CTL epitopes is a key goal in immunoinformatics (Yu et al. 2002; Flower 2003; Larsen et al. 2005). Common approaches to this problem include motif-based methods (Schuler et al. 2007), structure-based methods (Rognan et al. 1999; Brusica et al. 2004), and machine-learning algorithms (Nielsen et al. 2003; Zhao et al. 2003; Yanover and Hertz 2005; Heckerman et al. 2006). However, none of the tools proposed thus far consider immune-mediated evolution as a distinguishing feature of epitopes. Here, we develop a mutation–selection model of viral escape from host CTL responses, which is applicable when viral-coding sequence data as well as host HLA genotype data are available. For an alignment of HIV-1 sequences of the same subtype, we assume that there is a single amino acid at each site, which is fittest in the absence of a CTL

response targeting the site. We refer to this residue as the wild-type amino acid and approximate it with the most common amino acid observed at a site in the alignment. Versions of this assumption are commonly applied to distinguish immune escape and reversion mutations in HIV-1 (see, e.g., Goonetilleke et al. 2009). In the presence of an immune response targeting a specific site, selective pressure for immune escape can cause a reduction in the frequency of the wild-type amino acid. When a site is known to be within a CTL epitope bound by a specific HLA allele, we find that allowing the frequency of the wild-type amino acid to depend on whether or not the allele is expressed in the host often provides a much better description of the data. When it is not known whether a site is within an epitope for a particular HLA allele, comparison of the fit of a model that allows the wild-type amino acid frequency to depend on the presence of the allele (the epitope model) to the fit of a model that ignores HLA genotype (the non-epitope model) can provide evidence that the site is within or close to an epitope. A contiguous set of sites favoring the epitope model provide stronger evidence for an epitope. We show that this evidence can be combined using a phylogenetic hidden Markov model (phylo-HMM) to make probabilistic epitope predictions based on sequence evolution. Evolutionary information provides a means of epitope prediction, which is orthogonal to existing sequence- and structure-based methods. We demonstrate how this information can be integrated with other prediction methods within the phylo-HMM framework, taking a method based on anchor residue motifs as an example.

Materials and Methods

Data

We analyzed a multiple alignment of 506 HIV-1 subtype C Gag p17 and p24 sequences from a Durban, South Africa cohort, contributed to the Los Alamos HIV Molecular Immunology Database by Rousseau et al. (2008). The sequences and HLA genotypes of the subjects were determined as described in Rousseau et al. (2008). The data utilized in this study correspond to HXB2 Gag positions 1–358, excluding 16 amino acid sites at the end of p17 (positions 112–127) that appeared to be misaligned (most sites contained more than 90% gaps). Two HIV-1 subtype A and subtype B sequences as well as an intersubtype recombinant were removed from the original data set. The alignment was screened for intrasubtype recombination with GARD (Kosakovsky Pond et al. 2006), but we were unable to detect departures from the star phylogeny within estimated recombination breakpoints due to the large number of taxa relative to the number of codon sites between breakpoints. Consequently, we do not account for intrasubtype recombination in our analyses, although different phylogenies across the alignment could be accommodated straightforwardly in principle. HLA-binding motifs and lists of optimally defined epitopes were obtained from the Los Alamos HIV Molecular Immunology Database (<http://www.hiv.lanl.gov/content/immunology>).

Model of Immune Escape

We take account of immune-mediated selection pressure by developing a codon substitution model with site-specific equilibrium frequencies that reflect the dependence of the wild-type amino acid at a site on the HLA genotype of the host. Our model is based on that of Halpern and Bruno (1998) in which differences in the fitness of codons are modeled in terms of site-specific codon equilibrium frequencies. In this model, the instantaneous rate of substitution from codon I to codon J , where we require that I and J are both sense codons, is defined as follows:

$$q_{ij} = \frac{\mu_{ij}^k \ln \left(\frac{\nu_j \mu_{ij}^k}{\nu_i \mu_{ji}^k} \right)}{1 - \frac{\nu_i \mu_{ij}^k}{\nu_j \mu_{ji}^k}}, \quad (1)$$

where ν_j is the equilibrium frequency of codon J at a particular site in the alignment, μ_{ij}^k is the mutation rate from nucleotide i to nucleotide j in the k th codon position, and $\ln\{(\nu_j \mu_{ij}^k)/(\nu_i \mu_{ji}^k)\}/\{1 - (\nu_i \mu_{ij}^k)/(\nu_j \mu_{ji}^k)\}$ is proportional to the fixation probability of the mutation (Halpern and Bruno 1998). In our case, we wish to have this fixation probability depend on the HLA genotype of the infected host at a given point on the viral phylogeny whenever the mutation involves the replacement of a wild-type amino acid with a non-wild type residue (escape mutation) or vice versa (reversion mutation). In order to achieve this, we parameterize the site-specific codon equilibrium frequencies in terms of the equilibrium frequency of the wild-type residue, γ , at a site and define

$$\nu_j = \begin{cases} \frac{\gamma \pi_j}{\sum_{x \in W} \pi_x} & \text{if } j \in W, \\ \frac{(1-\gamma) \pi_j}{1 - \sum_{x \in W} \pi_x} & \text{if } j \notin W, \end{cases} \quad (2)$$

where W represents the set of codons that encode the wild-type amino acid and π_j is the equilibrium frequency of codon J estimated empirically from the entire alignment as described in Kosakovsky Pond and Muse (2005). An increase in the equilibrium frequency of the wild-type residue, γ , therefore translates into an increase in the equilibrium frequencies, $\nu_{j \in W}$, of codons that encode the wild-type residue. The factors involving the π_j parameters ensure that the effects of codon usage bias, estimated across the whole alignment, are retained in the model. To allow γ to depend on the HLA genotype of the host, we set $\gamma = p_0 \gamma_0 + p_1 \gamma_1$ with $p_0 + p_1 = 1$, where p_0 (p_1) denotes the probability that a specific HLA allele is absent (present) in the host, and γ_0 (γ_1) represents the equilibrium frequency of the wild-type residue given that the HLA allele is absent (present). We make the simplifying assumption that the presence ($p_1 = 1$) or absence ($p_0 = 1$) of a specific HLA allele during the entire period represented by a branch on the phylogeny is dictated by the HLA genotype of the host at the tip of that branch. When it is unknown whether or not the HLA allele is expressed in the host at a vertex in the phylogeny, such as at all ancestral nodes, we set p_0 and p_1 equal to the proportion of HLA-genotyped individuals in the study for which

the particular HLA allele is absent and present, respectively. We also constrain $\gamma_0 \geq \gamma_1$ because HLA-mediated selection pressure is expected to increase the frequency of escape variants over wild-type amino acids.

We parameterize the mutation rate μ_{ij}^k according to the MG94×HKY85 Dual GDD 3×3 model of Kosakovsky Pond and Muse (2005) and use an F3×4 estimator for π_j . Substituting these parameters into equation (1) yields the following:

$$q_{ij} = \begin{cases} \pi_j^k \alpha & \text{synonymous transversion,} \\ \kappa \pi_j^k \alpha & \text{synonymous transition,} \\ \pi_j^k \beta \times \frac{\ln(\zeta)}{1-\zeta^{-1}} & \text{transversion, } I \notin W, J \in W, \\ \pi_j^k \beta \times \frac{\ln(\zeta^{-1})}{1-\zeta} & \text{transversion, } I \in W, J \notin W, \\ \pi_j^k \beta & \text{other nonsynonymous} \\ & \text{transversion,} \\ \kappa \pi_j^k \beta \times \frac{\ln(\zeta)}{1-\zeta^{-1}} & \text{transition, } I \notin W, J \in W, \\ \kappa \pi_j^k \beta \times \frac{\ln(\zeta^{-1})}{1-\zeta} & \text{transition, } I \in W, J \notin W, \\ \kappa \pi_j^k \beta & \text{other nonsynonymous} \\ & \text{transition,} \\ 0 & >1 \text{ nucleotide difference,} \end{cases} \quad (3)$$

where π_j^k denotes the equilibrium frequency of target nucleotide j in codon position k , κ is the transition/transversion rate ratio, α is the synonymous substitution rate, β is the nonsynonymous substitution rate, and $\zeta = \{\gamma/(1-\gamma)\} \{(1-\sum_{x \in W} \pi_x)/\sum_{x \in W} \pi_x\}$. Notice that the factors involving γ only appear in the instantaneous rate matrix for substitutions that result in the replacement of a wild-type residue with a non-wild type residue or vice versa. Consequently, the selective influence of the host's HLA genotype is limited to these substitutions only as desired.

Equation (3) defines a time-reversible substitution process for each branch in the phylogeny. However, the equilibrium frequencies, and hence the instantaneous rate matrix describing this model, will differ from branch to branch depending on whether a specified HLA allele is present, absent, or unknown along that branch. Consequently, the substitution process is not time reversible when considered over the entire tree and the location of the root node and the initial codon frequencies must therefore be specified. The problem of identifying the root node is greatly simplified in our case if we assume that the initial codon frequencies are given by equation (2) with an indeterminate HLA genotype. Because the HLA genotypes of donors are always unknown, the subprocess along all interior branches is time reversible and hence any interior node other than the most recent common ancestors of the leaf nodes may be selected to root the tree. Indeed, the root of our HIV-1 subtype C phylogeny did fall within this region of the tree when the two subtype A and subtype B sequences in the original alignment were used as outgroups for tree construction.

Let ψ' denote all the parameters except γ and γ_0 in the codon substitution process defined by equation (3). Because γ_1 is defined in terms of γ and γ_0 , it is ignored here. The site likelihood $p(x_\ell | \psi', \gamma, \gamma_0)$ at column ℓ in an alignment x may then be computed for any parameter values in $\{(\gamma, \gamma_0) : 0 \leq \gamma \leq \gamma_0 \leq 1\}$ and nonnegative real-valued branch lengths along a phylogeny using Felsenstein's pruning algorithm (Felsenstein 1981). We adopt a random effects approach with respect to the parameters γ and γ_0 and integrate them out of the site likelihood to obtain

$$p_E(x_\ell | \psi') = \int_{\gamma} \int_{\gamma_0} p(x_\ell | \psi', \gamma, \gamma_0) p(\gamma_0 | \gamma, \xi_0) p(\gamma | \xi) d\gamma_0 d\gamma, \quad (4)$$

where $p(\gamma_0 | \gamma, \xi_0)$ and $p(\gamma | \xi)$ denote prior distributions on γ_0 and γ with hyperparameters ξ_0 and ξ , respectively. The subscript E in the likelihood $p_E(x_\ell | \psi')$ refers to the epitope state and is introduced here to highlight the fact that this site likelihood was computed by allowing the equilibrium frequency of the wild-type amino acid to depend on the HLA genotype of the host, as would be expected in epitope regions of the viral genome. In non-epitope regions, the absence of HLA-mediated selective pressure implies that $\gamma = \gamma_0 = \gamma_1$ and site likelihoods may therefore be computed as

$$p_N(x_\ell | \psi') = \int_{\gamma} p(x_\ell | \psi', \gamma) p(\gamma | \xi) d\gamma. \quad (5)$$

This is equivalent to equation (4) with $p(\gamma_0 | \gamma, \xi_0) = 1$ if $\gamma_0 = \gamma$ and $p(\gamma_0 | \gamma, \xi_0) = 0$ otherwise. A comparison of $p_E(x_\ell | \psi')$ and $p_N(x_\ell | \psi')$ over a region of the viral genome will reveal localized sets of sites that evolve under immune pressure and are thus indicative of CTL epitopes. We train a hidden Markov model with the state emission probabilities given by equations (4) and (5) and hence are able to infer the epitope and non-epitope states taking account of all the information available across the alignment.

Phylogenetic Hidden Markov Model

A phylo-HMM is defined as the four-tuple $\theta = (s, \psi, A, b)$ consisting of a set of M hidden states $s = \{s_1, \dots, s_M\}$ with associated phylogenetic models $\psi = \{\psi_y : y \in s\}$, a state transition probability matrix $A = \{a_{yz}\}$ for $y, z \in s$ and a vector of initial state probabilities $b = (b_{s_1}, \dots, b_{s_M})$ (Siepel and Haussler 2005). Here, all hidden states share a common phylogenetic model and differ only in the constraints imposed on the parameters γ and γ_0 in the computation of the state emission probabilities. This continuous-time Markov stochastic process is defined by an instantaneous codon substitution rate matrix Q with entries given by equation (3), codon equilibrium frequencies given by equation (2) and a binary tree topology with a corresponding set of branch lengths. The constraints on γ and γ_0 determine whether evolution in a particular state is immune mediated or not.

Let $\phi = (\phi_1, \dots, \phi_L)$ denote a path through the phylo-HMM with $\phi_\ell \in s$ for $1 \leq \ell \leq L$ in an alignment x with L

codon sites. The joint probability of the alignment and the path is

$$p(\mathbf{x}, \phi | \theta) = b_{\phi_1} p_{\phi_1}(\mathbf{x}_1 | \psi') \prod_{\ell=2}^L a_{\phi_{\ell-1}, \phi_{\ell}} p_{\phi_{\ell}}(x_{\ell} | \psi').$$

The emission probability $p_{\phi_{\ell}}(x_{\ell} | \psi')$ is given by equation (4) if ϕ_{ℓ} corresponds to the epitope state and equation (5) if it corresponds to the nonepitope state. The likelihood of the alignment $p(\mathbf{x} | \theta) = \sum_{\phi} p(\mathbf{x}, \phi | \theta)$ may be computed with the forward algorithm, whereas the maximum likelihood path $\phi_{ML} = \operatorname{argmax}_{\phi} p(\mathbf{x}, \phi | \theta)$ may be obtained via Viterbi decoding. The forward-backward algorithm may be used to obtain posterior state probabilities $p(\phi_{\ell} = y | \mathbf{x}, \theta)$ for all ℓ and y (Rabiner 1989).

Two-State Epitope Model

We first consider a model with $s = \{E, N\}$ where E denotes an epitope state and N denotes a nonepitope state. The emission probabilities for the epitope and nonepitope states are given by equations (4) and (5), respectively. To compute these probabilities, we define $p(\gamma | \xi)$ as a beta distribution on $[0, 1]$ and $p(\gamma_0 | \gamma, \xi_0)$ as a beta distribution with support $[\gamma, \min(\gamma/\hat{p}_0, 1)]$, where \hat{p}_0 is an empirical estimate of the proportion of individuals who do not express the HLA allele under consideration. The support of $p(\gamma_0 | \gamma, \xi_0)$ ensures that $0 \leq \gamma_1 \leq \gamma \leq \gamma_0 \leq 1$. In order to alleviate the computational burden of numerical integration, we utilize discretized versions of these distributions that allows the double integration to be approximated by summation. For each $\gamma \in \{0.05 + 0.15i : i = 1, \dots, 5\}$, we compute the likelihood $p(x_{\ell} | \psi', \gamma, \gamma_0)$ at seven equally spaced values of γ_0 in the interval $[\gamma, \min(\gamma/\hat{p}_0, 1)]$ for the epitope state. A better approximation to the integration may be attained with a finer discretization over the (γ, γ_0) space, particularly near the support boundaries when the beta densities are convex; however, our experience suggests that the incremental gain in predictive accuracy is not worth the increase in computational time needed to calculate the likelihood function for the additional (γ, γ_0) pairs at every site. Given estimates of the hyperparameters ξ and ξ_0 , the density of each (γ, γ_0) pair is computed and normalized to obtain its joint probability. The probability $p_E(x_{\ell} | \psi')$ of the epitope state emitting alignment column x_{ℓ} is then calculated as the sum of the conditional probabilities $p(x_{\ell} | \psi', \gamma, \gamma_0)$ weighted by the joint probabilities of γ and γ_0 . The emission probability $p_N(x_{\ell} | \psi')$ for the nonepitope state is computed analogously with respect to the discretized marginal distribution on γ . This model is illustrated in figure 1A.

Epitope Model with Duration

The two-state model ignores several known sequence and structural features of epitopes. The requirement that an epitope be between 8 and 11 amino acids in length in order to bind to an HLA class I molecule is perhaps the simplest constraint that is not considered in the two-state model. We incorporate this restriction by defining a phylo-HMM with duration (phylo-HMMwD) on the state space

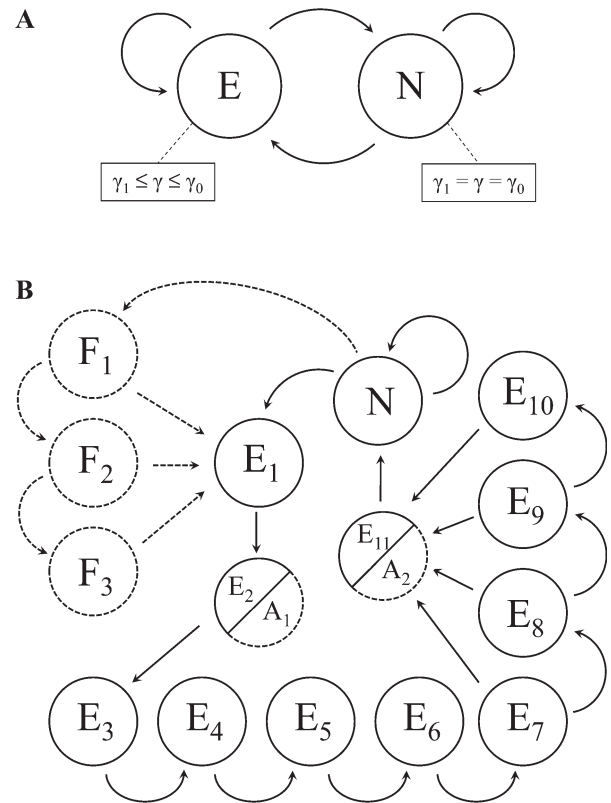


FIG. 1. (A) Two-state epitope model. (B) Phylo-HMMwD+M for HLA-B*57. Modifications to the phylo-HMMwD are indicated with dashed lines.

$s = \{E_1, \dots, E_{11}, N\}$, where E_i denotes the i th amino acid in the epitope state for $i = 1, \dots, 11$. The architecture of this model is illustrated with solid nodes and edges in figure 1B. The phylo-HMMwD provides the foundation for incorporating other spatial features relevant to epitope prediction. In what follows, we show how to integrate motif-based prediction with sequence evolution within this framework.

Phylo-HMMwD Integrated with Motif-Based Prediction

Motif-based methods of epitope prediction exploit the fact that the binding affinity of an antigenic peptide and HLA molecule is critically dependent on a few anchor residues in the peptide sequence. For example, ligands of HLA-B*57 molecules typically have the form X[ATS]XXXXXX[FW], where the amino acids in square brackets indicate dominant anchor residues and X denotes an indeterminate amino acid (Marsh et al. 2000). The region between the two anchors is of variable length and affects the conformational structure of the bound peptide-HLA complex (Stewart-Jones et al. 2005). We demonstrate how to extend our phylo-HMMwD to incorporate motif-based epitope prediction using HLA-B*57 as an example. We include the anchor motif in the model by replacing states E_2 and E_{11} with anchor states A_1 and A_2 in figure 1B. Because anchor motifs are HLA specific, the architecture of the graphical model describing the sequence of hidden states will differ between

HLA alleles. Epitope prediction may also be informed by processing escape mutations that have been reported in the N-terminal flanking regions of some epitopes (Draenert et al. 2004). We therefore introduce three optional flanking states F_1 , F_2 , and F_3 prior to entry into the first epitope state E_1 in figure 1B. We will refer to this model as a phylo-HMMwD+M, where the suffix “+M” indicates the combining of a phylo-HMMwD with motif-based epitope prediction.

To take account of the anchor motifs, we define an indicator variable $\mathbb{1}_{\ell i} (r \geq \lambda)$, or $\mathbb{1}_{\ell i}$ for short, which equals one if the proportion r of observed amino acids corresponding to anchor residue i at site ℓ is greater than or equal to a predefined threshold λ and zero otherwise. In the case of HLA-B*57, for example, $\mathbb{1}_{\ell 1}$ indicates whether the proportion of sequences with Ala, Thr, or Ser at site ℓ exceeds λ . The indicator variable $\mathbb{1}_{\ell 2}$ is similarly defined with respect to Phe and Trp. The probability of emitting column \mathbf{x}_ℓ under the i th anchor state is then

$$\begin{aligned} p_{A_i}(\mathbf{x}_\ell | \boldsymbol{\psi}') &= \sum_{j \in \{0,1\}} p_{A_i}(\mathbf{x}_\ell | \mathbb{1}_{\ell i} = j, \boldsymbol{\psi}') p_{A_i}(\mathbb{1}_{\ell i} = j | \boldsymbol{\psi}') \\ &= p_{A_i}(\mathbf{x}_\ell | \mathbb{1}_{\ell i} = 1, \boldsymbol{\psi}') \end{aligned}$$

because $p_{A_i}(\mathbb{1}_{\ell i} = 0 | \boldsymbol{\psi}') = 0$ and $p_{A_i}(\mathbb{1}_{\ell i} = 1 | \boldsymbol{\psi}') = 1$. Setting $p_{A_i}(\mathbf{x}_\ell | \mathbb{1}_{\ell i} = 1, \boldsymbol{\psi}') = p_E(\mathbf{x}_\ell | \mathbb{1}_{\ell i} = 1, \boldsymbol{\psi}')$, we obtain

$$p_{A_i}(\mathbf{x}_\ell | \boldsymbol{\psi}') = \begin{cases} \frac{p_E(\mathbf{x}_\ell | \boldsymbol{\psi}')}{p_E(\mathbb{1}_{\ell i} = 1 | \boldsymbol{\psi}')} & \text{if } \mathbb{1}_{\ell i} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

because $p_E(\mathbf{x}_\ell | \mathbb{1}_{\ell i} = 0, \boldsymbol{\psi}') = 0$ when $\mathbb{1}_{\ell i} = 1$ and thus $p_E(\mathbf{x}_\ell | \boldsymbol{\psi}') = p_E(\mathbf{x}_\ell | \mathbb{1}_{\ell i} = 1, \boldsymbol{\psi}') p_E(\mathbb{1}_{\ell i} = 1 | \boldsymbol{\psi}')$. The probability $p_E(\mathbf{x}_\ell | \boldsymbol{\psi}')$ is given by equation (4) and we approximate $p_E(\mathbb{1}_{\ell i} = 1 | \boldsymbol{\psi}')$ with the proportion of alignment sites at which $\mathbb{1}_{\ell i} = 1$. Note that if $\lambda = 0$, $\mathbb{1}_{\ell i} = 1$ at all sites and the anchor residues will not inform epitope prediction. Larger values of λ will result in anchor state emission probabilities of zero at some sites and thereby restrict epitope predictions to those regions where the anchor motifs are observed in at least some of the aligned sequences. In our analyses, we set $\lambda = 0.3$ and will therefore not detect epitopes when the anchor residues are only present in less than 30% of sequences. It is worth noting that at almost 90% of the sites in our alignment, the most common amino acid occurred at a frequency greater than 80% and our results are consequently robust to the choice of λ .

Implementation

Given an inferred topology (see below) and assuming $F_3 \times 4$ codon equilibrium frequencies, all other mutational parameters in equation (3) and the branch lengths were estimated by maximum likelihood in HyPhy (Kosakovsky Pond et al. 2005) under a standard MG94 \times HKY85 Dual GDD 3×3 model of codon substitution (i.e., without the factor involving ζ in equation (3)). Site likelihoods were then computed under the Halpern–Bruno model defined by equation (3) for discrete values of γ and γ_0 given the previously inferred mutational parameters and branch lengths. The hyperparameters of the prior distributions on γ and γ_0 and the

transition probabilities between hidden states were optimized with the Baum–Welch algorithm implemented in the R Language and Environment for Statistical Computing (R Development Core Team 2009). The initial state probabilities were constrained such that the Viterbi path always begins in the nonepitope state. The computer code is freely available from the corresponding author.

In order to determine the extent to which our results depend on the inferred phylogeny, we estimated several maximum likelihood topologies using PhyML (Guindon and Gascuel 2003) and GARLI (Zwickl 2006). Fifteen trees were estimated with different starting topologies in PhyML by specifying a general time reversible model of nucleotide substitution with variable substitution rates modeled by a four-category discrete gamma distribution with unit mean. Branch swaps were achieved by subtree pruning and regrafting and nearest neighbor interchange. A further 20 trees were inferred with GARLI under a GY94 \times HKY85 codon substitution model with nonsynonymous rates drawn from a three-category general discrete distribution. We trained our phylo-HMMs using several of the highest scoring topologies and did not find any substantive differences in our results. The phylogeny ultimately used here was resolved in PhyML and had the highest likelihood score computed in HyPhy under the MG94 \times HKY85 Dual GDD 3×3 model of codon substitution.

Results

We applied our models to a multiple alignment of 506 HIV-1 subtype C Gag p17 and p24 sequences described in Materials and Methods. All amino acid positions discussed below correspond to those of the HXB2 reference strain.

Predicted HLA-B*57-restricted epitopes

We begin by considering the specific case of HIV-1 evolution in the context of the HLA-B*57 allele for which immune-mediated escape and compensatory mutations have been extensively documented. Twenty-six of the 502 HLA-typed individuals in this study were HLA-B*57 positive (HLA-B*57⁺). The majority (15/26) of these individuals expressed the HLA-B*5703 molecule as expected in African populations (Cao et al. 2004), whereas 7 expressed HLA-B*5702, and the remainder were typed at two-digit resolution only. Molecular differences between HLA-B*57 alleles do not, however, significantly affect peptide-binding affinity and we therefore expect a considerable overlap in their bound peptide repertoires (Stewart-Jones et al. 2005), although differential T-cell receptor recruitment may result in subtype-specific escape (Yu et al. 2007).

A comparison of the integrated site likelihoods of the HLA-B*57-restricted epitope model and nonepitope model is presented in figure 2. For each site in Gag p17 and p24, we computed the Bayes factor

$$B = \frac{p_E(\mathbf{x}_\ell | \boldsymbol{\psi}')}{p_N(\mathbf{x}_\ell | \boldsymbol{\psi}')}$$

assuming flat priors on γ and γ_0 given γ and plotted $2 \ln(B)$, which is on the same scale as the likelihood ratio

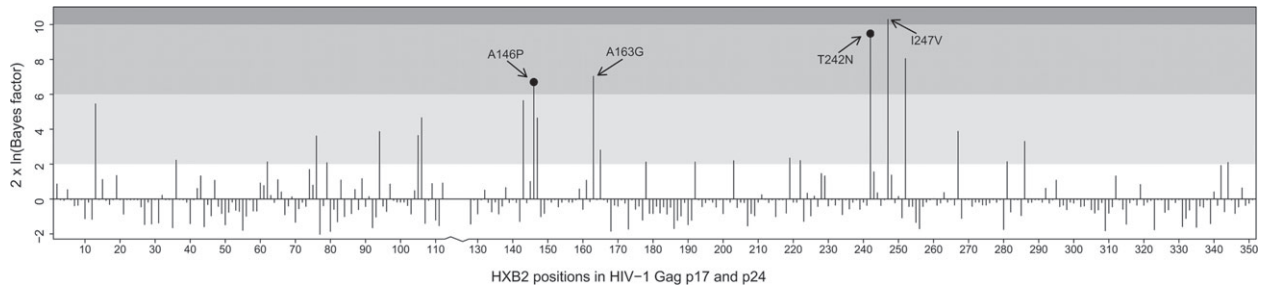


FIG. 2. Bayes factors comparing the HLA-B*57-restricted epitope model to the nonepitope model at each site in HIV-1 Gag p17 and p24. Shaded regions indicate the degree of evidence in favor of the epitope model: white indicates negligible evidence, light gray indicates positive evidence, intermediate gray indicates strong evidence, and dark gray indicates definitive evidence (Kass and Raftery 1995). Prominent escape mutations known to be associated with the HLA-B*57 allele are annotated. Black dots indicate sites with a significant ($q < 0.2$) association between host HLA and HIV mutations as determined by the phylogenetically corrected method described in Bhattacharya et al. (2007).

test statistic (Kass and Raftery 1995). The results of the two-state phylo-HMM and phylo-HMMwD+M are illustrated in [figure 3](#) for a region of the p24 protein known to contain immunodominant HLA-B*57-restricted epitopes. The results for the entire Gag p17 and p24 region are given in [supplementary figures S1 and S2](#), Supplementary Material online.

Strong HLA-B*57-mediated selection pressure has been described in and around three peptides in the Gag p24 capsid region (Chopera et al. 2008; Crawford et al. 2009), namely ISW9 (₁₄₇ISPRTLNAW₁₅₅), TW10 (₂₄₀TSTLQEQIAW₂₄₉), and KF11 (₁₆₂KAFSPEVIPME₁₇₂) (anchor residues are underscored). Escape from ISW9 occurs first during acute HIV-1 subtype C infection in HLA-B*57⁺ African populations (Crawford et al. 2009). The processing escape mutation A146P is often observed in the N-terminal flanking position of this epitope and prevents ER aminopeptidase I from cleaving an N-extended peptide to an optimal size for binding to the HLA-B*57 molecule (Draenert et al. 2004). A further escape substitution of Ile for Leu at site 147 has also been described (Draenert et al. 2004; Peters et al. 2008). [Figure 2](#) shows strong evidence in support of the epitope model at position 146 and positive evidence at position 147. The two-state phylo-HMM captures both escape mutations with sharp peaks in the posterior probability of the epitope state at sites 146 and 147 in [figure 3](#). Viterbi decod-

ing of the phylo-HMMwD+M correctly predicts the ISW9 epitope.

Two escape sites at positions 242 and 247 within the TW10 epitope are identified by large Bayes factors in [figure 2](#) and high posterior probabilities in [figure 3](#). The former position corresponds to the well-known T242N mutation observed in almost all HLA-B*57⁺ individuals (Brumme et al. 2008), whereas a mutation from Ile to Val at the latter position appears to be specific to HLA-B*57⁺ individuals infected with HIV-1 subtype C (Crawford et al. 2009). The T242N mutation disrupts critical interactions with the host protein cyclophilin A (CypA) resulting in a 10-fold reduction in viral replicative capacity (Borghans et al. 2007). This fitness loss may be partially restored by compensatory substitutions, particularly in or near the CypA-binding loop upstream of the TW10 epitope (Brockman et al. 2007). Indeed, [figure 3](#) indicates some, albeit weaker, evidence in favor of the epitope state at sites corresponding to the previously described H219Q and M228I/L compensatory mutations within this region.

The KF11 epitope is immunodominant during chronic infection (Leslie et al. 2005). Escape mutations have been described at positions 163 and 165 within this epitope, although their prevalence depends on the HIV-1 subtype and the specific HLA-B*57 allele, with more frequent escape

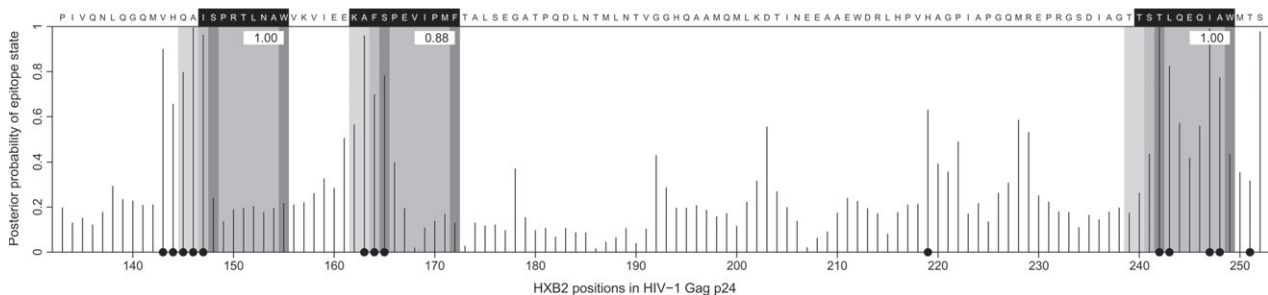


FIG. 3. Posterior probabilities of the epitope state for HLA-B*57 in HIV-1 Gag p24 based on the two-state epitope model. Gray-shaded regions indicate epitopes predicted by Viterbi decoding of the phylo-HMMwD+M with the posterior probability of each epitope given on the top right-hand corner. Predicted anchors are shaded in dark grey and flanking regions are shaded in light gray. Black dots indicate epitopes predicted by Viterbi decoding of the two-state model. The best defined or “A-list” epitopes described in Yusim et al. (2009) are indicated with white text against a black background.

observed in clade C sequences in the context of HLA-B*5703 (Yu et al. 2007). It is therefore not surprising that our two-state phylo-HMM predicts both these mutations. Indeed, the Bayes factors too favor the epitope model at these positions. A mutation from Ala to Gly at position 163 occurs first in HLA-B*57⁺ individuals and inhibits intracellular processing of the epitope precursor but compromises viral fitness substantially (Leslie et al. 2005). This fitness cost is partially restored by a compensatory substitution of Ser for Asn at position 165, which may explain the lower Bayes factor and posterior escape probability at this site relative to position 163.

Finally, a further B*57-restricted epitope QW9 (₃₀₈QATQDVKNW₃₁₆) has also been mapped to the p24 capsid protein (Stewart-Jones et al. 2005). Two mutations within this epitope, namely S310T and E312D, have been described (Navis et al. 2007). However, these mutations represent the most common variant of QW9 in HIV-1 subtype C and do not appear to be the result of CTL escape (Currier et al. 2005). Correspondingly, we too do not find any evidence of escape at these or any other sites within the QW9 epitope.

Epitope Predictions for Other HLA Alleles

We used our phylo-HMMs to predict epitopes for a further 31 HLA alleles that were observed at relatively high frequencies and for which anchor motifs were available in the Los Alamos HIV Molecular Immunology Database. For each HLA allele, the phylo-HMMwD+M was defined to accommodate epitopes of 8 to 11 amino acids in length with an N-terminal flanking region of 3 amino acids. Full plots of the epitopes predicted in HIV-1 Gag p17 and p24 are provided as [supplementary figures S1 and S2](#), Supplementary Material online, for all the HLA molecules considered. The anchor motifs used for each allele are given in [supplementary table S1](#), Supplementary Material online.

Overall, Viterbi decoding revealed 81 and 65 epitopes across all HLA alleles in the phylo-HMMwD and phylo-HMMwD+M, respectively, compared with 191 nonoverlapping epitopes identified by simply counting the number of segments with appropriately spaced anchor residues present in at least 30% of sequences (see [supplementary table S2](#), Supplementary Material online, for details). The results clearly indicate that the number of potential epitopes based solely on anchor motifs is much larger than that predicted by either phylo-HMM considering evolutionary information. Indeed, the phylo-HMMs limit the set of predictions to only those epitopes, which exhibit HLA allele-dependent immune escape. We also examined the proportion of predicted epitopes, which overlap with known epitopes documented in Yusim et al. (2009) for each of the three methods but did not find any notable differences (anchor motifs only, 0.33; phylo-HMMwD, 0.26; and phylo-HMMwD+M, 0.37). However, predicted epitopes that are unknown are not necessarily false positives, and, in fact, these are more likely to be valid epitopes under the phylo-HMMwD+M where predictions satisfy anchor motifs and exhibit immune-mediated escape

patterns. Moreover, we envisage that our model be combined with additional sequence and structural information so as to further reduce the number of false positives that may occur.

Many well-known epitopes do not conform to the documented HLA-binding motifs and will be missed by the phylo-HMMwD+M. For example, none of the known HLA-B*5802-restricted epitopes in HIV-1 Gag p17 and p24 documented in Yusim et al. (2009) contain the anchor motif X[ST]XXX[R]XX[F] suitable for binding to an HLA-B*5802 molecule (Marsh et al. 2000; Schuler et al. 2007). The two-state epitope model and phylo-HMMwD are likely to be more useful in such cases where there may be uncertainty regarding the binding motif. For example, the posterior probabilities of the two-state model for HLA-C*04 clearly indicate strong immune pressure within the known HLA-C*04-restricted epitope QW9 (₃₀₈QATQDVKNW₃₁₆), even though this epitope is undetected by the phylo-HMMwD+M, which specifies a Tyr, Phe, or Pro B-pocket anchor residue and a Leu, Phe, or Met F-pocket anchor residue.

Discussion

Pathogens, such as HIV-1, adapt to common immune responses in host populations (Kawashima et al. 2009) and to the immune repertoire of specific infected hosts (Bhattacharya et al. 2007). This phenomenon has attracted substantial interest in the HIV-1 field, partly because it provides insights into the future trajectory of the pandemic but also because the interplay between the immune response and the evolving virus is an essential component of HIV-1 disease dynamics. In this work, we have developed the first mutation–selection model of CTL escape, incorporating host HLA information explicitly into the evolutionary model. We can then make use of model comparison techniques to identify sites within the virus that are correlated with host HLA alleles. Amino acid positions that support the model of epitope evolution are consistent with viral polymorphisms that have been reported as associated with the host HLA genotype. Because CTLs recognize linear epitopes, we can incorporate this model into a phylo-HMM to predict epitopes on the basis of patterns of evolution at contiguous amino acid positions. Indeed, the possibility of combining evidence from evolutionary models over contiguous sites makes the problem of predicting linear epitopes a natural application of phylo-HMMs.

Current methods for predicting T-cell epitopes are based on the characteristics of individual amino acid sequences that affect peptide processing in the endogenous pathway or binding by HLA molecules. As described above, epitopes are also associated with characteristic patterns of evolution and we demonstrate here that the phylo-HMM framework allows us to integrate this evolutionary evidence into an existing sequence-based strategy to enhance epitope prediction. In doing so, we shift the focus from the prediction of epitopes in individual sequences to the prediction of epitopes that tend to occur in a set of viruses circulating in

a host population for which HLA data are available. Our approach is unlikely to identify all the potential epitopes in a given sequence; rather, it is designed to identify epitopes that are associated with consistent patterns of immune escape and reversion. This is the case for many well-studied and important epitopes in HIV-1. Epitopes that exhibit consistent patterns of immune escape and reversion are significant because they frequently correspond to immune responses for which there are a limited number of possible escape pathways and for which escape mutations incur a fitness cost, such that they tend to revert upon transmission to a host lacking the required HLA allele.

The evolutionary models are fitted separately for each HLA allele, followed by decoding of the phylo-HMM to predict the corresponding epitopes. However, HLA alleles do not occur independently of one another. Instead, it is possible for different HLA alleles to cooccur more often than would be expected by chance, either as a result of linkage disequilibrium or population structure. To identify HLA alleles that cooccur, we carried out Bonferroni-corrected Fisher's exact tests (supplementary table S3, Supplementary Material online) and subsequently examined the epitope predictions of the associated alleles to see if we could identify false positives resulting from linkage. In several cases, common peaks in the posterior probabilities of the epitope state based on the two-state epitope model are observed for HLA alleles, which exhibit significant linkage disequilibrium (see, e.g., the p17 region in supplementary figure S1, Supplementary Material online, for HLA-B*4201 and HLA-C*1701). Undocumented epitopes predicted by the phylo-HMMwD+M and presented by such HLA molecules were screened and highlighted as potentially spurious where appropriate. However, in most cases, the specification of HLA specific-binding motifs in the phylo-HMMwD+M eliminates these potential false positives.

The characteristic pattern of HLA-associated evolution provides a basis for epitope prediction, which is orthogonal to the characteristics of individual sequences. As a result, it is possible to combine this with any existing sequence- or structure-based epitope prediction strategy. We demonstrated this using the example of motif-based epitope prediction. Because the sequence motifs associated with HLA binding have a low-information content, an epitope prediction method that considers motifs exclusively has a very high false positive rate. Combining the phylo-HMM with motif-based prediction, we were able to substantially reduce the number of predicted epitopes. We argue that the resulting predictions are the most relevant because they are associated with a shared pattern of immune escape and reversion across different infections. The combination of the phylo-HMM with motif-based prediction also provided a fully probabilistic approach to epitope prediction and is easily interpretable. For example, we can calculate the posterior probability of an epitope at each amino acid site using the forward-backward algorithm, which is accurate given the explicit assumptions of the model. Consequently, we

present the phylo-HMM as a general epitope prediction framework, which may be readily adapted to combine evolutionary evidence of an epitope with any other epitope prediction method.

Supplementary Material

Supplementary figures S1 and S2 and tables S1, S2, and S3 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

The authors wish to thank Sergei L. Kosakovsky Pond for use of his Linux cluster and assistance with HyPhy.

References

- Bhattacharya T, Daniels M, Heckerman D, et al. (20 co-authors). 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* 315:1583–1586.
- Borghans JAM, Mølgaard A, Boer RJD, Keşmir C. 2007. HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. *PLoS One*. 2:e920.
- Brockman MA, Schneidewind A, Lahaie M, et al. (14 co-authors). 2007. Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A. *J Virol*. 81:12608–12618.
- Brumme ZL, Brumme CJ, Carlson J, et al. (21 co-authors). 2008. Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. *J Virol*. 82:9216–9227.
- Brusic V, Bajic VB, Petrovsky N. 2004. Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications. *Methods* 34:436–443.
- Cao K, Moormann AM, Lyke KE, et al. (18 co-authors). 2004. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*. 63: 293–325.
- Carlson JM, Brumme ZL. 2008. HIV evolution in response to HLA-restricted CTL selection pressures: a population-based perspective. *Microbes Infect*. 10:455–461.
- Chopera DR, Woodman Z, Mlisana K, et al. (12 co-authors). 2008. Transmission of HIV-1 CTL escape variants provides HLA-mismatched recipients with a survival advantage. *PLoS Pathog*. 4:e1000033.
- Crawford H, Lumm W, Leslie A, et al. (18 co-authors). 2009. Evolution of HLA-B*5703 HIV-1 escape mutations in HLA-B*5703-positive individuals and their transmission recipients. *J Exp Med*. 206: 909–921.
- Currier JR, Harris ME, Cox JH, McCutchan FE, Birx DL, Maayan S, Ferrari G. 2005. Immunodominance and cross-reactivity of B5703-restricted CD8 T lymphocytes from HIV type 1 subtype C-infected Ethiopians. *AIDS Res Hum Retroviruses*. 21:239–245.
- Draenert R, Le Gall S, Pfafferott KJ, et al. (12 co-authors). 2004. Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J Exp Med*. 199: 905–915.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.
- Flower DR. 2003. Towards in silico prediction of immunogenic epitopes. *Trends Immunol*. 24:667–674.

- Goonetilleke N, Liu MK, Salazar-Gonzalez JF, et al. (24 co-authors). 2009. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med*. 206:1253–1272.
- Guindon S, Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.
- Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*. 15:910–917.
- Heckerman D, Kadie C, Listgarten J. 2006. Leveraging information across HLA alleles/supertypes improves epitope prediction. *J Comput Biol*. 14:736–746.
- Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc*. 90:773–795.
- Kawashima Y, Pfafferott K, Frater J, et al. (43 co-authors). 2009. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*. 458:641–645.
- Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 21:676–679.
- Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. 22:2375–2385.
- Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics*. 22:3096–3098.
- Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M. 2005. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol*. 35:2295–2303.
- Leslie A, Kavanagh D, Honeyborne I, et al. (27 co-authors). 2005. Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *J Exp Med*. 201:891–902.
- Marsh SGE, Parham P, Barber LD. 2000. The HLA factsbook. San Diego (CA): Academic Press.
- Navis M, Matas DE, Rachinger A, Koning FA, van Swieten P, Kootstra NA, Schuitemaker H. 2008. Molecular evolution of human immunodeficiency virus type 1 upon transmission between human leukocyte antigen disparate donor-recipient pairs. *PLoS One*. 3:e2422.
- Navis M, Schellens I, van Baarle D, Borghans J, van Swieten P, Miedema F, Kootstra N, Schuitemaker H. 2007. Viral replication capacity as a correlate of HLA-B57/B5801-associated nonprogressive HIV-1 infection. *J Immunol*. 179:3133–3143.
- Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*. 12:1007–1017.
- Peters HO, Mendoza MG, Capina RE, et al. (13 co-authors). 2008. An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 Gag in the Pumwani Sex Worker Cohort. *J Virol*. 82:1980–1992.
- R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
- Rabiner L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*. 77:257–286.
- Rognan D, Lauemøller SL, Holm A, Buus S, Tschinke V. 1999. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem*. 42:4650–4658.
- Rousseau CM, Daniels MG, Carlson JM, et al. (23 co-authors). 2008. HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. *J Virol*. 82:6434–6446.
- Schuler MM, Nastke M, Stevanović S. 2007. SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol*. 409:75–93.
- Siepel A, Haussler D. 2005. Phylogenetic hidden Markov models. In: Nielsen R, editor. *Statistical methods in molecular evolution*. New York: Springer. p. 325–351.
- Stewart-Jones GBE, Gillespie G, Overton IM, Kaul R, Roche P, McMichael AJ, Rowland-Jones S, Jones EY. 2005. Structures of three HIV-1 HLA-B*5703-peptide complexes and identification of related HLAs potentially associated with long-term nonprogression. *J Immunol*. 175:2459–2468.
- Vyas JM, van der Veen AG, Ploegh HL. 2008. The known unknowns of antigen processing and presentation. *Nat Rev Immunol*. 8: 607–618.
- Yanover C, Hertz T. 2005. Predicting protein-peptide binding affinity by learning peptide-peptide distance functions. In: Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner P, Waterman M, editors. *Research in Computational Molecular Biology, 9th Annual International Conference, RECOMB 2005*. Cambridge, MA: Springer. p. 456–471.
- Yu K, Petrovsky N, Schonbach C, Koh JLY, Brusica V. 2002. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med*. 8:137–148.
- Yu XG, Lichterfeld M, Chetty S, et al. (26 co-authors). 2007. Mutually exclusive T-cell receptor induction and differential susceptibility to human immunodeficiency virus type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes. *J Virol*. 81:1619–1631.
- Yusim K, Korber BTM, Brander C, Haynes BF, Koup R, Moore JP, Walker BD, Watkins DI. 2009. HIV Molecular Immunology 2009. Los Alamos, NM: Los Alamos National Laboratory, Theoretical Biology and Biophysics.
- Zhao Y, Pinilla C, Valmori D, Martin R, Simon R. 2003. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*. 19:1978–1984.
- Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD dissertation]. [Austin (TX)]: University of Texas at Austin.