Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Detecting selection in DNA sequences: Bayesian Modelling and Inference

Daniel Merl & Raquel Prado
University of California Santa Cruz, USA
dmerl@ams.ucsc.edu, raquel@ams.ucsc.edu

Summary

Recent developments in Bayesian modelling of DNA sequence data for detecting natural selection at the amino acid level are presented. This article summarizes and discusses empirical model-based approaches. Key features of the modelling framework include the incorporation of biologically meaningful information via structured priors, posterior detection of sites under selection, and model validation via posterior predictive checks and/or estimation of gene and species trees. In addition, model selection is handled using a minimum posterior predictive loss criterion. The models presented here can incorporate relevant covariates such as amino acid properties, extending in this way previous approaches. Applications include the analysis of two DNA sequence alignments with different characteristics in terms of evolutionary divergences among the sequences: an abalone sperm lysin alignment with a strong underlying phylogenetic structure and a low divergence sequence alignment encoding the Apical Membrane Antigen-1 (AMA-1) in the human *P.falciparum* malaria parasite.

Keywords and Phrases: Bayesian Generalized Linear Models; DNA Sequence Data; Natural Selection; Model Comparison; Structured Priors; Tree Estimation.

1. INTRODUCTION

Determining the effect of natural selection in DNA sequence data is a key subject in the areas of computational biology and population genetics. Several approaches have been developed in recent years to detect positive selection at the amino acid level. Relevant references include, among others, Goldman and Yang (1994); Muse and Gaut (1994); Nielsen and Yang (1998); Suzuki and Gojorobi (1999); Yang et al. (2000); Huelsenbeck and Dyer (2004); Suzuki (2004); Yang et al. (2005); Kosakovsky Pond and Frost (2005a, 2005b). All these approaches have focused on analyzing "phylogenetic data", i.e., data in which each sequence in the alignment represents a unique species. When several sequences representing different

Daniel Merl is PhD candidate and Raquel Prado is Assistant Professor, Applied Mathematics and Statistics, Baskin School of Engineering, University of California Santa Cruz.

individuals from one or more populations of the same species are considered, the approaches mentioned above may produce unreasonable results due to the lack of evolutionary divergence among the sequences. In this paper we summarize and discuss recent modelling approaches specifically designed to analyze this latter type of data, referred to as "polymorphic data". This model-based methodology permits the incorporation of biologically meaningful prior information, while simultaneously allowing maximum flexibility in modelling substitution rates at the amino acid level. We therefore extend previous approaches presented in Prado et al. (2006) in order to include the following features: incorporation of relevant covariates, such as amino acid properties, and clustering functions on model parameters that can describe population, or geography specific effects. Additionally, we show how phylogenetic posterior estimation and posterior predictive checks can be used as model validation tools.

Section 2 summarizes the biological terminology that will be used throughout the paper and describes the models. Section 3 discusses different ways of identifying positively and negatively selected amino acid sites. The definitions are based on posterior distributions of the model parameters. In addition, a description of how to obtain estimates of phylogenies and gene trees is included. Although the main objective of the methodology presented here is detection of amino acid sites under selection, estimates of phylogenies — even if such estimates may only describe crude topological features underlying the sequence alignments — can be useful as model checking tools. In Section 4, model comparison and model validation procedures are discussed. Section 5 illustrates various aspects of the models and methodology in the analyses of two data sets with different evolutionary characteristics. First, analyses of an alignment coding a 122 residue region of the sperm lysin protein in 25 species of California abalone are presented. This data set has been extensively studied and it is considered to be a good example of how positive selection can act on individual amino acid sites. The alignment also displays a relatively strong phylogenetic signal. Then, analyses of sequences encoding AMA-1, a candidate antigen for malaria vaccine development, are presented. This data set is in some ways orthogonal to the lysin data set, as it consists of multiple sequences encoding AMA-1 within a single species, the human *P.falciparum* malaria parasite. Because of this the sequences display relatively low evolutionary divergence. Finally, Section 5 concludes with a summary of remarks, as well as current and future directions for research.

2. MODELS

We begin by summarizing some biological concepts that will be used throughout the paper.

Codon. This is the term given to a three nucleotide sequence codifying one of the 20 amino acids that serve as the building blocks of proteins. The universal genetic code has 64 possible codons of which 61 encode amino acids and the remaining 3 are stop codons, designating the end of the DNA transcription into RNA.

Synonymous and non-synonymous substitutions. Given that there are 20 amino acids and 61 possible codons, multiple condons codify the same amino acid. Synonymous substitutions are those between codons that specify the same amino acid; e.g., a substitution of TTA for CTC is a synonymous one since both codons encode the amino acid Leucine (L). Non-synonymous substitutions are those between codons specifying different amino acids; e.g., a substitution of TTC, encoding Phenylalanine

(F), for TTA (L).

Neutral, negative and positive selection. Synonymous substitutions are expected to be neutral since they do not affect the amino acid composition of proteins. Non-synonymous substitutions may have negative effects on the protein function and so, they are expected to be eliminated by negative selection. In the event that such substitutions are selectively favorable, the frequency of the gene containing the new amino acid is increased until it becomes fixed in the population. This process is known as positive natural selection or adaptive evolution.

Transitions and transversions. Transitions are nucleotide substitutions between purines (Adenine (A) and Guanine (G)) or between pyrimidines (Cytosine (C) and Thymine (T)). Transversions include all the other nucleotide substitutions.

Phylogeny and genealogy. Phylogenies are evolutionary trees that describe the pattern of divergences by which a single common ancestral sequence evolved, over time, into the descendant sequences comprising a given alignment. Phylogenies are used to represent evolutionary relationships among species using sequence data in which each sequence represents a species (phylogenetic data). If the sequences are from different individuals of the same species (polymorphic data), the information is genealogical and so, genealogies or gene trees can be used to show which sequences are most closely related.

In order to describe the general model formulation we follow the notation of Prado et al. (2006). Specifically, let $\mathbf{Y}$ denote the sequence alignment consisting of N sequences with $3 \times I$ nucleotides (i.e., I codons). Let $\mathbf{Z}$ denote the substitution count data obtained from $\mathbf{Y}$ as follows. Define $\mathbf{y}_{i,j}$ to be the pair of homologous codons at site i, $i = 1, \ldots, I$, for the sequence pair indexed by j, with $j = 1, \ldots, J$ and $J = \binom{N}{2}$, the total number of possible pairs of sequences. Typically, only the polymorphic sites, i.e., only those sites i that display at least one substitution in one pair of sequences, are included in the model. $\mathbf{Z}$ can be obtained in a number of ways, depending on how many types of substitutions will be represented in the model, and depending on whether or not phylogenetic or genealogical information will be used. For instance, Merl et al. (2005) used phylogenies to estimate ancestral nucleotide sequences, and then used the reconstructed sequences to count the total number of substitutions between two codons using a method similar to that proposed in Kosakovsky Pond and Frost (2005b). Prado et al. (2006) averaged the different numbers of substitutions per site over all possible one-substitution pathways that could have been followed between any two codons without allowing back-substitutions, self-canceling loops, and eliminating the pathways including stop codons. Regardless of which methodology is used to obtain $\mathbf{Z}$, $\mathbf{z}_{i,j}$ is a K-dimensional vector of counts, where each component, $z_{k,i,j}$, represents the number of substitutions of type k at site i between the two codons in the pairwise sequence comparison indexed by j.

For each $\mathbf{z}_{i,j}$ we define $\boldsymbol{\theta}_{i,j}$, a K-dimensional vector where each component, $\theta_{k,i,j}$, denotes the probability of substitution type k for site i and pair j. The model is then described by

$$\begin{aligned} \mathbf{z}_{i,j} &\sim \text{Multinomial}(n_{i,j}, \boldsymbol{\theta}_{i,j}), \quad n_{i,j} = \sum_k z_{k,i,j}, \\ \theta_{k,i,j} &= \frac{\exp(\eta_{k,i,j})}{\sum_l \exp(\eta_{l,i,j})}, \end{aligned} \tag{1}$$

$$\eta_{k,i,j} \quad = \quad \alpha_k + \beta_{k,i} + \gamma_{k,h(j)} + \boldsymbol{\delta}'_{k,i}\mathbf{x}_{i,j}.$$

The K-dimensional parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$ models the baseline effects for each substitution type. The K-dimensional vectors $\boldsymbol{\beta}_i = (\beta_{1,i}, \ldots, \beta_{K,i})'$ capture site-specific departures from baseline substitution effects, i.e., they account for different strengths of selective pressures expressed for each substitution type, for each amino acid site i. The K-dimensional vectors $\boldsymbol{\gamma}_{h(j)} = (\gamma_{1,h(j)}, \ldots, \gamma_{K,h(j)})'$ describe pairwise, or more generally groupwise, departures from baseline and site-specific effects. In the particular case of $h(j) = j$, $\boldsymbol{\gamma}_{h(j)}$ models pairwise effects due to evolutionary divergence between the two sequences indexed by j. These parameters would be associated with phylogenetic or gene tree effects. Other choices of $h(j)$ will be discussed later. Finally, the $\mathbf{x}_{i,j}$'s are C-dimensional vectors of covariates and the $\boldsymbol{\delta}_{k,i}$'s are C-dimensional parameter vectors. Covariates can include measures of amino acid properties such as polarity and hydrophobicity, or amino acid score matrices that measure distances between amino acids in terms of various properties (e.g., the Grantham matrix, see Grantham, 1974). The inclusion of these covariates may be useful in determining whether very radical substitutions are being encouraged by natural selection.

2.1. *Sub-Models*

Choosing K, the number of categories. Models with $K = 3$, where only synonymous, non-synonymous and no-change categories are considered, are useful in identifying amino acid sites under positive selection. Models with $K = 5$ in which synonymous transitions and transversions, non-synonymous transitions and transversions and no-substitutions are considered, are also used to detect sites under selection in sequences for which discriminating between transitions and transversions is key to determine if the observed rates of substitutions may be the result of codon bias. Codon bias is the tendency for a species to use a given set of codons more than others to encode a particular amino acid. Prado et al. (2006) used a 5-category model to analyze alignments of the AMA-1 antigen from the *P.falciparum* human malaria parasite. Accounting for transitions and transversions in these data is important to assess whether the increased rates of non-synonymous substitutions estimated at some amino acid sites are the result of the A+T richness in the genome (Escalante et al., 1998). Other model possibilities that would be useful for this purpose include, for example, $K = 4$, with a single category for synonymous substitutions and two categories of non-synonymous substitutions — e.g., those between G and C nucleotides, and all remaining substitutions — and a no-substitution category.

Pairwise and group effects. Models with $h(j) = j$ for all j specifically include all possible pairwise effects. These effects are often relevant in the analysis of phylogenetic sequences, i.e., in alignments for which the sequences were obtained from distinct species. This approach is followed in the analyses of the sperm lysin sequences presented here. In cases of polymorphic data, more parsimonious choices can be considered. For instance, the AMA-1 alignment analyzed in Prado et al. (2006), and revisited here, consists of 23 sequences in total, with 12 isolates from Kenya, 5 isolates from India, and 6 isolates from Thailand. In this case, it may be reasonable to assume that we have samples from two distinct populations: an African population represented by the 12 sequences from Kenya, and an Asian population represented by the 11 sequences from India and Thailand. Then, we can write $h(j) = o$, for $o = 1, 2, 3$ defined in such a way that $h(j) = 1$ if j indexes a pair of sequences from Africa, $h(j) = 2$ if j indexes a pair of sequences from Asia, and

$h(j) = 3$ if j indexes a pair of sequences from different populations, i.e., one from Africa and the other one from Asia. Another potentially interesting model is that in which $\boldsymbol{\gamma}_{h(j)} = \mathbf{0}$ for all j. This model does not include any group effect (pairwise or other type) and it is sometimes useful in modelling sequences for which the genetic variability is mostly explained by the site-specific effects and so, phylogenetic or population effects are considered negligible.

Covariates. Sub-models include those with $\boldsymbol{\delta}_{k,i} = \boldsymbol{\delta}_k$ for all i, models with $\boldsymbol{\delta}_{k,i} = \boldsymbol{\delta}_i$ for all k, and those with $\boldsymbol{\delta}_{k,i} = \boldsymbol{\delta}$ for all k, i. Note that the inclusion of covariates that carry information on amino acid properties only affects non-synonymous substitutions and so, $\boldsymbol{\delta}_{k,i} = \mathbf{0}$ for all k indexing a synonymous or no-substitution category in any model that includes covariates.

2.2. *Prior Structure*

The prior structure used here is similar to that proposed in Prado et al. (2006). We now summarize the main features of such structure, directing reader to the previous reference for details on the prior construction and a discussion about its biological implications.

The prior distributions on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are all Gaussian. This is, $N(\boldsymbol{\alpha}|\mathbf{m}_\alpha, \sigma_\alpha^2\mathbf{I})$, $N(\boldsymbol{\beta}_k|\mathbf{m}_{\beta_k}, \sigma_\beta^2\mathbf{I})$ and $N(\boldsymbol{\gamma}_k|\mathbf{m}_{\gamma_k}, \sigma_\gamma^2\mathbf{W}_{\gamma_k})$, where $\mathbf{I}$ denotes the identity matrix with the appropriate dimension in each case. In addition, constraints on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are set to guarantee identifiability (see Prado et al., 2006).

In absence of prior information about site-specific effects we set $\mathbf{m}_{\beta_k} = \mathbf{0}$ for all k. When phylogenetic/genealogical information is available, it may be incorporated in the prior structure by first translating the phylogeny/genealogy into a matrix of distances $\mathbf{D}$, as well as a matrix of "distances between distances" $\tilde{\mathbf{D}}$ (see Prado et. al., 2006). Then, $\mathbf{m}_\alpha$, $\mathbf{m}_{\gamma_k}$ and $\mathbf{W}_{\gamma_k}$, are expressed as functions of the elements of $\mathbf{D}$, $\tilde{\mathbf{D}}$ and hyper-parameters $\mathbf{w}$, χ, χ^*, ζ and ζ^*. Here $\mathbf{w}$ is a vector of dimension $K-1$ containing prior estimates for the relative frequencies of the first $K-1$ substitution types, while χ, χ^*, ζ and ζ^* are used to control the strength of the phylogenetic/genealogy effects in the prior structure. These hyper-parameters can be given fixed values a priori or estimated a posteriori. In the latter case, the hyper-parameters are assumed independent a priori with χ and χ^* following uniform priors and ζ, ζ^*, as well as each element of $\mathbf{w}$ following exponential priors. In addition, the parameters σ_α^2, σ_β^2 and σ_γ^2 can be fixed a priori or estimated a posteriori under inverse-gamma priors. Alternatively, if the alignment includes sequences with relatively low evolutionary divergence, for which none or only very weak phylogenetic information is available, $\mathbf{m}_{\gamma_k} = \mathbf{0}$ and $\mathbf{W}_{\gamma_k} = \mathbf{I}$ are typically used.

Finally, Gaussian priors are also specified for the C-dimensional parameter vectors $\boldsymbol{\delta}_{k,i}$. Specifically, $N(\boldsymbol{\delta}_{k,i}|\mathbf{0}, \sigma_\delta^2\mathbf{I})$ are used for all the categories k that model non-synonymous substitutions and all polymorphic sites indexed by i. As it was the case with other variance parameters, σ_δ^2 can be set to some fixed value or estimated a posteriori under an inverse-gamma prior.

3. POSTERIOR ESTIMATION

Posterior estimation is achieved via standard MCMC methods using the Poisson formulation of the multinomial model (see Baker, 1994).

3.1. *Identifying Sites Under Selection*

Once samples from the posterior distribution of the model parameters are obtained, sites under negative or positive selection can be identified by investigating the behavior of specific functions of such parameters.

3.1.1. *Definitions based on* $\boldsymbol{\theta}$

Let $\mathcal{I}^*$ be a specific set of sites in the alignment. For instance, $\mathcal{I}^*$ can be the set of all the polymorphic sites in the alignment, or the set of all the sites that display at least one non-synonymous substitution. Let $\theta_S^{\mathcal{I}^*}$ and $\theta_{NS}^{\mathcal{I}^*}$ be the average synonymous and non-synonymous substitution probabilities, respectively, for the sites in $\mathcal{I}^*$. This is,

$$\theta_S^{\mathcal{I}^*} = \frac{1}{(|\mathcal{I}^*| \times J)} \sum_{i \in \mathcal{I}^*} \sum_{j=1:J} \sum_{l \in \mathcal{C}_S} \theta_{l,i,j}, \text{ and } \theta_{NS}^{\mathcal{I}^*} = \frac{1}{(|\mathcal{I}^*| \times J)} \sum_{i \in \mathcal{I}^*} \sum_{j=1:J} \sum_{l \in \mathcal{C}_{NS}} \theta_{l,i,j},$$

where $\mathcal{C}_S$ and $\mathcal{C}_{NS}$ are the sets of indexes of all the categories that involve synonymous and non-synonymous substitutions, respectively. Then, we say that there is evidence of positive selection in the gene if

$$P(\omega^* > \omega_0 | \mathbf{Z}) \equiv P\left(\frac{\theta_{NS}^{\mathcal{I}^*}}{\theta_S^{\mathcal{I}^*}} > \omega_0 \middle| \mathbf{Z} \right) \geq (1 - \alpha_1), \tag{2}$$

with $\alpha_1 \in [0, 1)$ and typically, $\alpha_1 \in [0, 0.05]$. The value of ω_0 is fixed and often set at $\omega_0 = 1$. A non-synonymous to synonymous substitution probabilities ratio equal to one is considered indicative of neutral selection, ratios smaller than one indicate negative selection, while ratios greater than one are linked to positive selection (e.g., Yang et al., 2005). When (2) holds, we can proceed to identify sites under positive selection. Specifically, we say that a site i is a positively selected site if

$$P(i^+ | \mathbf{Z}) \equiv P\left(\frac{\theta_{NS,i}}{\theta_S^{\mathcal{I}^*}} > \frac{\theta_{NS}^{\mathcal{I}^*}}{\theta_S^{\mathcal{I}^*}} \middle| \mathbf{Z} \right) = P(\theta_{NS,i} > \theta_{NS}^{\mathcal{I}^*} | \mathbf{Z}) \geq (1 - \alpha_2), \tag{3}$$

with $\alpha_2 \in [0, 1)$ and usually $\alpha_2 \leq 0.05$. If (2) does not hold then we say that the alignment is under neutral and/or negative selection.

Sites under negative selection can also be identified. Specifically, a site i is said to be under negative selection if

$$P\left(\frac{\theta_S^{\mathcal{I}^*}}{\theta_{NS}^{\mathcal{I}^*}} > \frac{1}{\omega_1} \middle| \mathbf{Z} \right) \geq (1 - \alpha_4), \tag{4}$$

and if

$$P(i^- | \mathbf{Z}) \equiv P\left(\frac{\theta_{S,i}}{\theta_{NS}^{\mathcal{I}^*}} > \frac{\theta_S^{\mathcal{I}^*}}{\theta_{NS}^{\mathcal{I}^*}} \middle| \mathbf{Z} \right) = P\left(\theta_{S,i} > \theta_S^{\mathcal{I}^*} | \mathbf{Z} \right) \geq (1 - \alpha_3), \tag{5}$$

with $\alpha_3, \alpha_4 \in [0, 1)$ and typically α_3, $\alpha_4 \in [0, 0.05]$. The value of ω_1 is set by the practitioner. For instance, if sites under very strong negative selection must be identified, $1/\omega_1$ is fixed at a value greater than 2.0.

3.1.2. *A definition of positive selection based on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$*

Prado et al. (2006) considers another definition for detecting sites under positive selection by writing $P(i^+|\mathbf{Z})$ in terms of the $\boldsymbol{\beta}$ parameters. In general, it has been found that such definition is more conservative that the definition in (3). Simulation studies suggest that a definition based on $\boldsymbol{\beta}$ produces less false positives but also has less power than the definition in (3) (see Prado et al., 2006).

Here, we extend the definition of Prado et al. (2006) to account for possible covariates added to the model. Once again, we first determine whether there is evidence of positive selection in the alignment by looking at $P(\omega^* > \omega_0|\mathbf{Z})$. In other words, if (2) holds for some fixed values ω_0 and α_1, we then proceed to identify which sites are under positive selection. Then, a site i is said to be under positive selection if

$$P(i^+|\mathbf{Z}) \equiv P\left(\beta^*_{NS,i} + f(\boldsymbol{\delta}'_{NS,i}\mathbf{x}_i) > \beta^*_{S,i}|\mathbf{Z}\right) \geq (1-\alpha_2), \tag{6}$$

with

$$\beta^*_{k,i} = \beta_{k,i} - \frac{1}{|\mathcal{I}^*|}\sum_{i\in\mathcal{I}^*}\beta_{k,i}, \quad \text{and} \quad f(\boldsymbol{\delta}'_{NS,i}\mathbf{x}_i) = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\delta}'_{NS,i}\mathbf{x}_{i,j}.$$

3.2. *Phylogenetic and Gene Tree Estimation*

Although the models considered here were not developed for the purpose of phylogenetic/genealogic inference, it is possible to obtain posterior estimates of phylogenies or gene trees based on posterior distance matrices as follows.

Let $d_{h(j)}$ be an estimate of the distance between the pair of sequences indexed by j computed as follows

$$d_{h(j)} = \frac{1}{|\mathcal{I}^*|}\sum_{i\in\mathcal{I}^*}\sum_{k\in\mathcal{C}}\theta_{k,i,h(j)}, \tag{7}$$

with $\mathcal{C}$ including a particular set of substitution categories. For example, $\mathcal{C}$ can be the set containing all the substitution types (e.g., synonymous and non-synonymous), only the synonymous substitutions, or only the non-synonymous substitutions. Here, $\theta_{k,i,o}$, for a given o, is computed as

$$\theta_{k,i,o} = \frac{1}{|o|}\sum_{j;h(j)=o}\theta_{k,i,h(j)},$$

where $|o|$ is the number of indexes j such that $h(j) = o$. The distances (pseudo-distances) in (7) can be used to build a matrix $\mathbf{D}_h$ whose dimension depends on the structure defined by the function $h(j)$. For example, if $h(j) = j$ for all pairs indexed by j, the matrix $\mathbf{D}_h$ will have dimension $N \times N$, with off diagonal elements computed as $g(d_j)$, with g a particular function, such as the identity or the exponential. The diagonal elements of $\mathbf{D}_h$ can be computed using $g(0)$ (the distance between a sequence and itself is zero). In the example of the malaria sequences from two populations discussed above, $h(j) = o$ with $o = 1, 2, 3$, and so, $\mathbf{D}_h$ would be a 2×2 matrix. The off-diagonal elements would measure the average distances between populations (or a function of such distances), while the diagonal entries would contain average within population distances (or a function of such distances).

The matrix $\mathbf{D}_h$ can then be used as an input to one of the distance-based algorithms often used in practice to estimate phylogenies or gene trees such as the neighbor joining algorithm (see for example Felsenstein, 2004 for a detailed explanation of this and other related algorithms for phylogenetic estimation). Therefore, posterior estimates/samples of genealogies based on the distances defined in (7) can be obtained.

4. MODEL SELECTION AND MODEL VALIDATION

4.1. *Model Selection*

Model selection among different models, such as the various sub-models discussed in Section 2, is performed via the minimum posterior predictive loss approach of Gelfand and Ghosh (1998). This criterion can be computed easily using MCMC output and it has a decision theoretical justification given that it is obtained by minimizing a posterior predictive loss function within a particular family of models, and then, selecting the model that minimizes such criterion.

For each model $\mathcal{M}_m$ from a collection of M models, $\mathcal{M}_1, \ldots, \mathcal{M}_M$, the following quantity is computed,

$$D_\kappa(m) = \sum_{i,j} \min_{\mathbf{a}_{i,j}} \left\{ E_{\mathbf{z}_{i,j}^{rep}|\mathbf{Z}^{obs},m} \left[L(\mathbf{z}_{i,j}^{rep}, \mathbf{a}_{i,j}) + \kappa L(\mathbf{z}_{i,j}^{obs}, \mathbf{a}_{i,j}) \right] \right\}, \quad \kappa \geq 0, \tag{8}$$

where $\mathbf{Z}^{obs} \equiv \mathbf{Z}$ are the observed count data, $\mathbf{z}_{i,j}^{rep}$ is a K-dimensional vector of counts that replicates $\mathbf{z}_{i,j}^{obs}$, $L(\cdot,\cdot)$ is a loss function, $\mathbf{a}_{i,j}$ is a "guess", representing a compromise between $\mathbf{z}_{i,j}^{rep}$ and $\mathbf{z}_{i,j}^{obs}$ and κ is a constant that weights the discrepancy between $\mathbf{a}_{i,j}$ and $\mathbf{z}_{i,j}^{obs}$. In other words, when $\kappa = 0$, $\mathbf{a}_{i,j}$ is chosen as a guess for $\mathbf{z}_{i,j}^{rep}$ and if $\kappa \neq 0$ the closeness of $\mathbf{a}_{i,j}$ to $\mathbf{z}_{i,j}^{obs}$ is also rewarded, and so, a compromised choice is required.

Various loss functions can be considered. Prado et al. (2006) computed (8) for the model formulation (2) with five categories using two loss functions: a squared error loss function and a loss function written in terms of the logarithm of a ratio of likelihoods, i.e.,

$$L(\mathbf{z}_{i,j}, \mathbf{a}_{i,j}) = 2 \log \frac{f(\mathbf{z}_{i,j}|\mathbf{q}(\mathbf{z}_{i,j}))}{f(\mathbf{z}_{i,j}|\mathbf{q}(\mathbf{a}_{i,j}))},$$

for some function $\mathbf{q}$. This loss function takes into account the GLM structure of the model. Details regarding the specific form of $\mathbf{q}$ and the calculation of $D_\kappa(m)$ appear in Prado et al. (2006).

4.2. *Model Checking*

We follow a posterior predictive approach to model checking. After obtaining R posterior samples of the K-dimensional probability vectors $\boldsymbol{\theta}_{i,j}$, for each i, j, i.e., $\boldsymbol{\theta}_{i,j}^{(r)}$ for $r = 1, \ldots, R$, we can obtain R replicates $\mathbf{z}_{i,j}^{rep}$ for each i, j. This is

$$\mathbf{z}_{i,j}^{rep,r} \sim \text{Multinomial}(n_{i,j}, \boldsymbol{\theta}_{i,j}^{(r)}). \tag{9}$$

Then, we can summarize posterior distributions of relevant functions of $\mathbf{z}_{i,j}^{rep}$ and compare them with functions of the actual count values $\mathbf{z}_{i,j}$. For instance, we

could derive the distribution of the number of transitions, transversions, synonymous and/or non-synonymous substitutions based on the replicates $\mathbf{Z}^{rep}$ for all, or a few, of the sites indexed by i, and determine whether the corresponding observed values are plausible values under such distributions.

In order to appropriately simulate $\mathbf{z}_{i,j}^{rep}$ in (9), we need to take into account the process used to obtain the count data $\mathbf{Z}$ from the alignment. For instance, when the procedure described in Prado et al. (2006) is followed to obtain $\mathbf{Z}$ in a 5-category model for which the categories are synonymous transitions and transversions, non-synonymous transitions and transversions and no substitutions, the resulting $z_{5,i,j}$ is a binary entry. In this case $\mathbf{z}_{i,j}^{rep,r}$ is simulated as follows. First $z_{5,i,j}$ is simulated from a Bernoulli distribution with probability $\boldsymbol{\theta}_{5,i,j}^{(r)}$. Then, $\mathbf{z}_{1:4,i,j}^{rep,r}$ is simulated from a multinomial distribution with parameters $(n_{i,j} - z_{5,i,j}^{rep,r})$ and $\boldsymbol{\theta}_{1:4,i,j}^{*,(r)}$, with $\theta_{k,i,j}^{*,(r)} = \theta_{k,i,j}^{(r)} / \sum_{k=1}^{4} \theta_{k,i,j}^{(r)}$. This will be illustrated in Section 5.

Another way of assessing model fit in phylogenetic data is by looking at the phylogenies obtained from the posterior estimates (or posterior samples) of the distance matrix $\mathbf{D}_h$. This can only be done if no phylogenetic structure has been included in the prior specification. Close inspection of posterior tree estimates/samples can then be used as a tool for model validation. We can compare such estimates to substantive knowledge about the evolutionary process underlying the sequences whenever such information is available.

5. DATA ANALYSES

5.1. *Abalone Sperm Lysin Alignment*

This alignment codes for a 122 residue region of the sperm lysin protein for 25 species of California abalone. Abalone reproduction involves species specific sperm-egg recognition in which the sperm lysin binds and dissolves a complementary vitelline envelope (VE) surrounding the egg cell. This species-specific interaction is subjected to positive selection of some 23 residues in the lysin protein, as it compensates for ongoing genetic drift in the VE receptor (Galindo et al, 2003; Lee et al, 1995; Yang et. al., 2000). These data have been extensively studied and provide a good example of positive selection acting on individual amino acid sites. The data are included in the `PAML` software distribution (Yang, 1997). The sequences are sufficiently divergent, with a total tree length of 8.2 nucleotide substitutions per codon. Additionally, the crystalline structure of the molecule can be used to support or refute claims of positively selected amino acid sites (Yang et al., 2000).

Various models were fit to the count data obtained from the alignment using the procedure described in Prado et al. (2006). We focus on the results drawn from two 3-category models: a model with a prior specification that includes the phylogenetic structure shown in Figure 3, and another model where such structure is not incorporated. The three categories correspond, respectively, to synonymous substitutions, non-synonymous substitutions and no substitutions. The phylogeny in Figure 3 is that of Lee et al. (1995). For both models, the results presented here are based on 1,000 MCMC samples obtained after convergence. Also, both models incorporate a single covariate $x_{i,j}$, where $x_{i,j}$ corresponds to the normalized Grantham matrix score (see Grantham, 1974) between the two amino acids indexed in the pairwise comparison j at site i. For each amino acid substitution, the Grantham score represents a physicochemical distance between the two amino acids involved in such substitution. The normalization of the matrix involved dividing each entry in the

matrix by the maximum score value.

The phylogenetic prior was specified using the procedure described in Prado et al. (2006), taking into account that we are fitting a 3-category model instead of a 5-category one, and setting $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\gamma^2 = 10$, $\sigma_\delta^2 = 1$, $\mathbf{m}_\beta = \mathbf{m}_\delta = \mathbf{0}$, and the hyperparameters needed to define $\mathbf{m}_\alpha$, $\mathbf{m}_{\gamma_k}$ and $\mathbf{W}_{\gamma_k}$ to $\mathbf{w} = (1, 1)$, $\chi = \chi^* = 1$, $\zeta = 1$ and $\zeta^* = 1$. The independent prior was specified by setting $\mathbf{m}_\alpha = \mathbf{m}_\beta = \mathbf{0}$, $\mathbf{m}_\gamma = \mathbf{m}_\delta = \mathbf{0}$, $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\gamma^2 = 10$ and $\sigma_\delta^2 = 1$, and all the prior variance-covariance matrices equal to identities. Posterior results, in terms of which sites were detected as sites under selection, were not very sensitive to changes in the prior values of these parameters.

Table 1 shows various model selection criteria values for the two types of models fitted to the count data. Three criteria were considered, two of which correspond to the posterior predictive criteria of Gelfand and Ghosh (1998) under the log-likelihood ratio and the squared error loss functions, denoted by D_{dev}^{κ} and D_{se}^{κ}, respectively. The values displayed on the table correspond to $\kappa = 100$. Other values of κ were considered, leading to the same conclusions. The deviance information criteria (DIC), was also computed for both models (Spiegelhalter et al., 2002). Based on these values, the model with structured phylogenetic priors is preferred by all the criteria.

Table 1: *Model selection criteria. Lysin alignment.*

Model	D_{dev}^{100}	D_{se}^{100}	DIC
Independent	65413	37558	65142
Phylogenetic	65349	37495	64929

Even though the preferred model, according to the model selection criteria discussed above, is the one with phylogenetic priors, we now focus on the results obtained from the model with independent priors in order to illustrate some modelling features, particularly those related to model validation. Figure 1 shows the posterior distributions of the non-synonymous to synonymous probabilities ratios for the sites identified as positively selected by both of the GLM-based definitions described in Section 3. The dark boxplots correspond to sites that were also identified as positively selected by at least one of the methods implemented in `PAML` (Yang, 1997) or in `HYPHY` (Kosakovsky Pond et al, 2005). Specifically, model M8b in `PAML` was fitted to the alignment. This is a model in which a discretized beta distribution is used to describe the non-synonymous to synonymous rates ratio –denoted by ω in the computational biology literature– between 0 and 1, and an additional category with $\omega > 1$. Three different methodologies were considered in `HYPHY`: the single ancestor counting method (SLAC), the fixed-effects likelihood method (FEL) and the random-effects likelihood method (REL). Details about how these models are used to detect sites under positive selection can be found in Kosakovsky Pond and Frost (2005a, 2005b). Figure 1 shows that there is good agreement between the proposed GLM-based methodology and the existing methodologies based on stochastic models of molecular evolution and phylogenetic-based models implemented in `PAML` and `HYPHY`.

Figure 2 displays the lysin crystal structure for one of the abalone species in

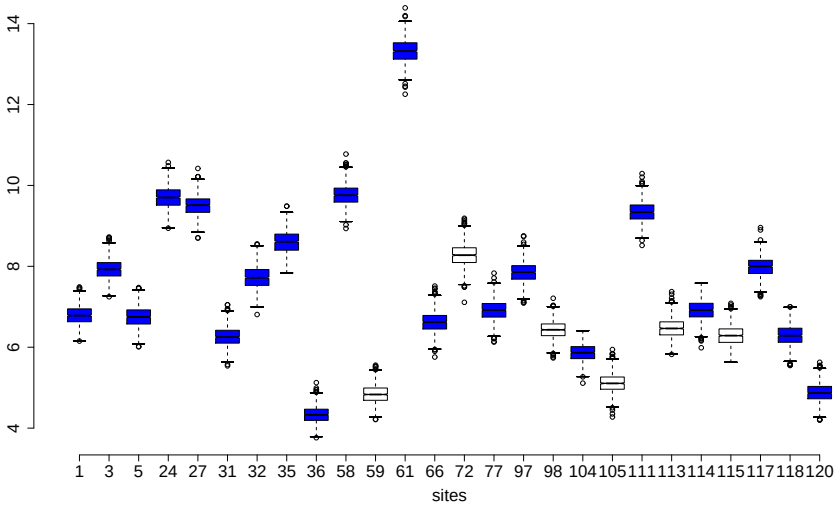

Figure 1: *Posterior distributions of the non-synonymous to synonymous probabilities ratios for the sites positively selected by GLM-based definitions. Dark boxplots indicate that those sites were also identified by at least one of the* `PAML` *or* `HYPHY` *methods.*

the alignment (*H. Rufenses*). This is the same crystal structure shown in Yang et al. (2000). The sites positively identified by the GLM-based methodology do not exactly match those identified by model M8b in `PAML`, however, as shown in Figure 2, there is a great deal of agreement in terms of the locations of positively selected sites in the folded protein. Figure 2 shows that the GLM positively selected sites cluster at the top and at the bottom of the molecule. These findings are in agreement with the results presented in Yang et al. (2000). Most of the conserved sites lie in the internal portions of the alpha helices of the protein. Such sites are involved in interhelical interactions and are functionally constrained in all lysins.

Figure 4 shows an estimated phylogeny constructed using the posterior mean estimate of a distance matrix $\mathbf{D}_{h(j)} = \mathbf{D}_j$, whose entries are given by

$$d_j = \frac{1}{|\mathcal{I}^*|} \sum_{i \in \mathcal{I}^*} \theta_{1,i,j} + \theta_{2,i,j},$$

where $\theta_{1,i,j}$ and $\theta_{2,i,j}$ are the probabilities of synonymous and non-synonymous substitutions, respectively, for the pair of sequences indexed by j and the residue indexed by i. The estimated phylogeny displayed in the figure is based on posterior samples from the model with independent priors. A neighbor joining algorithm was

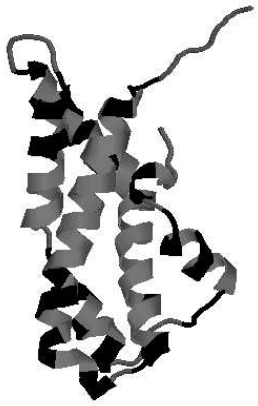

Figure 2: *Lysin crystal structure from the red abalone* H. Rufenses. *Sites identified as positively selected are in black.*

used for the tree construction. Although the phylogenies in figures 3 and 4 are not identical, they show many topological similarities, indicating that the GLM-based approach provides a good model fit. In particular, most of the species that appear clustered in Figure 3 are also clustered in the phylogeny displayed in Figure 4.

5.2. *AMA-1 Alignment*

Prado et al. (2006) presents analyses of alignments comprising 23 sequences of the AMA-1 antigen in the human *P.falciparum* malaria parasite. We consider additional analyses of this alignment here, extending the approach of Prado et al. (2006) in order to add covariates related to amino acid properties. In addition, a collection of models that make various assumptions about the underlying evolutionary characteristics of the sequences are fitted to the data, and compared via posterior predictive model selection criteria.

Malaria is a major public health problem, with approximately 300 to 500 million clinical cases and 1 to 3 million deaths estimated per year (Sachs and Malaney, 2002) and so, vaccine development against the parasites that produce the disease is a priority. AMA-1 is one of the candidate antigens currently being considered for use in vaccine development. AMA-1 has been extensively studied, and there is convincing evidence that this antigen elicits a protective immune response against malaria (Polley et al., 2004). In addition, some residues have been associated with

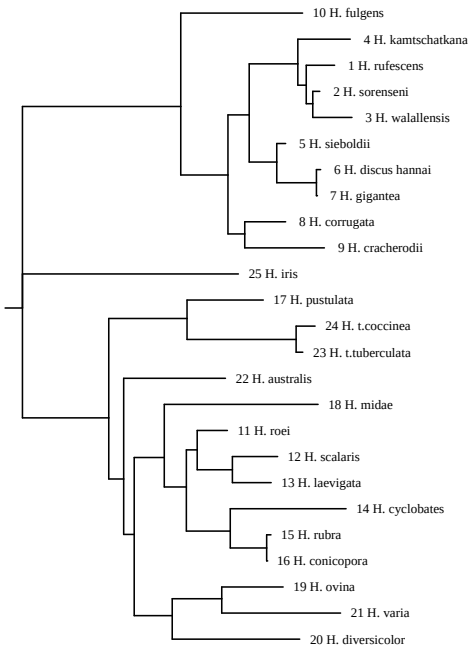

Figure 3: *Abalone sperm lysin phylogeny from Lee et al., 1995.*

various clinical manifestations of the disease (Cortes et al., 2003). Genetic evolutionary studies have also shown that there is evidence that the gene is under positive selection (Escalante et al., 2001; Polley et. al., 2003).

The alignment considered here consists of 23 sequences, each encoding a total number of 620 residues. From these 23 sequences, 12 sequences were taken from subjects in Kenya, 5 from subjects in India and 6 from subjects in Thailand. The sequences display 84 polymorphic sites and are available in GenBank. Prado et al. (2006) analyzed this alignment using 5-category models with synonymous transitions and transversions, non-synonymous transitions and transversions and no-substitutions. Here, we consider 3-category models accounting for synonymous substitutions, non-synonymous substitutions and no-substitutions. In addition, three types of models were chosen by considering three different choices of the function $h(j)$ that defines the groupwise effects among the sequences. First, we consider models with $h(j) = j$ for all j, i.e., models in which the $\boldsymbol{\gamma}$ parameters describe

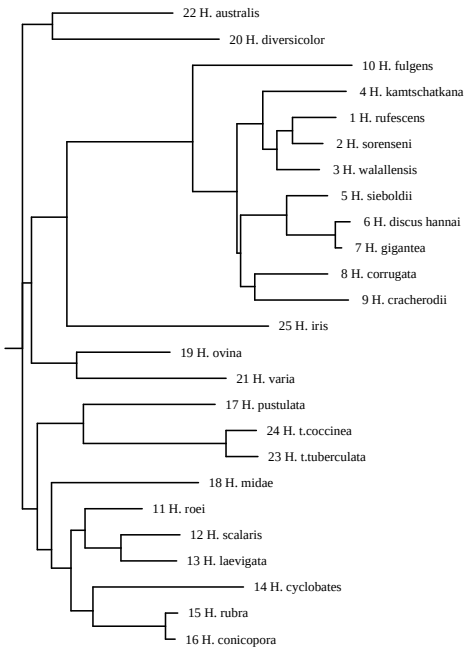

Figure 4: *Posterior estimation of an abalone sperm lysin phylogeny obtained from GLM approach.*

evolutionary distances among pairs of sequences. We refer to these models as "full" models. Prado et al. (2006) only fitted these types of models. Second, we consider models in which $h(j) = 1$ for all the pairs j in which both sequences were from Africa, $h(j) = 2$ for all the pairs j in which both sequences were from Asia and $h(j) = 3$ for all the pairs j in which one sequence was from Africa and the other one was from Asia. We refer to these models as "continent" models. Finally, we also fit models in which $h(j) = 1$ for all the pairs j in which both sequences are from Kenya, $h(j) = 2$ for all the pairs in which both sequences are from India, $h(j) = 3$ for all the pairs in which both sequences are from Thailand. Additionally, $h(j) = o$, with $o = 4, 5, 6$, for all the pairs j in which the sequences are from different countries (Kenya and India, Kenya and Thailand and India and Thailand, respectively). We refer to these models as the "country" models. For each of the three types of models described above we fit all possible combinations of covariate-specific and

site-specific effects, namely: models with and without a covariate that represents amino acid substitutions scores ($\boldsymbol{\delta}$) – again, the normalized Grantham matrix was used as a covariate – and models with and without the site-specific effects $\boldsymbol{\beta}$. Finally, models that do not include any group effects $\boldsymbol{\gamma}$ are also considered.

Table 2 shows various model selection criteria values for the all the model types described above, denoted by ($\boldsymbol{\alpha}$), ($\boldsymbol{\alpha}+\boldsymbol{\beta}$), ($\boldsymbol{\alpha}+\boldsymbol{\gamma}$), ($\boldsymbol{\alpha}+\boldsymbol{\delta}$), ($\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}$), ($\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\delta}$), ($\boldsymbol{\alpha}+\boldsymbol{\gamma}+\boldsymbol{\delta}$) and ($\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta}$). Notice that four of these models, ($\boldsymbol{\alpha}$), ($\boldsymbol{\alpha}+\boldsymbol{\beta}$), ($\boldsymbol{\alpha}+\boldsymbol{\delta}$) and ($\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\delta}$), are equivalent under any choice of $h(\cdot)$, since they do not involve a group-specific $\boldsymbol{\gamma}$ term. For purposes of model selection, we compare minimum posterior predictive deviance (D^{κ}_{dev}), minimum posterior predictive squared-error loss (D^{κ}_{se}), and the deviance information criterion (DIC). For D^{κ}_{dev} and D^{κ}_{se}, various values of κ were used, yielding virtually identical results. Values reported here were obtained using $\kappa = 100$. Superior model fit is indicated by smaller values of each model selection criterion. The value corresponding to the best-fitting model under each criterion is shown in bold. All model selection criteria indicate that the full model provides the best fit to the data. Adding the site-specific effects $\boldsymbol{\beta}$ to a given model produces the largest decrease in the model selection criteria values. For instance, compare the decrease in D^{100}_{dev}, D^{100}_{se} and DIC values for the models ($\boldsymbol{\alpha}+\boldsymbol{\beta}$) and ($\boldsymbol{\alpha}+\boldsymbol{\delta}$) with respect to ($\boldsymbol{\alpha}$), and the decrease in the criteria values of models ($\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}$) and ($\boldsymbol{\alpha}+\boldsymbol{\gamma}+\boldsymbol{\delta}$) with respect to ($\boldsymbol{\alpha}+\boldsymbol{\gamma}$) in the country, continent and full models. Adding the group-specific effects $\boldsymbol{\gamma}$ to a given model produces the smallest decrease in the model selection criteria values. In fact, the model selection criteria values for the continent and country models are virtually equivalent. These findings are important since they imply that the variables that have the largest effects in the subsitution probabilities are the site-specific effects, while those related to various assumptions on the evolutionary distances among the sequences have the least impact. This indicates that, if clustering of the sequences in terms of their evolutionary distances is feasible, such clustering would not be related to geographical location.

We now discuss some posterior results obtained from the full model with the Grantham scores covariate. Figure 5 displays the posterior predictive distributions of the number of synonymous substitutions $z_{34,1}$, the number non-synonymous substitutions $z_{34,2}$, and the number of no-substitutions $z_{34,3}$, adding over all the pairwise sequences for site 34. The dots in the histograms correspond to actual count values in $\mathbf{Z}$. These types of graphs are helpful for determining which features of the data are not well captured by the model. We looked at several of these graphs, focusing in particular on sites that displayed above average numbers of non-synonymous substitutions, as those are important for assessing the effect of positive selection in these antigen sequences. In all cases the graphs did not suggest major discrepancies between the posterior predictive distributions and the observed values.

Table 3 lists the positively selected sites detected by the two methods described in Section 3. Method 1 uses a definition of positive selection based on the $\boldsymbol{\theta}$ parameters, while Method 2 uses a definition based on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Most of the sites listed here were previously identified as positively selected using a 5-category model without covariates (see results in Prado et al., 2006), except for those sites marked with (*). The full 3-category model that includes the amino acid scores detects the same sites previously detected by the 5-category model with no covariate when Method 2 is used. Four additional sites are identified at the level $\alpha_2 = 0.05$ when Method 1 is used. Further analyses need to be performed in order to assess whether these additional sites appear due to the reduction in the number of categories, or due to

Table 2: *Model selection criteria. AMA-1 alignment*

Model	D_{dev}^{100}	D_{se}^{100}	DIC
$(\boldsymbol{\alpha})$	42961	19039	30063
$(\boldsymbol{\alpha}+\boldsymbol{\beta})$	38156	16713	25433
$(\boldsymbol{\alpha}+\boldsymbol{\delta})$	40601	17921	27675
$(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\delta})$	36161	15817	23675
continent $(\boldsymbol{\alpha}+\boldsymbol{\gamma})$	42966	19033	30065
continent $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma})$	38118	16700	25439
continent $(\boldsymbol{\alpha}+\boldsymbol{\gamma}+\boldsymbol{\delta})$	40574	17915	27682
continent $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta})$	36124	15803	23681
country $(\boldsymbol{\alpha}+\boldsymbol{\gamma})$	42950	19030	30063
country $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma})$	38108	16703	25441
country $(\boldsymbol{\alpha}+\boldsymbol{\gamma}+\boldsymbol{\delta})$	40579	17915	27679
country $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta})$	36123	15802	23681
full $(\boldsymbol{\alpha}+\boldsymbol{\gamma})$	41722	18528	29857
full $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma})$	37027	16258	25249
full $(\boldsymbol{\alpha}+\boldsymbol{\gamma}+\boldsymbol{\delta})$	39377	17436	27540
full $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}+\boldsymbol{\delta})$	**35158**	**15439**	**23631**

the incorporation of the amino acid scores.

Sites in bold correspond to residues located in previously reported epitopes (Escalante et al., 2001). Epitopes are antigenic portions to which antibodies bind, and are therefore immunologically relevant. Further analyses that include several more sequences of AMA-1 in *P.falciparum* are needed to determine if the increased non-synonymous substitutions for the residues listed in Table 3 – particularly for those residues located in epitopes – are associated with specific clinical manifestations of the disease or with specific immune responses.

Figures 6, 7 and 8 depict site-specific effects related to particular amino acid substitutions. Sites displaying large δ_i effects that are also associated with relatively high normalized scores are of particular interest. Radical substitutions between two amino acids may be associated with the effects of positive selection. Each figure shows two sets of boxplots. The light boxplots summarize the posterior distributions of the δ_i effects for the polymorphic sites in the AMA-1 alignment, while the dark boxplots summarize the posterior of $\delta_i \bar{x}_i^*$, with $\bar{x}_i^*$ the average normalized Grantham score over all the non-synonymous substitutions for a particular site i. Some of the sites display relatively large δ_i effects, however, they are associated with low Grantham scores, indicating that the observed substitutions in such sites are between largely similar amino acids, and are less likely to produce changes to the protein morphology that would be targeted by positive selection. This is the case of site 187, for example. Other sites, such as 175, 225, 302, 263 and 302, display moderately large δ_i effects associated with large or moderately large Grantham scores, indicating that the observed substitutions in these sites were between relatively different amino

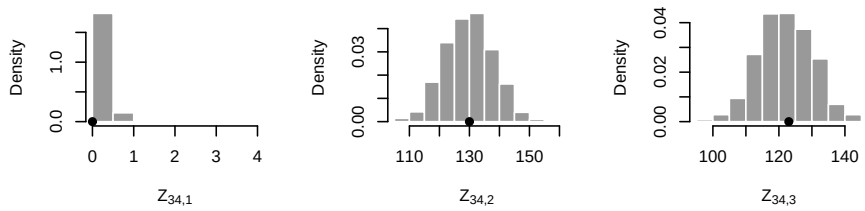

Figure 5: *Posterior predictive distributions for site 34. Dots indicate observed values.*

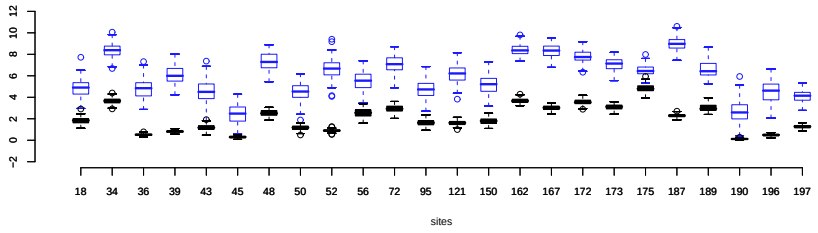

Figure 6: *Effects of the covariate in the AMA-1 antigen sequences. Sites 18 to 197.*

acids in terms of a distance based on their physicochemical properties.

6. CURRENT RESEARCH AND FUTURE DIRECTIONS

The class of GLMs for substitutions count data derived from a given alignment constitutes a novel approach to modelling and describing genetic variability at the molecular level in DNA sequence data with relatively low evolutionary divergence. These models provide an empirical framework for detecting molecular adaptation at the amino acid level by expressing the observed genetic variability in a DNA sequence alignment in terms of species/population effects, residue-specific effects and possible covariates. The methods and models of Prado et al. (2006) are extended here to incorporate covariates and population and/or geographic-specific effects, as well as model validation tools that are based on comparing posterior results to substantial biological information. The GLM-based methodology also provides a way to include structured prior information on the underlying evolutionary processes that describe

Table 3: *Positively selected sites.*

Method	Sites identified
Method 1 ($p > 0.99$, $\alpha_2 = 0.01$)	**34 39 52** 162 167 172 187 190 197 200 201 204 225* 230 242 243 **267** 282 **283 285** 296 300 308 393 404 405 435 439 485 493 496 503 512 544 **581 584** 589
Method 1 ($p > 0.95$, $\alpha_2 = 0.05$)	+ 36* 175* 196* 245*
Method 2 ($p > 0.99$, $\alpha_2 = 0.01$)	187 200 243 405
Method 2 ($p > 0.95$, $\alpha_2 = 0.05$)	+ **34 39 52** 167 172 190 204 230 242 **267 282 285** 296 308 393 404 435 485 493 496 503 512 **581 584**

the substitution patterns in the sequences, whenever such information is available.

Future research will involve incorporating the **Z**'s as latent variables in the models. The analyses of Prado et al. (2006), as well as some simulation studies included in Merl et al. (2005), suggest that the posterior results, at least in terms of which sites are identified as positively selected a posteriori, are not sensitive to the various methods used to obtain **Z** from the sequence alignments in the data analyzed here. However, a fully Bayesian approach that can quantify the uncertainty related to these underlying evolutionary processes should be considered. Simulation studies have also been performed to compare simpler versions of the GLM model in (1) to currently available methods for detecting positive selection such as those implemented in `PAML`, `HYPHY` and `MrBayes` (Huelsenbeck and Dyer, 2004). Such studies (see Merl et al., 2005) suggest that the GLM-based methodology has higher power than other methods to detect sites under selection in sequences with relatively low evolutionary divergences. Further simulation studies will be performed in the future to assess the effect of covariates in detecting positive selection.

The models presented here allow us to determine, from a statistical viewpoint, which sites are more likely to be under positive selection in malaria antigens. Many more sequences of AMA-1 for human *P.falciparum* and *P.vivax* are available, as well as alignments of two more candidate antigens for both species of the parasite. We expect to carry out extensive GLM-based analyses for these data. The preliminary results presented here, as well as future results from GLM-based analyses, will assist immunologists in the identification of specific residues that may be relevant to determining if, in fact, the candidate antigens are appropriate for vaccine development.

Other future research directions relate to developing a variable selection approach to detecting which amino acid properties are significant. Recently, `Tree SAAP` (Woolley et al., 2003), a software that measures selective influences on 31 structural and biochemical amino acid properties during phylogenesis, was developed. Many additional amino acid properties can be included. We expect to extend the approach presented here to appropriately tackle the problem of selecting relevant

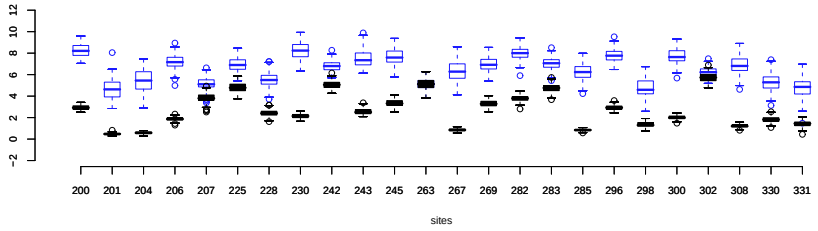

Figure 7: *Effects of the covariate in the AMA-1 antigen sequences. Sites 200 to 331.*

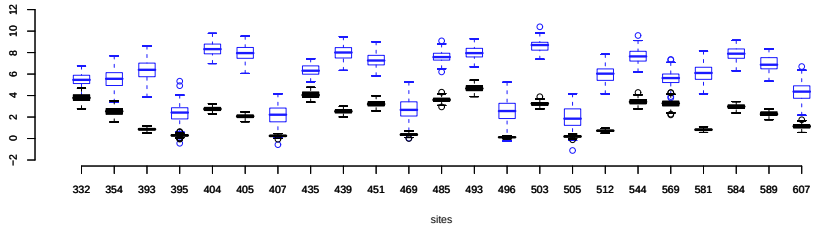

Figure 8: *Effects of the covariate in the AMA-1 antigen sequences. Sites 332 to 607.*

covariates from a large pool of covariates that describe amino acid properties.

ACKNOWLEDGEMENTS

The work was supported by grant R01GM072003-02 from the National Institutes of Health/National Institute of General Medical Sciences. The authors acknowledge useful discussions with Ananías Escalante, School of Life Sciences, Arizona State University.

REFERENCES

Baker, S. (1994). The multinomial-Poisson transformation. *The Statistician* **43**, 495–504.

Escalante, A. A., Grebert, H. M., Chaiyaroj, S. C., Magris M., Biswas S., Nahlen, B. L. and Lal, A .A. (2001). Polymorphism in the gene encoding the apical membrane antigen 1 (AMA-1) of Plasmodium falciparum. X Asembo Bay Cohort Project. *Mol. Biochem. Parasitol.* **113**, 279–287.

Escalante, A. A., Lal, A. A. and Ayala, F .J. (1998). Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. *Genetics* **149** 189–202.

Felsenstein J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc.

Galindo, B., Vacquier V. and Swanson, W. J. (2003). Positive selection in the egg receptor for abalone sperm lysin. *PNAS* **100**, 4639–4643.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1-11.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding

Grantham, R. (1974) Amino acid different formula to help explain protein evolution. *Science* **185**, 862-864.

Huelsenbeck, J. P. and Dyer, K. A. (2004). Bayesian estimation of positively selected sites. *J. Mol. Evol.* **58**, 661–672.

Kosakovsky Pond, S. L. and Frost, S. D. W. (2005a) A simple hierarchical approach to modelling substitution rates. *Mol. Biol. and Evol.* **22**, 223–234.

Kosakovsky Pond, S. L., Frost, S. D. W and Muse, S. V. (2005) HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679. *Mol. Biol. and Evol.* **22**, 1208–1222.

Kosakovsky Pond, S. L. and Frost, S. D. W. (2005b) Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. and Evol.* **22**, 1208–1222.

Lee, Y. H., Ota T. and Vacquier V. D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. and Evol.* **12**, 231–238.

Merl, D., Prado, R. and Escalante, A. A. (2005). Bayesian estimation of differential selection using generalized linear models. *Tech. Rep.*, Applied Mathematics and Statistics, University of California Santa Cruz.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the Chloroplast genome. *Mol. Biol. and Evol.* **11**, 715–724.

Nielsen, R. and Yang, Z. (1998). Simple methods for estimating the numbers of synonymous substitutions

Polley, S. D., Chokejindachai, W. and Conway, D. J. (2003). Allele frequency based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics* **165**, 555-561.

Polley, S. D., Mwangi T., Kocken C. H., Thomas A. W., Dutta S., Lanar D. E., Remarque E., Ross A., Williams T. N., Mwanbingu G., Lowe B., Conway D. J. and Marsh K. (2004). Human antibodies to recombinant protein constructs of Plasmodium falciparum Apical Membrane Antigen 1 (AMA1) and their associations with protection from malaria. *Vaccine* **23**, 718–728.

Prado, R., Merl, D. and Escalante, A. A. (2006). Detecting selection in DNA sequences: A model-based approach. *Tech. Rep.*, Applied Mathematics and Statistics, University of California Santa Cruz.

Schwartz, R. M. and Dayhoff, M. O. (1978). *Atlas of Protein Sequence and Structure, 5 suppl.* (Vol. 3:353-358). Washington D.C.: Nat. Biomed. Res. Found.

Spiegelhalter, D. J., Best, N. G., Carlin B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. B* **44**, 377–387.

Suzuki, Y. (2004). New methods for detecting positive selection at single amino acid sites. *J. Mol. Evol.* **59**, 11–19.

Suzuki, Y. and Gojorobi, T. (1999). A method for detecting positive selection at single amino acid sites. *Mol. Biol. and Evol.* **16**, 1315–1328.

Woolley S., Johnson J., Smith M. J., Keith A. Crandall K. A. and McClellan D. A. (2003). TreeSAAP: Selection on amino acid properties using phylogenetic trees. *Bioinformatics* **19**, 671-672.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.***13**, 555–556.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. K. (2000). Codon substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449.

Yang, Z., Swanson, W. J. and Vacquier V. D. (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. and Evol.* **17**, 1446–1455.

Yang, Z., Wong, W. S. W. and Nielsen R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. and Evol.* **22**, 1107–1118.