Bayesian Semiparametric Regression Models to Characterize Molecular Evolution

Saheli Datta*[1], Abel Rodriguez[2], Raquel Prado[2]

[1] Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[2] Department of Applied Mathematics and Statistics, University of California Santa Cruz, Santa Cruz, CA, USA

Email: Saheli Datta*- saheli@ams.ucsc.edu; Abel Rodriguez - abel@ams.ucsc.edu; Raquel Prado - raquel@ams.ucsc.edu;

*Corresponding author

Abstract

Background: Statistical models and methods that associate changes in the physicochemical properties of amino acids with natural selection at the molecular level typically do not take into account the correlations between such properties. We propose a Bayesian hierarchical regression model with a generalization of the Dirichlet process prior on the distribution of the regression coefficients that describes the relationship between the changes in amino acid distances and natural selection in protein-coding DNA sequence alignments.

Results: The Bayesian semiparametric approach is illustrated with simulated data and the abalone lysin sperm data. Our method identifies groups of properties which, for this particular dataset, have a similar effect on evolution. The model also provides nonparametric site-specific estimates for the strength of conservation of these properties.

Conclusions: The model described here is distinguished by its ability to handle a large number of amino acid properties simultaneously, while taking into account that such data can be correlated. The multi-level clustering ability of the model allows for appealing interpretations of the results in terms of properties that are roughly equivalent from the standpoint of molecular evolution.

1 Background

The structural and functional role of a codon in a gene determines its ability to freely change. For example, nonsynonymous (amino acid altering) substitutions may not be tolerated at certain codon sites due to strong negative selection, while at other sites some nonsynonymous substitutions may be allowed if they do not affect key physicochemical properties associated with protein function [1]. Thus, at such preferentially changing sites, more frequent substitutions occur between physicochemically similar amino acids (or codons which lead to those amino acids) than dissimilar ones [2–4]. Methods which use changes in physicochemical amino acid properties have thus been proposed in the study of evolution. For example, [5–7] use distances to calculate deviations from neutrality for a particular amino acid property. Alternative approaches model the evolution of protein coding sequences as continuous-time Markov chains with rate matrices that distinguish

between property-altering and property-conserving mutations as in [8] and [9]. More recently, [10] proposed a Bayesian hierarchical regression model that compares the observed amino acid distances to the expected distances under neutrality for a given set of amino acid properties and incorporates mixture priors for variable selection. The hierarchical mixture priors enable the model in [10] to identify neutral, conserved and radically changing sites, while automatically adjusting for multiple comparisons and borrowing information across properties and sites.

A common feature of all the methods listed above is the implicit assumption that properties are independent from each other in terms of their effect on evolution. A review of the amino acid index database (available for example at `http://www.genome.jp/dbget/aaindex.html`), which lists more than 500 amino acid properties, shows that a large number of them are highly correlated. Although the correlations we observe in the data can be different from those computed from the raw amino acid scores due to the influence of factors such as codon bias, by ignoring these correlations we are also ignoring the fact that correlated properties may affect a particular site in similar ways. Hence, approaches that do not take into account the correlations in the rates of mutations on different codons do not make use of key information about the relative importance of different physicochemical properties on molecular evolution.

A natural way to account for correlations in the data is by considering a factor structure, see for example [11]. However, selecting the number and order of the factors can be a difficult task in this type of factor models. In addition, the particular structure of the model in [11] makes it difficult to incorporate the effect of the factors on regions that are very strongly conserved. This paper extends the Bayesian hierarchical regression model in [10] by placing a nonparametric prior on the distribution of the regression coefficients describing the effect of properties on molecular evolution. The prior is an extension of the well known Dirichlet process prior [12, 13] to model separately exchangeable arrays [14, 15]. As in [10], the main goal of the model described in this paper is to identify sites that are either strongly conserved or radically changing. In order to account for correlations across properties, our model clusters properties with similar effects on evolution, and within each such group, clusters sites with similar regression coefficients and nonparametrically estimates their distribution. In addition to accounting for correlations across properties, this structure allows us to dramatically reduce the number of parameters in the model and generate interpretable insights about molecular evolution at the codon level.

Although the clusters of properties can in principle be considered nuisance parameters that are of no direct interest, in practice posterior inference on the clustering structure can provide interesting insights about the molecular evolution process of a given gene. Indeed, as will become clear in the following sections, our approach incorporates the effect of codon-usage bias. Hence, any significant differences between the cluster structure estimated from the observed protein-coding sequence alignment and the correlation structure derived from the raw distances between the properties in such cluster can be interpreted a signal of extreme codon-usage bias in that particular region of the genome.

The rest of the paper is organized as follows. A brief review of DP mixture models along with the details of our model is provided in Section 2. This section also includes a review of some of the currently available methods for characterizing molecular evolution that take into account changes amino acid properties. The model is then evaluated via simulation studies and illustrated through a real data example. The simulated and real data analyses, as well as comparisons between the proposed semiparametric regression approach and other methods, are presented in Section 3. Finally, Section 4 provides our concluding remarks.

2 Methods

Dirichlet Process Mixture Models

The Dirichlet process (DP) was formally introduced by [12] as a prior probability model for random distributions G. A DP(ρ, G_0) prior for G is characterized by two parameters, a positive scalar parameter ρ, and a parametric base distribution (or centering distribution) G_0. ρ can be interpreted as the precision parameter, with larger values of ρ resulting in realizations of G that are closer to the base distribution G_0.

One of the most commonly used definitions of the DP is its constructive definition [13], which characterizes

DP realizations as countable mixtures of point masses. Specifically, a random distribution G generated from DP(ρ, G_0) is almost surely of the form

$$G(\cdot) = \sum_{l=1}^{\infty} w_l \delta_{\phi_l}(\cdot),$$

where $\delta_{\phi_l}(\cdot)$ denotes a point mass at ϕ_l. The locations ϕ_l are i.i.d. draws from G_0, while the corresponding weights w_l are generated using the following "stick-breaking" mechanism. Let $w_1 = v_1$ and define $w_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$ for $l = 2, 3, \ldots$, where $\{v_l : l = 1, 2, \ldots\}$ are i.i.d. draws from a $\mathsf{Beta}(1, \rho)$ distribution. Defining the weights in this way ensures $\sum_{l=1}^{\infty} w_l = 1$. Furthermore, the sequences $\{v_l : l = 1, 2, \ldots\}$ and $\{\phi_l : l = 1, 2, \ldots\}$ are independent.

The DP is most often used to model the distribution of random effects in hierarchical models. In the simplest case where no covariates are present, these models reduce to nonparametric mixture models (e.g., [16–18]). Assume that we have an independent sample of observations $y_1, y_2, \ldots, y_n$ such that $y_i|\theta_i \overset{ind}{\sim} k(\cdot; \theta_i)$, where $k(\cdot; \theta_i)$ is a parametric density. Then, the DP mixture model places a DP prior on θ_i as

$$\begin{aligned} \theta_i|G &\overset{i.i.d.}{\sim} G, \quad i = 1, \ldots, n \\ G|\rho &\sim \mathrm{DP}(\rho, G_0) \end{aligned}$$

The almost sure discreteness of realizations of G from the DP prior allows ties in θ_i, making DP mixture models appealing in applications where clustering is expected. The clustering nature is easier to see from the Pólya urn characterization of the DP [19] which gives the induced joint distribution for the θ_is, by marginalizing G over its DP prior. Under that representation, we can write $\theta_i = \theta^*_{\xi_i}$ where $\theta^*_1, \theta^*_2, \ldots$ is an independent and identically distributed sample from G_0 and the indicators $\xi_1, \ldots, \xi_n$ are discrete indicators sequentially generated with $\xi_1 = 1$ and

$$\Pr(\xi_{i+1} = k|\rho, \xi_i, \ldots, \xi_1) = \begin{cases} \frac{r^i_k}{i+\rho} & k \leq \max_{j \leq i}\{\xi_i\} \\ \frac{\rho}{i+\rho} & k = \max_{j \leq i}\{\xi_i\} + 1, \end{cases}$$

where $r^i_k = \sum_{j=1}^{i} I(\xi_j = k)$ and

$$I(\xi_j = k) = \begin{cases} 1 & \xi_j = k \\ 0 & \text{otherwise.} \end{cases}$$

One advantage of DP mixture models over other approaches to clustering and classification is that they allow us to automatically estimate the number of components in the mixture. Indeed, from the Pólya urn representation of the process it should be clear that, although the number of *potential* mixture components is infinite, the model implicitly places a prior on the number of components that, for moderate values of ρ, favors the data being generated by an effective number of components $K^* = \max_{i \leq n}\{\xi_i\} < n$.

The Model

Our data consist of *observed* and *expected amino acid distances* derived from a DNA sequence alignment, a specific phylogeny, a stochastic model of sequence evolution, and a predetermined set of physicochemical amino acid properties. In the analyses presented here, we disregard uncertainty in the alignment/phylogeny/ancestral sequence level since our main focus is the development and implementation of models that allow us to make inferences on the latent effects that several amino acid properties may have on molecular evolution for a given phylogeny and an underlying model of sequence evolution. Extensions of these analyses that take into account these uncertainties are briefly described in Section 4. For further discussion on this issue, see also [10].

In order to calculate the observed distances, we first infer the ancestral sequences under a specific substitution model and a given phylogeny. In our applications, we use PAML version 3.15 [20] and the codon

substitution model of [21], which accounts for the possibility of multiple substitutions at a given site. Nonsynonymous substitutions are then counted by comparing DNA sequences between two neighboring nodes in the phylogeny. The observed mean distance, denoted as $y_{i,j}$ for site i and property j, is obtained as the mean absolute difference in the property scores due to all nonsynonymous substitutions at site i. Only those sites with at least one nonsynonymous change from the ancestral level are retained for further analysis.

To compute the expected distances, note that each codon can mutate to one of at most nine alternative codons through a single nucleotide substitution [5], only some of which are nonsynonymous (changes to stop codons are ignored). Let N_k be the number of nonsynonymous mutations possible through a single nucleotide change, corresponding to a particular codon k ($k = 1, \ldots, 61$). Let $D_{k,l}^{i,j}$ be the absolute difference in property j between nonsynonymous codon pairs at site i differing at one codon position, where $l = 1, \ldots, N_k$. The frequency of codon k at a particular site i in the DNA sequence under study is denoted by F_k^i. Then, the expected mean distance for a particular site i and a given property j is given by

$$x_{i,j} \equiv D_E^{i,j} = \frac{\sum_{k=1}^{61} F_k^i \sum_{l=1}^{N_k} D_{k,l}^{i,j}}{\sum_{k=1}^{61} F_k^i N_k}.$$

We consider a hierarchical regression model that relates $x_{i,j}$ to $y_{i,j}$ and allows us to compare the expected and observed distances at the codon level for several properties simultaneously with the following rationale. If a given site i is neutral with respect to property j, then $y_{i,j} \approx x_{i,j}$. If property j is conserved at site i, then $y_{i,j} << x_{i,j}$ and finally, if property j is radically changing at site i, then $y_{i,j} >> x_{i,j}$.

To construct our model, we first standardize the distances $x_{i,j}$ and $y_{i,j}$ by dividing them by the maximum possible distance for each property. This enables us to use priors with the same scale for all the regression coefficients. Our regression model for the standardized distances $y_{i,j}^*$ and $x_{i,j}^*$, for sites $i = 1, \ldots, I$ and properties $j = 1, \ldots, J$, can be written as

$$y_{i,j}^* | \beta_{i,j}, \sigma_{i,j}^2 \sim \begin{cases} \mathsf{N}(\beta_{i,j} x_{i,j}^*, \sigma_{i,j}^2) & \text{if } \beta_{i,j} = 0 \\ \mathsf{N}(\beta_{i,j} x_{i,j}^*, \sigma_{i,j}^2 / n_i^O) & \text{if } \beta_{i,j} \neq 0, \end{cases} \tag{1}$$

where n_i^O is the observed number of nonsynonymous changes at a particular site i and $\beta_{i,j}$ and $\sigma_{i,j}^2$ are the regression coefficient and variance parameter associated with site i and property j. The mixture model accounts for the fact that some of the y_{ij}^*s can be equal to zero as some nonsynonymous changes do not change the value of the property being measured (e.g., Aspargine, Aspartic acid, Glutamine, Glutamic acid all have the same hydropathy score).

To complete the model, we need to describe a model for the matrix of regression coefficients $[\beta_{i,j}]$. There are a number of possible models for this type of data which utilize Bayesian nonparametric methods; some recent examples include the infinite relational model (IRM) [22, 23], the matrix stick breaking process (MSBP) [24], and the nested infinite relational model (NIRM) [14, 15].

In this paper we focus on the NIRM, which is constructed by partitioning the original matrix into groups corresponding to entries with similar behavior. This is done by generating partitions in one of the dimensions of the matrix (say, rows) that are nested within clusters of the other dimension (columns). This structure allows us to identify groups of (typically correlated) properties with similar pattern and then, within each such group, identify clusters of sites with similar values of $\beta_{i,j}$ (Figure 1 provides a graphical representation of this idea). In our setting, we take $[\boldsymbol{\theta}_{ij}] = [\beta_{i,j}, \sigma_{i,j}^2]$ and employ a NIRM to generate a prior for $[\boldsymbol{\theta}_{ij}]$.

More specifically, we denote by $\boldsymbol{\theta}_j = (\theta_{1,j}, \ldots, \theta_{Ij})'$ the vector of regression coefficients and the associated variances corresponding to property (column) j. To obtain clusters for the properties, we assume that $\boldsymbol{\theta}_j \sim F$, where

$$F = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k^*} \tag{2}$$

is a random distribution such that $\pi_k = v_k \prod_{s<k}(1 - v_s)$, $v_k \sim \mathsf{Beta}(1, \rho)$, and $\boldsymbol{\theta}_k^* \sim H_k$. Indeed, the discrete nature of F ensures that ties among the $\boldsymbol{\theta}_j$ happen with non-zero probability.

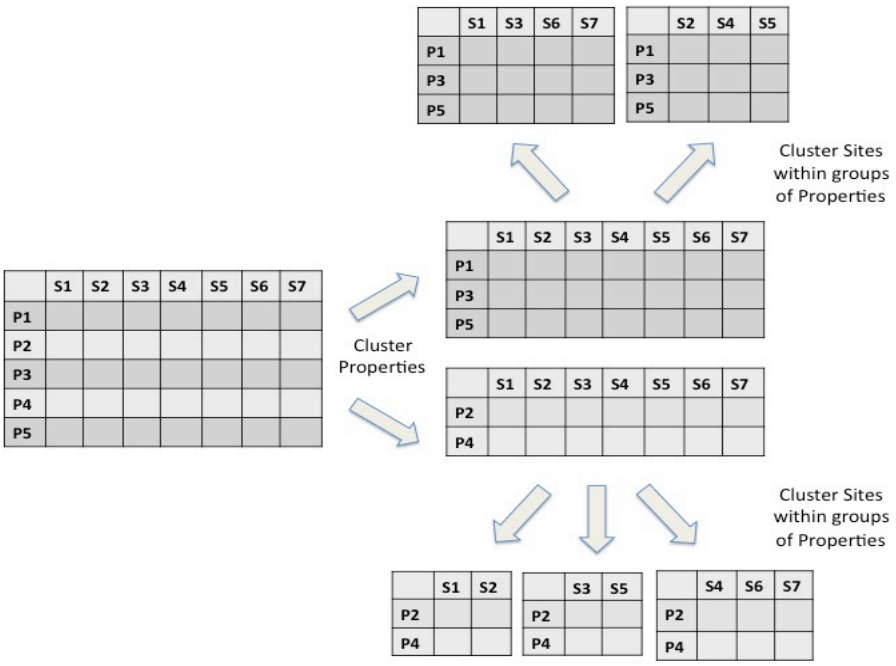

Figure 1: Stylized representation of our model. Each sub table at the second level of clustering shares a common value for the regression coefficient $\beta_{i,j}$. Rows correspond to properties, while columns correspond to sites.

To obtain cluster-specific partitions for the sites (rows), H_k (the joint distribution associated with all sites for a given cluster of properties) has to be chosen carefully. In particular, we write $\boldsymbol{\theta}^*_k = (\theta^*_{1,k}, \ldots, \theta^*_{I,k})'$ for any specific specific cluster of properties k and let

$$\theta^*_{i,k} \sim \sum_{l=1}^{\infty} w_{l,k}\delta_{\varphi_{l,k}}, \tag{3}$$

with $w_{l,k} = u_{l,k}\prod_{r<l}\{1-u_{r,k}\}$, $u_{l,k} \sim \mathsf{Beta}(1,\gamma_k)$ for every k, and $\varphi_{l,k}$ are independently drawn from the baseline measure $G_{0,l,k}$.

The baseline measure $G_{0,l,k}$ is chosen to accommodate the fact that some $y^*_{i,j}$s can be zero, since some nonsynonymous changes can keep the value of the property being measured unchanged. Thus, $G_{0,l,k}$ is a mixture with a point mass at zero and a continuous density otherwise. To allow for a more flexible model we assume that different prior variances are associated with the $y^*_{i,j}$s which are zero and those $y^*_{i,j}$s that are different from zero, with the specific form of G_{0lk} as below.

$$\varphi_{l,k} = (\phi_{l,k}, \vartheta^2_{l,k}) | G_{0lk} \sim G_{0lk}$$

with

$$G_{0lk} = \lambda 1_{\{\phi_{l,k}=0\}} p_1(\vartheta^2_{l,k}) + (1-\lambda) p(\phi_{lk}|\vartheta^2_{l,k}) p_2(\vartheta^2_{l,k}), \tag{4}$$

where $p_1(\vartheta^2_{l,k}) \sim \mathsf{Inv\text{-}Ga}(a_\kappa, b_\kappa)$, $p(\phi_{l,k}|\vartheta^2_{l,k}) \sim \mathsf{N}(\alpha_k, \vartheta^2_{l,k}/V_0)$ and $p_2(\vartheta^2_{l,k}) \sim \mathsf{Inv\text{-}Ga}(a^*_\sigma, b^*_\sigma)$. Here $\phi_{l,k}$ and $\vartheta^2_{l,k}$ respectively denote the unique values $\beta_{i,j}$ and $\sigma^2_{i,j}$ can take, whereas λ is the prior probability that $\phi_{l,k}$ has the value zero (i.e., the properties associated with this cluster are strongly conserved at this cluster of sites).

Note that our model implies that both sites and properties are exchangeable a priori. If no additional prior information is available, this type of assumption seems reasonable. However, a posteriori, it is possible to have sites behave differently in different clusters.

To complete the model we place hyperpriors on all parameters of the resulting model. Conjugate priors are chosen for ease of computation. α_k denotes the mean for the $\phi_{l,k}$s that are different from zero belonging to a specific cluster of properties k and is assumed to have a $\mathsf{N}(m_\alpha, C_\alpha)$ prior for all k. The DP concentration parameters ρ and γ_k are assumed to follow $\mathsf{Ga}(a_\rho, b_\rho)$ with mean a_ρ/b_ρ, and $\mathsf{Ga}(a_\gamma, b_\gamma)$ with mean a_γ/b_γ for all k, respectively. λ, which is the prior probability for the point mass at 0 in G_{0lk}, follows a $\mathsf{Beta}(a_\lambda, b_\lambda)$. The specific choice of hyperparameters is discussed later as part of each data analysis. In general, we use $\mathsf{Ga}(1,1)$ priors for the DP concentration parameters and a $\mathsf{N}(1, C_\alpha)$ prior for α_k to correspond to our assumption of neutrality a priori for the properties.

Related Work

We compare results from our proposed method with results from a few currently available methods that aim to characterize molecular evolution while also taking into account changes in amino acid properties, namely, the regression model in [10], `TreeSAAP` [25], and `EvoRadical` [9].

In [10], the first level of the model is the regression equation on $y^*_{i,j}$ as in equation (1), but it implicitly assumes independence among properties and independence among sites unlike our current model. The model in [10] is suitable for use when a few mostly independent amino acid properties are being analyzed whereas the new semiparametric model is better suited to the analysis of a large number of possibly correlated properties.

`TreeSAAP` uses the methods of [6] to classify nonsysnonymous substitutions into one of M categories, with higher numbered categories corresponding to sites showing radical changes and lower numbered categories used for sites showing conserved changes for a given property. For the analysis considered here, we used 8 categories where categories 6, 7, and 8 corresponded to sites showing radical changes, and categories 1 and 2 to sites showing conserved changes. Nonsynonymous changes are inferred from the ancestral reconstruction using the nucleotide substitution models in `baseml` implemented in `PAML`. We used a Bonferroni correction to correct for multiple comparisons.

`EvoRadical` implements the models of [9], which use partitions of amino acids to parameterize the rates of property-conserving and property-altering codon substitutions in a maximum likelihood framework. The model considers three types of substitutions: synonymous, property-conserving nonsynonymous and property-altering nonsynonymous which is a slight improvement from [8]. For analyses with multiple properties, one has to create different partitions for the different properties and run `EvoRadical` for each property.

Posterior Simulation

Various algorithms exist for posterior inference of DP mixtures - some of the most popular ones use (i) the Pólya urn characterization to marginalize out the unknown distribution(s) [26, 27], (ii) a truncation approximation to the stick-breaking representation of the process which paves the way for the use of methods employed in finite mixture models [28,29], (iii) reversible jump MCMC or split-merge methods [30,31]. Some other recent approaches have also used variational methods [32] and slice samplers [33].

We use an extension of the finite mixture approximation discussed in [28] for its ease of implementation. Truncating F at a sufficiently large K, we write $F^{(K)} = \sum_{k=1}^{K} \pi_k \delta_{\theta^*_k}$, with the weights π_k and locations θ^*_k generated as described earlier in this Section. Next we introduce configuration variables $\{\zeta_j\}$ such that, for $k = 1, \ldots, K$, $\zeta_j = k$ if and only if $\boldsymbol{\theta}_j = \theta^*_k$. Similarly for G_k, we truncate at a sufficient level L, and introduce another set of configuration variables $\{\xi_{i,k}\}$ where $\xi_{i,k} = l$, with $l = 1, \ldots, L$, if and only if $\theta^*_{i,k} = \varphi_{l,k}$. Additional details about the algorithm are provided in the Appendix.

To determine the truncation levels K and L, we follow [29]. In particular, note that conditional on ρ (the DP concentration parameter), the tail probability $\sum_{k=K}^{\infty} \pi_k$ has expectation $\{\rho/(1+\rho)\}^{K-1}$. Using prior

guesses for ρ and acceptable tolerance levels for the tail probability to be small, one can then solve for the truncation level K. In our analyses, we used K and L in the range of 25 to 35. These values are in line with those used in other applications (for example, see [34]).

3 Results

Empirical Exploration via Simulation Studies

We present two simulation studies to check the performance of the model under different scenarios. Additional simulation scenarios that may be of interest are available as an online supplement.

Simulation Study 1

The setup for the first simulation is as follows. We generate values for the distinct regression coefficients ($\phi_{l,k}$) from a $\mathsf{N}(1, 0.25)$. The number of distinct regression coefficients depends on the particular clustering structure for the corresponding simulation. Once we obtain the regression coefficients, we generate observations $y_{i,j}$ from $\mathsf{N}(\phi_{l,k}x_{i,j}, \sigma^2 = 0.001)$. The $x_{i,j}$s are obtained from the lysin data set described below with analyses for 32 properties, which implies $J = 32$ and $I = 94$.

We fitted the model in Section 2 to the $y^*_{i,j}$s and $x^*_{i,j}$s, with the following modifications: (i) the NIRM is imposed on $\beta_{i,j}$, so $\varphi_{l,k} = \phi_{l,k}$ and (ii) $\phi_{l,k} \sim G_0$ where $G_0 \sim \mathsf{N}(\alpha, \tau^2)$. We used $K = 25$ and $L = 25$ for the simulations. The MCMC algorithm was run with the following hyperpriors: $\rho \sim \mathsf{Ga}(1, 1)$, $\gamma_k \sim \mathsf{Ga}(1, 1)$ for all k, $\alpha \sim \mathsf{N}(1, 0.25)$. $\sigma^2 \sim \mathsf{Inv\text{-}Ga}(100, 10)$ and $\tau^2 \sim \mathsf{Inv\text{-}Ga}(2, 4)$ were chosen such that the prior means corresponded to the true values for these hyperparameters. Results are based on 15000 iterations, with the first 5000 discarded as burn-in. Convergence was assessed by running two chains where each chain was initialized by randomly assigning the $\beta_{i,j}$s to different partitions. Posterior summaries based on the two chains were consistent with each other.

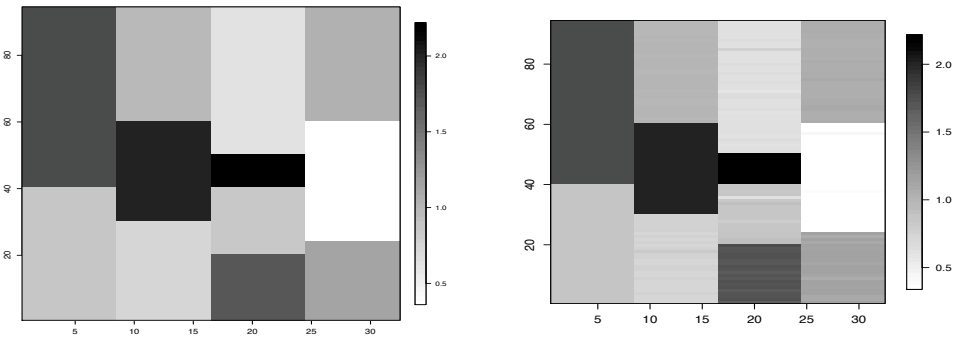

Figure 2: Image plots for true $\beta_{i,j}$ values (left panel) and posterior means $\hat{\beta}_{i,j}$s (right panel).

In this scenario, we had four clusters for the columns, each with differing number of groups, leading to twelve distinct cluster combinations for the entire matrix of $\beta_{i,j}$s (Figure 2, left panel). Figure 3 shows the marginal probabilities for any two columns (properties) of belonging to the same cluster. The model correctly identifies that there are 4 clusters for the columns and assigns each set of columns to its corresponding cluster with no uncertainty.

Similar graphical summaries obtained for the structure of rows within each cluster of columns show that the correct clustering structures for the rows, within each cluster of columns, are inferred (see Figure 4). For this level, however, there is some uncertainty about the membership of the clusters for a few rows. See, for

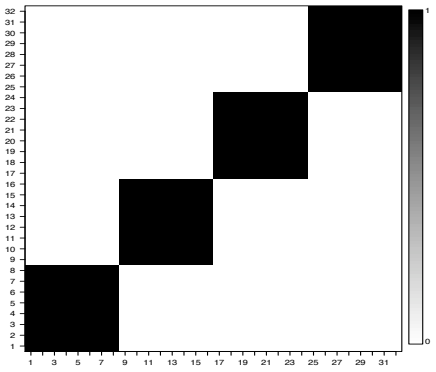

Figure 3: Marginal posterior probabilities of each pair of columns belonging to the same cluster.

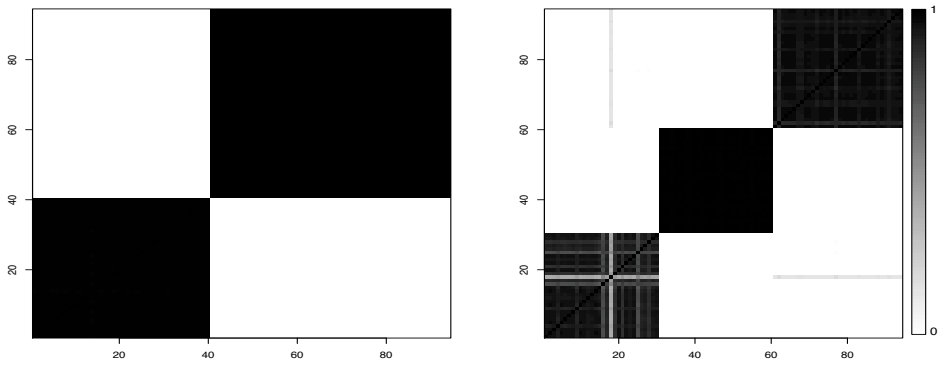

Figure 4: Marginal posterior probabilities of each pair of rows belonging to the same cluster for two different clusters of columns.

example, the right panel of Figure 4. Some rows in cluster 1 (in the lower left) are sometimes being assigned to cluster 3 (top right). The distinct values of ϕ used for these two clusters were 0.73 and 0.98, therefore, it does not seem unreasonable to see some uncertainty in the assignment of clusters. Posterior means of $\hat{\beta}_{i,j}$s agree closely with the true values as shown in Figure 2.

This scenario corresponds to the type of situation we expect on most real datasets: properties will cluster into groups and, within each group of properties, clusters of sites with similar responses can be clearly identified. Our results suggest that, as expected, the model is capable of identifying these multiple clusters with high accuracy and therefore accurately estimate the value of the regression coefficients. Other scenarios, including extreme cases where all properties belong to a common cluster while sites belong to one of several clusters, and cases where each property has a different effect on amino acid rates are available as supplementary material.

To investigate the effect of the truncation levels and the priors on our model, we performed sensitivity analysis by varying the truncation levels as well as the different hyperparameters. Increasing the truncation level to 35 did not affect the results and the estimated posterior means of the βs showed close agreement with the true values. The analyses was also fairly robust to the choice of the priors, since varying the hyperparameters had almost no effect on the results. Decreasing the prior variance of τ^2 makes the results marginally better, i.e., posterior means of the $\beta_{i,j}$s, $\hat{\beta}_{i,j}$s, are slightly closer to the true values.

Simulation Study 2 - Data Simulated From A Biological Model

In our second simulation study the model is evaluated in the context of biological sequences generated from an evolutionary model. In particular, a Markov model was used to generate 20 sequences of 90 codons each. For the first one-third of the sites (sites 1-30) we used transition probabilities obtained from the codon-substitution model of [21] with equal equilibrium probabilities for all 61 codons. For the second one-third of the sites (sites 31-60), we modified the transition probability matrix from the previous step by increasing the probabilities of transitions between codons that have small distances for volume and decreasing the probabilities of transitions between codons that have large distances for volume - this was done to encourage only those changes that conserve volume in this part of the sequences. Finally, for the last one-third of the sites (sites 61-90), we modified the original transition probability to encourage radical changes in hydropathy. Thus, we increased some transition probabilities between codons that have very different hydropathy scores and decreased a few of those that have similar hydropathy scores. Note that, since the equilibrium probabilities are either uniform or roughly uniform across all sites, the correlation structure across properties is retained in the expected distances, which simplifies the interpretation of the results.

Once we obtained the sequences, we generated ancestral sequences using `PAML`, version 3.15, [20] and calculated observed and expected distances y_{ij} and x_{ij} for five properties, namely, hydropathy (h), volume (M_v), polarity (p), isoelectric point (pH_i) and partial specific volume (V^0). Of these, h and p are correlated and so are M_v and V^0.

Our model was fitted with $K = 25$ and $L = 25$ as truncation levels. The prior distributions were the same as the ones used for our previous simulation. Results are based on 15000 iterations, of which the first 5000 were burn-in. There did not seem to be any obvious problems with convergence, which was assessed by visual inspection of trace plots of some of the parameters.

The analyses found that there were three clusters of properties - the first cluster has properties h and p, the second cluster comprised of properties M_v and V^0 and the third cluster only had property pH_i as shown in Figure 5. Figure 6 shows the posterior means of $\beta_{i,j}$s for representative properties of the three clusters in Figure 5. Sites 24, 65, 67, 71, 81, 82, and 89 have large posterior means $\hat{\beta}_{ij}$s for cluster 1 (h and p). These are also the same sites that show up in the small cluster at the top right in Figure 7. Specifically, Figure 7 shows how often any two sites in cluster 1 are grouped together. The sites in the lower left (16, 28, 46, 51) have small posterior means $\hat{\beta}_{i,j}$s for these properties (h and p) and are grouped together more often. The big group of sites in the middle mostly seem to have mean $\hat{\beta}_{i,j}$s around 1 while sites 81, 89, 71, and 65

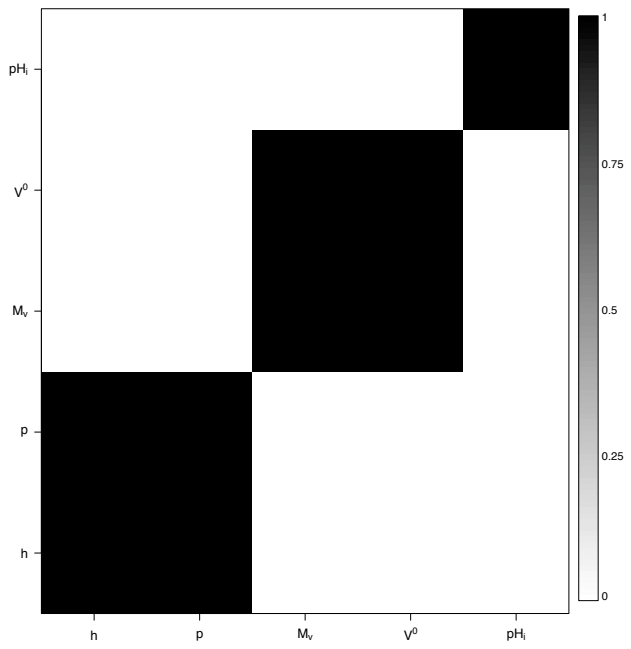

Figure 5: Marginal posterior probabilities of any two properties being in the same cluster for the simulated data.

have the largest $\hat{\beta}_{i,j}$ values and very large probabilities of being clustered together in cluster 1. Thus, the model successfully identifies sites that have similar $\beta_{i,j}$ values in a specific cluster and groups them together. Groups of sites that change a property can also be identified for clusters 2 and 3 in Figure 5. In particular, for cluster 2 (M_v and V^0), there is a big group of sites which conserve these properties. Most of these sites are in the central one-third portion (i.e., the portion that includes sites 31-60) which were simulated under a transition probability matrix that favors transitions that conserve volume. Finally, for cluster 3 (pH_i) there is one large group of sites which conserve the property and one group comprising sites 39 and 80 which change the property greatly.

To better understand the performance of our method, we also analyzed the sequences generated above with the parametric regression model in [10], `TreeSAAP` [25], and `EvoRadical` [9]. Table 1 lists the thirty sites with the largest posterior means $\hat{\beta}_{i,j}$s for h, and the thirty sites with the smallest posterior means $\hat{\beta}_{i,j}$s for M_v for the regression model of [10] and also for our new semiparametric approach. Many of the same sites are identified by both methods, however, our new method performs slightly better than the regression model in [10]. In particular the new method identifies two additional sites in the 61-90 region as sites that change h.

Table 2 lists sites that `TreeSAAP` finds significant for the different properties. All of the sites that `TreeSAAP` finds significant are also identified by our methods. However, note that once we correct for multiple comparisons in the `TreeSAAP` results, only one site (74) still remains significant. We note that the hierarchical specification of the priors in our models automatically accounts for multiple comparisons and no corrections are needed (see [10] for more discussion on this).

Finally, we analyzed the sequences generated previously with `EvoRadical` using two different partitions [8] - one for p and the other for M_v. We chose to run `Evoradical` with p instead of h, since a partition of the amino acids for polarity was already available in [8]. Additionally, given that h and p are correlated, we expect to see somewhat similar results for these two properties.

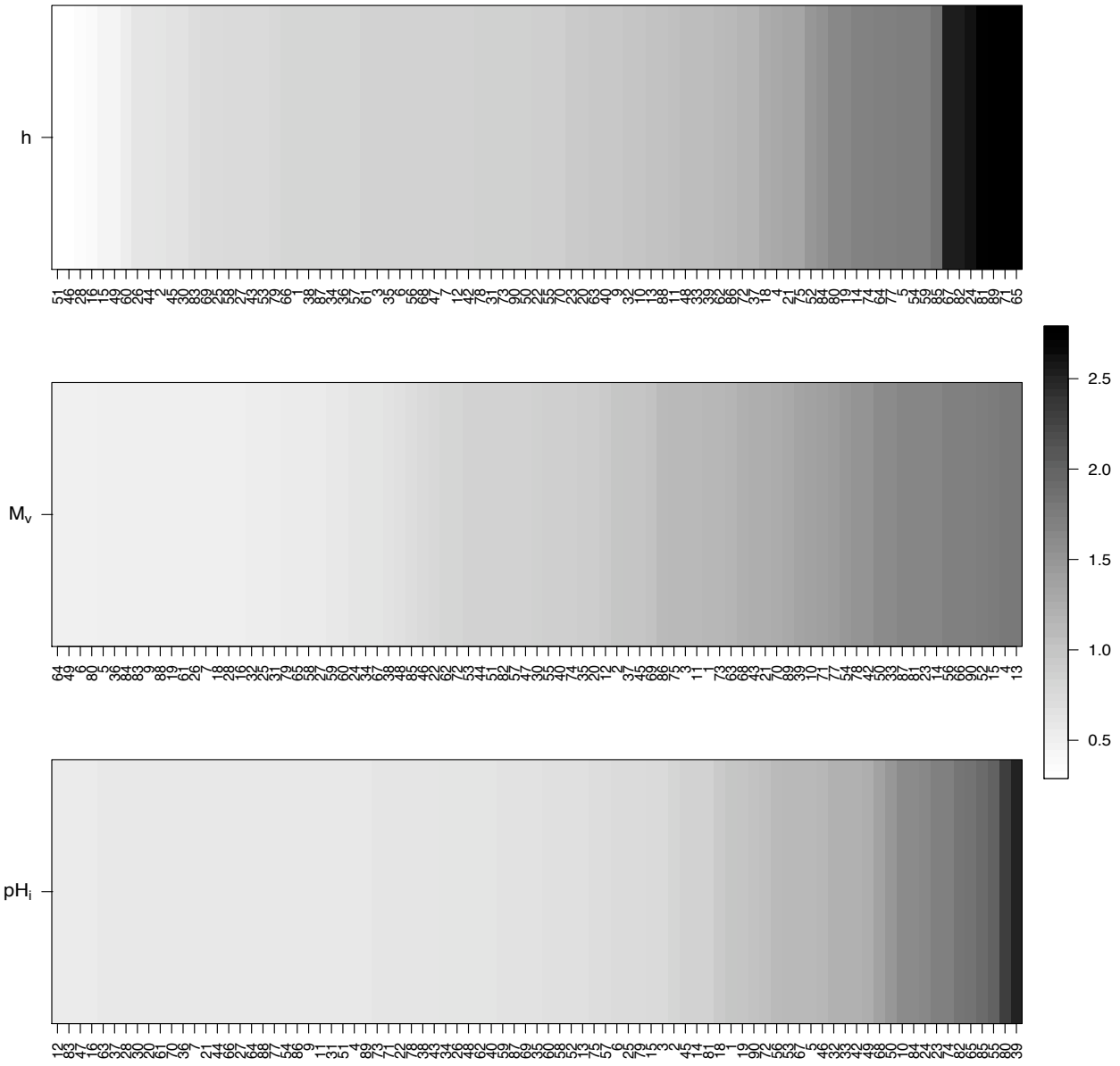

Figure 6: Posterior means of $\beta_{i,j}$s for the three clusters in Figure 5. The sites are sorted according to the increasing value of posterior means.

Table 3 lists site-specific results from `EvoRadical`. The sites listed have high posterior probabilities (> 0.95) of being in the different site classes. This was the criterion that was used to identify significant sites in [9]. The results presented here correspond to Model A1 in [9] which uses ω for the nonsynonymous to synonymous substitution rate ratio for codons encoding amino acids with properties in the same partition, and γ measures the nonsynonymous to synonymous substitution rate ratio between codons for properties

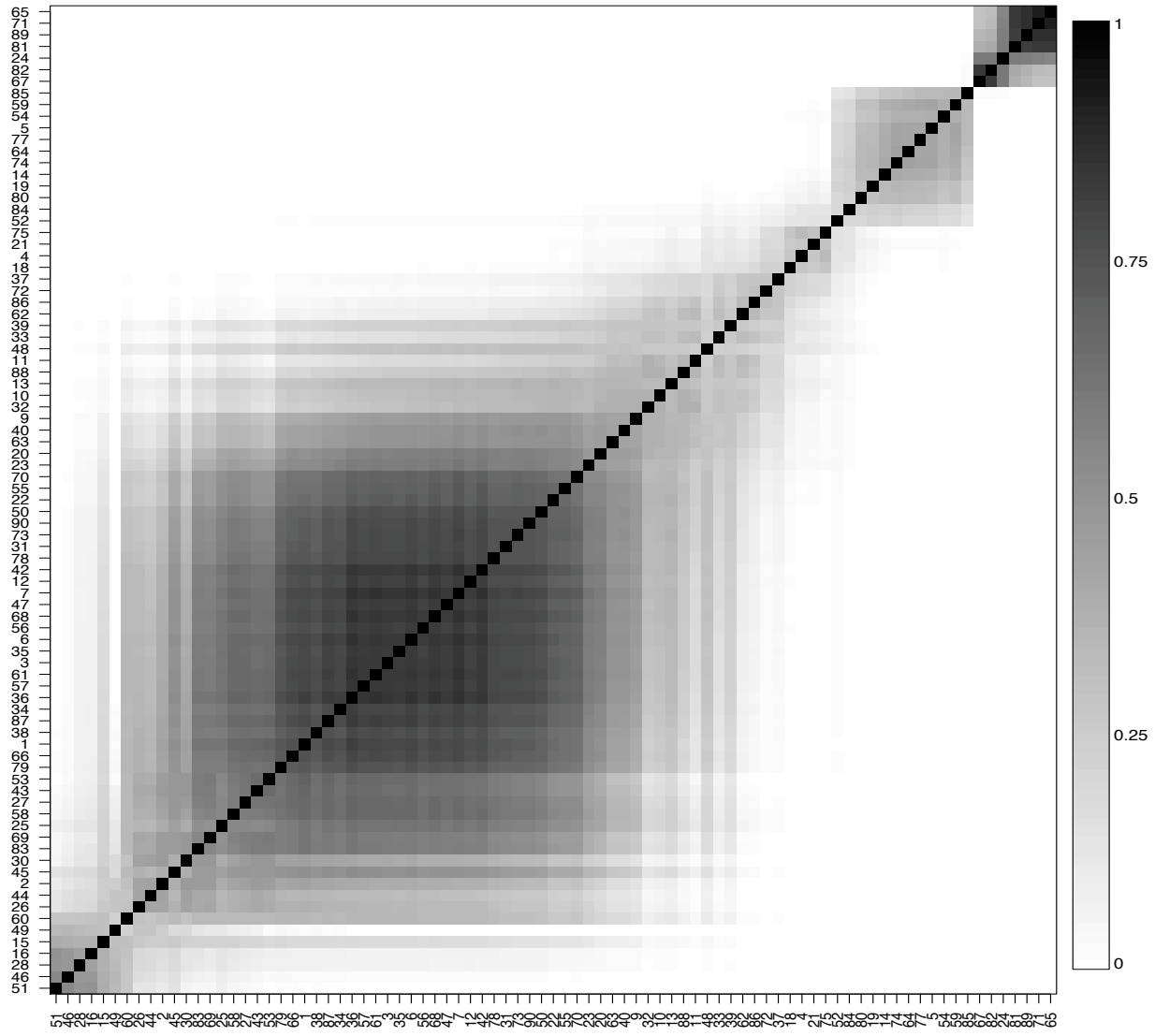

Figure 7: Marginal posterior probabilities of any two sites being grouped together in the first cluster in Figure 5. The sites are sorted according to the increasing value of posterior means of $\beta_{i,j}$s.

belonging to different partitions. While the sites listed for p somewhat match results from the other methods, the results for M_v are not in agreement. This is probably due to the fact that partitions are not always directly comparable with the amino acid distances. For example, under the volume partition of [8], both glycine and valine are small and glutamine is large, while looking at the volume scores glycine is very different from valine and glutamine. Thus, our models would consider a change from glycine to valine as radical, whereas for the partition-based method of [9], there would be no change. The fact that the user has to define a property-specific partition in advance, as opposed to directly working with the physicochemical distances, is one of the disadvantages of partition-based methods.

Table 1: Comparing results from models in [10] and the new semiparametric model. Sites marked in bold are the ones which are in the region of interest - for h this is where radical changes were encouraged and for M_v where small changes were encouraged while generating the sequences. Underlined sites are identified by both methods.

	Parametric regression	Semiparametric regression
30 sites with largest posterior mean $\hat{\beta}_{i,j}$ for h	4, 5, 6, 10, 14, 18, 19, 21, 22, 23, 24, 33, 48, 52, 54, 59, **62**, **64**, **65**, **67**, **71**, **74**, **75**, **77**, **80**, **81**, **82**, **84**, **85**, **89**	4, 5, 14, 18, 19, 21, 24, 33, 37, 39, 48, 52, 54, 59, **62**, **64**, **65**, **67**, **71**, **72**, **74**, **75**, **77**, **80**, **81**, **82**, **84**, **85**, **86**, **89**
30 sites with lowest posterior mean $\hat{\beta}_{i,j}$ for M_v	5, 6, 7, 9, 16, 19, 24, 25, 26, 27, 28, **31**, **32**, **36**, **44**, **49**, **51**, **58**, **59**, **60**, 61, 64, 65, 67, 79, 80, 82, 83, 85, 88	5, 6, 7, 9, 16, 18, 19, 24, 25, 26, 27, 28, **31**, **32**, **34**, **36**, **38**, **49**, **58**, **59**, **60**, 61, 64, 65, 67, 79, 80, 83, 84, 88

Table 2: Sites identified as significant by TreeSAAP for the different properties. Values in parentheses denote the cut-off values for the z-test statistic. Sites marked in bold are in the region of interest.

Property	Radically changing (1.645)	Radically changing (3.695)	Conserved (1.645)	Conserved (3.695)
h	5, 59, **65**, **67**, **71**, **74**, **81**, **82**, **89**	**74**	36, 83	None
p	21, 24, 37, **64**, **65**, **67**, **71**, **74**, **75**, **81**, **82**, **89**	None	7, 18, 36, 49, 55	None
M_v	10, 33, 66	None	5, 18, **36**, **49**	None
V^0	10, 13, 33, 66	None	18, **36**	None
pH_i	39, 55, 72	None	11, 64, 72	None

Illustration with Lysin Data

Our proposed model was applied to the sperm lysin data set which consisted of cDNA from 25 abalone species with 135 codons in each sequence [35]. Sites with alignment gaps were removed from all sequences, which resulted in 122 codons for the analysis presented here. The phylogeny of [35] and the codon substitution model `M8` in `PAML`, version 3.15, [20] was used to generate the ancestral sequences. The model `M8` uses a discretized beta distribution to model ω values between zero and one with probability p_0 and allows for an additional positive selection category with $\omega > 1$ and probability p_1.

The lysin data was analyzed with the model in Section 2 with the 32 amino acid properties listed in Table 5 in the Appendix. Only sites which showed at least one nonsynonymous change were retained for the final analysis, which led to a data set with 94 sites. We used $K = 25$ and $L = 35$ as truncation levels for this data. The prior distributions with the following hyperparameters were used in the analysis. The DP concentration parameters ρ and γ_k were assumed to follow a $\mathsf{Ga}(1, 1)$. λ, the prior probability for $\phi_{l,k}$ being 0, was assumed to follow a $\mathsf{Beta}(2, 8)$ which implied that about 20% of the unique $\beta_{i,j}$s were expected to be 0 *a priori*. a_κ and b_κ, the hyperparameters for the prior of $\vartheta^2_{l,k}$ when $\phi_{l,k}$ is 0, were chosen as 2 and 100

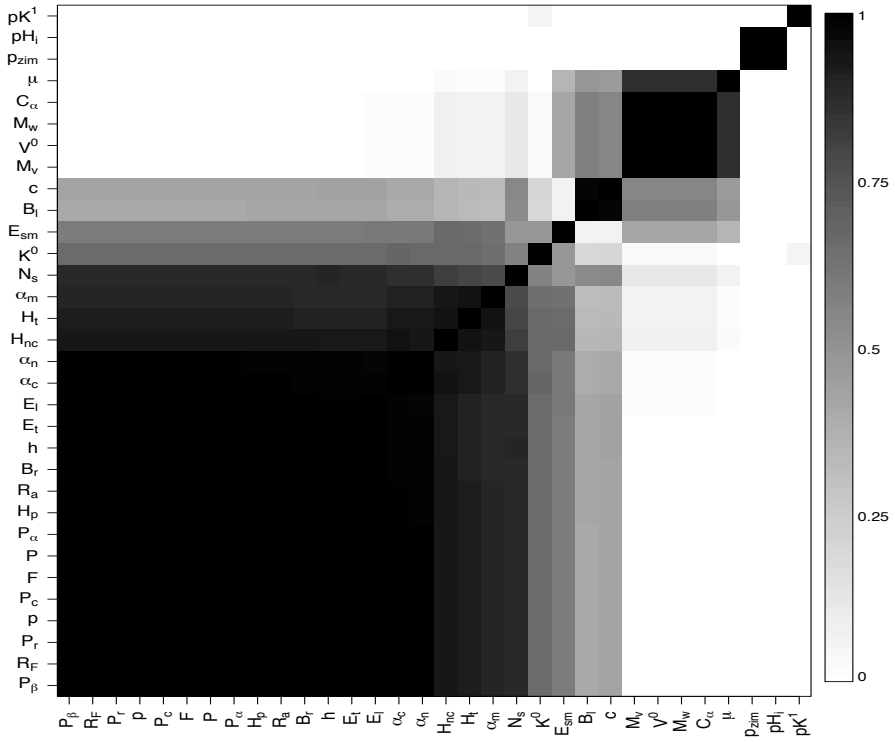

Figure 8: Marginal posterior probabilities of any two properties being in the same cluster for the lysin data.

Table 3: Sites that have high posterior probabilities (> 0.95) of belonging to each site class for the different partitions for `EvoRadical`. Sites marked in bold are in the region of interest.

Property	$\omega \leq 1, \gamma \leq 1$	$\omega \leq 1, \gamma > 1$	$\omega > 1, \gamma \leq 1$	$\omega > 1, \gamma > 1$
p	None	None	None	1, 2, 5, 7, 10, 11, 12, 13, 14, 18, 19, 20, 26, 27, 30, 32, 33, 34, 36, 37, 42, 43, 47, 53, 57, 59, **61**, **62**, **63**, **64**, **66**, **67**, **68**, **69**, **72**, **73**, **74**, **75**, **77**, **82**, **83**, **86**, **87**, **88**, **90**
M_v	None	None	None	2, 7, 9, 18, 19, 20, 22, 27, 31, 32, 36, 38, 53, 55, 61, 62, 64, 67, 72, 74, 86

which implied a prior mean of 0.01. When $\phi_{l,k}$ is different from zero, $a^*_\sigma = 2$ and $b^*_\sigma = 10$ control the prior for $\vartheta^2_{l,k}$. V_0, the scale factor for $\vartheta^2_{l,k}$, was fixed at the ratio of prior means of σ^2 and τ^2_i (the variance terms in the regression model in [10] for which we had used prior means of 0.1 and 0.01 respectively). Finally, the α_ks were assumed to follow a $\mathsf{N}(1, 0.25)$ to conform to our prior assumption of neutrality for the properties. Results are based on 20000 iterations, of which the first 10000 were burn-in. Convergence was assessed by visual inspection of trace plots of some of the parameters and there did not seem to be any obvious problems with convergence.

Figure 8 shows the marginal posterior probabilities of any two properties being assigned to the same cluster. There seem to be four mostly distinct clusters in the properties in our list. The biggest cluster consists of 20 properties that are related to polarity and hydropathy. All 20 properties are assigned to this cluster with very high probability. The next cluster is comprised of the properties B_l, and c. There is also a fairly big cluster whose members are related to volume (M_v, V^0, M_w, C_α, μ). p_{zim}, which is correlated with p to some extent, is clustered with pH_i with which it shows a large correlation value (about 0.9). There is some uncertainty regarding the membership of K^0 and E_{sm}, since both of them are assigned to the largest cluster about 50% of the time, while E_{sm} is clustered with properties related to volume to a lesser extent. pK^1 is the only property that is almost never clustered with other properties.

Site specific results based on the posterior means (denoted by $\hat{\beta}_{i,j}$s), for one representative property each from the four clusters in Figure 8 are shown in Figure 9. The sites are sorted according to the increasing value of mean $\hat{\beta}_{i,j}$ for each image. Sites on the far right radically change properties in each group. For example, most of the sites that appear on the far right for cluster 1 (represented by h) have $\hat{\beta}_{i,j}$ values of 1-1.4. There seem to be more sites radically changing properties in cluster 1 than in clusters 2 (represented by c) or 3 (represented by M_v). The first three clusters also have a fairly large number of sites with mean $\hat{\beta}_{i,j}$ between 0 and 1. This is different from what we see for cluster 4 (represented by p_{zim}), which corresponds to properties p_{zim} and pH_i. A large number of sites in cluster 4 strongly conserve the properties, as is evident by the very small mean $\hat{\beta}_{i,j}$s for sites in the far left, unlike in the other clusters.

Figure 10 shows the posterior summaries of $\beta_{i,j}$s different from zero for sites 82, 99, 120 and 127 for properties belonging to different clusters. Of these, sites 120 and 127 were found to be under positive

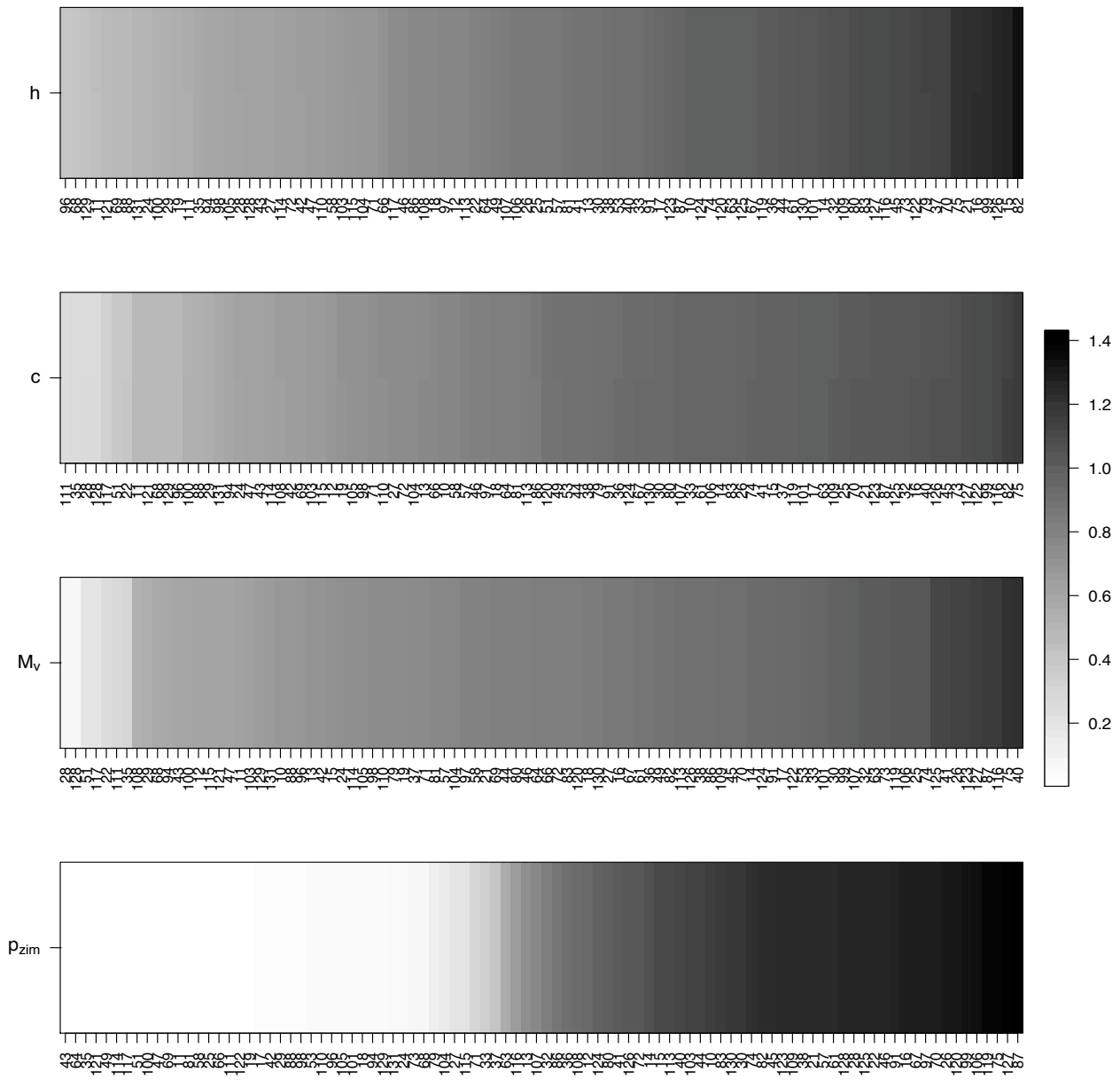

Figure 9: Posterior means $\hat{\beta}_{i,j}$s for the four clusters (denoted by representative properties) in Figure 8. The sites are sorted according to the increasing value of posterior means.

selection by `PAML`, while sites 82, 99 and 127 were identified as radically changing some of the properties by the regression model in [10]. The sites show different behavior for the different properties, for example, site 82 shows radical changes for h, while it conserves M_v. We can also see similarities in the posterior summaries across sites. For example, for property pK^1 sites 82, 120 and 127 have similar values for $\beta_{i,j}$.

Table 4 lists sites that are highly conserved with posterior mean $\hat{\beta}_{i,j}$s less than 0.4 for the different

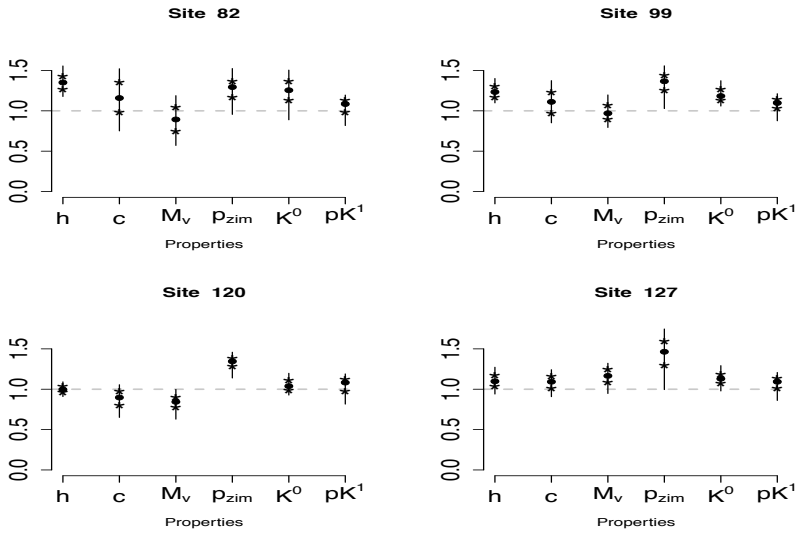

Figure 10: Posterior summaries of $\beta_{i,j}$s different from zero for sites 82, 99, 120 and 127. The first 4 properties on the x-axis belong to 4 different clusters and the next 2 do not belong to any specific cluster all the time. The vertical lines are 90% posterior intervals of the $\beta_{i,j}$s that are different from 0, the medians (filled circles) and the 25^{th} and 75^{th} percentiles (stars) are highlighted.

clusters. The largest number of highly conserved sites appears in cluster 4, which includes properties p_{zim} and pH_i, in agreement with Figure 9. Some of these sites like 35, 51, 111 and 117 also conserve properties in clusters 2 and 3. A number of them, such as sites 24, 28, 35, 51, 53, 58, 66, 94, 96, 104, 105, 111, 117, and 128 are also identified as sites under negative selection by methods that take into account the relative rate of nonsynonymous to synonymous rate ratio, such as those implemented in `PAML` [20].

Table 4: Strongly conserved sites ($\hat{\beta}_{ij} < 0.4$) for different clusters.

Cluster	Site Number
1	96
2 and 3	22, 28, 35, 51, 111, 117, 128
4	11, 17, 18, 19, 24, 25, 27, 29, 33, 35, 42, 43, 47, 49, 51, 53, 58, 64, 66, 68, 69, 71, 73, 79, 81, 88, 94, 96, 98, 100, 101, 104, 105, 110, 111, 114, 115, 117, 121, 122, 129, 131

The results are fairly robust to the choice of different hyperparameter values. Note that the scale factor for $\vartheta^2_{l,k}$ ultimately affects the variation in the $\beta_{i,j}$ values, and it is advisable to choose it so that the prior variance for the unique $\beta_{i,j}$s is not too large.

4 Conclusions

In this paper, we present a Bayesian hierarchical regression model with a nested infinite relational model on the regression coefficients. The model is capable of identifying sites which show radical or conserved amino acid changes. The (almost sure) discreteness of the DP realizations induces clustering at the level of

properties which is analogous to the factor model in [11], with the advantage being that the nonparametric method automatically determines the appropriate number of clusters. The multi-level clustering ability of the NIRM also induces clustering at the level of sites and allows us to capture skewness and heterogeneity in the distribution of the random effects distribution associated with each cluster of properties.

The main advantage of the models we have described is their ability to simultaneously handle multiple properties with potentially correlated effects on molecular evolution. Our simulations suggest that our models are flexible but robust, being capable of dealing with a range of situations including those where properties are perfectly correlated, as well as those where all properties are uncorrelated. Our semiparametric regression models also work well, particularly in comparison with the regression model in [10], TreeSAAP and `EvoRadical`, when applied to DNA sequence data generated from an evolutionary model. In addition, the analysis of the lysin data suggests that the model leads to reasonable results.

The NIRM that is the basis of our model defines a separately exchangeable prior on matrices. This means that the prior is invariant to the order in which properties and sites are included. This is due to the fact that the rows as well as the columns of the parameter of interest are independent draws from a DP. From the point of view of modeling multiple properties, this is a highly desirable property. However, assuming that DNA sites are exchangeable can be questionable. Although this is a potential limitation of our model, we should note that the assumption of independence across sites (which is a stronger assumption than exchangeability) underlies all the methods discussed in Section 1. If information about the 3-dimensional structure of the encoded protein or other sequence specific information that can guide the construction of the dependence model is available, our model could be easily extended to account for this feature. In the absence of such information, exchangeability across DNA sites seems to be a reasonable prior assumption. Indeed, in contrast to the most common independence assumption, our exchangeability assumption allows us to explain correlations at the level of sites.

Finally, it is important to note that the "observed" distances are not really directly observed, but are instead constructed from ancestral sequences and, therefore, subject to error. A simple way to account for this additional level of uncertainty is to modify the computation of expected distances by incorporating the ideas of [36]. This approach was previously employed in [10], with little impact on the final results.

Appendix: Details about the Gibbs sampler

The truncations and the introduction of the configuration variables imply that (2) and (3) can be written as

$$\zeta_j|\{\pi_k\} \sim \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*} \quad \xi_{i,k}|\{w_{l,k}\} \sim \sum_{l=1}^{L} w_{l,k} \delta_{\varphi_{l,k}} \tag{5}$$

with $\varphi_{l,k} \sim G_{0lk}$ and π_k and $w_{l,k}$ being the appropriate stick breaking weights. Writing the model as in (5) helps in obtaining the forms of the full conditionals as below.

The column indicators ζ_j for $j = 1, \ldots, J$ are sampled from a multinomial distribution with probabilities

$$P(\zeta_j = k|\cdots) = q_j^k \propto \sum_{l=1}^{L} \prod_{\{i:\xi_{i,k}=l\}} \pi_k \mathsf{N}(y_{i,j}^*|\phi_{l,k}x_{i,j}^*, \vartheta_{l,k}^2),$$

where $\vartheta_{l,k}^2$ is $\vartheta_{l,k}^2$ if $\phi_{l,k} = 0$ or is $\vartheta_{l,k}^2/n_i^O$ if $\phi_{l,k}$ is different from zero. π_k is sampled in two parts: first, by generating v_k from a $\mathsf{Beta}(1 + m_k, \rho + \sum_{s=k+1}^{K} m_s)$ for $k = 1, \ldots, K-1$ and $v_K = 1$, where m_k is the number of columns assigned to cluster k and then, by constructing $\pi_k = v_k \prod_{s=1}^{k-1}(1 - v_s)$.

For $i = 1, \ldots, I$ and $k = 1, \ldots, K$, the indicators $\xi_{i,k}$ are also sampled from a multinomial with probabilities of the form

$$P(\xi_{i,k} = l|\cdots) = p_{i,k}^l \propto \prod_{\{j:\zeta_j=k\}} w_{l,k} \mathsf{N}(y_{i,j}^*|\phi_{l,k}x_{i,j}^*, \vartheta_{l,k}^2).$$

The updated weights $w_{l,k}$ are sampled in a manner similar to the π_k, i.e., $u_{l,k}$ are generated from a $\mathsf{Beta}(1+n_{l,k}, \gamma_k + \sum_{r=l+1}^{L} n_{lr})$ for $l = 1, \ldots, L-1$ and $u_{Lk} = 1$, where $n_{l,k}$ is the number of $\beta_{i,j}$s assigned to atom l of cluster k and then, by constructing $w_{l,k} = u_{l,k} \prod_{r=1}^{l-1}(1-u_{r,k})$.

Following [18], the DP concentration parameters ρ and γ_k are sampled in two steps by introducing auxiliary variables η_1 and η_2. First, sample η_1 from

$$p(\eta_1|\rho, \cdots) = \mathsf{Beta}(\rho + 1, J)$$

and then ρ from

$$p(\rho|\eta_1, \cdots) = \frac{a_\rho + n^*_\zeta - 1}{a_\rho + n^*_\zeta - 1 + J(b_\rho - \mathsf{log}(\eta_1))}\, \mathsf{Ga}(a_\rho + n^*_\zeta, b_\rho - \mathsf{log}(\eta_1)) + \frac{J(b_\rho - \mathsf{log}(\eta_1))}{a_\rho + n^*_\zeta - 1 + J(b_\rho - \mathsf{log}(\eta_1))}\, \mathsf{Ga}(a_\rho + n^*_\zeta - 1, b_\rho - \mathsf{log}(\eta_1)),$$

where n^*_ζ is the number of unique column indicators ζ_j. Similarly, for each $k = 1, \ldots, K$,

$$p(\eta_2|\gamma_k, \cdots) = \mathsf{Beta}(\gamma_k + 1, I),$$

$$p(\gamma_k|\eta_2, \cdots) = \frac{a_\gamma + m^*_{\xi,k} - 1}{a_\rho + m^*_{\xi,k} - 1 + I(b_\gamma - \mathsf{log}(\eta_2))}\, \mathsf{Ga}(a_\gamma + m^*_{\xi,k}, b_\gamma - \mathsf{log}(\eta_2)) + \frac{I(b_\gamma - \mathsf{log}(\eta_2))}{a_\gamma + m^*_{\xi,k} - 1 + I(b_\gamma - \mathsf{log}(\eta_2))}\, \mathsf{Ga}(a_\gamma + m^*_{\xi,k} - 1, b_\gamma - \mathsf{log}(\eta_2)),$$

where $m^*_{\xi,k}$ is the number of unique row indicators $\xi_{i,k}$, for a specific cluster of columns k.

To sample the unique $\varphi_{l,k} = (\phi_{l,k}, \vartheta^2_{l,k})$s given in (4), we introduce a set of indicator variables $\psi_{l,k}$ which take the value 1 when $\phi_{l,k}$ is different from zero. For $l = 1, \ldots, L$ and $k = 1, \ldots, K$, $\psi_{l,k}$, $\vartheta^2_{l,k}$ and $\phi_{l,k}$ are jointly sampled in the following way - $\psi_{l,k}$ is sampled by integrating out $\phi_{l,k}$ and $\vartheta^2_{l,k}$ from its full conditional, $\vartheta^2_{l,k}$ is sampled conditional on $\psi_{l,k}$ and $\phi_{l,k}$ is sampled conditional on both the corresponding $\psi_{l,k}$ and $\vartheta^2_{l,k}$, i.e.,

$$p(\psi_{l,k}, \vartheta^2_{l,k}, \phi_{l,k}|\cdots) = p(\psi_{l,k}|\cdots)p(\vartheta^2_{l,k}|\psi_{l,k}, \cdots)p(\phi_{l,k}|\psi_{l,k}, \vartheta^2_{l,k}, \cdots)$$

with the individual expressions obtained as follows.

First, let $\Omega^{i,j}_{l,k} = \{(i,j) : \xi_{i\zeta_j} = l, \zeta_j = k\}$. Then,

$$p(\psi_{l,k}|\cdots) \propto \lambda \int \left[\prod_{\Omega^{i,j}_{l,k}} \mathsf{N}(y^*_{i,j}|0, \vartheta^2_{l,k})\right] \mathsf{IG}(\vartheta^2_{l,k}|a_\kappa, b_\kappa)\mathrm{d}(\vartheta^2_{l,k}) +$$

$$(1-\lambda) \int\int \left[\prod_{\Omega^{i,j}_{l,k}} \mathsf{N}(y^*_{i,j}|\phi_{l,k}x^*_{i,j}, \vartheta^2_{l,k}/n^O_i)\right] \mathsf{N}(\phi_{l,k}|\alpha_k, \vartheta^2_{l,k}/V_0)\mathsf{IG}(\vartheta^2_{l,k}|a^*_\sigma, b^*_\sigma)\mathrm{d}(\phi_{l,k})\mathrm{d}(\vartheta^2_{l,k}).$$

$$p(\vartheta^2_{l,k}|\psi_{l,k}, \cdots) = \begin{cases} \mathsf{IG}\left(\dfrac{I^*J^*}{2} + a_\kappa, \left[\dfrac{1}{b_\kappa} + \sigma_{1,scale}\right]^{-1}\right) & \text{if } \psi_{l,k} = 0 \\ \mathsf{IG}\left(\dfrac{I^*J^*}{2} + a^*_\sigma, \left[\dfrac{1}{b^*_\sigma} + \sigma_{2,scale}\right]^{-1}\right) & \text{if } \psi_{l,k} = 1, \end{cases}$$

where $I^*J^* = \sum_{i,j} 1_{\{\xi_{i\zeta_j}=l,\zeta_j=k\}}$ and the update terms are given by $\sigma_{1,scale} = \sum_{\Omega^{i,j}_{l,k}} \frac{y^{*2}_{i,j}}{2}$ and $\sigma_{2,scale} = \frac{\alpha_k^2 V_0}{2} + \sum_{\Omega^{i,j}_{l,k}} \frac{n_i^O y^{*2}_{i,j}}{2} - \frac{(\alpha_k V_0 + \sum_{\Omega^{i,j}_{l,k}} n_i^O y^*_{i,j} x^*_{i,j})^2}{2(V_0 + \sum_{\Omega^{i,j}_{l,k}} n_i^O x^{*2}_{i,j})}$.

$$p(\phi_{l,k}|\psi_{l,k}, \vartheta^2_{l,k}, \cdots) = \begin{cases} 0 & \text{if } \psi_{l,k} = 0 \\ \mathsf{N}(m_\phi, C_\phi) & \text{if } \psi_{l,k} = 1, \end{cases}$$

where $m_\phi = \left(\frac{\alpha_k V_0 + \sum_{\Omega^{i,j}_{l,k}} n_i^O y^*_{i,j} x^*_{i,j}}{V_0 + \sum_{\Omega^{i,j}_{l,k}} n_i^O x^{*2}_{i,j}} \right)$ and $C_\phi = \frac{\vartheta^2_{l,k}}{V_0 + \sum_{\Omega^{i,j}_{l,k}} n_i^O x^{*2}_{i,j}}$.

The full conditional of λ is given by

$$p(\lambda|\cdots) \sim \mathsf{Beta}(a_\lambda + \sum_{l,k} 1_{\{\psi_{l,k}=0\}}, b_\lambda + \sum_{l,k} 1_{\{\psi_{l,k}=1\}}).$$

Finally, for $k = 1, \ldots, K$, the full conditional of α_k is given by

$$p(\alpha_k|\cdots) \sim \mathsf{N}(m^*_\alpha, C^*_\alpha)$$

where $C^*_\alpha = \frac{1}{\left(\frac{1}{C_\alpha} + \sum_{\{l:\psi_{l,k}=1\}} \frac{V_0}{\vartheta^2_{l,k}} \right)}$ and $m^*_\alpha = C^*_\alpha \left(\frac{m_\alpha}{C_\alpha} + \sum_{\{l:\psi_{l,k}=1\}} \frac{V_0 \phi_{l,k}}{\vartheta^2_{l,k}} \right)$.

Software availability: The R code implementing the models in the paper are available from the authors on request.

Appendix: Properties used in the analysis

A few of the properties were chosen because of their functional importance. Some of the other properties have been previously used in analyses by [25].

Additional material

Additional simulations are available at the end of the article.

References

1. Pakula AA, Sauer RT: **Genetic analysis of protein stability and function**. *Annual Review Genetics* 1989, **23**:289–310.
2. Zuckerkandl E, Pauling L: **Evolutionary divergence and convergence in proteins**. In *Evolving Genes and Proteins*, Academic Press, New York 1965:97–166.
3. Sneath PHA: **Relations between chemical structure and biology**. *Journal of Theoretical Biology* 1966, **12**:157–195.
4. Miyata T, Miyazawa S, Yasunaga T: **Two types of amino acid substitution in protein evolution**. *Journal of Molecular Evolution* 1979, **12**:219–236.
5. Xia X, Li WH: **What amino acid properties affect protein evolution?** *Journal of Molecular Evolution* 1998, **47**:557–564.
6. McClellan DA, McCracken KG: **Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains**. *Molecular Biology and Evolution* 2001, **18**:917–925.

Table 5: List of 32 amino acid properties used in the analysis. Properties marked by * are from [37].

AAindex accession number (if available)	Property	Symbol	AAindex accession number (if available)	Property	Symbol
KYTJ820101	Hydropathy	h	*	Helical contact area	C_a
GRAR740103	Molecular volume	M_v	ZIMJ680104	Isoelectric point	pH_i
MANP780101	Surrounding hydrophobicity	H_p	OOBM770103	Long-range non-bonded energy	E_l
ZIMJ680103	Polarity(Zimmerman)	p_{zim}	*	Mean r.m.s. fluctuation displacement	F
CHOP780201	Alpha-helical tendencies	P_α	FASG760101	Molecular weight	M_w
GRAR740102	Polarity(Grantham)	p	*	Normalized consensus hydrophobicity	H_{nc}
PONP800108	Average number of surrounding residues	N_s	COHE430101	Partial specific volume	V^0
*	Power to be at the C-terminal	α_c	WOEC730101	Polar requirement	P_r
GRAR740101	Composition	c	*	Power to be at the middle of alpha-helix	α_m
*	Compressibility	K^0	*	Power to be at the N-terminal	α_n
FAUJ880113	Equilibrium constant (ionization of COOH)	$pK^{'}$	MCMT640101	Refractive index	μ
CHOP780202	Beta-structure tendencies	P_β	OOBM770102	Short and medium range non-bonded energy	E_{sm}
ZIMJ680102	Bulkiness	B_l	PONP800107	Solvent accessible reduction ratio	R_a
*	Buriedness	B_r	*	Thermodynamic transfer hydrophobicity	H_t
*	Chromatographic index	R_F	OOBM770101	Total non-bonded energy	E_t
CHAM830101	Coil tendencies	P_c	CHOP780101	Turn tendencies	P

7. McClellan D, Palfreyman E, Smith M, Moss J, Christensen R, Sailsbery J: **Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins**. *Molecular Biology and Evolution* 2005, **22**:437–455.
8. Sainudiin R, Wong WSW, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R: **Detecting site-specific physicochemical selective pressures: applications to the class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system**. *Journal of Molecular Evolution* 2005, **60**:315–326.
9. Wong WSW, Sainudiin R, Nielsen R: **Identification of physicochemical selective pressure on protein encoding nucleotide sequences**. *BMC Bioinformatics* 2006, **7**:148–157.
10. Datta S, Prado R, Rodriguez A, Escalante AA: **Characterizing molecular evolution: a hierarchical approach to assess selective influence of amino acid properties**. *Bioinformatics* 2010, **26**:2818–2825.
11. Datta S, Prado R, Rodriguez A: **Bayesian factor models in characterizing molecular adaptation**. Tech. rep., University of California, Santa Cruz 2012.
12. Ferguson T: **A Bayesian analysis of some nonparametric problems**. *The Annals of Statistics* 1973, **1**:209–230.
13. Sethuraman J: **A constructive definition of Dirichlet priors**. *Statistica Sinica* 1994, **4**:639–650.
14. Shafto P, Kemp C, Mansinghka V, Gordon M, Tenenbaum JB: **Learning cross-cutting systems of categories**. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* 2006.
15. Rodriguez A, Ghosh K: **Modeling relational data using nested infinite relational models**. Tech. rep., University of California, Santa Cruz 2011.
16. Lo AY: **On a class of Bayesian nonparametric estimates: I. density estimates**. *The Annals of Statistics* 1984, **12**:351–357.
17. Escobar MD: **Estimating normal means with a Dirichlet process prior**. *Journal of the American Statistical Association* 1994, **89**:268–277.
18. Escobar MD, West M: **Bayesian density estimation and inference using mixtures**. *Journal of the American Statistical Association* 1995, **90**:577–588.
19. Blackwell D, Macqueen JB: **Ferguson distribution via Pólya urn schemes**. *The Annals of Statistics* 1973, **1**:353–355.
20. Yang Z: **Phylogenetic analysis using parsimony and likelihood methods**. *Journal of Molecular Evolution* 1997, **42**:294–307.
21. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV–1 envelope gene**. *Genetics* 1998, **148**:929–936.
22. Kemp C, Tenenbaum JB, Griffiths TL, Yamada T, Ueda N: **Learning systems of concepts with an infinite relational model**. In *Proceedings of the 22nd Annual Conference on Artificial Intelligence* 2006.
23. Xu Z, Tresp V, Yu K, Kriegel HP: **Infinite hidden relational models**. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence* 2006.
24. Dunson DB, Xue Y, Carin L: **The matrix stick-breaking process: flexible Bayes meta-analysis**. *Journal of the American Statistical Association* 2008, **103**:317–327.
25. Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA: **TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees**. *Bioinformatics* 2003, **19**:671–672.
26. MacEachern SN: **Estimating normal means with a conjugate style Dirichlet process prior**. *Commnunications in Statistics, Part B - Simulation and Computation* 1994, **23**:727–741.
27. MacEachern SN, Muller P: **Estimating mixture of Dirichlet process models**. *Journal of Computational and Graphical Statistics* 1998, **7**:223–238.
28. Ishwaran H, James LF: **Gibbs sampling methods for stick-breaking priors**. *Journal of the American Statistical Association* 2001, **96**:161–173.
29. Ishwaran H, Zarepour M: **Dirichlet process sieves in finite normal mixtures**. *Statistica Sinica* 2002, **12**:941–963.

30. Green PJ, Richardson S: **Modelling heterogeneity with and without the Dirichlet process**. *Scandinavian Journal of Statistics* 2001, **28**:355–375.

31. Jain S, Neal RM: **A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model**. *Journal of Computational and Graphical Statistics* 2004, **13**:158–182.

32. Blei DM, Jordan MI: **Variational inference for Dirichlet process mixtures**. *Bayesian Analysis* 2006, **1**:121–144.

33. Walker SG: **Sampling the Dirichlet mixture model with slices**. *Communications in Statistics - Simulation and Computation* 2007, **36**:45.

34. Rodriguez A, Dunson DB, Gelfand AE: **The nested Dirichlet process**. *Journal of the American Statistical Association* 2008, **103**:534–546.

35. Yang Z, Swanson W, Vacquier V: **Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineage and sites**. *Molecular Biology and Evolution* 2000, **17**:1446–1455.

36. Minin VN, Suchard MA: **Counting labeled transitions in continuous-time Markov models of evolution**. *Journal of Mathematical Biology* 2008, **56**:391–412.

37. Gromiha MM, Oobatake M, Sarai A: **Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins**. *Biophysical Chemistry* 1999, **82**:51–67.

Online Supplement for Bayesian Semiparametric Regression Models to Characterize Molecular Evolution

Saheli Datta saheli@ams.ucsc.edu
Abel Rodriguez abel@ams.ucsc.edu
Raquel Prado raquel@ams.ucsc.edu

Additional Simulations

The setup for the simulations are as follows. We generate values for the distinct regression coefficients ($\phi_{l,k}$) from a $\mathsf{N}(1, 0.25)$. The number of distinct regression coefficients depends on the particular clustering structure for the corresponding simulation. Once we obtain the regression coefficients, we generate observations $y_{i,j}$ from $\mathsf{N}(\phi_{l,k}x_{i,j}, \sigma^2 = 0.001)$. The $x_{i,j}$s are obtained from the lysin data set described below with analyses for 32 properties, which implies $J = 32$ and $I = 94$, unless otherwise mentioned.

We fitted the model in Section 2 of the main paper to the $y^*_{i,j}$s and $x^*_{i,j}$s, with the following modifications: (i) the NIRM is imposed on $\beta_{i,j}$, so $\varphi_{l,k} = \phi_{l,k}$ and (ii) $\phi_{l,k} \sim G_0$ where $G_0 \sim \mathsf{N}(\alpha, \tau^2)$. We used $K = 25$ and $L = 25$ for the simulations. The MCMC algorithm was run with the following hyperpriors: $\rho \sim \mathsf{Ga}(1, 1)$, $\gamma_k \sim \mathsf{Ga}(1, 1)$ for all k, $\alpha \sim \mathsf{N}(1, 0.25)$. $\sigma^2 \sim \mathsf{Inv\text{-}Ga}(100, 10)$ and $\tau^2 \sim \mathsf{Inv\text{-}Ga}(2, 4)$ were chosen such that the prior means corresponded to the true values for these hyperparameters. For all the simulations, results are based on 15000 iterations, of which the first 5000 are burn-in. Convergence was assessed by running two chains where each chain was initialized by randomly assigning the $\beta_{i,j}$s to different partitions. Posterior summaries based on the two chains were consistent with each other.

Simulation Study 3

For this simulation, all the columns were assumed to belong to the same cluster. Six distinct $\phi_{l,k}$ values were used to generate the observations $y_{i,j}$ from the appropriate Normal density. Posterior summaries of the column cluster indicators revealed that the analysis indeed concludes all the columns belong to the same cluster. Clustering within columns was also inferred correctly. There was no uncertainty associated with the cluster memberships at either level. In this case also, posterior means of the $\beta_{i,j}$s showed very close agreement with true values of $\phi_{l,k}$ used to generate the data, as shown in Figure 1.

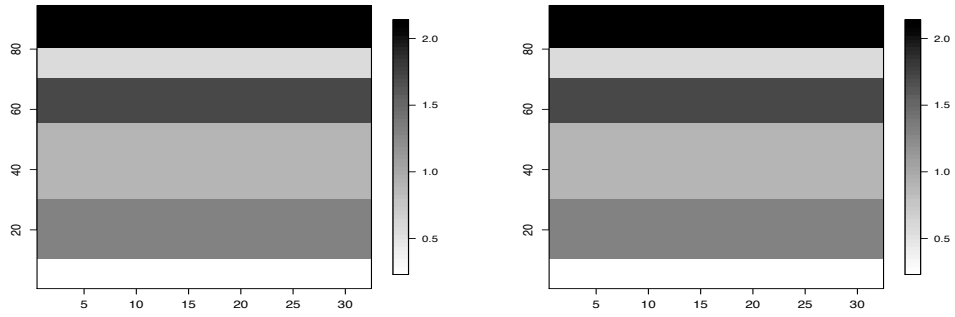

Figure 1: Image plots for true $\beta_{i,j}$ values (left panel) and posterior means $\hat{\beta}_{i,j}$s (right panel).

Simulation Study 4

For this simulation we chose a scenario where each column was different from the other. The number of columns for this simulation was 10 (so, $J = 10$). The number of rows in this simulation was 30. The first 18 rows were assumed to have the same $\phi_{l,k}$ for all the columns while the remaining $\phi_{l,k}$s were generated independently from $\mathsf{N}(1, 0.25)$.

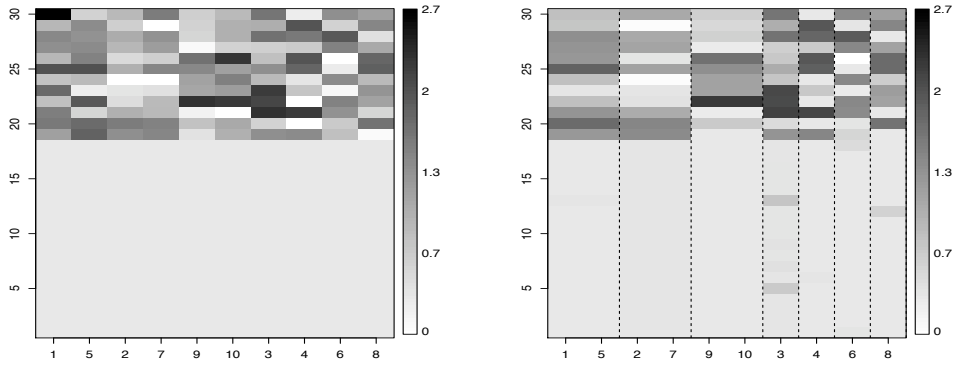

Figure 2: Image plots for true $\beta_{i,j}$ values (left panel) and posterior means of the $\beta_{i,j}$s (right panel). The columns have been arranged according to the clustering inferred by the model.

In this case, the model infers that there are 7 clusters for the columns, with very little uncertainty about the cluster memberships. For each cluster of column, the model assigned the first 18 rows to the same cluster with fairly high probability (> 0.83). In case of the remaining 12 rows, rows were assigned to the same cluster if the corresponding true ϕs were close. Figure 2 shows the true $\beta_{i,j}$ values and the estimated posterior means. While the images seem reasonably close, small differences do exist. For example, in Figure 2 (right panel) since columns 1 and 5 were assigned to the same cluster with very high probability,

we have $\mathsf{E}\{\beta_{30,1}|\text{Data}\} = \mathsf{E}\{\beta_{30,5}|\text{Data}\} = 0.8$, while the true values were $\beta_{30,1} = 2.65$ and $\beta_{30,5} = 0.56$ respectively. In spite of these differences, for each of the columns that were clustered together more than 2/3 of the true $\beta_{i,j}$s were very close (less than 0.1 difference).

Simulation Study 5

Our final simulation study was designed to investigate the extreme case where the rows and the columns were all generated independently. We considered 10 columns and 30 rows for this scenario. All 300 $\beta_{i,j}$s were generated independently from $\mathsf{N}(1, 0.25)$.

The model correctly infers that all 10 columns are independent. In cases where the true $\beta_{i,j}$ values are close for different i for a fixed j, a few of the rows are sometimes clustered together. As in the previous simulations, the posterior means of the $\beta_{i,j}$s are good estimates of the true $\beta_{i,j}$s as evident from Figure 3.

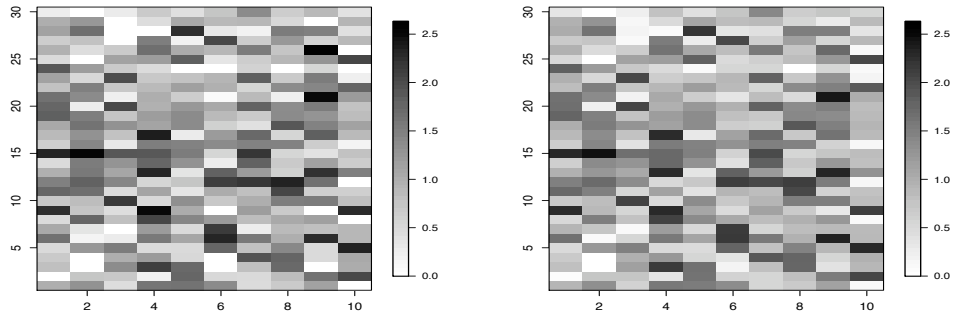

Figure 3: Image plots for true $\beta_{i,j}$ values (left panel) and posterior means of the $\beta_{i,j}$s (right panel).

Discussion

The additional simulations evaluate the performance of the model for extreme cases. In the third simulation, all properties belong to a common cluster, while sites belong to one of several clusters. The results suggest that the model is parsimonious and does not induce unnecessary clusters that are not supported by the data. Scenario 4 was constructed so that the effect of each property on amino acid substitution rates is different. Simulation study 5 was more extreme as both the effect of each property and each site was different. This is the most challenging scenario for our model, as our prior favors the clustering of properties. The fact that the reconstruction of the regression coefficient matrices is reasonably close to the true values suggests that the model, in spite of allowing for a potentially very large number of parameters, will not overfit the data.