BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–8

Characterizing molecular adaptation: a hierarchical approach to assess the selective influence of amino acid properties

Saheli Datta [1], Raquel Prado [1]*, Abel Rodríguez [1] and Ananías A. Escalante [2]

[1]Department of Applied Mathematics and Statistics, University of California Santa Cruz, CA, USA
[2]School of Life Sciences, Arizona State University, Tempe, AZ, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: A number of methods for detecting positive selection in protein coding DNA sequences are based on whether each site/region has a nonsynonymous to synonymous substitution rates ratio ω greater than one. However, a site/region may show a relatively large number of nonsynonymous mutations that conserve a particular property. Recent methods have proposed to consider as evidence for molecular adaptations how conserving, or radically different, nonsynonymous mutations are with respect to some key amino acid properties. While such methods have been useful in providing a qualitative assessment of molecular adaptation, they rely on independent statistical analyses for each amino acid property and typically do not properly adjust for multiple comparisons when selection needs to be assessed at several sites.

Results: We consider a Bayesian hierarchical model that allows us to jointly determine if a set of amino acid properties are being conserved or radically changed while simultaneously adjusting for multiple comparisons at the codon level. We illustrate how this model can be used to characterize molecular adaptation in two datasets: an alignment from 6 class I alleles of the human MHC and a sperm lysin alignment from 25 abalone species. We compare the results obtained with the proposed hierarchical models to those obtained with alternative methods. Our analyses indicate that a more complete quantitative and qualitative characterization of molecular adaptation is achieved by taking into account changes in amino acid properties.

Contact: raquel@ams.ucsc.edu

Availability: The R code for implementing the hierarchical models is freely available at http://www.ams.ucsc.edu/∼raquel/software/.

Supplementary information: Supplementary Data is available at *B*ioinformatics online.

1 INTRODUCTION

There are a variety of tests for detecting departures from neutrality. Such tests could be classified into two groups: those based on the distribution of allele frequencies and/or segregating sites (e.g., Tajima's test, Tajima, 1989), and those that explicitly study patterns of polymorphisms/divergence in genes encoding proteins. In this second category, most methods compare nonsynonymous to synonymous substitution rates at some level. Such methods usually estimate nonsynonymous to synonymous rates ratios to study intra-specific variation (e.g., Li, 1993; Pamilo and Bianchi, 1993), provide a formal comparison between intra- and inter-specific genetic variation (e.g., McDonald and Kreitman, 1991; Sawyer and Hartl, 1992), or investigate character evolution in a phylogeny by using codon-based models (e.g., Suzuki and Gojobori, 1999; Yang *et al.*, 2000a; Yang and Nielsen, 2002; Yang and Swanson, 2002; Suzuki, 2004). In all cases, an excess of nonsynonymous over synonymous mutations is considered as indicative that some form of positive selection may be taking place if a set of assumptions is met (see e.g., Nei and Kumar, 2000; Anisimova and Kosiol, 2009).

Several investigators have tried to enrich the approaches listed above by considering methods that take into account changes in the physicochemical properties of the amino acids; the rationale behind this is that a replacement between two amino acids with similar physicochemical characteristics may not change the phenotype in the same way that a replacement between two amino acids that are radically different does (Hughes *et al.*, 1990; Zhang, 2000). While the value of comparing radical versus non-radical changes as a formal test for positive selection is still under discussion (Dagan *et al.*, 2002; Sainudiin *et al.*, 2005; Hanada *et al.*, 2007), having a flexible framework that allows us to explore changes in amino acid properties can provide useful insight in comparative studies that aim to assess the effect of positive selection (Hanada *et al.*, 2007; Popadin *et al.*, 2007). Among the methods that take into account amino acid properties are those in Xia and Li (1998) and McClellan *et al.* (2005) that use calculations of expected random distributions of possible amino acid changes based on fixed differences between residues given a particular property. A related approach is that of Pupko *et al.* (2003), which tests significant deviations of the mean physicochemical distance from the random expectation along a lineage or across a subtree. The idea is to detect significant deviations from a process consistent with neutrality. For instance, a trend toward radical changes in a particular property is, in principle, inconsistent with neutrality. More recent methods also include those in Sainudiin *et al.* (2005) and Wong *et al.* (2006) which consider codon substitution models that incorporate changes in physicochemical properties. These methods proceed by partitioning the codons on the basis of a specific property, and then categorize

*to whom correspondence should be addressed

substitutions as property-conserving or property-altering depending on whether there is a change in the partition.

The methods listed above have been very useful in providing a qualitative assessment of changes in amino acid properties that may be indicative of selection acting on the divergence of a protein-coding gene. However, such methods perform inferences independently for each amino acid property or a prespecified (property-driven) partition of the amino acids. In addition, some of these methods do not properly adjust for multiple comparisons when tests are performed at several sites. We propose a Bayesian hierarchical model that allows us to determine if a set of amino acid properties are being conserved or radically changed at the codon level. This approach can be used to jointly analyze a set of properties and automatically adjusts for multiple comparisons in cases where several amino acid sites are considered.

2 METHODS

2.1 Bayesian hierarchical model

We develop a Bayesian hierarchical model that compares amino acid distances inferred from ancestral sequences derived from a given phylogeny to distances expected under neutrality for a given set of amino acid properties. We treat the distances as continuous variables, and so in this sense our model is more related to the methods of Pupko *et al.* (2003) and McClellan *et al.* (2005), than to the approaches of Sainudiin *et al.* (2005) and Wong *et al.* (2006), since the latter require a predetermined partition of the amino acids based on a property or a set of properties. The proposed hierarchical model is site-specific, allowing us to determine which sites show nonsynonymous mutations that either conserve or radically change a particular property while properly adjusting for multiple comparisons if several sites are considered.

More specifically, the "expected" distances (under neutrality) and the "observed" distances — inferred from the ancestral sequences derived from the observed DNA alignment and a given phylogeny — are calculated for each property and for each site showing nonsynonymous substitutions. Following Xia and Li (1998) we assume that each codon can mutate to one of at most nine alternative codons through a single nucleotide change. Some of these mutations are nonsynonymous (changes to stop codons are ignored). The number of nonsynonymous mutations possible through a single nucleotide change, corresponding to a particular codon k, is denoted by N_k ($k = 1 : 61$). The absolute difference in property j between nonsynonymous codon pairs at site i differing at one codon position is denoted by $D_{k,l}^{i,j}$, with $l = 1 : N_k$. Let F_k^i denote the frequency of codon k at site i in the alignment under study. Then, the *expected* mean distance for site i and property j is defined by:

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

which is the weighted average of the differences due to possible nonsynonymous changes for a codon by looking at substitutions that result from a single nucleotide mutation.

Once the ancestral sequences at all the internal nodes are obtained under a specific substitution model and a prespecified tree, the observed distances for each physicochemical property are calculated as follows. First, nonsynonymous substitutions are counted by comparing DNA sequences between two neighboring nodes in the phylogeny. Then, the observed mean distance,$y_{i,j} \equiv D_O^{i,j}$, for site i and property j is taken as the mean absolute difference in property j due to all observed nonsynonymous substitutions at that site. Only sites which show at least one nonsynonymous change at the ancestral level are retained for further analysis. Then, equation (1) is used to compute the $x_{i,j}$s only for such sites.

The hierarchical regression model relates $x_{i,j}$ to $y_{i,j}$. The underlying idea is that if site i is neutral with respect to property j, then the observed mean distance at such site, $y_{i,j}$, should not be very different from the expected mean distance at the same site, $x_{i,j}$. On the other hand, if $y_{i,j} << x_{i,j}$, property j is being conserved at site i, and if $y_{i,j} >> x_{i,j}$ property j is radically changing at site i. Additionally, in order to compare different properties and sites, $x_{i,j}$ and $y_{i,j}$ are standardized by dividing them by the maximum possible distance for each property. This allows us to use a common prior distribution on the parameters associated to the properties as will be described below. Let $x_{i,j}^*$ and $y_{i,j}^*$ be the standardized expected and observed mean distances, respectively. Then, for each site i and property j with $i = 1 : I$ and $j = 1 : J$, we consider the model

$$y_{i,j}^* = \beta_{i,j} x_{i,j}^* + \epsilon_{i.j},$$

with

$$\epsilon_{i,j} \sim \begin{cases} \mathsf{N}(0, \kappa^2) & \text{if } \beta_{i,j} = 0 \\ \mathsf{N}(0, \sigma^2/n_i^O) & \text{otherwise.} \end{cases}$$

To understand why we use this mixture for the error distribution, note that some of the $y_{i,j}^*$s can be equal to zero since some of the nonsynonymous changes can result in no difference in the value of the property being measured (for example, both Phenylalanine and Isoleucine have the same value under Grantham Polarity). We therefore approximate a point mass at zero with a suitably tight prior on κ^2 = Var($y_{i,j}^* | \beta_{i,j} = 0$). Also, since the number of observed nonsynonymous changes can be very different for different sites, Var($y_{i,j}^* | \beta_{i,j} \neq 0$) $= \sigma^2/n_i^O$, where n_i^O is the observed number of nonsynonymous changes at site i.

The next level of the hierarchy considers a mixture prior on $\beta_{i,j}$ with point masses at 0 and 1, and a normal density. The point mass at 0 is used to describe sites which are strongly conserved, i.e., those whose observed mean distances are zero or very close to zero. The point mass at 1 is used to model sites that are believed to be neutral with respect to a particular property. Finally, those $\beta_{i,j}$s that are not 0 or 1 can be less than 1 or greater than 1, and so a normal density is chosen to model such sites a priori. Thus, the structure on $\beta_{i,j}$ is

$$\beta_{i,j} \sim \pi_{i,0} 1_{\{\beta_{i,j}=0\}} + \pi_{i,1} 1_{\{\beta_{i,j}=1\}} + \pi_{i,2} \mathsf{N}(\alpha_\mathrm{j}, \tau_\mathrm{i}^2),$$

with $\pi_{i,2} = 1 - \pi_{i,0} - \pi_{i,1}$. The mixture probabilities and the variances on the Gaussian components are assumed to be the same across properties but are site-specific, while the means are assumed to depend exclusively on the properties. This structure is used for two reasons. First, at least a priori and even under the assumption of neutrality, the variability of the changes for a given amino acid property may be viewed as site-dependent due to, for example, functional constraints, while the mean of the Gaussian distribution is thought as the average coefficient for a given property. Second, using the same α_j across sites for a given property and the same mixture probabilities and variance, $\pi_{i,0}, \pi_{i,1}, \pi_{i,2}$ and τ_i^2, across properties for each site allows us to borrow strength across sites and

properties for estimation purposes, improving our ability to detect changes. In other words, data from all the sites are used to obtain estimates of α_j and data from all the properties are used to estimate the site-specific parameters. This allows us to consider several properties within a single model instead of conducting separate analyses for each one.

The final level of the model corresponds to the priors on the mixture probabilities, the scale parameters, and the means of the Gaussian priors. In order to simplify computations these priors are chosen to be conditionally conjugate, i.e., $(\pi_{i,0}, \pi_{i,1}, \pi_{i,2}) \sim \mathsf{Dirichlet}(a_0, a_1, a_2)$, $\alpha_j \sim \mathsf{N}(m_\alpha, C_\alpha)$, $\tau_i^2 \sim \mathsf{IG}(\alpha_\tau, \beta_\tau)$, $\kappa^2 \sim \mathsf{IG}(\alpha_\kappa, \beta_\kappa)$ and $\sigma^2 \sim \mathsf{IG}(\alpha_\sigma, \beta_\sigma)$, with hyperparameters $\mathsf{a}_0, \mathsf{a}_1, \mathsf{a}_2, \mathsf{m}_\alpha, \mathsf{C}_\alpha, \alpha_\tau, \beta_\tau, \alpha_\kappa, \beta_\kappa, \alpha_\sigma$, and β_σ set as follows. We assume a priori that most of the sites are neutral ($\beta_{i,j} = 1$), so we assign a fairly large weight to this category, and equal weights to the probabilities of $\beta_{i,j}$ being zero and being different from zero and one. Thus, we use a $\mathsf{Dirichlet}$(1, 4, 1) prior for the πs which assigns about 67% weight to the probability of a site being neutral and about 16.5% weight to each of the other two categories. We also assume all the properties to be neutral a priori, so all the α_js were given the same prior mean m_α = 1 and the same prior variance C_α = 0.25. κ^2 was believed to be smaller a priori than σ^2, so we assign prior means of 0.01 and 0.1 and use IG(2,100) and IG(2,10), respectively, as the priors for these parameters. The τ_i^2s are given IG(2,100) priors with prior means of 0.01.

An additional advantage of our hierarchical specification is that it allows us to automatically control the error rate associated with multiple comparisons. Two conditions need to be satisfied before a Bayesian approach can claim that posterior inference adjusts for multiplicities (Scott and Berger, 2003). In the context of our model, these are the following: (i) The model needs to assign a positive prior probability of site i showing $\beta_{i,j}$ = 0 or $\beta_{i,j}$ = 1 and (ii) These prior probabilities cannot be fixed, rather they have to come from some distribution. However, some care is necessary when eliciting hierarchical priors. The results from our sensitivity analyses in Section 3.2 suggest that the prior that affects the results most is the prior on πs. Increasing the prior probability that a site is not neutral (i.e., increasing the prior probabilities for $\pi_{i,0}$ and $\pi_{i,2}$) increases the posterior probability of the site being not neutral. We suggest the use of a prior that assigns most of the probability mass to each site being neutral, such as the $\mathsf{Dirichlet}$(1, 4, 1), unless there is strong prior evidence suggesting otherwise. For the variance terms (σ^2, κ^2 and τ_i^2), we suggest using priors with small values for the mean, since the variance in the data is usually small. Increasing the prior mean for the variance terms means we add more uncertainty in our model, which might not be feasible for all scenarios.

Posterior estimation of the model parameters is achieved via Gibbs sampling (e.g., Gamerman and Lopes, 2006). Chains were run for 20,000 burn-in iterations, followed by 200,000 sampling iterations. The sampling iterations were thinned by a factor of 20 to reduce correlation, resulting in effective sample sizes of 10,000 posterior draws. Convergence was assessed by monitoring parameter trace plots. No convergence problems were detected in any of the runs. For details about the MCMC method please refer to the Appendix.

2.2 Data and other methods

2.2.1 Data

We apply the proposed models to the following datasets: an alignment of six class I alleles of the human major histocompatibility complex (MHC) and an alignment of the sperm lysin protein for 25 abalone species. In addition, a simulation study and the analysis of lysozyme gene sequences in primates are included in the Appendix.

MHC. These data comprise six class I MHC alleles from HLA-A and HLA-B loci with 365 codons. Sites with gaps were removed, finally resulting in 362 codons. These sequences were previously shown to be under positive selection by Swanson *et al.* (2001).

Lysin. These data consist of cDNA from 25 abalone species with 135 codons. Several sites have been labeled as positively selected using codon substitution models that allow ω to vary among amino acid sites (Yang *et al.*, 2000b). Sites with gaps were removed which resulted in 122 codons for the analysis presented here.

We compare the results obtained via the Bayesian hierarchical models to those obtained using `TreeSAAP` and the methods of Sainudiin *et al.* (2005) and Wong *et al.* (2006). We consider five properties: Hydropathy, Isoelectric Point, two measures of Polarity (Grantham and Zimmerman) and Volume. These were obtained from the amino acid index database at http://www.genome.jp/aaindex. Many more properties can be included, however, our aim is to illustrate the methodology using a small number of properties.

In order to account for uncertainty in the tree and in the ancestral sequences we analyze data derived from 25 neighbor-joining (NJ) trees for each of the two alignments.The 25 NJ phylogenies — obtained from the R package `markovjumps` — are based on the methods of O'Brien *et al.* (2009) that calculate labeled distances (e.g., synonymous and nonsynonymous distances) between sequences by "robust counting", i.e., by building on a reversible continuous-time Markov chain model of substitution. Robustness is achieved by conditioning on pairwise site patterns to obtain the conditional mean numbers of labeled substitutions and then by averaging the conditional expectations over the empirical distribution of site patterns from the observed sequences. For datasets with more sequences, one might need to look at a larger number of trees. However, in our case, MHC and lysin have 6 and 25 sequences, respectively. Phylogenies obtained via the Bayesian methods implemented in `MrBayes` (Ronquist and Huelsenbeck, 2003) were also considered. The Bayesian approach to phylogeny estimation leads to samples from the posterior distribution of trees and so, we considered the 5 phylogenies with the largest posterior probabilities and the consensus tree reported by `MrBayes`. MHC had only 6 different topologies out of the 25 NJ trees. Bayesian phylogenetic inference in these data provided a single topology with 0.99 posterior probability. Furthermore, this topology is precisely one the 6 different NJ topologies. For lysin, all the 25 NJ topologies were different but a large number of them were very similar. The 5 trees with the highest posterior probabilities and the consensus tree obtained from `MrBayes` were also similar to some of the NJ trees. Given that the `MrBayes` phylogenies are identical or very similar to the NJ phylogenies for the two data sets we only report analyses based on the latter. The hierarchical models are not tied to a particular method to estimate the phylogeny and the ancestral sequences, and so the user can simply apply the models to data derived from any prespecified set of phylogenies and ancestral sequences.

The NJ trees obtained from `markovjumps` (based on nonsynonymous distances) were used to generate 25 sets of ancestral sequences under model M8 in `codeml` implemented in PAML, Version 3.15 (Yang, 1997). The 25 sets of ancestral sequences were then taken as inputs for the hierarchical regressions, i.e., observed distances were computed for each of the 25 sets of ancestral sequences and separate regressions were fit. For each of the two alignments we also performed regression analyses that combined the observed distances by taking a weighted average of the $y_{i,j}$s from the 25 sets of ancestral sequences. Finally, hierarchical analyses of the ancestral sequences obtained using the maximum likelihood (ML) phylogeny for the MHC data and the phylogeny of Lee *et al.* (1995) (used in Yang *et al.*, 2000b) for lysin were also performed.

For both datasets, and for each set of ancestral sequences, positively and negatively selected sites were identified using the codon substitution model M8 in PAML, in which a discretized beta distribution models ω values between zero and one with probability p_0, while an additional positive selection category with $\omega > 1$ and probability p_1 models positive selection. Sites were identified as positively selected if the Bayes-empirical-Bayes posterior probability (Yang *et al.*, 2005) of belonging to this last category was greater than 0.95 and as negatively selected if $\Pr(\omega > 1|\text{ data}) < 0.5$ and $\hat{\omega} < 0.3$, where $\hat{\omega}$ is the estimated posterior mean.

2.2.2 *Other methods*

In addition to looking at results obtained via the Bayesian hierarchical models for sites classified as positively selected using ω-based methods, we also compare such results to two approaches that take physicochemical properties into account. Specifically, we applied the methods of McClellan *et al.* (2005) implemented in `TreeSAAP` version 3.2 (Woolley *et al.*, 2003) and the methods of Sainudiin *et al.* (2005) and Wong *et al.* (2006) implemented in `codemlz` in PAML, 3.14z and `EvoRadical`, respectively.

`TreeSAAP` implements the modified MM01 model of McClellan and McCracken (2001), which is outlined in McClellan *et al.* (2005). In model MM01 each nonsynonymous substitution is assigned to one of M categories, with categories indexed by lower numbers corresponding to sites with more conservative changes for a given property, and those indexed by higher numbers corresponding to sites displaying radical changes. In the analyses presented here we considered $M = 4$ categories indicating, respectively, conservative changes ($m = 1$), moderate changes ($m = 2$), radical changes ($m = 3$) and very radical changes ($m = 4$). Nonsynonymous changes were inferred from the ancestral reconstruction obtained via the reversible nucleotide substitution model in `baseml` (no codon substitution models are implemented in the current version of `TreeSAAP`). The ML phylogeny and the phylogeny of Lee *et al.* (1995) were used for the MHC and lysin data, respectively. Each of the nonsynonymous changes was assigned to one of the magnitude categories for each of the five properties being considered. In order to test the hypothesis of neutrality, McClellan *et al.* (2005) divide the number of inferred amino acid replacements per magnitude category for a given property by the number of evolutionary pathways assigned to that category to obtain a set of proportions, p_m for $m = 1 : M$. Under neutrality, it is expected that these proportions are equal to the overall mean. McClellan *et al.* (2005) test neutrality for each property using z-scores. Site-specific z-scores are also provided, however, they do not adjust for multiple comparisons.

Table 1. Posterior estimates of the proportion of sites with $\beta = 0$, $\beta = 1$, $0 < \beta < 1$, and $\beta > 1$ for MHC data (combined distances).

Property	Proportions			
	$\beta = 0$	$\beta = 1$	$0 < \beta < 1$	$\beta > 1$
Hydropathy (H)	0.183	0.652	0.001	0.164
Isoelectric Point (IP)	0.259	0.570	0.014	0.154
Polarity-G (PG)	0.161	0.658	0.001	0.180
Polarity-Z (PZ)	0.398	0.401	0.000	0.201
Volume (V)	0.095	0.717	0.064	0.124

More recently, Sainudiin *et al.* (2005) and Wong *et al.* (2006) developed codon substitution models that incorporate changes in amino acid properties. In these approaches amino acids are first partitioned on the basis of a particular property. This partition is then used to parameterize the rates of property-conserving and property-altering codon substitutions using a maximum likelihood framework. Sainudiin *et al.* (2005) divide substitutions into two groups: synonymous and property-conserving nonsynonymous substitutions, and property-altering nonsynonymous substitutions. Wong *et al.* (2006) generalize the idea to allow three classes: synonymous, property-conserving nonsynonymous, and property-altering nonsynonymous substitutions. The latter model can determine the type of selective pressure acting on a particular property of interest, while accounting for the non-specific selective pressure at the amino acid level. The results reported in both these papers are based on posterior probabilities computed via the naïve empirical Bayes method (Nielsen and Yang, 1998).

3 RESULTS

3.1 MHC

Table 1 displays the posterior means of the proportions of sites where each of the five properties are strongly conserved (β = 0), conserved ($0 < \beta < 1$), neutral (β = 1) or radically changed ($\beta > 1$). These results were obtained by fitting the hierarchical model to the combined distances from the 25 sets of ancestral sequences. Very similar results are obtained from fitting 25 separate hierarchical models, one per set of ancestral sequences, and also from fitting a hierarchical model to the ancestral sequences obtained from the ML tree. For all the properties except Polarity-Z — which shows essentially equal proportions in the first two categories — the majority of the sites show substitutions that are neutral. In addition, non-neglible proportions of sites either do not alter the property (β = 0) or radically change it.

Table 2 lists the sites in MHC in domains I, II and III for which $\Pr(\beta_{i,j} \neq 0 \text{ and } \beta_{i,j} \neq 1|\text{data})$ is maximized in at least one of the 25 regression analyses. We choose to label sites according to the categories for which the posterior probabilities are maximized because it is the standard rule in classification problems and it leads to good sensitivity and specificity rates (see the simulation study and further discussion in the Appendix). Sites in bold also maximized such probability in the combined regression. Furthermore, all of the sites listed are such that $\Pr(\beta_{i,j} > 1|\text{ data})$ was maximum. No sites

Table 2. Sites for which $\Pr(\beta_{i,j} \neq 0 \text{ and } \beta_{i,j} \neq 1|\text{data})$ is maximized in at least one of the 25 regressions for the MHC data. The numbers in parentheses indicate for how many of the 25 data sets this was the case. Sites in bold also had maximum probability of being in this category in the combined data. Underlined sites were identified as property altering by Sainudiin *et al.* (2005).

Property	Sites
H	**45**(25), **76**(25), **116**(12), **152**(25), **156**(25)
IP	76(6), **163**(12)
PG	**45**(25), **76**(25), **90**(25), **116**(25), **152**(25)
PZ	**30**(25), **45**(25), **63**(25), **76**(25), **77**(25), **90**(25), **113**(25), **127**(25), **152**(25), **163**(25), **171**(25), **253**(25)

were found to maximize the probability of belonging to the third category for Volume. Comparing Table 2 to the results in Sainudiin *et al.* (2005) and Wong *et al.* (2006), we see that these authors found the following sites to have significant probability of showing radical changes: site 116 under their Polarity partition, and sites 63, 67 and 97 under their Volume partition. These were the sites identified by the naïve empirical Bayes method. However, on using the Bayes-empirical-Bayes method only site 45 had a significant probability of showing radical changes under the Polarity partition, and no sites were significant under the Volume partition. In addition, sites 114 and 156 were found to be under positive selection using ω-based methods. Using the hierarchical approach we find that sites 45, 63, 116 and 156 radically change Hydropathy, Polarity-G or Polarity-Z. Therefore, even though there are similarities between our results and those in Sainudiin *et al.* (2005) and Wong *et al.* (2006), there are also some discrepancies. However, it should be emphasized that the methods of Sainudiin *et al.* (2005) and Wong *et al.* (2006) strongly depend on the partition chosen by the user and so, they are not directly comparable to the hierarchical approaches presented here since the latter use amino acid distances directly.

Analyzing the MHC data with `TreeSAAP`, which uses ancestral sequences obtained from `baseml` and the ML phylogeny, we found that the hypothesis of neutrality could not be rejected (i.e., none of the z-scores were significant) for any of the five properties. In order to compare our method to `TreeSAAP` we performed an additional hierarchical regression analysis using distance data from these ancestral sequences. Our results are very similar to those obtained with the codon-based models and shown in Table 1, i.e., it was found that although a majority of the sites show neutral changes, each property has a considerable proportion of sites in the other categories ($\beta = 0$, $0 < \beta < 1$ and $\beta > 1$) taken together. The fact that `TreeSAAP` does not find any property to be conserved or radically changing is probably related to the fact that the z-scores are based on all the sites. In contrast, one of the advantages of the hierarchical approach is that it is site-specific and so, even if the changes are neutral or conserve a given property on average, some of the sites may show mutations that radically change such property.

3.2 Lysin

Table 3 shows posterior estimates of the proportions of sites with $\beta = 0$, $\beta = 1$, $0 < \beta < 1$ and $\beta > 1$, respectively, based on the combined distances from the 25 sets of ancestral sequences. About 64%-70% of the sites lie in the neutral category ($\beta = 1$) for Hydropathy, Isoelectric Point, Polarity-G and Volume, while Polarity-Z has 53% of sites in this category. Hydropathy, Isolectric Point, Polarity-G and Volume are also fairly comparable in terms of the percentages of sites (15%-20%) that show nonsynonymous mutations that strongly conserve these properties ($\beta = 0$), while approximately 29% of the sites are showing nonsynonymous mutations that strongly conserve Polarity-Z.

Table 3. Posterior estimates of the proportion of sites with $\beta = 0$, $\beta = 1$, $0 < \beta < 1$ and $\beta > 1$ for the lysin data (combined data).

Property	Proportions			
	$\beta = 0$	$\beta = 1$	$0 < \beta < 1$	$\beta > 1$
Hydropathy (H)	0.164	0.702	0.000	0.134
Isoelectric Point (IP)	0.205	0.637	0.061	0.097
Polarity-G (PG)	0.166	0.672	0.069	0.093
Polarity-Z (PZ)	0.289	0.529	0.001	0.181
Volume(V)	0.147	0.676	0.163	0.014

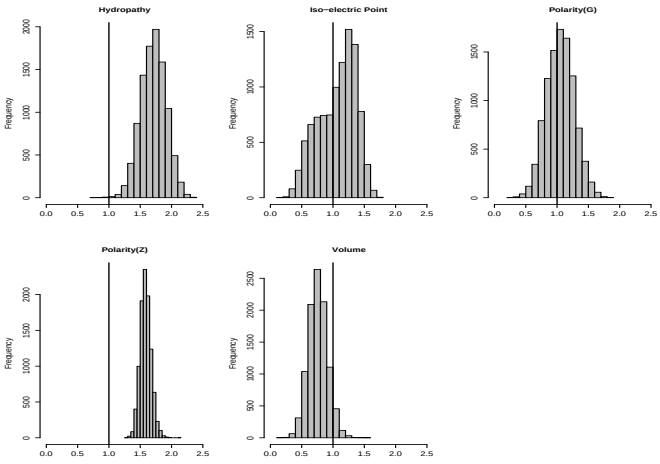

Fig. 1. Posterior summaries of α_js from the combined analysis for lysin.

The posterior densities of α_j for the five properties are shown in Figure 1. Recall that α_j is mean of the Gaussian distributions used to model the $\beta_{i,j}$s that are different from zero and one and so, α_j is a measure of the average behavior for the property based on sites whose changes conserve it or radically change it ($0 < \beta < 1$ or $\beta > 1$). The figure shows that the distributions of α_j for Isoelectric Point and Polarity-G are roughly centered at one, and so on average changes are neutral with respect to these properties. The distributions for Hydropathy and Polarity-Z have most of the mass above 1, consistent with radical changes. Finally, the distribution for Volume is skewed to the right, with more mass below 1, consistent with changes that conserve the property. Results from the analysis done using the ancestral sequences obtained from model M8 in `codeml` and the tree of Lee *et al.* (1995) are essentially the same as those summarized in Table 3 and Figure 1. Results obtained from the 25 separate regressions are also very similar for Volume in that the proportions of sites in the different categories are comparable to those shown in Table 3 and Figure 1. This suggests that tree uncertainty does not affect the overall

Table 4. Sites that maximize $\Pr(\beta_{i,j} \neq 0, 1|\text{data})$ in at least one of the 25 regressions and the combined regression for the lysin data, with numbers in parentheses indicating for how many of the 25 regressions this was the case. Sites in bold were also identified as under positive selection by ω-based methods.

Property	Sites
H	15(10), 16(7), 21(11), **70**(18), 82(9), 99(16), **127**(17)
PZ	15(11), 16(11), 21(11), 57(11), **70**(9), 75(7), **87**(5), 91(6), 97(4), 99(10), 106(11), 119(11), **127**(15)

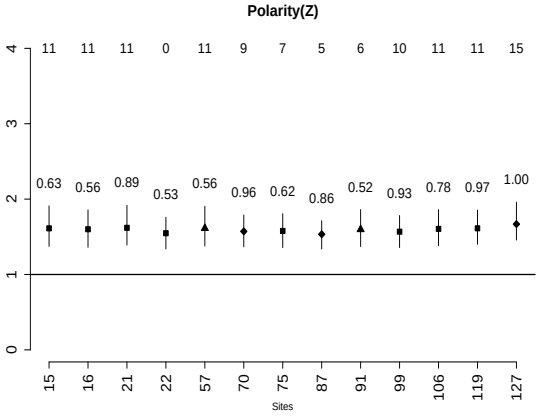

Fig. 2. Posterior summaries of all the sites in the lysin data that maximize $\Pr(\beta_{i,j} \neq 0$ and $\beta_{i,j} \neq 1|\text{data})$ in the combined data for Polarity-Z. The intervals are 95% posterior intervals of the $\beta_{i,j}$s that are different from zero and one; the medians are also highlighted. The numbers right at the top of the intervals are the estimated posterior probabilities that the $\beta_{i,j}$s are different from zero and one for the combined analysis. The numbers at the very top ($y = 4$) correspond to the number of times the site maximized $\Pr(\beta_{i,j} \neq 0$ and $\beta_{i,j} \neq 1|$ data) in the 25 regressions. Diamonds, squares and triangles label, respectively, positively selected sites, neutral sites, and negatively selected sites under the ω-based measure.

trends in lysin for Volume. For the remaining four properties, the proportions of sites in the strongly conserved and neutral categories remain fairly similar to the combined analysis for all 25 regressions. However, the results for the categories describing conserved or radical changes seem to be affected by phylogenetic and ancestral sequence uncertainty. In particular, for Hydropathy 16 of the 25 regressions have results similar to Table 3 but in the remaining 9 regressions, it had 9%-11% of sites in the conserved category. In the cases of Isoelectric Point and the two measures of Polarity, 9 and 11 of the 25 regressions respectively had results similar to the combined analysis. For the remaining regressions, about 15%-16% sites showed changes conserving Isoelectric Point, 14%-17% sites showed changes conserving Polarity-G and 15%-18% sites showed changes conserving Polarity-Z.

In terms of site-specific results, Table 4 lists the sites that maximize the probability of $\beta_{i,j}$ being different from 0 or 1, in at least one out of the 25 regressions and in the combined regression.

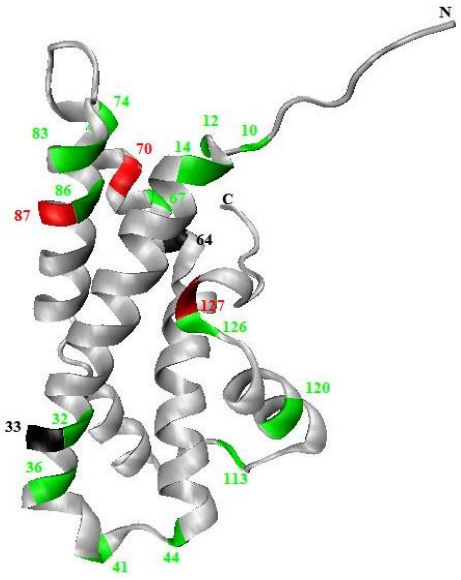

Fig. 3. Lysin crystal structure from the red abalone *H.rufescens* (Shaw *et al.*, 1993). The structure starts from amino acid 4. Positively selected sites using model M8 in PAML and the phylogeny of Lee *et al.* (1995) are highlighted. Sites 70, 87, and 127 display mutations that radically change Polarity-Z ($\beta > 1$) and sites 33 and 64 strongly conserve the the property ($\beta = 0$). The remaining green sites are neutral with respect to Polarity-Z ($\beta = 1$).

Isoelectric point, Polarity-G and Volume did not show any sites that maximize such probability in the combined data. Sites identified as positively selected using ω-based methods appear in bold. In this case all the sites listed also maximize the probability of $\beta > 1$ and so, these sites show substitutions that radically change Hydropathy and Polarity-Z. It is also possible to obtain informative graphical displays that summarize site-specific behavior for a particular property. For instance, Figure 2 plots the information obtained from the posterior samples of $\beta_{i,j}$ for the combined analysis under Polarity-Z for those sites which had significant posterior probability of being different from 0 or 1. The medians and the central 95% credible intervals (vertical lines) from the posterior samples of $\beta_{i,j}$ which are different from 0 and 1 are shown. All of the sites show radical changes (posterior intervals are above 1) using the hierarchical approach, but only sites 70, 87, and 127 — which are clustered at the top of the molecule in Figure 3 — are identified as positively selected by model M8 in PAML. Also, some of the sites identified as positively selected using ω-based methods are neutral or strongly conserve Polarity-Z. Specifically, Figure 3 highlights the sites found to be under positive selection according to model M8 in PAML using the phylogeny of Lee *et al.* (1995). Among these sites, sites 70, 87 and 127, show mutations that radically change Polarity-Z. Sites 33 and 64 were found to show nonsynonymous mutations that strongly conserve Polarity-Z, and the remaining sites were identified as neutral with respect to this property.

Wong *et al.* (2006) also analyzed the lysin data using different partitions for Polarity and Volume and identified sites that were under specific selective pressures for the different properties. Once again we emphasize that the results presented here and those in Wong *et al.* (2006) are not directly comparable because the

Table 5. `TreeSAAP` output for lysin. z-scores are compared to the cut-off values from the right tail of a $N(0,1)$. The presence of * indicates that the property is significant at the corresponding α value of testing.

Property	Category	Value	.05	.01	.001
Hydropathy (H)	4	-2.892	*	*	
Isoelectric Point (IP)	3	-2.399	*	*	
	2	1.783	*		
Polarity-G (PG)	3	-2.645	*	*	
	4	-3.888	*	*	*
	1	3.761	*	*	*
Volume (V)	2	-5.336	*	*	*
	4	-2.464	*	*	

latter approach uses partitions and the hierarchical models use distances, and so, it is possible that a large difference in the actual value of the property is ignored or a small difference magnified while constructing the categories. For example, under the Volume partition, Wong *et al.* (2006) classify both Glycine and Valine as small and Glutamine as large, whereas for us their Volume scores were 3, 84 and 85 respectively. Thus, Wong *et al.* (2006) would say that a change between Glutamine and Valine alters the property while a mutation between Glycine and Valine does not, even though the difference between Glutamine and Valine Volume scores is much smaller than the difference between the Glycine and Valine scores.

The lysin sequences were also analyzed using `TreeSAAP`. Table 5 shows a `TreeSAAP` output that lists the significant amino acid properties based on z-scores. From this table we can conclude that Hydropathy and Isoelectric Point show significantly smaller proportions of very radical and radical changes, respectively. Polarity-G shows a significantly higher proportion of moderate changes, and lower proportions of radical and very radical changes. Volume shows a significantly higher proportion of changes that conserve the property, and lower proportions of moderate and very radical changes. Therefore, Volume seems to be conserved. The absence of a property (or category) implies that no significant departures from neutrality are found, and so there is no significant deviation from the assumption of selective neutrality for Polarity-Z. To compare our results with those of `TreeSAAP`, we performed a hierarchical analysis using data derived from the same ancestral sequences used by `TreeSAAP` (obtained using `baseml`). Posterior estimates of the proportions of sites with $\beta = 0$, $\beta = 1$, $0 < \beta < 1$ and $\beta > 1$ for the five properties (not shown) are similar to those in Table 3. Specifically, Volume shows the largest percentage of sites (about 30%) lying in the strongly conserved and conserved categories, and the smallest percentage of sites (about 1%) in the radically changing category. These average results agree with those obtained from `TreeSAAP`.

4 DISCUSSION

We present a Bayesian hierarchical regression model that allows us to detect radical amino acid changes that could facilitate the identification of adaptations by quantifying the magnitude of changes in amino acid properties. The model is flexible, with the following main features: (i) It provides global results, i.e., point estimates and associated uncertainties of the posterior probabilities of the proportions of sites that are neutral, highly conserved, conserved or radically changing with respect to a particular property can be obtained; (ii) It provides site-specific results, i.e., the approach can identify sites showing mutations that radically change or strongly conserve a particular property, even in cases where the mutations are neutral on average. This is an important improvement over models that detect selection only in cases where it leads to an excess in the total number of radical changes, since essentially these methods average over all sites; (iii) The hierarchical specifications of the priors ensure that the models and methods presented here properly account for multiple comparisons when a large number of amino acid sites are analyzed. (iv) The prior specification allows borrowing of information across sites and properties for estimation purposes which leads to higher power with respect to approaches that analyze sites or properties individually (Zhang and Cao, 2009). (v) The hierarchical structure also preserves the parsimony of the model; indeed, hierarchical priors induce correlations across model parameters that reduce the number of effective parameters in the model and prevent overfitting without hindering model flexibility (for further discussion and two examples see Gelman *et al.*, 1996 and Zhu and Hero, 2007).

The methods in Sainudiin *et al.* (2005) and Wong *et al.* (2006) are also site-specific but, unlike those proposed here, they rely on prespecified partitions of the amino acids according to some physicochemical property and so, the results are highly dependent on the partition being used. Our hierarchical models are more similar to methods based on amino acid distances, such as those of McClellan *et al.* (2005). However, as mentioned above, the hierarchical models properly handle the multiple comparison problem when several sites are considered.

The proposed hierarchical approach assumes that the phylogeny underlying the sequences is known, and that a specific model of sequence evolution is used to generate the ancestral sequences. This is also true for the methods we compared to, such as those implemented in `TreeSAAP`, or those in Pupko *et al.* (2003) and Sainudiin *et al.* (2005). When there is uncertainty regarding the phylogeny and/or the evolutionary model used to generate the ancestral sequences, such uncertainty needs to be taken into account. One way of doing so is by performing hierarchical analyses with data obtained from several of the most likely phylogenies and/or evolutionary models and seeing if the results are sensitive with respect to these parameters. This was the route we took here by considering hierarchical analyses of data obtained under different phylogenies. A more robust way of dealing with this would be extending the regression models to consider phylogenetic uncertainty. This will be explored in the future, however, it is also worth mentioning that such approaches would be computationally intensive (Huelsenbeck *et al.*, 2000).

The regression methods presented here complement traditional ω-based methods for detecting molecular adaptation by providing a qualitative assessment of the amino acid changing mutations with respect to a specific set of physicochemical properties. When both methods are combined, it is possible to determine which sites among those labeled as positively selected using ω-based methods are showing radical changes with respect to a given property. Based on the analyses presented here we see that in some cases, sites that have a large number of nonsynonymous substitutions, and are therefore identified as positively selected, may conserve a given

property. Conversely, sites showing a relatively small number of substitutions, possibly classified as neutral or negatively selected, may show substitutions that radically change a given property. If such changes are advantageous to the protein function, then these sites could be targets for natural selection (Sainudiin *et al.*, 2005). Thus, the regression methods can be viewed as complementary to any codon-based methods, providing additional information about putative molecular adaptations that could be further explored.

The analyses presented here use a small number of properties that do not appear to be highly correlated in terms of the observed (and expected) distances for the alignments considered. However, the amino acid index database has a very large number of properties, many of which are highly correlated. The hierarchical regression framework can be extended to handle a large number of possibly correlated amino acid properties by incorporating a factor structure. Factor models explain the variability in the data in terms of a smaller number of unobserved variables called factors. The idea is similar to that of principal component analysis. We will consider such extensions in the near future. For the time being, we suggest using the hierarchical regression model with a small number of relevant properties. The properties can be chosen based on preliminary principal component analysis or other relevant biological information.

ACKNOWLEDGEMENTS

Funding: SD, RP and AE were supported by the NIH/NIGMS grant R01GM072003-02. AR was supported by the NIH/NIGMS grant R01GM090201-01.

REFERENCES

Anisimova, M. and Kosiol, C. (2009). Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.*, **26**, 255–271.

Dagan, T., Talmor, Y., and Graur, D. (2002). Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Mol. Biol. Evol.*, **19**, 1022–1025.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, 2nd edition.

Gelman, A., Bois, F., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Amer. Stat. Assoc.*, **91**, 1400–1412.

Hanada, K., Shiu, S.-H., and Li, W.-H. (2007). The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Mol. Biol. Evol.*, **24**, 2235–2241.

Huelsenbeck, J., Rannala, B., and Masly, J. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, **288**, 2349–2350.

Hughes, A., Ota, T., and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class i major-histocompatibility-complex molecules. *Mol. Biol. Evol.*, **7**, 515–524.

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. Evol.*, **12**, 231–238.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96 – 99.

McClellan, D. A. and McCracken, K. G. (2001). Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. *Mol. Biol. Evol.*, **18**, 917 – 925.

McClellan, D. A., Palfreyman, E. J., Smith, M. J., Moss, J. L., Christensen, R. G., and Sailsbery, J. K. (2005). Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *Mol. Biol. Evol.*, **22**, 437–455.

McDonald, J. and Kreitman, M. (1991). Adaptive protein evolution at the *adh* locus in *Drosophilia*. *Nature*, **351**, 652 – 654.

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. New York: Oxford University Press.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV–1 envelope gene. *Genetics*, **148**, 929 – 936.

O'Brien, J., Minin, V., and Suchard, M. A. (2009). Learning to count: Robust estimates for labeled distances between molecular sequences. *Mol. Biol. Evol.*, **26**, 801–814.

Pamilo, P. and Bianchi, N. O. (1993). Evolution of the *Zfx* and *Zfy* genes : rates and interdependence between the genes. *Mol. Biol. Evol.*, **18**, 917 – 925.

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc. Natl. Acad. Sci. USA*, **104**, 13390–13395.

Pupko, T., Sharanb, R., Hasegawac, M., Shamird, R., and Graur, D. (2003). Detecting excess radical replacements in phylogenetic trees. *Gene*, **319**, 127–135.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572 – 1574.

Sainudiin, R., Wong, W. S. W., Yogeeswaran, K., Nasrallah, J. B., Yang, Z., and Nielsen, R. (2005). 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. *J. Mol. Evol.*, **60**, 315–326.

Sawyer, S. A. and Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics*, **1132**, 1161 – 1176.

Scott, J. G. and Berger, J. O. (2003). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, **136**, 2144–2162.

Shaw, A., McRee, D. E., Vacquier, V. D., and Stout, C. D. (1993). The crystal structure of lysin, a fertilization protein. *Science*, **262**, 1864–1867.

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

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

Swanson, W. J., Yang, Z., Wolfner, M. F., and Aquadro, C. F. (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. USA*, **98**, 2509–2514.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Wong, W. S. W., Sainudiin, R., and Nielsen, R. (2006). Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics*, **7**, 148.

Woolley, S., Johnson, J., Smith, M. J., Crandall, K. A., and McClellan, D. A. (2003). TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. *Bioinformatics*, **19**, 671 – 672.

Xia, X. and Li, W. H. (1998). What amino acid properties affect protein evolution? *J. Mol. Evol.*, **47**, 557 – 564.

Yang, Z. (1997). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, **42**, 294 – 307.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908 – 917.

Yang, Z. and Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.*, **19**, 49 – 57.

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

Yang, Z., Swanson, W. J., and Vacquier, V. D. (2000b). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineage and sites. *Mol. Biol. 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. Evol.*, **22**, 1107–1118.

Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.*, **50**, 56–68.

Zhang, S. and Cao, J. (2009). A close examination of double filtering with fold change and t test in microarray analysis. *BMC Bioinformatics*, **10**, 402.

Zhu, D. and Hero, A. O. (2007). Bayesian hierarchical model for large-scale covariance matrix estimation. *J. Comput. Biol.*, **14**, 1311–1326.