Bayesian Methods for Phylogeny Independent Detection of Positively Selected Amino Acid Sites

Daniel Merl (dmerl@ams.ucsc.edu), Raquel Prado, UCSC
and Ananías Escalante, CDC

Key Words: Bayesian inference, positive selection, sequence evolution.

Abstract:

A positively selected amino acid site is one for which natural selection encourages diversification. The identification of such sites is of biomedical importance, as diversifying sites cannot act as reliable binding sites for location-specific drugs. We introduce a new method for detecting positive selection based on a class of Bayesian generalized linear models (GLMs). This method does not require explicit assumptions about phylogeny and offers relatively reduced time to Markov chain Monte Carlo (MCMC) convergence. We compare our Bayesian GLM approach with three current methods for detecting positive selection: Nei and Gojobori's `ADAPTSITE`, Yang's `PAML`, and Huelsenbeck and Ronquist's `MrBayes`.

1. Introduction

Molecular sequences are said to experience purifying, neutral, or positive selection depending on the ratio of nonsynonymous (amino acid changing) nucleotide substitutions to synonymous (amino acid preserving) nucleotide substitutions. A positively selected amino acid site is one for which natural selection encourages the fixation of nonsynonymous substitutions. The identification of such sites is of biomedical importance; for instance, an antigen with many positively selected sites would be an unsuitable candidate for vaccine development. Our research has been motivated by the need to recognize this property in several recently sequenced antigenic regions of human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*.

Detection of positively selected sites is usually accomplished by inferring the rate of nonsynonymous substitutions per synonymous substitution (ω) for each codon–the nucleotide triplets that codify a particular amino acid–in a given protein coding DNA sequence alignment, with $\omega > 1$ being the criterion for positive selection [5, 6] . Substitution rates are traditionally inferred using codon-based stochastic models of sequence evolution, however, the computational complexity associated with fitting a stochastic model to real data usually precludes the simultaneous estimation of substitution rates and sequence phylogeny, even though the terms are highly dependent on one another. Computationally efficient methods for detecting positive selection, such as Yang's maximum likelihood based `PAML` [5] and Nei and Gojobori's `ADAPTSITE` [3, 4], assume fixed phylogenies, and in the latter case, fixed ancestral sequences, to ease the computational burden of estimating substitution rates. This assumption may be reasonable for certain distantly related sequences with a clear phylogenic structure, but is generally inappropriate for closely related sequences for which any existing phylogenetic information should be regarded as probabilistic at best. Hierarchical Bayesian treatment of the stochastic models of sequence evolution exist in Huelsenbeck and Ronquist's `MrBayes` [2]. In Bayesian tradition, all model parameters, including phylogeny, are given prior distributions, and MCMC is used to sample from the joint posterior distribution. By integrating over tree structures, `MrBayes` is fairly robust with respect to phylogenetic assumptions. However, MCMC convergence becomes a problem for the codon-substitution models implemented in `MrBayes` when dealing with large numbers of closely related sequences. In preliminary analyses of the plasmodium sequences, we found that `MrBayes'` MCMC convergence was both slow and difficult to assess.

We propose a new method based on a class of hierarchical Bayesian GLMs capable of inferring substitution ratios and allowing the identification of sites under positive selection without requiring explicit assumptions about phylogeny. This latter feature is key to the analysis of many (> 50) very closely related sequences for which no particular phylogenic structure is available, as is the case with the plasmodium alignments.

In Section 2 we describe the new Bayesian models for detection of positively selected sites. Section 3 presents a comparison of the results obtained with `ADAPTSITE`, `MrBayes`, `PAML` and the new methodology, in the study of two simulated data sets. Finally, Section 4 presents the conclusions and future work.

2. Model Description

Our model assumes that underlying nucleotide substitutions occur at fixed probabilities for a given amino acid site within a given pair of sequences. In the general case, we consider 5 types of substitution: synonymous or nonsynonymous, transition (purine-purine and pyrimidine-pyrimidine substitutions) or transversion (all others), and no substitution. The data consist of a gapless alignment of N protein coding DNA sequences of I codons ($3 \times I$ nucleotides). We first transform the sequences into vectors of the form $(z^i_{1,j}, z^i_{2,j}, z^i_{3,j}, z^i_{4,j}, z^i_{5,j})$, where each component respectively indicates the number of synonymous transitions, synonymous transversions, nonsynonymous transitions, and nonsynonymous transversions, or no substitution, occurring at each codon i ($i = 1, \ldots, I$), for each pair of sequences j ($j = 1, \ldots, N(N-1)/2$). If codons y^i_j (sites i in comparison j) differ at 0 or 1 nucleotide positions, $\mathbf{z}^i_j$ is multinomially distributed as $\mathbf{z}^i_j \sim Multinomial(1, \boldsymbol{\theta}^i_j)$, where $\boldsymbol{\theta}^i_j = (\theta^i_{1,j}, \theta^i_{2,j}, \theta^i_{3,j}, \theta^i_{4,j}, \theta^i_{5,j})$ are the underlying substitution probabilities for synonymous transitions, synonymous transversions, nonsynonymous transitions, nonsynonymous transversions, and no substitution at site i for sequence pair j. If the codons differ at more than 1 position, we consider the different pathways that could result in the observed substitutions, each of which may include different numbers of component substitutions. In general, if y^i_j differ at n positions, there are at most $n!$ substitution pathways, and each pathway results in a unique $\mathbf{z}^i_j$ consisting of the sum totals of each substitution type along the pathway. In this case, each $\mathbf{z}^i_j$ is distributed as $\mathbf{z}^i_j \sim Multinomial(n, \boldsymbol{\theta}^i_j)$. Pathways containing intermediary stop codons should not be allowed. Our model allows for a mixture over possible pathways by defining the probability mass function $p(\mathbf{z}^i_j | \mathbf{y}^i_j)$. This allows the exploration of different hypotheses regarding pathways, such as those which suggest the increased probability of pathways consisting primarily of synonymous or transitionary substitutions. Thus, we have the following

$$p(\boldsymbol{\theta}^i_j | y^i_j) \propto p(\boldsymbol{\theta}^i_j) \sum_{\mathbf{z}^i_j} \binom{\sum_{l=1}^5 z^i_{l,j}}{z^i_{1,j} \ldots z^i_{5,j}} \prod_{l=1}^5 (\theta^i_{l,j})^{z^i_{l,j}} \times p(\mathbf{z}^i_j | \mathbf{y}^i_j).$$

Finally for each site i and each pair j, we relate the mean of the outcome variable to a linear predictor using the following link function:

$$g\left(\theta^i_{l,j}\right) = \alpha_l + \beta_{l,i} + \gamma_{l,j},$$

for $l = 1, \ldots, 5$. After imposing the appropriate constraints on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and defining a prior distribution on the model parameters $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, we can use MCMC to obtain samples from the joint posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{y})$. Choosing a logit link function g and Gaussian priors on $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ allows the use of an iteratively reweighted least squares algorithm (IRLS) within the Metropolis sampler, thereby producing efficient Gaussian proposal mechanisms that enhance convergence [1].

In our model, the $\boldsymbol{\alpha}$ terms model mutation-level effects, the $\boldsymbol{\beta}$ terms model site-level effects, and the $\boldsymbol{\gamma}$ terms model cross-sequence effects. The nonsynonymous/synonymous substitution ratio for each codon i can be estimated as a function of the $\theta^i_{l,j}$'s by averaging over pairwise comparisons j. These ratios can be used to estimate the positive selection probabilities for each codon. Similarly, by averaging over all sites i, it is possible to define a measure of the evolutionary distance between the two sequences represented by j, based on the estimated probability of substitution between the sequences. A phylogeny tree can be recovered from these distances using an algorithm for tree reconstruction such as neighbor joining.

3. Examples

A preliminary version of this model has been implemented in C++ and used to analyze two simulated data sets. The data sets *sim1* and *sim2* were simulated using Yang's `evolver` program, included in the `PAML` distribution [5]. In `evolver`, the user specifies each parameter of the codon-substitution model *M3* [6]. The M3 discrete model is characterized by three nonsynonymous/synonymous substitution rates, ω_1, ω_2 and ω_3, having probabilities p_1, p_2 and p_3. Other user specified parameters include the phylogeny, the underlying codon frequencies, and the transition/transversion ratio κ. The program then simulates sequence evolution along the given phylogeny according to the parameter values using the Nielsen-Yang model of sequence evolution. In the simulations we use $\omega_1, \omega_2 \leq 1$ and $\omega_3 > 1$, therefore only the codons evolved under ω_3 are positively selected.

The data set *sim1* consists of 15 sequences of 100 codons with $\omega_1 = 0.1$, $\omega_2 = 0.8$, $\omega_3 = 3$, $Pr(\omega_1) = 0.5$, $Pr(\omega_2) = 0.4$ and $Pr(\omega_3) = 0.1$. The codon frequencies for the 61 codons were assumed to be equal, i.e., $Pr(AAA) = \ldots = Pr(TTT) = 0.01693$. The transition/transversion ratio κ was set to 2. Data set *sim2* was similarly prepared, but smaller. The *sim2* alignment contains 8 sequences of 25 codons.

Here the nonsynonymous/synonymous substitution rates were set at $\omega_1 = 0.1$, $\omega_2 = 1$, and $\omega_3 = 10$, with probabilities 0.4, 0.4, and 0.2 respectively. Codon frequencies and the transition/transversion ratio remained as above. A phylogenetic tree was chosen for each simulation such that the tree's branch lengths would be of the same order of magnitude as those of the best estimates for the malaria parasite phylogenies. The total branch length for the *sim1* tree was 1.47 expected nucleotide substitutions per codon, and 1.06 for the *sim2* tree. The tree topologies were chosen arbitrarily.

Each simulated data set was analyzed for positive selection by four different methods: `PAML`'s `codeml`, `MrBayes`, `ADAPTSITE`, and the new Bayesian method. Analysis with `codeml` was conducted assuming the same M3 evolutionary model used previously to generate the sequences. We also provided `codeml` with the true generative tree topologies. All other model parameters and branch lengths were then estimated by maximum likelihood. `codeml` identifies a site i as positively selected when empirical Bayes calculations indicate $Pr(\omega_i > 1) > 0.9$. Analysis in `MrBayes` builds upon that of `codeml` by assuming prior distributions for the M3 model parameters and then using MCMC to sample from the joint posterior distribution of parameters. For these analyses, `MrBayes'` default priors were used, and the chains were started with random phylogenies rather than the true generative phylogenies provided to `codeml`. The *sim1* MCMC was run for $100,000$ iterations, sampled every 100, with a 100 sample burn-in. The *sim2* MCMC ran for $50,000$ iterations, also sampled every 100 with a 100 sample burn-in. Convergence in each case was assessed by means of parameter traces. During each iteration of MCMC, `MrBayes` calculates the posterior probability of each w for each site i in order to determine $Pr(\omega_i > 1)$ for each site. A site is identified as positively selected when the posterior mean $Pr(\omega_i > 1)$ is greater than 0.9.

`ADAPTSITE` detects and measures positive selection differently than the previous two methods. It requires a distance-based phylogeny created using the `njboot` program provided as part of the `LINTREE` package, and the complete 4×4 nucleotide substitution rate matrix. Our trials were conducted using a Kimura 2-parameter rate matrix, utilizing the true transition/transversion ratio used during sequence generation ($\kappa = 2$). Ancestral sequences are then estimated by maximum parsimony in order to count the numbers of nonsynonymous and synonymous substitutions. These counts are used to fit a binomial substitution model. Positive selection is identified when the binomial model's two-tailed p-value rejects a neutral selection hypothesis, and the observed substitutions counts favor nonsynonymous substitutions.

	sim1	*sim2*
actual sites	6, 16, 26, 60, 62, 70, 78, 81	4, 5, 11, 13, 15
`codeml`	6, 16, 26, 60, 62, 70	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25
`MrBayes`	16	11, 13, 15
`ADAPTSITE`	none	none
Bayesian GLM method	6, 16, 26, 60, 89	4, 8, 11, 13, 15, 20

Table 1: Sites identified as positively selected by each method. Top row indicates sites known to be positively selected.

For the Bayesian GLM-based method, we assigned Gaussian prior distributions for each parameter block, i.e., $\boldsymbol{\alpha} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_I \overset{\text{IID}}{\sim} N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J \overset{\text{IID}}{\sim} N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, where $\boldsymbol{\mu} = vec(0)$ and $\sigma^2 = 10$. We also assumed a uniform mixture over substitution pathways, i.e., $p(\mathbf{z}_j^i|\mathbf{y}_j^i) \propto 1$. A Metropolis-Hastings algorithm was used to to sample from the appropriate posterior distributions. In this preliminary version, relatively simple independent random walk proposal distributions were used. In our trials the *sim1* MCMC ran for $5,000$ iterations, sampled every 10, with a 50 sample burn-in. The *sim2* MCMC ran for $2,500$ iterations, also sampled every 10 with a 50 sample burn-in. At this point, both chains exhibited convergence. During each iteration, we computed the value $\omega_i^* = \frac{\theta_3^i + \theta_4^i}{\theta_1^i + \theta_2^i + \theta_5^i}$ for each site i by averaging over j. A site is identified as positively selected when $Pr(\omega_i^* > 1) > 0.9$.

Table 1 summarizes the sites identified as positively selected by each method according to the criteria described above. These results demonstrate important drawbacks to each method. `PAML`'s identification of all sites as positively selected in the case of *sim2* occurred as a result of the ML estimates for ω_1, ω_2, and ω_3 all being greater than 1. `MrBayes` conservative–but accurate–predictions are likely due to phylogeny-related MCMC convergence

problems. Despite having met convergence criteria for scalar parameters, it was not possible to assess stationarity of the phylogeny parameter in the same manner. The phylogeny parameter in `MrBayes` consists of a tree topology and the associated branch lengths, each of which are sampled separately during MCMC, but written to output as a single tree structure . Assessing convergence of the topology component might involve parsing the tree output in order to abstract the topology, enumerating all possible topologies, and monitoring a trace based on the enumerated values. However, an enumeration of topologies would likely obfuscate many important similarities between tree topologies which should be considered when assessing parameter convergence. Furthermore, it does not make sense to speak of branch length convergence until *after* topology convergence has been achieved. Detecting positive selection via stochastic evolutionary models requires highly accurate recovery of branch lengths. The sensitivity of these inferences to small discrepancies in branch length is especially strong when the branch lengths are sufficiently small. Thus if the phylogeny parameter (topology and branch lengths) had not yet achieved stable mixing by the time all other scalar parameters appeared to have reached stability, the probabilities of positive selection would no doubt be affected. `ADAPTSITE`'s inability to detect any instances of positive selection can also be attributed to phylogenetic uncertainty. Suzuki and Gojobori previously remarked that their method's ability to detect positive selection is diminished when the total tree length falls below approximately 2.5 [3]. Below this threshold, the parsimony-based substitution counts are generally too low for `ADAPTSITE` to resolve the dominant selective pressure at a site. The new method performed consistently well for both data sets. The false positive rate may be seen as a cause for concern, but it is worth noting that 60% of the substitutions at *sim1* site 89 were nonsynonymous, as were 68% and 64% of the substitutions at *sim2* sites 8 and 20. We expect our inferences to improve as additional attention is given to model assumptions such as the substitution pathway mixture distribution $p(\mathbf{z}_j^i|\mathbf{y}_j^i)$ and the prior distributions.

4. Conclusions and Future Work

We consider these preliminary results of this new methodology to be very promising. As mentioned in Section 2, the new model also allows for estimation of measures of evolutionary distances between sequences, which can then be used to reconstruct a phylogeny tree via neighbor joining. Separate analyses (not shown) indicate our method to be quite adept at recovering pairwise distances for data of varying degrees of speciation. In particular, we were able to infer the correct tree topologies for the simulated data of Section 3. These results should serve to reassure that our models, while not necessarily based on explicit assumptions about the phylogeny, are able to capture the underlying phylogenetic structure.

Our future work will address issues such as sensitivity to the prior distributions, as well as how to include phylogenetic information via the prior distributions for the model parameters when such information is available. We expect to devote considerable time to investigating issues of model validation via simulation studies. In particular, we are interested in studying the impact of branch lengths and number of sequences on the predictive capabilities of our models.

From a biological point of view, our interest lies in the analyses of *P.falciparum* and *P.vivax* gene sequences encoding various candidate malaria antigens. These data sets consist of many (¿100) closely related DNA sequences for which no phylogenetic information is available, and therefore for which most traditional methods for detecting positive selection are not appropriate.

References

[1] Dani Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. Chapman Hall/CRC, 1997.

[2] JP Huelsenbeck and F Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–5, Aug 2001.

[3] Yoshiyuki Suzuki and Takashi Gojobori. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol*, 16(10):1315–1328, 1999.

[4] Yoshiyuki Suzuki, Takashi Gojobori, and Masatoshi Nei. ADAPTSITE: Detecting natural selection at single amino acid sites. *Bioinformatics*, 17(7):660–661, 2001.

[5] Z Yang. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–6, Oct 1997.

[6] Z Yang, R Nielsen, N Goldman, and AM Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–49, May 2000.