PhyloCSF's 'Omega Test'

Michael F. Lin*

Introduction

We have previously described PhyloCSF, a method that analyzes a multi-species nucleotide sequence alignment to determine whether it is likely to represent a protein-coding region [1]. PhyloCSF is based on a statistical comparison of empirical codon models (ECMs), which use thousands of parameters to model the rates of all possible codon substitutions on the phylogenetic tree relating the aligned species [2]. Estimating these ECMs requires alignments of many thousands of known coding and non-coding regions as training data. This presents an obstacle for using PhyloCSF in genomes that do not already have high-quality protein-coding gene annotations.

In this note, we describe an alternative mode implemented in the PhyloCSF software that performs a comparison of much simpler codon models, which do not require such training data. The approach is very similar to the d_N/d_S likelihood ratio test mentioned in our previous publications [1, 3] and implemented in PAML [4], but with a few useful tweaks. This alternative mode is not as accurate as the full ECM-based PhyloCSF method, but it can be applied immediately to alignments of any closely related species, requiring only a reasonable estimate of the phylogenetic tree relating them.

Brief review of the d_N/d_S test

The d_N/d_S test uses phylogenetic codon models (recently reviewed in [5] and [6]) to test for evidence that non-synonymous codon substitutions have occurred at significantly lower rates than synonymous substitutions in a given alignment. Specifically, the codon rate matrix is parameterized by the d_N/d_S ratio ω, the transition/transversion rate ratio κ, and the vector of codon frequencies π (further details given below). In addition to the rate matrix parameters, PAML takes an assumed tree topology as input and estimates each individual branch length.

In alignments of conserved coding regions, we expect $\omega < 1$, while under neutral or non-coding evolution, we expect $\omega \approx 1$; κ, π, and the branch lengths are essentially nuisance parameters. To evaluate a given alignment, we compute its probability under maximum likelihood estimates (MLEs) of all parameters, and its probability under the constraint that $\omega = 1$ and MLEs of the other parameters. We then report the log-ratio of these two likelihoods as the score for the alignment (if the estimated $\omega < 1$).

The new test implemented in PhyloCSF differs from this approach in three relatively minor ways. First, while PAML's models exclude stop codons and require them to be censored in the input alignments, we explicitly model the expected absence of stop codons in coding regions. As a result, the appearance of a stop codon in an input alignment leads to a penalty in its score. Second, like the full PhyloCSF method, we avoid estimating each individual branch length (which is difficult in a short alignment), instead estimating a scale factor for a predefined tree "shape." This tree shape is the only additional

*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. 32 Vassar St. 32-D510, Cambridge, MA 02139. mlin@mit.edu

information needed to evaluate a given alignment, and can be estimated using standard phylogenetic tools based on alignments of just a few known genes. Third, since the MLEs of some of the parameters can be poorly behaved in very short alignments, we regularize our estimates using weakly-informative prior distributions.

Codon model parameterization

We now give a detailed specification for the phylogenetic codon models used in our new test. The rate matrix uses a typical GY94 formulation [7, 5, 6], with the exception that we explicitly model substitution rates for the three stop codons, resulting in a 64×64 rate matrix rather than the 61×61 typically used in codon models.

$$q_{ij} \propto \begin{cases} \omega\pi_j & \text{if } i \text{ and } j \text{ are non-synonymous sense codons and differ by one transversion,} \\ \omega\kappa\pi_j & \text{if } i \text{ and } j \text{ are non-synonymous sense codons and differ by one transition,} \\ \pi_j & \text{if } i \text{ and } j \text{ are synonymous codons (or either is stop) and differ by one transversion,} \\ \kappa\pi_j & \text{if } i \text{ and } j \text{ are synonymous codons (or either is stop) and differ by one transition,} \\ -\sum_{k \neq i} q_{ik} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The full rate matrix $\mathbf{Q}$ is scaled to unity mean rate of replacement at equilibrium. Note that ω does not apply to nonsense substitution rates; instead, these rates are penalized through the codon frequencies π.

Our parameterization for π is also a slight variant of the typical F3×4 approach [8]. Let ϕ_a^p represent codon position-specific nucleotide frequencies, with $\sum_{a \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \phi_a^p = 1$ for each $p \in \{1, 2, 3\}$. For example, $\phi_{\mathtt{C}}^2$ is the frequency of $\mathtt{C}$ in the second codon position. We also define a new parameter σ to model a reduction in the frequency of the three stop codons, relative to expectation under F3×4.

$$\pi_{xyz} \propto \begin{cases} \sigma\phi_x^1\phi_y^2\phi_z^3 & \text{if } xyz \text{ is a stop codon,} \\ \phi_x^1\phi_y^2\phi_z^3 & \text{otherwise} \end{cases}$$

and $\sum_{xyz} \pi_{xyz} = 1$. The scale factor to ensure this is $1/\left(1 - (1-\sigma)\left(\phi_{\mathtt{T}}^1\phi_{\mathtt{A}}^2\phi_{\mathtt{A}}^3 + \phi_{\mathtt{T}}^1\phi_{\mathtt{A}}^2\phi_{\mathtt{G}}^3 + \phi_{\mathtt{T}}^1\phi_{\mathtt{G}}^2\phi_{\mathtt{A}}^3\right)\right)$.

Lastly, as previously mentioned, we avoid estimating each individual branch length in the assumed phylogenetic tree. Instead, we assume a fixed tree "shape" specifying a *relative* length of each branch, and use a single alignment-specific parameter ρ, which operates as a scale factor on the tree shape, to determine the absolute branch lengths. That is, if branch i has length t_i in the tree shape, it has length ρt_i in the model where ρ is variable, and shared throughout the tree.

Omega Test: Bayesian formulation

We now describe a Bayesian hypothesis-testing approach to distinguish coding and non-coding regions based on this codon model. To evaluate a given alignment A, we wish to formally test for evidence that $\omega < 1$ and $\sigma < 1$. Specifically, we define a "null hypothesis" H_0 corresponding to $\omega = 1$, $\sigma = 1$, and a composite alternative hypothesis H_1 specifying prior distributions for ω and σ, strongly preferring $\omega < 1$ and $\sigma < 1$. Additionally, both H_0 and H_1 specify diffuse prior distributions for $\kappa \in [1, \infty)$ and $\rho \in [0, \infty)$.

We compute the Bayes factor for H_1 against H_0,

$$K = \frac{\Pr(A|H_1)}{\Pr(A|H_0)}$$
$$= \frac{\iiiint \Pr(\rho)\Pr(\kappa)\Pr(\omega)\Pr(\sigma)\Pr(A|\rho,\kappa,\omega,\sigma)\ d\rho\ d\kappa\ d\omega\ d\sigma}{\iint \Pr(\rho)\Pr(\kappa)\Pr(A|\rho,\kappa,\omega=1,\sigma=1)\ d\rho\ d\kappa}$$

If $\log K > 0$, then under our model's assumptions the evidence in the alignment favors H_1. K can also be combined with prior probabilities for H_1 and H_0 to determine a posterior probability $\Pr(H_1|A)$ [9].

For the parameter priors, we used:

$$\begin{aligned}
\rho - 1 &\sim \text{half-Cauchy}(1) \\
\kappa - 1 &\sim \text{Gamma}(4, \tfrac{1}{2}) \\
\omega &\sim \text{Beta}(1, 8) \\
\sigma &\sim \text{Beta}(1, 8)
\end{aligned}$$

The "half-Cauchy" prior we suggest for ρ (the Cauchy distribution defined only on non-negative reals) has a nice property of entertaining arbitrarily small values of ρ, including $\rho = 0$ [10]. As a practical matter, we use F3×4 point estimates for the codon frequency parameters [8].

Although we have successfully experimented with this approach using reversible-jump MCMC, we do not consider it computationally cost-effective for genome-wide application to many thousands of candidate regions, as we intend our software to be used. We next describe a more practical approximation implemented in our software, but it is useful to have the fully-Bayesian formulation in mind.

Omega Test: implemented formulation

Our implemented version computes an approximate Bayes factor using predefined point estimates of ω and σ and maximum *a posteriori* (MAP) estimates of ρ and κ.

$$\widetilde{K} = \frac{\max_{\rho,\kappa} \Pr(\rho)\Pr(\kappa)\Pr(A|\rho,\kappa,\omega=0.2,\sigma=0.01)}{\max_{\rho,\kappa} \Pr(\rho)\Pr(\kappa)\Pr(A|\rho,\kappa,\omega=1,\sigma=1)}$$

The priors for the nuisance parameters ρ and κ mainly serve to regularize the MAP estimates, which is useful in very short alignments. For example, if we examine a short alignment that happens to show no transversions, then the MLE of κ is infinite. This creates some practical difficulties in implementation, which we can avoid by instead computing the MAP estimate with the diffuse prior.

The use of small point estimates for ω and σ reflects the fact that we are mainly interested in the very large reductions in non-synonymous rates and stop codon frequencies that any conserved coding region should exhibit. In particular, we are fine with returning a low score for an alignment in which the "true" ω is 0.9, even if we could in principle detect a statistically significant reduction. Furthermore, an alignment with a "true" ω much smaller than 0.2 will still be much more probable under this assumed value than under $\omega = 1$. The small assumed value of σ effectively leads to a large penalty for the appearance of stop codons in the alignment. It would be straightforward to also compute MAP estimates of ω and σ, but the method as described is already considerably slower than the full PhyloCSF method (which only searches for an MLE of ρ). The exact assumed values of ω and σ can be adjusted by the user.

References

[1] Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27: i275–i282.

[2] Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Molecular Biology and Evolution 24: 1464–1479.

[3] Lin MF, Deoras AN, Rasmussen MD, Kellis M (2008) Performance and scalability of discriminative metrics for comparative gene identification in 12 drosophila genomes. PLoS Computational Biology 4: e1000067.

[4] Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586 –1591.

[5] Anisimova M, Kosiol C (2008) Investigating Protein-Coding sequence evolution with probabilistic codon substitution models. Molecular Biology and Evolution 26: 255–271.

[6] Delport W, Scheffler K, Seoighe C (2008) Models of coding sequence evolution. Briefings in Bioinformatics 10: 97–109.

[7] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11: 725–736.

[8] Kosakovsky Pond S, Delport W, Muse SV, Scheffler K (2010) Correcting the bias of empirical frequency parameter estimators in codon models. PLoS ONE 5: e11230.

[9] Goodman SN (1999) Toward evidence-based medical statistics. 2: The bayes factor. Annals of Internal Medicine 130: 1005–1013.

[10] Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1: 119.