MBE Advance Access published February 6, 2012

An Improved Likelihood Ratio Test for Detecting Site-Specific Functional Divergence among Clades of Protein-Coding Genes

Cameron J. Weadick[†,1] and Belinda S.W. Chang[*,1,2,3]

[1]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

[2]Department of Cell & Systems Biology, University of Toronto, Toronto, Ontario, Canada

[3]Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada

†Present address: Department of Evolutionary Biology, Max Plank Institute for Developmental Biology, Tuebingen, Germany

***Corresponding author:** E-mail: belinda.chang@utoronto.ca.

Associate editor: Andrew Roger

Abstract

Maximum likelihood codon substitution models have proven useful for studying when and how protein function evolves, but they have recently been criticized on a number of fronts. The strengths and weaknesses of such methods must therefore be identified and improved upon. Here, using simulations, we show that the Clade model C versus M1a test for functional divergence among clades is prone to false positives under simple evolutionary conditions. We then propose a new null model (M2a_rel) that better accounts for among-site variation in selective constraint. We show that the revised test has an improved false-positive rate and good power. Applying this test to previously analyzed data sets of primate ribonucleases and mammalian rhodopsins reveals that some conclusions may have been misled by the original method. The improved test should prove useful for identifying patterns of divergence in selective constraint among paralogous gene families and among orthologs from ecologically divergent species.

Key words: codon substitution model, nonsynonymous-to-synonymous substitution rate ratio, *d*N/*d*S, clade model, maximum likelihood, gene family evolution.

Changes in protein function contribute to adaptive phenotypic diversity; determining when, how, and why protein function evolves constitute major goals within the field of molecular evolution (Dean and Thornton 2007). In recent years, likelihood-based codon models of evolution have become widely used for identifying signatures of functional diversification following gene duplication or niche shifts (Anisimova and Kosiol 2009). However, the design and implementation of many of these methods have attracted controversy (Nei et al. 2010), and it is important that these techniques be subject to rigorous critical evaluation (e.g., Yang and dos Reis 2011).

Codon substitution models provide estimates of ω, the nonsynonymous-to-synonymous substitution rate ratio, which speaks to the form and strength of selection operating on protein-coding DNA; $\omega < 1$, $\omega = 1$, and $\omega > 1$ indicate purifying selection, neutrality, and positive selection, respectively (Anisimova and Kosiol 2009). Clade models are a class of codon models that accommodate site-specific divergence in selective constraint among clades (Forsberg and Christiansen 2003; Bielawski and Yang 2004). Most notably, Clade model C (CmC) of Bielawski and Yang (2004) has recently been used to study ω divergence in a number of systems (reviewed in Chang et al. forthcoming). However, the statistical properties of this method have not been extensively evaluated, a shortcoming we address through analyses of simulated and biological data sets (for methods, see supplementary text, Supplementary Material online).

CmC accommodates divergence by estimating separate ω ratios for two or more clades (ω_2, $\omega_3 > 0$; assuming two clades) (table 1). Only some sites are fit with divergent ω ratios; the remainder are assumed to experience selection consistently across clades, evolving under purifying selection ($0 < \omega_0 < 1$) or neutrality ($\omega_1 = 1$), with the proportion (p) of sites in each site class (SC) estimated from the data (Bielawski and Yang 2004; Yang et al. 2005). Comparing the fit of CmC against the null model M1a (which assumes no divergence) is intended to test for site-specific divergence among clades, with significance established via a likelihood ratio test (LRT) against a χ^2_3 distribution. However, CmC uses three SCs to describe among-site variation, whereas M1a uses only two (table 1), introducing a possible confound to the LRT that could lead to false positives. To test this, we simulated data assuming three SCs but without among-clade variation (SC0: $\omega_0 = 0.0$, $p_0 = 0.5$; SC1: $\omega_1 = 1.0$, $p_1 = 0.2$; SC2: $\omega_2 = 0.4$, $p_2 = 0.3$; fig. 1*a* inset). The CmC versus M1a LRT, if working as intended, should generate few positive test results, but we found that 99% were significant using the standard $\alpha = 5\%$ (50% at the 0.01% level; supplementary fig. S1*a*, Supplementary Material online). The CmC versus M1a LRT is thus highly unreliable when faced with moderate among-site ω variation.

Letter

Table 1. Clade Model C (CmC) and Its Null Models M1a and M2a_rel.

SC / Model	0: Purifying ω	0: Purifying Proportion	1: Neutral ω	1: Neutral Proportion	2: Divergent ω	2: Divergent Proportion
CmC	$0 < \omega_0 < 1$	p_0	$\omega_1 = 1$	p_1	$\omega_2, \omega_3 > 0$	$1 - p_0 - p_1$
M1a	$0 < \omega_0 < 1$	p_0	$\omega_1 = 1$	$1 - p_0$	—	—
M2a_rel	$0 < \omega_0 < 1$	p_0	$\omega_1 = 1$	p_1	$\omega_2 (=\omega_3) > 0$	$1 - p_0 - p_1$

We devised a new null model for comparison against CmC, the M2a_rel model, which is formed by applying a single nonboundary constraint to CmC such that $\omega_2 = \omega_3$. As a result, ω no longer varies among clades (table 1), and the LRT's null distribution should follow a χ^2_1 distribution (Goldman and Whelan 2000). Applied to the same simulated data sets described above, the CmC versus M2a_rel LRT only produced significant test results 4% of the time (supplementary fig. S1*b*, Supplementary Material online). Furthermore, the distribution of observed test statistics closely followed the expected χ^2_1 null distribution, and parameter estimates generated using the M2a_rel model closely matched simulated values (fig. 1*a* and *b*). This LRT still fared well when the assumption of a neutral SC was violated (SC0: $\omega_0 = 0.0$, $p_0 = 0.5$; SC1: $\omega_1 = 2.0$, $p_1 = 0.2$; SC2: $\omega_2 = 0.4$, $p_2 = 0.3$), with a 6% false-positive rate (supplementary fig. S1*c*, Supplementary Material online). However, the error rate was slightly elevated when four SCs were assumed (SC0: $\omega_0 = 0.0$, $p_0 = 0.25$; SC1: $\omega_1 = 0.33$, $p_1 = 0.25$; SC2: $\omega_2 = 0.66$, $p_2 = 0.25$; SC3: $\omega_3 = 1.0$, $p_3 = 0.25$), with 9% of the LRTs producing false positives (supplementary fig. S1*d*, Supplementary Material online); analysis of complex data sets may thus require parametric bootstrapping to ensure an appropriate null distribution, and future refinement of these models should apply continuous β distributions for describing among-site ω variation (Yang et al. 2000). We note that the likelihood surface was uneven for both CmC and M2a_rel; and suboptimal M2a_rel analyses could lead to false positives. This problem is known to be an issue with certain data sets (Bielawski JP, personal communication; see supplementary text, Supplementary Material online).

To evaluate the power of the CmC versus M2a_rel LRT, we simulated data assuming moderate divergence among clades (fig. 1*a* inset), with some sites switching from strong to weak purifying selection (SC0: $\omega_0 = 0.0$, $p_0 = 0.5$; SC1: $\omega_1 = 1.0$, $p_1 = 0.2$; SC2: $\omega_2 = 0.15$ and $\omega_3 = 0.65$, $p_2 = 0.3$). Reassuringly, the CmC versus M2a_rel LRT identified the signature of divergent constraint in 100% of the simulated data sets. The test still displayed fair power when very weak divergence was assumed (SC0: $\omega_0 = 0.0$, $p_0 = 0.5$; SC1: $\omega_1 = 1.0$, $p_1 = 0.2$; SC2: $\omega_2 = 0.3$ and $\omega_3 = 0.5$, $p_2 = 0.3$), with 62% of the tests yielding significant results (75% given $\alpha = 10\%$). This LRT is thus powerful enough to detect slight patterns of ω divergence under ideal conditions (supplementary fig. S1*e* and *f*, Supplementary Material online). Other properties beyond the magnitude of among-clade ω variation will also affect the LRT's power (e.g., divergent SC size), but we did not explore them here.

Several recent studies have employed the CmC versus M1a LRT (reviewed in Chang et al. forthcoming), and we suspected that some of their results might have been misled by this test. Thus, we compared the conclusions generated by the original and revised LRTs for two of these studies. We first analyzed a data set of primate RNases that accompanies PAML as an example data set for CmC (Yang 2007) (see also Bielawski and Yang 2004). Compared with the paralogous EDN RNases, ECP RNases have divergent antiviral and antimicrobial properties (Dyer and Rosenberg 2006). Applying CmC suggested that much of this data set ($p_2 = 0.38$) evolved divergently ($\omega_{ECP} = 3.67$; $\omega_{EDN} = 1.94$), and an LRT against the M1a null indicated significance ($P < 10^{-5}$). However, comparing CmC against the M2a_rel null yielded

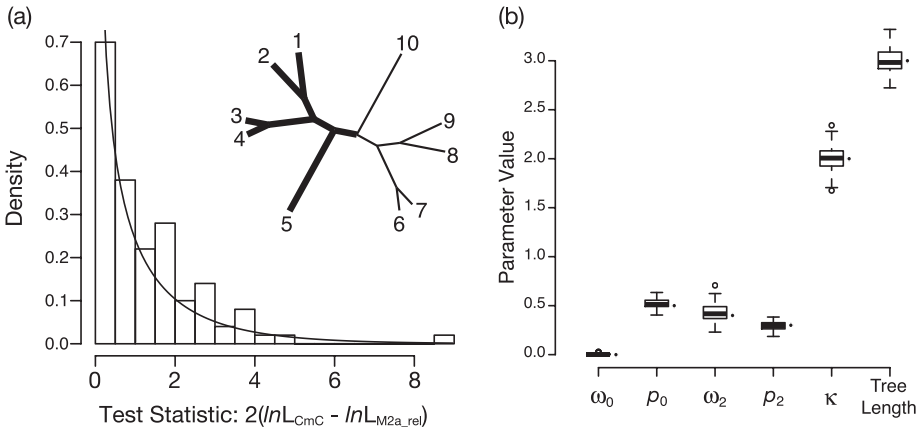

Fig. 1. (*a*) Histogram of CmC versus M2a_rel LRT test statistics given data simulated under the null model using a ten taxa tree (inset), with χ^2_1 density curve shown. Divergent clades indicated by thick versus thin branches. (*b*) Parameter estimates from M2a_rel analyses of the same data sets. Filled circles indicate the values used to simulate the data.

Table 2. LRT Results for Biological Data Sets.

Data Set	Model: Divergent Clade	Ln Likelihood	n.p.	LRT *P* Values	
				CmC versus M1a	CmC versus M2a_rel
	CmC: ECP	−2049.506	36	2.5×10^{-6}	0.0786
	M2a_rel	−2051.053	35	—	—
Primate RNases	M1a	−2063.958	33	—	—
	CmC: Mole rats	−9528.166	117	3.9×10^{-77}	0.3433
	CmC: Bats	−9518.779	117	3.3×10^{-81}	9.2×10^{-6}
	CmC: Cetaceans	−9512.950	117	9.9×10^{-84}	2.2×10^{-8}
	CmC: Pinnipeds	−9510.250	117	6.7×10^{-85}	1.4×10^{-9}
	M2a_rel	−9528.615	116	—	—
Mammalian Rhodopsins	M1a	−9706.830	114	—	—
	CmC: Fruit bats	−2828.207	33	0.0010	0.0697
	CmC: Yangochiroptera	−2828.858	33	0.0019	0.1585
	CmC: Rhinolophidae	−2828.921	33	0.0020	0.1723
	M2a_rel	−2829.853	32	—	—
Bat Rhodopsins	M1a	−2836.319	30	—	—

NOTE.—n.p., number of parameters.

a substantially larger *P* value that failed to breach the 5% threshold ($P = 0.08$; table 2). The choice of null model can therefore qualitatively affect conclusions based on CmC analyses.

Next, we examined the findings of Zhao et al. (2009), who used CmC to study divergence among mammalian rhodopsins. Rhodopsin mediates dim-light vision in vertebrates, and the invasion of light-limited niches may alter the nature of selection on this gene. Zhao et al. (2009) applied CmC to a diverse rhodopsin data set with bats, pinnipeds, cetaceans, or African mole rats considered, in turn, as the divergent clade. We reanalyzed this data set using the original CmC versus M1a LRT, and, like Zhao et al. (2009), found overwhelming support in favor of the alternative model (all $P < 10^{-75}$). By contrast, applying our revised LRT yielded much larger *P* values (all $P > 10^{-10}$) and no longer indicated significance when mole rats were considered the divergent clade ($P = 0.34$; table 2). Zhao et al. (2009) interpreted their significant results for the mole rat clade as indicating degenerative evolution or possibly positive selection, but our analyses suggest that such speculation might be unwarranted, as the difference in ω between the mole rats ($\omega \approx 0.25$) and other mammals ($\omega \approx 0.19$) is not statistically meaningful. Zhao et al. (2009) carried out further analyses on a reduced data set composed solely of bats that vary in whether and how they echolocate; Yangochiropteran bats distinguish calls and echoes by time, Rhinolophid bats by frequency, and Old World fruit bats do not echolocate at all (Teeling 2009). Zhao et al. (2009) applied the original CmC LRT with each group considered, in turn, as the divergent clade. Consistent with their analyses, we found that the CmC versus M1a LRT indicated significant differences for all three cases (all $P < 0.01$), but none remained significant when we reanalyzed the data with our revised LRT (all $P > 0.05$; table 2).

Overall, we have shown that the CmC versus M1a LRT for site-specific divergence among clades fails to distinguish among-site and among-clade ω variation, whereas our revised CmC versus M2a_rel LRT has an improved false-positive rate and good power. Moreover, we have shown that the choice of null model has implications for the analysis of biological data sets. Researchers that employed the M1a null model in past studies should revisit their data using our improved LRT. More broadly, the M2a_rel model will be available to those interested in testing for variation in selection pressure among clades via future versions of PAML (Yang Z, personal communication). We believe that the development and use of clade models will foster increased understanding of the processes that generate functional variation among genes and species.

Supplementary Material

Supplementary text and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank H. Rodd for feedback and support, Z. Yang for assistance with the codeml source code, S. Rossiter and H. Zhao for providing sequence data, and J. Bielawski, B. Fraser, A. Cutter, members of the Chang lab, and an anonymous reviewer for comments on earlier drafts. This work was supported by NSERC (to C.J.W., B.S.W.C., and H. Rodd) and a U. Toronto VSRP fellowship (to C.J.W.).

References

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

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol.* 59:121–132.

Chang BSW, Du J, Weadick CJW, Muller J, Bickelmann C, Yu DD, Morrow JM. Forthcoming. The future of codon models in studies of molecular function: ancestral reconstruction, and clade models of functional divergence. In: Cannarozii GM, Schneider A, editors. Codon evolution: mechanisms and models. Oxford: Oxford University Press.

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet.* 8:675–688.

Dyer KD, Rosenberg HF. 2006. The RNase a superfamily: generation of diversity and innate host defense. *Mol Divers.* 10:585–597.

Forsberg R, Christiansen FB. 2003. A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Mol Biol Evol.* 20:1252–1259.

Goldman N, Whelan S. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol Biol Evol.* 17:975–978.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet.* 11:265–289.

Teeling EC. 2009. Hear, hear: the convergent evolution of echolocation in bats? *Trends Ecol Evol.* 24:351–354.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol.* 28:1217–1228.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22:1107–1118.

Zhao H, Ru B, Teeling EC, Faulkes CG, Zhang S, Rossiter SJ. 2009. Rhodopsin molecular evolution in mammals inhabiting low light environments. *PLoS One* 4:e8326.