Looking for Darwin in Genomic Sequences—Validity and Success of Statistical Methods

Weiwei Zhai,[1] Rasmus Nielsen,[2,3] Nick Goldman,[4] and Ziheng Yang*[,5]

[1]Center for Computational Biology and Laboratory of Disease Genomics and Individualized Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China

[2]Department of Integrative Biology, University of California, Berkeley

[3]Department of Statistics, University of California, Berkeley

[4]EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[5]Department of Biology, Galton Laboratory, University College London, Gower Street, London, UK

***Corresponding author:** E-mail: z.yang@ucl.ac.uk.

Associate editor: Arndt von Haeseler

Abstract

The use of codon substitution models to compare synonymous and nonsynonymous substitution rates is a widely used approach to detecting positive Darwinian selection affecting protein evolution. However, in several recent papers, Hughes and colleagues claim that codon-based likelihood-ratio tests (LRTs) are logically flawed as they lack prior hypotheses and fail to accommodate random fluctuations in synonymous and nonsynonymous substitutions Friedman and Hughes (2007) also used site-based LRTs to analyze 605 gene families consisting of human and mouse paralogues. They found that the outcome of the tests was largely determined by irrelevant factors such as the GC content at the third codon positions and the synonymous rate d_S, but not by the nonsynonymous rate d_N or the d_N/d_S ratio, factors that should be related to selection. Here, we reanalyze those data. Contra Friedman and Hughes, we found that the test results are related to sequence length and the average d_N/d_S ratio. We examine the criticisms of Hughes and suggest that they are based on misunderstandings of the codon models and on statistical errors. Our analyses suggest that codon-based tests are useful tools for comparative analysis of genomic data sets.

Key words: codon model, Darwinian selection, likelihood-ratio test.

Introduction

It has long been realized that synonymous and nonsynonymous substitution rates can be compared to infer the direction and strength of natural selection acting on the protein (Kimura 1977). A number of methods were developed to estimate the numbers of synonymous and nonsynonymous substitutions per site (d_S and d_N) between two sequences (for review, see Yang 2006a). Applications of those methods to real data painted a picture of protein evolution dominated by purifying selection eliminating deleterious nonsynonymous mutations, with $d_N < d_S$. However, such pairwise comparisons lack power to detect positive selection as they average rates over all codons in the gene and over the whole time period separating the two sequences. Simultaneous comparison of many sequences using codon models (Goldman and Yang 1994; Muse and Gaut 1994) allows more sophisticated likelihood ratio tests (LRTs) to be constructed (Anisimova and Kosiol 2009). To improve power, the models focus on particular branches on the tree or individual codons in the gene, such as the "branch models" (Yang 1998), the "site models" (Nielsen and Yang 1998; Yang et al. 2000; Kosakovsky Pond and Muse 2005; Massingham and Goldman 2005; Rubinstein et al. 2011), and the "branch-site models" (Yang and Nielsen 2002; Yang et al. 2005). In several recent papers, Hughes and colleagues (Hughes 2007; Friedman and Hughes 2007; Hughes and Friedman 2008; Hughes 2012) raised a number of philosophical criticisms on the use of codon models to detect positive selection on the protein. They claim that the site-based tests lack prior hypothesis and fail to account for random fluctuations of substitutions in the gene. We examine those criticisms below. Additionally, Friedman and Hughes (2007), referred to later as "FH07," applied site-based LRTs (see table 1) to analyze human and mouse paralogous genes and found that the outcome of the LRTs was largely determined by the sequence length, the GC content at the third codon position, and d_S, factors that should have little to do with selection on the protein, but not by d_N or the d_N/d_S ratio, factors that should reflect the action of selection. Those results are surprising, and if true, may cast doubts on the performance of the codon-based LRTs.

Reanalysis of the Data of FH07

We reanalyzed the 605 gene families of FH07, with the results summarized in table 2. FH07 found 329 (54.5%) alignments in which at least one of the M1–M2 and M7–M8 tests (see table 1) detected positive selection at the 5% level. This proportion is very high and does not seem convincing, as suggested by FH07. In contrast, the corresponding figure from

Letter

Table 1. Site Models Implemented in Different Versions of the CODEML Program.

Model code	p	Parameters	Software	References
M0 (one-ratio)	1	ω		
M1 (neutral)	1	p_0, p_1 $\omega_0 = 0, \omega_1 = 1$	PAML 3.13	Nielsen and Yang (1998)
M2 (selection)	3	p_0, p_1, p_2 $\omega_0 = 0, \omega_1 = 1, \omega_2$	PAML 3.13	Nielsen and Yang (1998)
M1a (neutral)	2	p_0, p_1 $\omega_0 < 1, \omega_1 = 1$	PAML 3.14 or later	Wong et al. (2004) and Yang et al. (2005)
M2a (selection)	4	p_0, p_1, p_2 $\omega_0 < 1, \omega_1 = 1, \omega_2 > 1$	PAML 3.14 or later	Wong et al. (2004) and Yang et al. (2005)
M7 (beta)	2	p, q	PAML 3.13 or later	Yang et al. (2000)
M8a (beta and $\omega = 1$)	3	$p_0, p, q, \omega_s = 1$	PAML 3.13 or later	Swanson et al. (2003) and Wong et al. (2004)
M8 (beta and ω)	4	$p_0, p, q, \omega_s > 1$	PAML 3.13 or later	Yang et al. (2000) and Yang et al. (2005)

NOTE.—p is the number of free parameters. Three LRTs are constructed to compare the following pairs of models: M1–M2, M1a–M2a, and M7–M8. For each test, positive selection is inferred if $2\Delta\ell > 5.99$, $\hat{\omega} > 1$, and $\hat{p}_2 = 1 - \hat{p}_0 - \hat{p}_1 > 0$. A fourth LRT compares M8a against M8, with the 1:1 mixture of 0 and χ^2_1 used to conduct the test so that positive selection is inferred if $2\Delta\ell > 2.71$. The differences between the two PAML versions are as follows: (1) In M1 and M2 (PAML 3.13, released in 2002), $\omega_0 = 0$, while M1a and M2a (versions 3.14, September 2004, or later) have $0 < \omega_0 < 1$. (2) In version 3.13, only the naïve empirical Bayes (NEB) method (Nielsen and Yang 1998) is available, while version 3.14 added the BEB method (Yang et al. 2005). (3) In version 3.13, ω_2 in M2 and ω_s in M8 (beta and ω) are estimated over the range (0, ∞). In version 3.14 or later, they are estimated over the range (1, ∞).

our reanalysis using the same models in PAML 3.13 was only 63 (10%). Our efforts to reproduce the results of FH07 have not been successful. The large differences are not due to different versions of the PAML software and are unlikely to be due to numerical problems. We suspect that FH07 failed to apply the correct criterion in declaring positive selection by the tests (see table 1). At any rate the criticisms of FH07 on the codon models are based on incorrect results and are unfounded.

We also analyzed the data of FH07 using the M1a–M2a and the M8a–M8 tests using PAML version 4.4 (table 1). The M1a–M2a test identified 18 gene alignments with significant signal of positive selection, compared with 7 by the M1–M2 test (table 2). The M7–M8 test identified 64 genes to be under positive selection while the M8a–M8 test identified 63, both including as a subset the 18 genes identified by the M1a–M2a test. As noted previously (Wong et al. 2004), the M1a–M2a test tends to be more stringent than the M7–M8 test. In 14 gene alignments, M2a identified at least one codon with the Bayes empirical Bayes (BEB) posterior probability for positive selection, $P > 95\%$. For M8, this number is 31, including the 14 cases from M2a. For those data, the results for PAML versions 3.13 and 4.4 are very similar, and both are very different from those of FH07.

Following FH07, we examined the impact of various factors on the outcome of the LRTs. The results are summarized in supplementary table S1 (Supplementary Material online). FH07 found that the test outcome was significantly correlated with sequence length, GC_3 and mean synonymous substitution rate $(\bar{d}_S)$, but not with mean nonsynonymous substitution rate $(\bar{d}_N)$ or mean d_N/d_S ratio $(\bar{\omega})$. In contrast, we found that the test outcome is significantly influenced by the sequence length ($P < 0.05$, Kruskal–Wallis test), as in FH07, and by the mean d_N/d_S ratio $\bar{\omega}$ ($P < 0.05$, Kruskal–Wallis test), but not with GC_3, $\bar{d}_S$, or $\bar{d}_N$. The impact of sequence length is as expected, as the test tends to be more powerful in larger data sets (Anisimova et al. 2001). The correlation with $\bar{\omega}$ is not hard to explain since on average high d_N/d_S ratios may indicate positive selection driving amino acid changes, although a strong correlation is not automatically expected. The LRTs considered here accommodate variable ω ratios across codons so that the test may be significant even if most codons in the gene are under strong purifying selection while a few codons experience accelerated nonsynonymous substitutions driven by positive selection without elevating substantially the average ratio $\bar{\omega}$.

Table 2. Numbers of Genes Showing Significant (at the 5% level) Evidence of Positive Selection by Different Tests.

	M7/M8 Not Significant	M7/M8 Significant
FH07 (From Friedman and Hughes 2007)		
M1/M2 not significant	275 (45.5%)	120 (19.9%)
M1/M2 significant	102 (16.9%)	107 (17.7%)
PAML 3.13		
M1/M2 not significant	542 (89.6%)	56 (9.3%)
M1/M2 significant	1 (0.17%)	6 (1.0%)
PAML 4.4		
M1a/M2a not significant	541 (89.4%)	46 (7.6%)
M1a/M2a significant	0 (0%)	18 (3.0%)
PAML 4.4[a]		
M1a/M2a not significant	542 (89.6%)	45 (7.4%)
M1a/M2a significant	0 (0%)	18 (3.0%)

NOTE.—[a]The M8a–M8 comparison was used instead of the M7–M8 comparison.

As in FH07, we examined whether the set of genes showing evidence of positive selection tend to have certain tree topologies (supplementary table S2, Supplementary Material online). Like FH07, we found no strong and clear association between the outcome of the LRTs and the tree topology. However, the results from the two studies are in all other aspects quite different.

Do the LRTs Based on the Site Models Lack a Prior Hypothesis?

Hughes (2007) criticized the site-based LRTs (Nielsen and Yang 1998; Yang et al. 2000) and the parsimony method of Suzuki and Gojobori (1999) for lacking prior biological hypotheses about which codons are potentially under positive selection, likening them to an undirected "fishing expedition" (Hughes and Friedman 2008). Hughes values the approach taken by Hughes and Nei (1988) to analyze the Major Histocompatibility Complex (MHC) genes, where the overdominance theory of Doherty and Zinkernagel (1975) and the availability of a crystal structure (Bjorkman et al. 1987) led to the hypothesis that the antigen recognition site (ARS) may be the target of positive selection. This hypothesis was confirmed when Hughes and Nei showed that the ARS domain experienced accelerated nonsynonymous substitutions. Hughes (2007) criticized the site-based methods for lacking such a prior biological hypothesis.

Nevertheless, the LRTs based on the site models are based on a priori well-specified statistical hypotheses, derived from our biological knowledge of the effects of natural selection. For example, the null hypothesis M1a (neutral) assumes two site classes with $0 < \omega_0 < 1$ and $\omega_1 = 1$, but does not allow for sites with $\omega > 1$. The alternative hypothesis M2a (selection) adds another site class with $\omega_2 > 1$. Strong preference for M2a over M1a will be statistical evidence for the presence of codons at which nonsynonymous mutations have elevated fixation probabilities relative to synonymous mutations, the hallmark of positive selection driving amino acid changes. To test whether there exist amino acid residues under positive selection, one does not have to know where such residues may be. Such hypothesis testing need not be based on knowledge of the mechanisms of adaptive evolution. Nevertheless, such information, if available, can be used in the test (Yang and Swanson 2002).

In most cases, we lack prior knowledge of where in each gene positive selection may be acting. LRTs can still be validly used to identify genes showing previously unsuspected signals of selection—"fishing expeditions" may actually find and catch fish. Given a significant test result, statistical methods exist to identify sites under positive selection (e.g., Yang et al. 2005). Indeed, this is a major strength of those methods: by narrowing down amino acid changes that must be examined in the laboratory using site-directed mutagenesis, those methods may provide useful guidance for experimental work. Numerous examples now exist in which the results of such statistical analysis were validated by functional assays in the laboratory (e.g., Bielawski et al. 2004; Sawyer et al. 2005; Deng et al. 2010; Moury and Simon 2011). Even when structural information is available, it is rarely precise enough to pinpoint codons under positive selection, and the site-based model may still be very useful. This appears to be precisely the case with the MHC, the exemplar of positive selection, as demonstrated by Yang and Swanson (2002).

Does Random Fluctuation of Synonymous and Nonsynonymous Substitutions in a Codon Invalidate the Site-Based LRTs?

Hughes (2007; see also Hughes and Friedman 2005, 2008) claims that tests based on identifying codons with $d_N/d_S > 1$ are logically flawed because "all molecular data sets, even in the absence of positive selection, are likely to include some codons with $\omega > 1$." We suggest that such an argument confuses statistics with parameters. A parameter is a fixed constant that characterizes the population, while a statistic is a summary or estimate from the data and fluctuates among data sets. Part of the confusion may have resulted from the practice of using d_N/d_S to refer to both the parameter (ω) and its estimate ($\hat{\omega}$ or d_N/d_S). Statistical hypotheses underlying the LRTs are formulated using parameters (the true unobserved ω ratios), while the d_N/d_S (or $\hat{\omega}$) ratios that Hughes and Friedman (2005, 2008) calculated from the data are statistics (Yang 2006b). The error is most apparent in the calculation of Hughes and Friedman (2008: equations 1–6) of the probability of observing a codon with no synonymous differences and one or more nonsynonymous differences, as if the LRT would be significant in every data set containing such codons. The authors are correct to claim that such codons can occur by chance even if selection is purifying; they are wrong to believe that in every such data set the LRT will be significant. One can easily design a simulation experiment in which such codons occur in nearly every data set so that the argument of Hughes and Friedman would suggest the false positive rate of the LRT to be ~100%, whereas in reality it is <5% (Supplementary Material online). Even if the estimated ω ratio is infinity, if the estimate is based on very few changes the LRT will not be significant. Dealing with random fluctuations in the data is indeed the essence and purpose of a statistical test.

The Impact of Model Violation

A further criticism we seek to lay to rest is that "LRTs are invalid if neither of the two models compared is true." Criticisms in this area are old and have been answered before. As the statistician Box (1979) stated, "Models are never true, but fortunately it is only necessary that they be useful. For this it is usually needful only that they not be grossly wrong." Although model violation is a concern, a number of simulation studies have examined the statistical properties (such as robustness) of codon-based tests, including the site-based tests (Anisimova et al. 2001, 2002; Wong et al. 2004), the branch-site test (Yang et al.

2005; Zhang et al. 2005; Yang and dos Reis 2011), and the BEB identification of sites (Anisimova et al. 2002). For example, the site-based tests were found to be quite robust to misspecification of the functional form used to model variable ωs among sites (Anisimova et al. 2002; Wong et al. 2004), to the tree topology, and to among-site variability in other aspects of the substitution process (such as the transition/transversion rate ratio, codon usage, and d_S) (Bao et al. 2008). They are nevertheless sensitive to excessive recombinations (Anisimova et al. 2003; Shriner et al. 2003; Wilson and McVean 2006) and to sequence and alignment errors (Schneider et al. 2009; Fletcher and Yang 2010; Markova-Raina and Petrov 2011; Jordan and Goldman 2012). Furthermore, substantial progress has been made in improving codon models in the past 10–15 years (Rubinstein et al. 2011; Cannarozzi and Schneider 2012).

In summary, statistical tests throughout the biological sciences are not perfect, do not have perfect power and zero error, and cases can always be found where they mislead or fail. Nevertheless, they give us a means to progress and learn. Furthermore, those tests are constantly being improved upon, to become more powerful and robust.

Materials and Methods

The alignments of FH07 for 605 gene families with two human and two mouse paralogues were retrieved from the *Molecular Phylogenetics and Evolution* Web site. Alignment gaps were removed in FH07. The phylogeny was easily inferred and was either tree I: $(H_1, M_1) - (H_2, M_2)$ or tree II: $(H_1, H_2) - (M_1, M_2)$, with gene duplication either before or after species split, respectively.

Following FH07, we applied two LRTs comparing site models M1 (neutral) against M2 (selection), and M7 (beta) against M8 (beta and ω). FH07 appears to have used the CODEML program in PAML version 3.13, and we used the same version. In addition, we applied the LRTs comparing the modified models M1a (neutral) against M2a (selection), and M8a against M8, using the current PAML version 4.4.

We also followed FH07 to examine possible relationships between the LRT results and five features of the gene alignment: the sequence length (the number of codons), mean GC content at third codon positions (GC_3), mean d_S ($\bar{d}_S$), mean d_N ($\bar{d}_N$), and mean d_N/d_S ($\bar{\omega}$), all calculated by averaging over all pairwise comparisons.

Supplementary Material

Supplementary material and tables S1 and S2 are available at *Molecular Biology and Evolution online* (http://www.mbe.oxfordjournals.org/).

References

Anisimova M, Bielawski JP, Yang Z. 2001. The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. *Mol Biol Evol.* 18:1585–1592.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol.* 19:950–958.

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

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–1236.

Bao L, Gu H, Dunn KA, Bielawski JP. 2008. Likelihood-based clustering (LiBaC) for codon models, a method for grouping sites according to similarities in the underlying process of evolution. *Mol Biol Evol.* 25:1995–2007.

Bielawski JP, Dunn KA, Sabehi G, Beja O. 2004. Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc Natl Acad Sci U S A.* 101:14824–14829.

Bjorkman PJ, Saper SA, Samraoui B, Bennet WS, Strominger JL, Wiley DC. 1987. Structure of the class I histocompatibility antigen, HLA-A2. *Nature* 329:506–512.

Box GEP. 1979. Some problems of statistics and everyday life. *J Am Stat Assoc.* 74:1–4.

Cannarozzi G, Schneider A. 2012. Codon evolution: mechanisms and models. New York: Oxford University Press.

Deng C, Cheng C-HC, Ye H, He X, Chen L. 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc Natl Acad Sci U S A.* 107:21593–21598.

Doherty PC, Zinkernagel RM. 1975. A biological role for the major histocompatibility antigens. *Lancet* 1:1406–1409.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions and alignment errors on the branch-site test of positive selection. *Mol Biol Evol.* 27:2257–2267.

Friedman R, Hughes AL. 2007. Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal non-conservative and anomalous properties of widely used methods. *Mol Phylogenet Evol.* 42:388–393.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99:364–373.

Hughes AL. 2012. Evolution of adaptive phenotypic traits without positive Darwinian selection. *Heredity.* 108:347–353.

Hughes AL, Friedman R. 2005. Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. *Mol Biol Evol.* 22:1320–1324.

Hughes AL, Friedman R. 2008. Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. *Immunogenetics* 60:495–506.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol.* 29:1125–1139.

Kimura M. 1977. Prepondence of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267:275–276.

Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol.* 22:2375–2385.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res.* 21:863–874.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169:1753–1762.

Moury B, Simon V. 2011. dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus Y. *Mol Biol Evol.* 28:2707–2717.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11:715–724.

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

Rubinstein ND, Mayrose I, Doron-Faigenboim A, Pupko T. 2011. Evolutionary models accounting for layers of selection in protein coding genes and their impact on the inference of positive selection. *Mol Biol Evol.* 28:3297–3308.

Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A.* 102:2832–2837.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol.* 1:114–118.

Shriner D, Nickle DC, Jensen MA, Mullins JI. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res.* 81:115–121.

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

Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol.* 20:18–20.

Wilson DJ, McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172:1411–1425.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15:568–573.

Yang Z. 2006a. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2006b. On the varied pattern of evolution of two fungal genomes: a critique of Hughes and Friedman. *Mol Biol Evol.* 23:2279–2282.

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. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol.* 19:908–917.

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

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.

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