Ratios of Radical to Conservative Amino Acid Replacement are Affected by Mutational and Compositional Factors and May Not Be Indicative of Positive Darwinian Selection

Tal Dagan, Yael Talmor, and Dan Graur

Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University

The ratio of radical to conservative amino acid replacements is frequently used to infer positive Darwinian selection. This method is based on the assumption that radical replacements are more likely than conservative replacements to improve the function of a protein. Therefore, if positive selection plays a major role in the evolution of a protein, one would expect the radical-conservative ratio to exceed the expectation under neutrality. Here, we investigate the possibility that factors unrelated to selection, i.e., transition-transversion ratio, codon usage, genetic code, and amino acid composition, influence the radical-conservative replacement ratio. All factors that have been studied were found to affect the radical-conservative replacement ratio. In particular, amino acid composition and transition-transversion ratio are shown to have the most profound effects. Because none of the studied factors had anything to do with selection (positive or otherwise) and also because all of them (singly or in combination) affected a measure that was supposed to be indicative of positive selection, we conclude that selectional inferences based on radical-conservative replacement ratios should be treated with suspicion.

Introduction

Nonsynonymous substitutions are far more likely than synonymous substitutions to improve the function of a protein. Because advantageous mutations undergo fixation much more rapidly than neutral mutations and also because the rate of synonymous mutation per synonymous site is the same as the rate of nonsynonymous mutation per nonsynonymous site, the rate of nonsynonymous substitution is expected to exceed that of synonymous substitution, if positive Darwinian selection plays a major role in the evolution of a protein. Nei and Gojobori (1986) were the first to take advantage of this rationale to infer purifying selection. In their method, the ratio between nonsynonymous and synonymous rates is used; if the ratio is significantly larger than 1, advantageous selection is inferred. This method was used in a large number of studies, e.g., most recently by Bielawski and Yang (2001), Ford (2001), Johnson and Seger (2001), Lukens and Doebley (2001), Swanson et al. (2001), and Welch and Meselson (2001). Indeed, Endo et al. (1996) used this method to estimate the prevalence of positive selection and concluded that advantageous selection is a rare phenomenon, being detectable in their set of protein-coding genes in only ~0.5% of the cases. One problem with the nonsynonymous-synonymous ratio is that synonymous substitutions tend to become saturated; therefore, they are underestimated more quickly than nonsynonymous substitutions. In such cases, the nonsynonymous-synonymous ratio may artifactually exceed 1, and positive selection may be inferred where none exists.

Hughes, Ota, and Nei (1990) proposed to circumvent the saturation problem by using the ratio of radical to conservative amino acid replacements. The rationale of this method is very similar to that used in the nonsynonymous-synonymous ratio case. That is, radical replacements are assumed to be more likely than conservative replacements to improve the function of a protein. Therefore, if positive selection plays a major role in the evolution of a protein, we should expect the radical-conservative ratio to exceed the expectation under no selection. There are several methods to estimate the radicalism or conservatism of a particular amino acid replacement. One, for example, may decide that the property of interest is electric charge, and therefore, all replacements that result in charge changes are radical, whereas all replacements that do not affect charge are conservative. Alternatively, several properties may be considered simultaneously through the use of a physicochemical measure, such as Grantham's (1974) distance. The radical-conservative replacement ratio has also been used extensively to infer positive selection (e.g., Hughes, Ota, and Nei 1990; Hughes 1992; Rand, Weinreich, and Cezairliyan 2000; Hughes 2000, 2002).

In this study, we investigate the possibility that factors unrelated to selection influence the radical-conservative replacement ratio values. For example, it is known that transversions result in more dramatic changes than do transitions. That is, transversions are more likely than transitions to be nonsynonymous in protein-coding regions, and nonsynonymous transversions are more likely to result in radical replacement than nonsynonymous transitions (Zhang 2000). It is, therefore, possible that differences in radical-conservative replacement ratios may be caused by mutations factors, such as the transition-transversion ratio, rather than selectional forces. In this study, we simulated DNA-sequence evolution and resulting radical-conservative replacement ratios by varying transition-transversion ratios, codon usage, genetic code, and amino acid composition. In the simulation we introduced no hint of positive selection.

Key words: positive Darwinian selection, conservative replacement, radical replacement, transition bias, codon usage, genetic codes, amino acid composition.

Address for correspondence and reprints: Tal Dagan, Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel. E-mail: tali@kimura.tau.ac.il.

Mol. Biol. Evol. 19(7):1022–1025. 2002

Table 1
Amino Acid Frequencies in the Different Compositions

Amino Acid	Dayhoff Equilibrium	JTT Equilibrium	Nuclear Proteins	Intracellular Proteins	Membranal Proteins	Anchored Proteins	Extracellular Proteins	Proline-rich Proteins
Alanine.	0.087	0.077	0.083	0.079	0.081	0.076	0.086	0.045
Cysteine	0.033	0.020	0.016	0.019	0.020	0.022	0.029	0.045
Aspartic acid	0.047	0.052	0.047	0.055	0.038	0.052	0.049	0.045
Glutamic acid	0.050	0.062	0.065	0.071	0.046	0.062	0.051	0.045
Phenylalanine	0.040	0.040	0.027	0.039	0.056	0.040	0.037	0.045
Glycine.	0.089	0.074	0.063	0.071	0.070	0.069	0.078	0.045
Histidine.	0.034	0.023	0.021	0.021	0.020	0.021	0.021	0.045
Isoleucine.	0.037	0.052	0.037	0.052	0.067	0.051	0.046	0.045
Lysine	0.080	0.059	0.079	0.067	0.044	0.058	0.063	0.045
Leucine	0.085	0.091	0.074	0.086	0.110	0.094	0.088	0.045
Methionine.	0.015	0.024	0.023	0.024	0.028	0.021	0.025	0.045
Aspargine.	0.040	0.043	0.037	0.040	0.037	0.044	0.046	0.045
Proline	0.051	0.051	0.069	0.053	0.047	0.054	0.049	0.136
Glutamine	0.038	0.041	0.047	0.044	0.031	0.041	0.040	0.045
Arginine	0.041	0.051	0.087	0.049	0.046	0.050	0.042	0.045
Serine	0.070	0.069	0.088	0.066	0.073	0.072	0.073	0.045
Threonine.	0.058	0.059	0.051	0.053	0.056	0.061	0.060	0.045
Valine	0.065	0.066	0.053	0.068	0.077	0.067	0.067	0.045
Triptophan	0.010	0.014	0.007	0.012	0.018	0.014	0.014	0.045
Tyrosine	0.030	0.032	0.024	0.031	0.033	0.032	0.036	0.045

Methods

Simulated Protein Evolution

Each virtual protein-coding gene was 300 nucleotides long, resulting in a protein 100 amino acids in length. Genetic code, codon usage, and amino acid composition were fixed at the beginning of each simulation. Each virtual gene was used as the ancestor sequence in the simulated-evolution program of ROSE software (Stoye, Evers, and Meyer 1998). In each run, fixed transition-transversion ratio values were used. Each combination of variables was run 50 times. The number of substitutions between the ancestor sequence and the resulting sequence was 50.

Radical-Conservative Ratios

All the 190 possible amino acid replacements were classified using three independent criteria: (1) charge, (2) volume and polarity, and (3) Grantham's (1974) physico-chemical distance.

Classification by charge was made by dividing the amino acids into three categories: positive (R, H, K), negative (D, E), and uncharged (A, N, C, Q, G, I, L, M, F, P, S, T, W, Y, V).

Classification by volume and polarity was made by dividing the amino acids into six categories: special (C), neutral and small (A, G, P, S, T), polar and relatively small (N, D, Q, E), polar and relatively large (R, H, K), nonpolar and relatively small (I, L, M, V), and nonpolar and relatively large (F, W, Y).

The two classifications above were taken from Zhang (2000). We did not use an additional classification in Zhang (2000), i.e., polarity, in order to keep the divisions independent of one another. Within each of the two classifications above, amino acid replacements were deemed conservative if they involved exchanges within a category and radical if the exchanges occurred among categories.

As far as Grantham's (1974) distances are concerned, an amino acid replacement was deemed conservative if the distance value was smaller than 100 and radical otherwise.

Codon Usage

Three patterns of codon usage were used: random, GC biased, and AT biased. In the random pattern, each codon frequency was calculated as the frequency of the amino acid specified by the codon divided by the number of possible codons for the amino acid. In the GC- and AT-biased patterns of codon usage, each codon frequency was calculated as the frequency of the amino acid specified by the codon divided by the number of possible codons ending in GC or AT, respectively.

Amino Acid Composition

Eight amino acid compositions were used. Two compositions were the theoretical equilibrium expectations of two replacement matrices, i.e., Dayhoff's (1978, p. 345) and JTT (Jones, Taylor, and Thornton 1992). Five compositions were derived from mean amino acid frequencies in different protein classes: (1) extracellular proteins, (2) anchored proteins, (3) membranal proteins, (4) intracellular proteins, and (5) nuclear proteins. The values were taken from Cedano et al. (1997). The eighth composition was of a proline-rich protein as an example of extreme amino acid bias. In this case, the frequency of 19 amino acids was set at 0.045, whereas the frequency of proline was 0.136. All amino acid frequencies are shown in table 1.

Transition-Transversion Ratios

Transition-transversion ratios inferred from real data range widely, depending among others on divergence time, lineage, and DNA origin (e.g., Lanave et al.

Table 2
***P* Values (left column) and Percent Variation (right column) Explained for Multiway Analyses of Variance of the Effects of Transition-Transversion Ratio, Amino Acid Composition, Codon Usage, Genetic Code, and Their Interactions on Radical-Conservative Ratio Measures Based on Amino Acid Classifications by Charge, Volume, and Polarity and Grantham's Distances**

Source of Variation	Charge		Volume and Polarity		Grantham's Distances	
Transition-transversion ratio	<0.0001	11.44	<0.0001	47.06	<0.0001	59.01
Amino acid composition	<0.0001	49.02	<0.0001	20.05	<0.0001	6.50
Codon usage	<0.0001	1.04	<0.0001	2.82	<0.0001	1.14
Genetic code	<0.0001	12.96	<0.0001	6.79	<0.0001	3.84
Transition-transversion ratio × amino acid composition	<0.0001	1.04	<0.0001	3.24	<0.0001	5.51
Transition-transversion ratio × codon usage	<0.0001	0.41	0.5104	0.34	<0.0001	0.35
Transition-transversion ratio × genetic code	<0.0001	0.45	<0.0001	0.50	<0.0001	1.43
Amino acid composition × codon usage	<0.0001	21.12	<0.0001	8.64	<0.0001	16.93
Amino acid composition × genetic code	<0.0001	0.13	<0.0001	0.74	<0.0001	0.17
Codon usage × genetic code	0.2497	0.00	0.0359	0.02	<0.0001	0.44
Transition-transversion ratio × amino acid composition × codon usage	<0.0001	1.12	<0.0001	5.36	<0.0001	1.85
Transition-transversion ratio × amino acid composition × genetic code	0.0049	0.35	0.4520	1.20	0.0003	0.75
Transition-transversion ratio × codon usage × genetic code	0.0041	0.11	0.0433	0.42	<0.0001	0.35
Amino acid composition × codon usage × genetic code	<0.0001	0.23	<0.0001	0.44	<0.0001	0.39
Transition-transversion ratio × amino acid composition × codon usage × genetic code	0.5813	0.57	0.5245	2.37	0.0132	1.32

1986; Purvis and Bromham 1997; Yang and Yoder 1999). In our simulation we varied the ratio from 0.017 to 29. We studied 58 ratios, the probability for transition ranging from 0.001 to 0.0295 at 0.0005 intervals and the probability for transversion ranging from 0.029 to 0.0005 at 0.0005 intervals. These simulated values contain the range of ratios reported in the literature.

Insertion and deletion frequencies were set to zero in order to keep the length of the sequences constant and prevent gaps in the alignment.

Genetic Code

Two genetic codes were used: the standard (so-called universal) code and the vertebrate mitochondrial code.

Statistical Analyses

The effects of various variables and the interactions among them on the three radical-conservative replacement ratios were tested by a multiway analysis of variance (ANOVA). All the effects were considered as fixed.

Reality check

In order to establish that compositional and mutational factors may indeed produce false positive inferences of Darwinian selection, we simulated the evolution of several human protein-coding genes in which positive selection has never been reported, e.g., β hemoglobin, interleukin 2, ribosomal protein S21 (accession numbers NM_000518.3, NM_001024.2, and NM_000586.1, respectively) under the substitution matrix of pseudogenes (presumably a completely neutral matrix of substitution reflecting the pattern of mutation without selection). The neutral substitution matrix was taken from Graur and Li (1999, p. 126)

Results and Discussion

The results of the multiway ANOVA are shown in table 2. Regardless of the measure used to estimate the radical-conservative replacement ratio, all four factors that have been studied were found to affect the radical-conservative replacement ratio. The transition-transversion ratio and the amino acid composition, as well as the interaction between these two factors, were found to have the most pronounced affect on the radical-conservative ratio. All three radical-conservative measures are affected by mutational and compositional factors. When the amino acid replacements are classified by charge, most of the variation in the radical-conservative ratio is explained by amino acid composition. When the amino acid replacements are classified by either volume and polarity or by Grantham's distance, most of the variation in the radical-conservative ratio is explained by the transition-transversion ratio. These results were unaffected by either length of protein or divergence time between the proteins.

We tested the frequency of false positive inferences of Darwinian selection by simulating neutral evolution in β hemoglobin, interleukin 2, and ribosomal protein S21. When the radical-conservative ratio was calculated on the basis of volume and polarity, 100% of estimates were false positives. When the radical-conservative ratio was calculated on the basis of Grantham's distances for β hemoglobin, interleukin 2, and ribosomal protein S21, 17%, 21%, and 13% of the estimates, respectively, were false positives. With these three proteins, we obtained no false positives when the radical-conservative ratio was calculated on the basis of electric charge. We note, however, that false positive inferences of Darwinian selection with electric charge as the yardstick for computing radical-conservative ratio were especially abundant in our simulations when the amino acid composition was that of proteins located in the nucleus. None of the three proteins used in the reality check part are nuclear.

We conclude that many factors that have nothing to do with selection (positive or otherwise) either singly or in combination affect measures that were supposed to be indicative of positive selection. Therefore, selectional inferences based on radical-conservative replacement ratios should be treated with utmost caution. In fact, we recommend that these measures not be used at all.

LITERATURE CITED

BIELAWSKI, J. P., and Z. YANG. 2001. Positive and negative selection in the *DAZ* gene family. Mol. Biol. Evol. **18**:523–529.

CEDANO, J., P. ALOY, J. A. PEREZ-PONS, and E. OUEROL. 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. **266**:594–600.

DAYHOFF, M. O. 1978. Atlas of protein sequence and structure, Vol. 5 (Suppl.3). National Biomedical Research Foundation, Silver Spring, Md.

ENDO, T., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

FORD, M. J. 2001. Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol. Biol. Evol. **18**: 639–647.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. Science **85**:862–864.

GRAUR, D., and W.-H. LI. 1999. Fundamentals of molecular evolution. Sinauer Associates, Inc., Sunderland, Mass.

HUGHES, A. L. 1992. Coevolution of the vertebrate integrin α- and β-chain genes. Mol. Biol. Evol. **9**:216–234.

———. 2002. Origin and evolution of viral interleukin-10 and other dna virus genes with vertebrate homologous. J. Mol. Biol. **54**:90–101.

HUGHES, A. L., J. A. GREEN, J. M. GARBAYO, and R. M. ROBERTS. 2000. Adaptive diversifications within a large family of recently duplicated, placentally expressed genes. Proc. Natl. Acad. Sci. USA **97**:3319–3323.

HUGHES, A. L., T. OTA, and M. NEI. 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.

JOHNSON, K. P., and J. SEGER. 2001. Elevated rates of nonsynonymous substitution in island birds. Mol. Biol. Evol. **18**: 874–881.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

LANAVE, C., S. TOMMASI, G. PREPARATA, and C. SACCONE. 1986. Transition and transversion rate in the evolution of animal mitochondrial DNA. Biosystems **19**:273–283.

LUKENS, L., and J. DOEBLEY. 2001. Molecular evolution of the teosinte branched gene among maize and related grasses. Mol. Biol. Evol. **18**:627–638.

NEI, M., and T. GOJOBORI. 1986. Simple method for estimating the number of synonymous and non-synonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

PURVIS, A., and L. BROMHAM. 1997. Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. J. Mol. Evol. **44**:112–119.

RAND, D. M., D. M. WEINREICH, and B. O. CEZAIRLIYAN. 2000. Neutrality tests of conservative–radical amino acid changes in nuclear and mitochondrially encoded proteins. Gene **291**:115–125.

STOYE, J., D. EVERS, and F. MEYER. 1998. ROSE: generating sequence families. Bioinformatics **14**:157–163.

SWANSON, W. J., A. G. CLARK, H. M. WALDRIP-DAIL, M. F. WOLFNER, and C. F. AQUADRO. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. USA **98**:7375–7379.

WELCH, D. B., and M. S. MESELSON. 2001. Rates of nucleotide substitution in sexual and anciently asexual rotifers. Proc. Natl. Acad. Sci. USA **98**:6720–6724.

YANG, Z., and A. YODER. 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. **48**:274–283.

ZHANG, J. 2000. Rates of conservative and radical nucleotide substitutions in mammalian genes. J. Mol. Evol. **50**:56–68.

PEKKA PAMILO, reviewing editor

Accepted January 31, 2002