Higher Intensity of Purifying Selection on >90% of the Human Genes Revealed by the Intrinsic Replacement Mutation Rates

Sankar Subramanian†[1] and Sudhir Kumar*†*

*Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University; and
†School of Life Sciences, Arizona State University

For over 3 decades, the rate of replacement mutations has been assumed to be equal to, and estimated from, the rate of "strictly" neutral sequence divergence in noncoding regions and in silent-codon positions where mutations do not alter the amino acid encoded. This assumption is fundamental to estimating the fraction of harmful protein mutations and to identifying adaptive evolution at individual codons and proteins. We show that the assumption is not justifiable because a much larger fraction of codon positions is involved in hypermutable CpG dinucleotides as compared with the introns, leading to a higher expected replacement mutation rate per site in a vast majority of the genes. Consideration of this difference reveals a higher intensity of purifying natural selection than previously inferred in human genes. We also show that a much smaller number of genes are expected to be evolving with positive selection than that predicted using sequence divergence at intron and silent positions in the human genome. These patterns indicate the need for using new approaches for estimating rates of amino acid–altering mutations in order to find positively selected genes and codons in genomes that contain hypermutable CpG's.

Introduction

Following the lead of Kimura (1977), biologists have routinely assumed that non-synonymous (replacement) 'mutations' in codons occur at the same rate as silent mutations in molecular evolutionary analyses (Miyata and Yasunaga 1980; Nei and Kumar 2000; Bustamante et al. 2005; Nielsen et al. 2005). This assumption is made because replacement mutation rates often cannot be measured experimentally due to technical limitations or by comparative sequence analysis because natural selection prevents the fixation of a majority of replacement mutations between species. The equality assumption is at the heart of most widely used methods aimed at testing for selection on genes, identifying codons with adaptive evolutionary changes, estimating the number of deleterious mutations, and even finding coding regions in genomic sequences (Nei and Kumar 2000; Yang and Bielawski 2000; Nekrutenko et al. 2003; Yampolsky et al. 2005). Recently, the comparison of the chimpanzee genome with the human genome has established that there has been extensive selection against suboptimal synonymous codons in the human genome (Hellmann et al. 2003; Urrutia and Hurst 2003; Chamary and Hurst 2005; Lu and Wu 2005; Mikkelsen et al. 2005; Parmley et al. 2006). This fact has prompted the use of sequence divergence in introns and intergenic regions (excluding the regulatory) for establishing mutation rates at replacement positions in codons (Hellmann et al. 2003; Lu and Wu 2005; Mikkelsen et al. 2005; Parmley et al. 2006).

Should the actual mutation rate at replacement positions (called intrinsic mutation rate) be expected to be the same as the mutation rate inferred from sequence divergence at silent sites and in noncoding regions (called extrinsic mutation rate)? The context dependence of mutation rates suggests that it is unlikely. For example, positions involved in CpG dinucleotides are known to be hypermutable when methylated and are estimated to accumulate divergence at 5–20 times higher rate than other positions (Krawczak et al. 1998; Bird 1999; Hellmann et al. 2003; Subramanian and Kumar 2003). This higher mutation rate accounts for 25–50% of the interspecific variation between human and chimpanzee (Subramanian and Kumar 2003; Mikkelsen et al. 2005), single-nucleotide polymorphism diversity (Halushka et al. 1999), and replacement mutations associated with human diseases. Although this context dependence has been known for over 2 decades, severity of its effect on the disturbance of the assumed equality of mutation rates in different codon positions, introns, and intergenic regions remains to be seriously examined, particularly on the tests of selection in molecular evolutionary analyses.

The proportion of replacement positions in the CpG configuration is a function of the frequency of the codons that contain C's and G's, which is highly correlated with the frequency of amino acids in a given protein. By contrast, the fraction of CpG positions in introns is determined by the mutation-decay balance in their genomic milieu (Hwang and Green 2004; Fryxell and Moon 2005). These two fractions are rarely identical, as is evident from an analysis of all available functional human genes with at least one intron (10,196 genes; fig 1*a*). The distributions of the fraction of replacement and intronic positions involved in CpG dinucleotide configurations (CpG contents) are distinctly different in shapes and central tendencies. The CpG contents of replacement positions show a larger dispersion than introns, and the replacement positions of a vast majority of the genes have a higher CpG content compared with their introns. Over all genes, replacement positions are involved in CpG configurations 2 times more often than intron positions ($P < 0.0001$). This CpG content difference is statistically significant for 49% of the genes (Fisher's exact test; $P < 0.05$), with 92% of the proteins showing an excess of replacement positions involved in CpG dinucleotides as compared with the introns.

What effect does this larger proportion of hypermutable replacement positions have on the estimates of natural selection? This question can be examined by devising a formula that predicts the intrinsic mutational divergence per

[1] Present address: Allan Wilson Centre for Molecular Ecology and Evolution, Institute for Molecular Biosciences, Massey University, Auckland, New Zealand.

Key words: adaptive evolution, mutation rate, test of selection, comparative genomics.

E-mail: s.kumar@asu.edu.

Mol. Biol. Evol. 23(12):2283–2287. 2006
doi:10.1093/molbev/msl123
Advance Access publication September 18, 2006

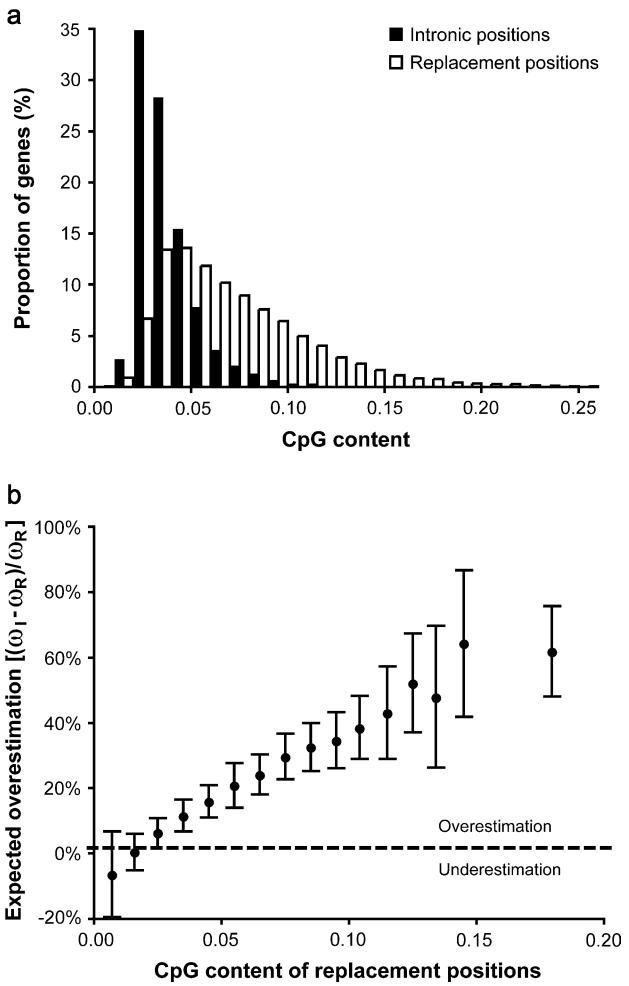


FIG. 1.—(*a*) The differential distribution of the fraction of intronic (filled bars) and replacement positions (open bars) involved in CpG dinucleotide configuration from 10,196 functional human genes containing at least one intron. On average, the replacement positions involved in CpG configurations are 2 times higher than those of intronic positions (6.2% and 2.9%, respectively). The dispersion indices (the ratio of variance to mean) of the distributions of intron and replacement positions are 0.015 and 0.024, respectively. (*b*) Overestimation of the coefficient of selection obtained using intronic divergence ($\omega_I = D_R/D_I$) as compared with the replacement mutation rate ($\omega_R = D_R/M_R$) is shown for 6,435 human genes that are devoid of CpG islands, have at least one intron, and show at least one replacement difference between human and chimpanzee (see Methods). The average CpG contents (fraction of replacement positions involved in CpG dinucleotides) shown on the *x* axis were estimated from the genes with CpG contents of (numbers of genes in parenthesis) 0–0.01 (56), 0.01–0.02 (461), 0.02–0.03 (892), 0.03–0.04 (842), 0.04–0.05 (741), 0.05–0.06 (649), 0.06–0.07 (567), 0.07–0.08 (494), 0.08–0.09 (431), 0.09–0.10 (333), 0.10–0.11 (268), 0.11–0.12 (188), 0.12–0.13 (143), 0.13–0.14 (101), 0.14–0.15 (71), and >0.15 (198). The average ω_I and ω_R were computed for each category. The overestimations of ω_R plotted on the *y* axis were computed as $(\omega_I - \omega_R)/\omega_R \times 100$. The error bars show the standard error of the mean. The relationship between the CpG content and the overestimation of ω_R is highly significant ($R^2 = 0.99$, $P < 0.01$).

replacement position (M_R) by accounting for the elevated mutation rate in the CpG dinucleotide positions. It is given by $M_R = [f_R + (1 - f_R)\,k](\mu_A + \mu_B) \times t$, where f_R is the fraction of replacement positions not involved in the CpG dinucleotides, μ_A and μ_B are the baseline mutation rates at non-CpG positions in species A and B, respectively, t is the time of species divergence, and k is the factor by which the per site CpG sequence divergence exceeds the baseline mutation rate. This equation can be further modified in such a way that the estimation of the expected replacement divergence between species (M_R) does not require the knowledge about the baseline mutation rates and the species divergence time (see Methods).

The fraction of replacement mutations eliminated by purifying selection is given by one minus the coefficient of selection (ω), where ω is equal to the observed replacement sequence divergence per site (D_R) divided by the expected number of replacement mutations per site (M_R). On the genomic scale, the average coefficient of selection ($\omega_I = D_R/D_I$) estimated using the sequence divergences at intron positions (D_I) as a proxy for replacement mutation rate is significantly higher than the estimates obtained by using the intrinsic replacement mutation rates ($\omega_R = D_R/M_R$) (0.24 and 0.20, respectively; $P < 10^{-20}$). Although the magnitude of the difference between the average estimates appears small (20%) on a genomic scale, the overestimation of the coefficient of selection ($\Delta\omega$) varies tremendously, with many genes showing >60% overestimation (fig. 1*b*). The coefficient of selection estimated using the intrinsic replacement mutation rate is higher for most of the human genes (94%) than that estimated using the extrinsic rate of their corresponding introns due to the greater fraction of replacement positions involved in CpG dinucleotides. The opposite pattern is true for a small fraction (~6%) of the genes, where ω is underestimated.

Significant overestimation of the coefficient of selection ($\omega_S = D_R/D_S$) results when the sequence divergence in silent positions is used as a proxy for mutation rate at replacement sites ($P < 10^{-32}$). However, the combined effect of differences in the fraction of hypermutable sites in silent and replacement positions and the varying degrees of purifying selection on interspecies divergence in silent positions leads to an unpredictable relationship between the estimates of ω based on intrinsic and extrinsic mutation rates. Still, the impact of using silent substitution rates on genes evolving with positive selection is striking when we examine genes for which $\omega_S > 1$, which is traditionally used as an indicator of positive selection.

The use of ω_S predicts that amino acid substitutions in 608 genes are positively selected, as compared with the use of ω_R that identifies a much smaller number of genes (173). Only 102 genes are shared by the predictions based on intrinsic and extrinsic rates, and a large fraction of the genes (83%) predicted by the use of silent divergence may fall into false-positive category (fig. 2*a*). At the same time, ω_S is too conservative for many other genes, as it fails to detect tracks of positive selection in 41% of the genes (71 out of 173). An examination of the biological processes performed by the positively selected genes indicates that a greater proportion of the genes identified using intrinsic mutation rates at replacement positions are involved in immunity and signal transduction (table 1). The use of silent divergence fails to identify over 50% of these genes, but identifies a large number of genes including those performing functions such as DNA replication and metabolism. The latter are unlikely candidates for positive selection as they are often involved in the housekeeping. Although the use of

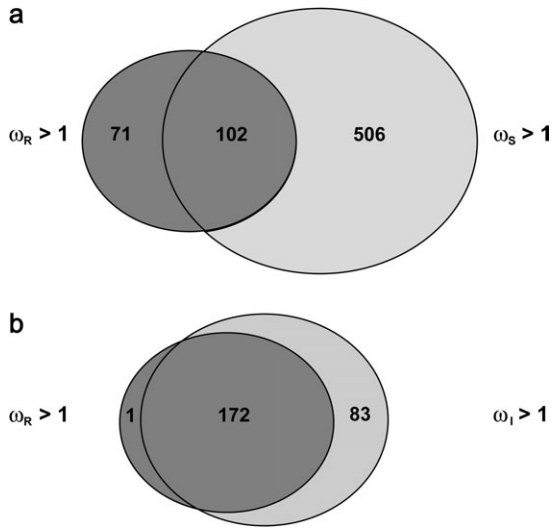


FIG. 2.—Venn diagrams showing the numbers of human genes under adaptive evolution ($\omega > 1.0$) predicted by 3 different methods that use silent divergence ($\omega_S = D_R/D_S$), intron divergence ($\omega_I = D_R/D_I$), and replacement mutation rate ($\omega_R = D_R/M_R$) to estimate coefficient of selection. The use of intrinsic replacement mutation rate was compared with the use of (*a*) silent divergence and (*b*) intron divergence in identifying the positively evolving genes. The numbers in the overlapping area indicate the genes that were identified by both the methods, and the numbers in the individual circles represent the genes that were predicted exclusively by one of the methods.

Table 1
Numbers of human genes identified to be positively selected ($\omega > 1$) by using predicted replacement mutation rates, intron sequence divergences, and silent sequence divergences

		Using Extrinsic Mutation Rate (false negatives[b], false positives[c])	
Biological Process[a]	Using Intrinsic Mutation Rate	Intron divergence	Silent divergence
Metabolism	20	25 (1, 6)	95 (10, 85)
Transport	9	14 (0, 5)	43 (2, 36)
Defense/immunity	71	89 (0, 18)	136 (34, 99)
Development	4	12 (0, 8)	20 (2, 18)
Signal transduction	32	50 (0, 18)	89 (17, 74)
Cellular process	23	39 (0, 16)	81 (6, 64)
Replication/transcription	10	11 (0, 1)	59 (4, 53)
Unknown biological process	80	119 (0, 39)	268 (31, 219)

[a] Genes belonging to more than one biological process are included in multiple categories, so the column sums may not equal those in figure 2.

[b] False negative refers to genes showing $\omega_R > 1$ by using intrinsic mutation rate from replacement positions, but missed when using intron or silent divergence ($\omega_I < 1$ or $\omega_S < 1$, respectively).

[c] False positive refers to genes that are identified by using intron or silent divergence to be positively selected ($\omega_I > 1$ or $\omega_S > 1$, respectively), but not by the use of predicted mutation rate from replacement positions ($\omega_R < 1$).

sequence divergence in introns predicts almost all the positively evolving genes that are identified by the intrinsic replacement mutation rate (fig. 2*b*), still 1 out of 3 predictions may be false positives. These false positives do not seem to be abundant in any specific functional category.

Our results demonstrate the presence of significantly more purifying selection on coding regions than previously thought because the replacement mutation rate is much higher than the observed substitution rate in introns and silent positions. These results indicate that the numbers of adaptive replacement changes in the genome of species with hypermutable CpG contents (e.g., primates) have been previously overestimated because a higher mutation rate is now implicated in producing the same observed replacement divergence between species (human and chimpanzee in the current case). These results are concerned with the strength of selection on the entire polypeptide and do not dispute existence of adaptive evolution on one or a few codons, which may occur even in highly conserved housekeeping genes (Yang and Nielsen 2002). In conclusion, however, it is clear that the estimates of mutation rates at replacement positions need to be based on the neighboring context of the replacement positions in analyses aimed at determining the proportion of replacement mutations eliminated by selection, predicting the fraction of disease-causing replacement mutations, and scanning the genome to find positively selected genes and codons.

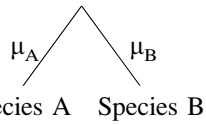
Methods

Protein coding and intron sequences of the human genome (Build 34) were obtained from GenBank (ftp://ftp.ncbi.nih.gov/genomes/). Because we compared mutations in introns and silent positions, only the functional genes with at least one intron were included in the analysis, which resulted in 10,196 genes. CpG content of replacement positions of a gene was estimated as the number of 0-fold degenerate positions (i.e., positions in which any mutation changes the amino acid coded by the codon) that are involved in the CpG configuration divided by the total number of 0-fold degenerate positions in the gene. Similarly, the CpG content of intron is the proportion of intron positions involved in CpG dinucleotide configuration (Krawczak et al. 1998; Bird 1999; Hellmann et al. 2003; Subramanian and Kumar 2003). The biological processes of human genes were obtained from Protein Analysis Through Evolutionary Relationships classification system (http://www.pantherdb.org/). Genes belonging to more than one biological process were included in multiple categories. For simplicity, we combined the biological processes that are known to be related and created 8 categories (table 1). For instance, cell structure, motility, proliferation, and adhesion were grouped into one category called cellular process (table 1). Similarly the biological processes such as DNA replication, repair, recombination, mRNA transcription, splicing, and pre–mRNA processing were combined to form the replication/transcription category.

The intrinsic mutability of the replacement positions is largely determined by the presence of CpG dinucleotides that mutate much faster than non-CpG nucleotides. Therefore, the mutation rate at replacement sites is the sum of the rates at CpG and at non-CpG positions. If f_R is the fraction of non-CpG positions and μ_A and μ_B are the baseline mutation rates at non-CpG nucleotides in lineages A and B, then the expected mutational divergence per replacement site (M_R) between the species A and B is given by

$$M_R = [f_R \times \mu_A \times t + (1 - f_R) \times k \times \mu_A \times t] + [f_R \times \mu_B \times t + (1 - f_R) \times k \times \mu_B \times t], \quad (1)$$

where t is the time of species divergence, k is the ratio of CpG to non-CpG mutation rate, and f_R is assumed to be equal in both the species compared (as the species considered are extremely closely related).



The equation (1) can be simplified as

$$M_R = [f_R + (1 - f_R) \times k] \times (\mu_A + \mu_B) \times t. \quad (2)$$

The factor $(\mu_A + \mu_B) \times t$ can actually be estimated from the sequence divergence observed in pairwise orthologous sequence comparison of intronic (or genomic noncoding) positions. If f_I is the fraction of non-CpG positions in the introns, then the intronic sequence divergence per site (D_I) between the same 2 species is given by

$$D_I = [f_I + (1 - f_I) \times k] \times (\mu_A + \mu_B) \times t,$$

$$(\mu_A + \mu_B) \times t = D_I/[f_I + (1 - f_I) \times k]. \quad (3)$$

Therefore, equation (2) can be modified as

$$M_R = D_I \times [f_R + (1 - f_R) \times k]/[f_I + (1 - f_I) \times k]. \quad (4)$$

Using equation (4), the expected mutational divergence at the replacement positions can be estimated without the knowledge of species divergence times as well as the baseline mutation rate in species A and B. For this purpose, we used the intronic sequence divergence (D_I) between human and chimpanzee for each gene as presented in Mikkelsen et al. (2005). The use of intronic divergence from human and chimpanzee will also incorporate the effects of other determinants of instantaneous mutation rate, such as the recombination rate and region-specific G + C content (Hardison et al. 2003; Hellmann et al. 2005; Mikkelsen et al. 2005). In these calculations, the CpG positions were, at an average, assumed to accumulate divergence at 10 times the rate at other positions ($k = 10$) (Krawczak et al. 1998; Bird 1999; Hellmann et al. 2003; Subramanian and Kumar 2003).

It is well known that the CpG dinucleotides present in CpG islands are not hypermutable (as they are not methylated), and these islands extend to the first few exons (and introns) (Bird 1999; Rollins et al. 2006; Saxonov et al. 2006). Therefore, the methylation status of the CpG dinucleotides is ambiguous in the genes containing CpG islands. For this reason, we identified the CpG islands following the method described by Duret and Galtier (2000) and excluded the genes that contain those islands from further analysis.

For the estimation of the coefficients of selection, 6,435 human genes that are devoid of CpG islands and have at least one intron and at least one observed replacement difference between human and chimpanzee were used. The replacement (D_R), silent (D_S), and intronic (D_I) sequence divergences between human and chimpanzee were obtained from the previous study conducted by chimpanzee sequencing and analysis consortium (Mikkelsen et al. 2005). The coefficients of selection were estimated using the intron divergence ($\omega_I = D_R/D_I$), silent divergence ($\omega_S = D_R/D_S$), and replacement mutation rate ($\omega_R = D_R/M_R$).

Acknowledgments

We thank Drs Brian Verrelli, Philip Hedrick, George Zhang, and Yuseob Kim for providing insightful remarks on a preliminary version of this article. We also thank an anonymous reviewer for constructive comments. This work was supported by a grant from National Institutes of Health to S.K.

Literature Cited

Bird A. 1999. DNA methylation de novo. Science. 286:2287–2288.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. Nature. 437:1153–1157.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6:R75.

Duret L, Galtier N. 2000. The covariation between TpA deficiency, CpG deficiency, and G + C content of human isochores is due to a mathematical artifact. Mol Biol Evol. 17:1620–1625.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 22:650–658.

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet. 22:239–247.

Hardison RC, Roskin KM, Yang S, et al. (18 co-authors). 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? Genome Res. 15:1222–1231.

Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. 13:831–837.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA. 101: 13994–14001.

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

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet. 63:474–488.

Lu J, Wu CI. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci USA. 102:4063–4067.

Mikkelsen TS, Hillier LD, Eichler EE, et al. (11 co-authors). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol. 16:23–36.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nekrutenko A, Chung WY, Li WH. 2003. An evolutionary approach reveals a high protein-coding capacity of the human genome. Trends Genet. 19:306–310.

Nielsen R, Bustamante C, Clark AG, et al. (13 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23:301–309.
Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH. 2006. Large-scale structure of genomic methylation patterns. Genome Res. 16:157–163.
Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA. 103: 1412–1417.
Subramanian S, Kumar S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 13:838–844.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res. 13:2260–2264.
Yampolsky LY, Kondrashov FA, Kondrashov AS. 2005. Distribution of the strength of selection against amino acid replacements in human proteins. Hum Mol Genet. 14:3191–3201.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Yoko Satta, Associate Editor

Accepted September 12, 2006