Widespread Positive Selection in Synonymous Sites of Mammalian Genes

Alissa M. Resch, Liran Carmel,* Leonardo Mariño-Ramírez,* Aleksey Y. Ogurtsov,* Svetlana A. Shabalina,* Igor B. Rogozin,* and Eugene V. Koonin**

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

Evolution of protein sequences is largely governed by purifying selection, with a small fraction of proteins evolving under positive selection. The evolution at synonymous positions in protein-coding genes is not nearly as well understood, with the extent and types of selection remaining, largely, unclear. A statistical test to identify purifying and positive selection at synonymous sites in protein-coding genes was developed. The method compares the rate of evolution at synonymous sites (Ks) to that in intron sequences of the same gene after sampling the aligned intron sequences to mimic the statistical properties of coding sequences. We detected purifying selection at synonymous sites in ~28% of the 1,562 analyzed orthologous genes from mouse and rat, and positive selection in ~12% of the genes. Thus, the fraction of genes with readily detectable positive selection at synonymous sites is much greater than the fraction of genes with comparable positive selection at nonsynonymous sites, i.e., at the level of the protein sequence. Unlike other genes, the genes with positive selection at synonymous sites showed no correlation between Ks and the rate of evolution in nonsynonymous sites (Ka), indicating that evolution of synonymous sites under positive selection is decoupled from protein evolution. The genes with purifying selection at synonymous sites showed significant anticorrelation between Ks and expression level and breadth, indicating that highly expressed genes evolve slowly. The genes with positive selection at synonymous sites showed the opposite trend, i.e., highly expressed genes had, on average, higher Ks. For the genes with positive selection at synonymous sites, a significantly lower mRNA stability is predicted compared to the genes with negative selection. Thus, mRNA destabilization could be an important factor driving positive selection in nonsynonymous sites, probably, through regulation of expression at the level of mRNA degradation and, possibly, also translation rate. So, unexpectedly, we found that positive selection at synonymous sites of mammalian genes is substantially more common than positive selection at the level of protein sequences. Positive selection at synonymous sites might act through mRNA destabilization affecting mRNA levels and translation.

Introduction

It is well established that nonsynonymous sites in protein-coding sequences are subject to purifying selection caused by constraints operating at the level of protein structure and function and that positive selection that, at least in mammals, affects a minority of genes and/or sites is an important force of adaptive evolution (Li 1997; Vallender and Lahn 2004; Bustamante et al. 2005; Nielsen et al. 2005). Synonymous (silent) sites are often used as a proxy for neutral evolution. Under this premise, the traditional gauge of selection in nonsynonymous sites is the ratio of nonsynonymous (Ka) over synonymous (Ks) substitutions. Ka/Ks <1 is thought to indicate purifying selection, whereas Ka/Ks >1 is construed as the signature of positive selection (Li 1997; Hurst 2002). However, the neutrality of synonymous substitutions is only a rough and not necessarily valid approximation; the extent, range, and underlying causes of selection in synonymous sites remain subjects of intense debate (Chamary, Parmley, and Hurst 2006). The results of several studies suggest that efficient translation (Ikemura 1985; Akashi and Eyre-Walker 1996; Eyre-Walker and Keightley 1999) and mRNA stability (Duan and Antezana 2003; Chamary and Hurst 2005; Shabalina, Ogurtsov, and Spiridonov 2006) are substantial forces of purifying selection in synonymous sites. It has also been shown that synonymous substitutions are under purifying selection in mammalian exonic splicing enhancer motifs (ESEs) (Yeo et al. 2004; Parmley, Chamary, and Hurst 2006) and in alternatively spliced exons (Xing and Lee 2005).

By contrast, to the best of our knowledge, positive selection in synonymous sites has not been reported. However, this possibility has been brought up in the course of analysis of the SPANX family of mammalian cancer-testis genes (Kouprina et al. 2004) and the insect cyclin A inhibitor gene (Avedisov et al. 2001) that are characterized by exceptionally high and comparable rates of evolution in both synonymous and nonsynonymous sites.

We were interested in addressing the problem at a fundamental level: is positive selection in synonymous positions a common phenomenon, and if so, what could be the underlying causes of such selection? We reasoned that, to investigate selection in synonymous sites, the substitution rate in intronic sequences (Ki) was a logical choice of the proxy for neutral evolution. A method for estimating the neutral rate using Ki has recently been reported (Hoffman and Birney 2007). In principle, at least, cases of negative selection in synonymous sites were identified as Ks/Ki <1 whereas cases of positive selection were indicated by Ks/Ki >1. No part of the genome can be automatically assumed to evolve neutrally: the possibility of a hidden function that constrains evolution or an adaptive component in the evolution of a sequence always should be considered. However, apart from pseudogenes, internal regions of introns are among the best candidates for neutrally evolving sequences. The sequences of ~30 nucleotides at each end of an intron are thought to be subject to weak purifying selection that stems from the involvement of these sequences in splicing (Louie, Ott, and Majewski 2003; Yeo et al. 2004) and SAS (unpublished observations). In addition, some of the introns contain highly conserved sequences with various, often unknown functions including genes for noncoding RNAs (Washietl et al. 2005). However, in

Key words: synonymous sites, nonsynonymous sites, positive selection, purifying selection, introns.

E-mail: koonin@ncbi.nlm.nih.govor; rogozin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 24(8):1821–1831. 2007
doi:10.1093/molbev/msm100
Advance Access publication May 23, 2007

mammals, these functional regions have been estimated to comprise <5% of the intronic sequences (Waterston et al. 2002). In addition, it has been demonstrated that, even in conserved noncoding sequences such as those found in introns, the pressure of purifying selection tends to be substantially weaker than in coding regions (Kryukov, Schmidt, and Sunyaev 2005). Thus, it appears that, after discarding terminal regions, introns could serve as a reasonable approximation of neutrally evolving sequences.

We found that Ks and Ki distributions for mammalian genes have different statistical properties, which makes Ki suspect as the neutral baseline for the analysis of selection in synonymous sites. Therefore we developed a computational procedure to shuffle aligned intron sequences such that their statistical properties mimic those of nonsynonymous sites and used the corresponding substitution rate (Ki-pseudo) to assess the extent of negative (purifying) and positive (diversifying) selection in synonymous sites of mammalian genes. It is shown that both types of selection in synonymous sites are widespread and that positive selection at synonymous sites is much more common than positive selection on protein sequence. Positive selection at synonymous sites is unrelated to functional constraints at the protein level, but is linked to gene expression, probably through mRNA destabilization.

Materials and Methods

Identification of Orthologous Genes

We calculated rates of divergence in coding and noncoding DNA for mouse-rat orthologs taken from the May 2004 HomoloGene database (Wheeler et al. 2006). HomoloGene orthologs are defined as bidirectional best hits using the BLAST program for sequence comparison (Altschul et al. 1997). Protein and mRNA sequences were obtained from the Entrez protein and nucleotide databases (Wheeler et al. 2006). We started with a total 8,178 mouse-rat orthologs but removed over half (see below) to eliminate the potential bias estimates of the Ks estimates that could be introduced by alternative splice variants and other alignment ambiguities (see next section). The final gene set employed for all analyses contained 1,562 mouse-rat orthologs.

Coding and Noncoding DNA Alignments

Protein alignments for mouse and rat were generated using the MUSCLE alignment package (Edgar 2004). Protein alignments were then used to guide alignment of the corresponding mouse and rat coding sequences (CDS). It was required that each coding sequence contain a start and stop codon, in order to eliminate all partial sequences. All alignments that contained insertions/deletions with the total length >30 bp were removed in order to exclude potential effects of incorrect gene prediction and alternative splicing.

Intron alignments were generated using the OWEN alignment tool (Ogurtsov et al. 2002) with the following specifications: (1) an intron must be bound on 5′/3′ ends by exons that align across ≥80% of length, (2) the presence of constitutive splice sites at each intron/exon boundary was required, (3) a P value <0.001 for each intron alignment was required, (4) 30-nucleotide regions from the 5′/3′ ends of each intron were removed, and 5) the proximal, 5′-terminal introns in the compared orthologous genes were discarded, because these introns are known to be enriched for various regulatory elements and, consequently, could be subject to purifying selection (Majewski and Ott 2002). These requirements help ensure that accurate orthologous intron alignments are generated. The 30-nucleotide regions from the ends of each alignment were removed to eliminate splicing signals from the estimates of intron divergence.

Comparison of Substitution Rates in Coding and Intronic Sequences

The evolutionary rates for coding DNA were originally calculated using the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993), which takes into account transition and transversion rates. Evolutionary rates for noncoding DNA were measured using the Kimura's 2-parameter model (Kimura 1980). However, considering the difference in the statistical properties of the CDS and intron sequences (see Results and Discussion), a method was developed to shuffle the intron sequence alignments such that their statistical properties mimicked those of coding sequences. Mouse-rat pseudo-CDS alignments were generated from alignments of mouse and rat intronic sequences using the following procedure for each alignment (supplementary fig. 1S): (1) Start the pseudo-CDS from ATG for both mouse and rat sequences. (2) Take the next pentanucleotide starting from the first codon position from the mouse CDS sequence and find this pentanucleotide in the mouse intronic sequence; if there are several such pentanucleotides, 1 is chosen randomly. (3) Add the corresponding segment of the intronic alignment to the pseudo-CDS; if the length of the pseudo-CDS alignment >5 nucleotides, the overlapping 2 nucleotides are chosen randomly; if a pentanucleotide is not found in the mouse intronic sequence, the corresponding fragment of the CDS alignment is added to the pseudo-CDS alignment. 4) The procedure is repeated until the end of the CDS alignment is reached. The resulting pseudo-CDS alignment has the same length as the CDS alignment, and the base compositions of mouse CDS and pseudo-CDS are identical. The significance of the difference in the codon composition of the rat CDS and pseudo-CDS was tested using a Monte-Carlo modification of the χ^2 test (Adams and Skopek 1987).

Detecting Positive and Negative Selection in Synonymous Sites

For each CDS alignment, 10,000 pseudo-CDS alignments were generated. A score of divergence at synonymous sites Ks was calculated using the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993) or the fraction of mismatches at 4-fold degenerate sites. This score was calculated for the mouse-rat CDS alignment (Ks) and 10,000 pseudo-CDS alignments (Ki-pseudo). The distribution of Ki-pseudo was used to calculate

probabilities $P(\text{Ks} \leq \text{Ki-pseudo})$ and $P(\text{Ks} \geq \text{Ki-pseudo})$; the fractions of the pseudo-CDS with Ks $\leq$ Ki-pseudo and Ks $\geq$ Ki-pseudo were taken as approximations of the respective P values. The genes with $P(\text{Ks} \geq \text{Ki-pseudo}) \leq 0.05$ were considered positively selected, and the genes with $P(\text{Ks} \leq \text{Ki-pseudo}) \leq 0.05$ were considered negatively selected. These calculations were performed for the complete alignments and repeated after masking CG, TG, and CA dinucleotides. For the analysis of statistical properties of distributions and correlation analysis, a pseudo-CDS alignment was randomly drawn from the total sample of 10,000, and Ki-pseudo was calculated using the PBL method.

Microarray Expression Analysis

The GNF Gene Expression Atlas2 data (Su et al. 2004) for mouse was used as the source of data on genes (rat expression data was limited for the majority of genes and therefore was not included in this analysis). The GNF Atlas2 data contain 2 replicates for each of 61 mouse tissues. The data for redundant tissue types was combined to yield a final set of 55 mouse tissues. Average expression values for each probe were calculated using raw expression data. Average probe expression values for raw data were calculated by summing the expression values across each probe set and dividing that sum by the total number of tissues (55). Tissue breadth values for each gene (the number of tissues that each probe is expressed in) were obtained using raw expression data. The raw expression value for a given tissue had to be $\geq$350 in order for that tissue to be counted in the tissue breadth analysis (Jordan, Marino-Ramirez, and Koonin 2005). Thus, the final tissue breadth score for a probe represents the number of all tissues with the raw expression value $\geq$350.

Codon Usage

The effective number of codons (ENC) for the coding sequences in the analyzed gene sets was calculated using previously described methods (Wright 1990). The codon adaptation index (CAI) scores (Sharp and Li 1987) for the analyzed coding sequences were calculated using the EMBOSS bioinformatics suite (Rice, Longden, and Bleasby 2000). The CAI values were calculated by comparing the codon usage patterns of a given gene against the codon usage patterns of a reference set of highly expressed mouse genes. Specifically, the GNF Atlas2 mouse expression data were used to identify the top ~10% (1,479/15,007) most highly expressed mouse genes by calculating the average overall expression level of each probe from raw expression data. The average expression values were ranked, from largest to smallest, to obtain the top 10%.

Distance Between Distributions

In order to quantify the dissimilarities between the distribution functions of Ka, Ks, Ki, and Ki-pseudo, we have computed pairwise distances between these distributions using an information-theoretic measure known as the L-divergence (Lin 1991). This distance measure is a refined version of the widely used Kullback-Leibler distance.

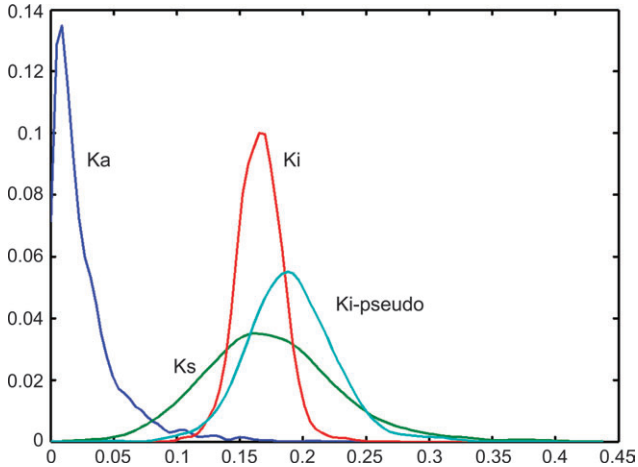

FIG. 1.—The distributions of Ka, Ks, Ki, and Ki-pseudo in the analyzed set of 1,562 rodent genes.

Results and Discussion

Using Intron Evolution Rate as the Baseline for Detecting Selection in Synonymous Sites

In order to avoid ambiguities of alignment, especially, in intron sequences, as well as substitution saturation effects, we limited the present analysis to orthologous genes from closely related rodents, mouse and rat. It has been recently shown that Ki is particularly prone to taxon-specific variation at longer evolutionary distances (Hoffman and Birney 2007). A critical issue is whether Ks/Ki is an adequate measure of selection in synonymous sites. We generated Ks and Ki distributions for a set of 1,562 reliable (see Materials and Methods) alignments of intronic and coding sequences from orthologous mouse and rat genes in order to assess the suitability of Ki as the baseline for detecting selection in synonymous sites. First, we compared the statistical properties of the distributions of Ki and Ks. The distribution of Ki had almost precisely the same mean and median as the Ks distribution but was quite narrow compared to the latter: the standard deviation of Ks was more than twice greater than that of Ki (fig. 1 and supplementary table 1S). Furthermore, the skewness of the distribution was also much greater for the Ks distribution than for the Ki distribution (fig. 1 and supplementary table 1S). We also compared the nucleotide compositions of introns and synonymous sites and found substantial differences between these two (supplementary table 2S). These observations showed that Ks and Ki distributions had distinct statistical properties and suggested that introns and synonymous positions in exons are subject to different evolutionary forces.

Previous studies that compared Ks and Ki do not seem to arrive to a consensus. Some reports have claimed that synonymous substitution rates are approximately equal to those in introns despite differences in the patterns of substitution (Hughes and Yeager 1997; Chamary and Hurst

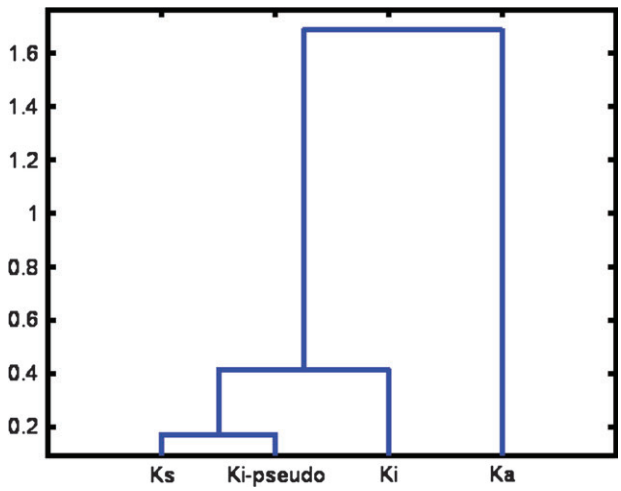

FIG. 2.—A dendrogram based on the pairwise distances between the distributions of Ka, Ks, Ki, and Ki-pseudo. The distances between the distributions are in bits.

Table 1
The Number of Genes with Significant Positive and Negative Selection in Synonymous Sites

Method	Positive Selection	*P*	Negative Selection	*P*
1) PBL, all sites	188	4×10^{-26}	517	$<10^{-50}$
2) PBL, CG, TG, CA removed	175	6×10^{-21}	279	$<10^{-50}$
3) PBL, CX, XG (X = A,T,G,C) removed	185	5×10^{-25}	218	6×10^{-40}
4) 4F, all sites	189	10^{-27}	438	$<10^{-50}$
5) 4F, CG, TG, CA removed	151	2×10^{-12}	227	2×10^{-44}

NOTE.—Selection in synonymous sites was measured using the Pamilo-Bianchi-Li (PBL) method and the fraction of mismatches at 4-fold degenerate sites (4F). The probability of finding this many or more cases of apparent positive or negative selection by chance was calculated using the binomial test; the expected number of genes in the positive and the negative set each is 78 (5% of 1,562 genes). The dramatic loss of sites caused by the CX/XG masking procedure made the 4F analysis (unlike the PBL method) inapplicable for this method. Therefore, the results obtained with this stringent filtering were not used for constructing the final sets of genes with apparent negative and positive selection in synonymous sites.

2004), whereas others have suggested that intron rates of divergence are greater than those in synonymous sites (Hellmann et al. 2003), or conversely, that synonymous substitution rates exceed those found in introns (Subramanian and Kumar 2003). Our present observation that Ks and Ki are (nearly) identical on average but are very differently distributed suggests that these diverging conclusions might be attributed to different evolutionary models and data sets used in the respective studies. It has been argued that the apparent increase in synonymous substitution rates of some genes over those of introns is due to the context-dependence of mutation in synonymous sites, in particular, the high mutation rate of CpG dinucleotides (Hughes and Yeager 1997; Kondrashov, Ogurtsov, and Kondrashov 2006).

Our observations, together with those in previous studies, suggest that Ki might not be a proper null model for Ks due to different nucleotide compositions of coding and noncoding DNA and distinct statistical properties of the Ks and Ki distributions. Thus, we developed a computational procedure to account for these differences between introns and synonymous sites. Under this approach, alignments of pseudo-coding sequence (pseudo-CDS) were generated by sampling alignments of intronic sequences such as to mimic the base composition of the synonymous sites for each respective gene and thus eliminate potential artifacts caused by differences in the CpG content and other compositional differences between synonymous sites and introns. The pseudo-CDS alignments were used to calculate Ki-pseudo values (see Material and Methods and Supplementary fig. 1S for details). Using an information-theoretical measure of divergence (see Materials and Methods for details), we computed distances between the distributions and found that, unlike Ki, Ki-pseudo had statistical properties highly similar to those of Ks (fig. 2 and supplementary table 3S). In particular, the distribution of Ki-pseudo was shifted to the right compared to the Ki distribution such that the right tail of the Ki-pseudo distribution behaved more similarly to that of the Ks distribution (fig. 1). Accordingly, Ki-pseudo values were used as the baseline to detect purifying and positive selection acting on synonymous sites of mammalian genes.

Partitioning Rodent Genes into Negatively Selected, Neutral, and Positively Selected Sets Using Synonymous Sites As the Criterion

We examined positive and negative selection in synonymous sites, using the Ks/Ki-pseudo ratio as the criterion under 2 distinct estimation schemes, the Pamilo-Bianchi-Li (PBL) method (Li 1993; Pamilo and Bianchi 1993) and the fraction of mismatches at 4-fold degenerate sites (4F method). These calculations were performed either before or after masking CG, TG, and CA dinucleotides (the highly mutable CpG sites and the ''highly CpG-prone'' sites, i.e., those convertible to CpG via a single transition [Kondrashov, Ogurtsov, and Kondrashov 2006]) or, finally, after removing all CpX and XpG dinucleotides (all CpG-prone sites). Starting with a set of 1,562 reliably aligned mouse-rat orthologs (see Materials and Methods), we identified a significant excess (compared to the random expectation) of genes with both negative and positive selection in synonymous sites in all 5 tests. Masking the mutable dinucleotides did not substantially affect the results (table 1). In order to obtain conservative estimates of positively and negatively selected genes, we required agreement between the 2 evolutionary models: only genes found to be positively or negatively selected in 3 or 4 tests were included in the final sets. The results of test #3 (table 1, PBL, CpX, and XpG sites removed) were not used in this selection procedure because of the dramatic loss of sites (>50%) that was caused by the masking procedure and made the 4F method inapplicable. However, the results of the PBL test show that this masking had but a small effect on the number of genes with apparent negative and positive selection in synonymous sites (table 1).

With this approach, 185 cases of positive (diversifying) selection (positive set) and 438 cases of negative (purifying) selection (negative set) were identified. The

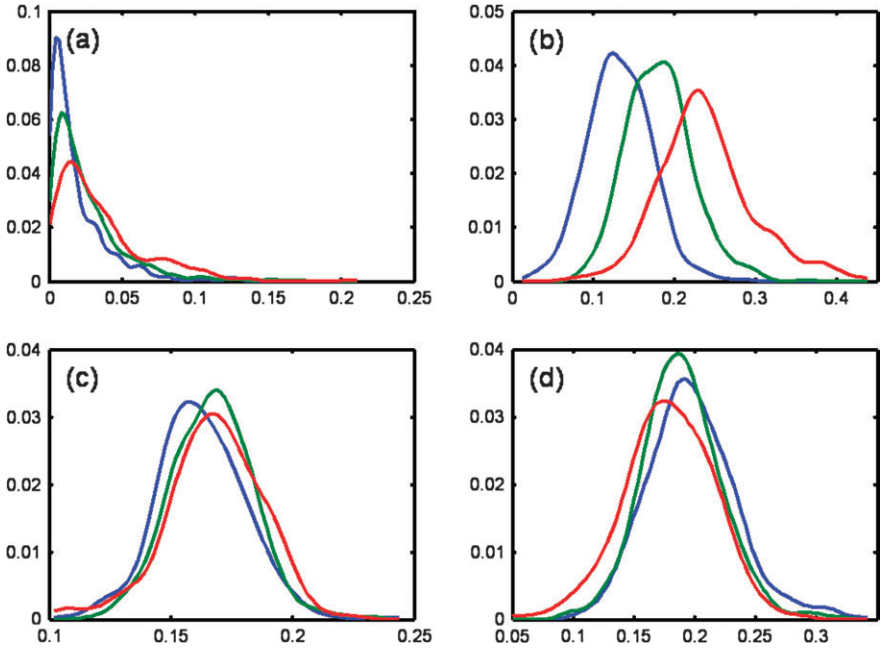

FIG. 3.—Distributions of Ka (*a*), Ks (*b*), Ki (*c*), and Ki-pseudo (*d*) for the 3 sets of rodent genes. In each panel, the blue curve corresponds to the negative set, the green curve to the neutral set, and the red curve to the positive set.

remaining 939 genes were conservatively assigned to the neutral set. Thus, ~28% of the analyzed rodent genes were found to be subject to substantive negative selection in synonymous sites, whereas roughly half as many genes (~12%) appeared to be subject to relatively strong positive selection. A comparison of the content of CpG sites involving synonymous position in the 3 gene sets did not reveal significant differences, suggesting that the observed distinct modes of evolution are not caused by effects of mutagenic contexts (supplementary table 4S).

We further compared the Ks, Ki, Ki-pseudo, and Ka distributions for the positive, negative, and neutral datasets. The distributions of Ka, Ki, and Ki-pseudo were very similar, in both shape and parameter values in all 3 sets (figs. 3*a*,*c*,*d*), although there was a statistically significant difference between the distributions of all 3 of these variables (supplementary table 5S). In a sharp contrast, the Ks distributions differed much more significantly between the 3 gene sets than the Ka, Ki-pseudo, or Ki distributions (see supplementary table 5S, and compare fig. 3*b* to figs. 3*a*,*c*,*d*). Although all 3 Ks distributions were, approximately, equally broad, the negative and positive set distributions were significantly shifted to the left and to the right, respectively, compared to the neutral set distribution (fig. 3). In particular, the mean Ks in the positive set was substantially greater than the mean Ks of the negative set (mean = 0.239 for positive; mean = 0.132 for negative set) as expected of sites evolving under positive selection.

It should be emphasized that the existence of a major difference between the Ks distributions but not Ki distributions for the negative and positive sets (compare fig. 3*b* and fig. 3*c*) all but rules out a potential alternative explanation of these results, namely, that the apparent positive selection in synonymous sites is an artifact caused by an anomalously high sequence conservation, due to purifying selection, in the intronic sequences of the respective genes. In order to further ascertain that purifying selection in introns was not a significant factor accounting for the detected positive selection in synonymous sites, we applied stringent filtering to remove potential functional elements from the intronic sequences used in the Ki and Ki-pseudo calculations. For that purpose, longer exon-flanking sequences were trimmed off the intron alignment, and short introns that could be enriched for functional elements were discarded. This procedure reduced the set of orthologous gene pairs available for the analysis to 952 but did not substantially change the fractions of genes subject to positive and negative selection in synonymous sites (table 6S). These results indicate that Ki-pseudo is, indeed, an appropriate baseline for measuring selection acting at other classes of sites in orthologous genes from closely related species.

Ks and Ka Are Correlated in the Negative and Neutral Sets but Not in the Positive Set

Does the relationship between Ks, Ki-pseudo, and Ka reveal anything about the evolutionary forces that affect the positive and negative sets? We addressed this question by checking whether any of the variables were correlated, and whether the strength of such correlations differed between the sets. We observed a moderate but statistically highly significant positive correlation between Ks and Ka for the negative set (table 2 and fig. 4), which could be expected, given that genes in this set are under strong purifying selection; similar observations have been reported previously in several independent studies (Lipman and Wilbur 1984; Wolfe and Sharp 1993; Mouchiroud, Gautier, and Bernardi 1995; Makalowski and Boguski 1998; Smith and Hurst 1999). A somewhat weaker but also highly significant correlation between Ka and Ks was seen in the

Table 2
The Correlations Between the Analyzed Variables

Negative Set

	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.27	0.29	0.13	−0.10	−0.20	0.13
Ks		1.00	0.46	0.67	−0.14	−0.15	−0.11
Ki			1.00	0.55	−0.08	−0.18	0.09
Ki-pseudo				1.00	−0.08	−0.11	−0.12
EL					1.00	0.69	0.13
EB						1.00	0.09
ΔG							1.00

Neutral Set

	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.19	0.24	0.14	−0.08	−0.21	0.09
Ks		1.00	0.53	0.67	−0.06	−0.08	0.05
Ki			1.00	0.58	0.03	−0.03	0.08
Ki-pseudo				1.00	0.01	−0.03	−0.01
EL					1.00	0.66	0.06
EB						1.00	−0.02
ΔG							1.00

Positive Set

	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.08	0.02	−0.01	−0.17	−0.25	0.25
Ks		1.00	0.48	0.72	0.15	0.18	0.17
Ki			1.00	0.54	0.10	−0.04	0.14
Ki-pseudo				1.00	0.16	0.09	0.05
EL					1.00	0.53	−0.02
EB						1.00	−0.20
ΔG							1.00

Complete Set

	Ka	Ks	Ki	Ki-pseudo	EL	EB	ΔG
Ka	1.00	0.25	0.23	0.08	−0.10	−0.22	0.14
Ks		1.00	0.46	0.44	−0.05	−0.07	0.10
Ki			1.00	0.54	0.01	−0.08	0.10
Ki-pseudo				1.00	0.01	−0.03	−0.06
EL					1.00	0.65	0.07
EB						1.00	−0.01
ΔG							1.00

NOTE.—The table shows the Pearson correlation coefficient (R) values for each pair of variables. Statistically significant ($P \leq 0.05$) values are indicated by shading.

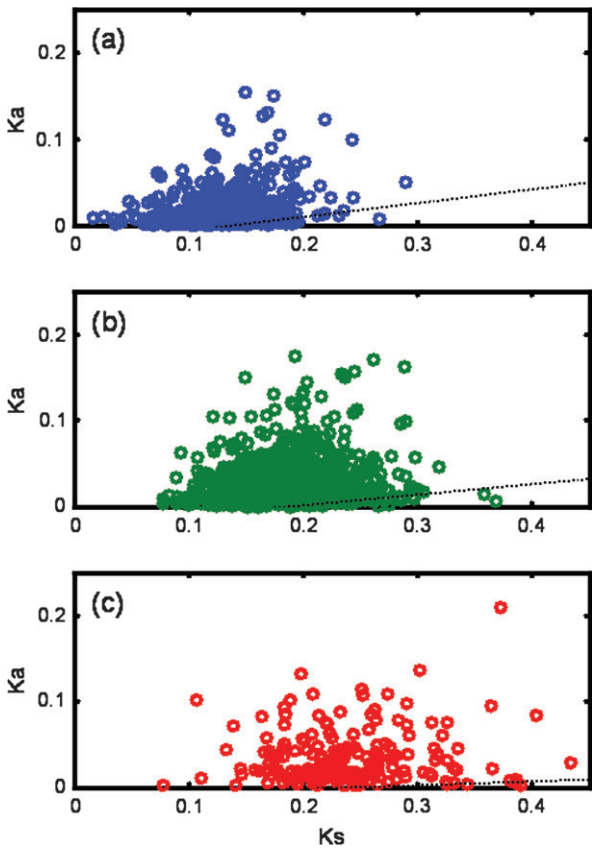

FIG. 4.—The correlations between Ka and Ks for the negative (*a*), neutral (*b*), and positive (*c*) gene sets.

neutral set (table 2 and fig. 4). The significant correlation between Ks and Ka in the negative and neutral sets implies that the evolutionary forces that exert purifying selection on synonymous sites in the negative set are linked to the evolution of protein structure and function. In particular, it seems likely that the negative selection acting on synonymous sites has to do with the high level of expression that is characteristic of genes encoding highly conserved proteins (Pal, Papp, and Hurst 2001; Krylov et al. 2003; Wolf, Carmel, and Koonin 2006). By contrast, in the positive set, Ks and Ka showed a much weaker and not significant correlation ($r = 0.08$, $P = 0.28$; table 2 and fig. 4). To control for the possible effect of the smaller sample size of the positive set, we generated 10,000 random samples of 185 genes each from the total dataset and found only 93 sampled sets with $r \leq 0.08$ ($P < 0.01$). Thus, the absence of a significant correlation between Ka and Ks in the positive gene set is not a sample size artifact. This observation suggests that, in sharp contrast with both the negative and the neutral sets, the forces that affect the evolution of synonymous sites in the positive-set genes are uncoupled from any selection acting at the level of protein structure and function.

We then examined the relationship between Ks and Ki-pseudo and observed strong positive correlations for all 3 gene sets; slightly weaker but also highly significant correlations were found for Ks and Ki (table 2). At first glance, this result suggests the possibility that evolution of introns might not be neutral, and accordingly, Ki-pseudo might not be a robust null model for measuring selection at synonymous sites. However, this does not seem to be the case, because the strength of the correlation was nearly identical among the 3 gene sets. It appears most likely that the correlations between Ks and Ki (and Ki-pseudo) reflect regional mutational biases across the genome. Such biases have been reported previously (Matassi, Sharp, and Gautier 1999), and for the rodent genes analyzed here, we observed a highly significant anticorrelation between the differential of the Ki and Ks values and the distance separating the respective genes on the chromosome: closely spaced genes, typically, had similar Ks, Ki, and Ki-pseudo values; no such effect was seen for Ka (supplementary figs. 2S and 3S).

Given that Ks is correlated with Ka in the negative and neutral sets, and with Ki in all 3 sets, we performed a partial correlation analysis in an attempt to disentangle these correlations. In the negative and neutral sets, the correlations between Ka and both Ks and Ki became smaller but remained highly significant after the removal of the effect of the other variable (supplementary table 7S). Thus, in the negative and neutral sets, the correlation between Ka and Ks appears to be valid in itself and might reflect similar selective pressures at synonymous and nonsynonymous sites. The correlation between Ka and Ki, which was, in

Table 3
Codon Bias (ENC and CAI), Gene Expression (EL and EB), and mRNA Stability (ΔG) in the 3 Sets of Analyzed Rodent Genes

		Positive		Neutral		Negative	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Codon Bias	ENC	49.618	4.245	49.890	4.081	49.199	4.400
	CAI	0.748	0.046	0.754	0.044	0.759	0.045
Gene Expression	EL	464.450	752.336	470.100	595.625	508.509	605.628
	EB	17.219	19.126	18.232	20.491	19.583	20.429
mRNA Stability	ΔG	–0.329	0.044	–0.338	0.042	–0.347	0.045

NOTE.—ΔG values are normalized for length.

Table 4
Statistics of the Comparison of the Negative, Neutral, and Positive Gene Sets with Respect to Codon Bias (ENC and CAI), Gene Expression (EL and EB) and mRNA Stability (ΔG)

	Codon Bias		Gene Expression		mRNA Stability
	ENC	CAI	EL	EB	ΔG
Positive versus Neutral*	1.000	0.259	1.000	1.000	0.0117
Positive versus Negative*	0.818	0.012	1.000	0.747	2.7×10^{-16}
Negative versus Neutral*	0.013	0.115	1.000	1.000	3.96×10^{-10}
Combined**	0.017	0.009	0.63	0.468	3.69×10^{-18}

NOTE.—Bonferroni adjusted *P* values computed using Student's *t*-test (*) and *P* values for combined data were computed using ANOVA (**).

part, independent of the Ks-Ki correlation and was greater in the negative set than in the neutral set (supplementary table 7S), is harder to explain. It cannot be ruled out that there is some pressure of purifying selection on intron sequences, the nature of which remains obscure. Should such a selective component, indeed, affect evolution of introns in the negative set, this would make our estimate of genes subject to purifying selection at synonymous sites even more conservative. By contrast, in the positive set, there was no significant correlation between Ka and either Ks or Ki, indicating that, in these genes, evolution of the protein sequence is completely decoupled from the evolution of noncoding sequences. Taken together, these results indicate that there are at least 2 distinct components in the evolution of synonymous sites, a selective one and a mutational one. The nature of the mutational component is the same across all analyzed genes. By contrast, the selective component is linked to the protein evolution in the negative set but apparently is of a different nature in the positive set.

Potential Driving Forces of Selection in Synonymous Sites: Significant Differences in Expression and mRNA Stability Between the Positive and Negative Sets

What factors contribute to positive and negative selection in synonymous sites? Perhaps even more importantly, can we identify the probable causes of the lack of correlation between Ks and Ka in the positive set? We evaluated the roles of gene function, codon bias, gene expression, and mRNA stability as potential driving forces of selection in synonymous sites. There were no significant differences in the distribution across the Gene Ontology (GO) categories between the genes of the negative, neutral, and positive sets (data not shown), hence no straightforward explanation for the observed differences in the selection regimes through the biological functions of the respective genes. We then compared the codon bias [determined either as the effective number of codons (ENC) or as the codon adaptation index (CAI)] between the 3 gene sets. Moderate but statistically significant differences in CAI were detected between the negative and positive sets, with the highest value observed for the negative set (tables 3 and 4). Thus, the pattern of codon bias exhibited by the genes in the negative set is more similar to the pattern found among highly expressed genes (the reference set) than to the pattern found within the positive set. This result is consistent with the expectation that genes that are more biased in their choice of synonymous codons tend to be more conserved. A significant difference in ENC was also observed between the negative and neutral sets (tables 3 and 4). Given that a tighter control over codon usage would be a side effect of strong purifying selection within the negative set, this result is not surprising. The nonsignificant differences between the positive set and the other 2 sets are likely to result from the fast evolution in synonymous sites of positively selected genes. A comparison of gene expression, determined either as expression level (EL) or as expression breadth (EB), revealed no significant differences between the positive, neutral, and negative sets (table 3 and 4). Some reports have suggested that purifying selection on synonymous sites affects the efficiency and accuracy of translation in certain model organisms such as *Escherichia coli* and *Drosophila melanogaster* (Ikemura 1985; Eyre-Walker 1996; Akashi and Eyre-Walker 1998); however, no strong indications of such selective pressures in mammalian genomes have been detected (Smith and Hurst 1999; Duret and Mouchiroud 2000; Iida and Akashi 2000). Furthermore, these findings were in agreement with previous reports indicating that rates of synonymous divergence are not correlated with patterns in gene expression (Lercher, Chamary, and Hurst 2004).

However, examination of the correlations between the rates of evolution in synonymous and nonsynonymous sites and characteristics of expression in the 3 gene sets produced more informative and, partly, unexpected results. In the negative set, there was a relatively low but statistically significant anticorrelation between Ks and expression (both EL and EB); these anticorrelations paralleled those between Ka and expression (table 2) and were compatible with the previous observations on slow evolution of highly and broadly expressed genes (Duret and Mouchiroud 2000; Pal, Papp, and Hurst 2001; Krylov et al. 2003; Wolf, Carmel, and Koonin 2006). The positive set showed a strikingly different pattern, with Ks being positively correlated with both EL and EB, whereas for Ka the correlation was negative and of roughly the same magnitude as in the other 2 gene sets (table 2). Thus, in a pattern that is the diametrical opposite of what is seen in the negative set (and, less pronouncedly, in the neutral set), fast evolution in synonymous sites that appear to be subject to positive selection is associated with higher and broader expression of the corresponding gene.

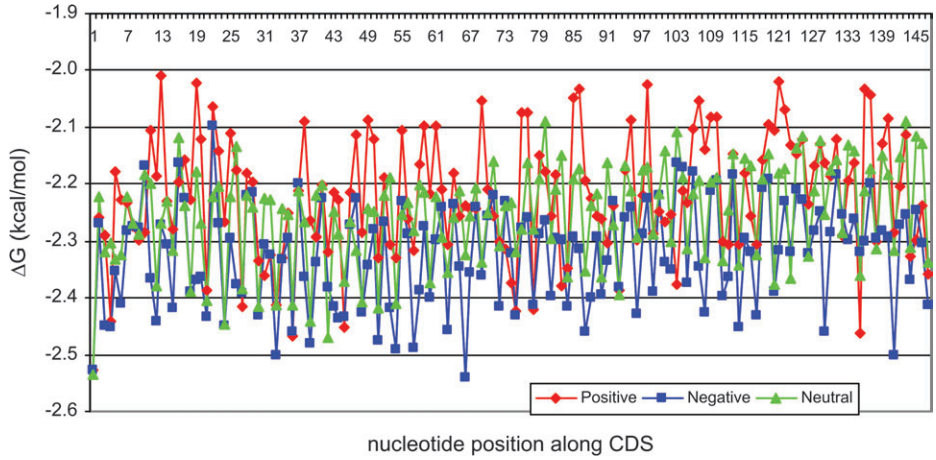

FIG. 5.—Plot of ΔG values (kcal/mol) calculated for base pairs along the 150-nucleotide stretch of coding sequence starting from the codon immediately following the start ATG codon. Values are averaged across the CDS in the negative set (blue), neutral set (green), and positive set (red).

It has been proposed that purifying selection on synonymous sites is linked to increased mRNA stability (Duan and Antezana 2003; Chamary and Hurst 2005; Shabalina, Ogurtsov, and Spiridonov 2006). Thus, we looked for differences in patterns of mRNA stability between the positive, neutral, and negative sets. Using previously published methods (Shabalina, Ogurtsov, and Spiridonov 2006), we found that the average predicted mRNA stability (kcal/mol) was significantly greater in the negative set than in the neutral or positive sets (fig. 5 and tables 5 and 6). It has been shown previously that the contributions of nucleotides to mRNA stability followed a periodic pattern dictated by the structure of the genetic code, with the third, degenerate position being the primary contributor (Shabalina, Ogurtsov, and Spiridonov 2006). This pattern was, indeed, apparent in all 3 gene sets analyzed here (fig. 5). Notably, the difference in RNA stability (estimated free energy) between the negative, neutral, and positive sets was consistent and significant over all 3 codon positions (fig. 5 and tables 5 and 6). This suggests that positive selection for mRNA destabilization affects all codon positions, although in nonsynonymous sites this relatively weak effect is overshadowed by selection acting at the protein level. No significant differences in nucleotide content within the codon positions were observed between the positive and negative sets (supplementary table 8S), indicating that the differences in mRNA stability are not artifacts caused by different base compositions.

We further assessed correlations between ΔG and Ks in the 3 gene sets and found the results to be consistent with selection acting to maintain or establish the optimal stability of the mRNA secondary structure. The correlation between ΔG and Ks was not significant in the neutral set, whereas the correlations of opposite signs were observed in the negative and the positive sets. In the negative set, there was a low but significant anticorrelation between ΔG and Ks, whereas the positive set showed a somewhat greater, positive correlation (table 2). In other words, in the negative set, the genes that evolve relatively slowly tend to possess less stable mRNA secondary structure than faster evolving genes; conversely, in the positive set the faster evolving genes are less stable than slowly evolving ones. Thus, it appears that purifying selection in the negative set prevents the formation of excessively stable secondary structure, whereas in the positive set diversifying selection might drive mRNA destabilization.

To summarize, the correlations between Ks, gene expression, and predicted mRNA stability were all negative in the negative set but positive in the positive set (fig. 6*A*). These coherent but contrasting correlation structures, certainly, do not prove a cause-and-effect relationship between mRNA stability and expression in mammalian evolution, but are compatible with the hypothesis that both purifying and positive selection in synonymous sites act at the level of mRNA secondary structure, which affects stability and, through mRNA degradation process, the expression levels measured in microarray experiments. In a sharp contrast, the correlation structures for Ka were identical for the positive and negative set (fig. 6*B*), suggesting similar patterns of (purifying) selection in nonsynonymous sites. In this case, acceleration of evolution, on average, seems to result in mRNA destabilization and lower expression.

Conclusions

We developed and applied a robust statistical test to identify purifying and positive selection acting on synonymous sites of mammalian genes using shuffled intron

Table 5
Free Energy (ΔG) of Base-Pairing for Individual Codon Positions for the 3 Gene Sets

	Positive		Neutral		Negative	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
pos_1	–2.182	0.092	–2.205	0.043	–2.275	0.066
pos_2	–2.175	0.089	–2.197	0.054	–2.276	0.063
pos_3	–2.293	0.091	–2.351	0.051	–2.411	0.06

NOTE.—ΔG values are normalized for the predicted number of base-paired nucleotides for each of the codon positions within the 150 upstream nucleotides of the CDS (AUG start codon removed).

Table 6
Statistics of the Comparison of the Base-Pairing Free Energies (ΔG) for Individual Codon Positions in the 3 Analyzed Gene Sets

	pos_1	pos_2	pos_3	Total CDS
Positive versus Neutral*	0.387	0.456	9.3×10^{-4}	0.0117
Positive versus Negative*	7.38×10^{-7}	3.6×10^{-8}	3.51×10^{-10}	2.7×10^{-16}
Negative versus Neutral*	1.19×10^{-7}	1.0×10^{-8}	2.95×10^{-6}	3.96×10^{-10}
Combined**	3.54×10^{-9}	1.45×10^{-10}	1.09×10^{-12}	3.69×10^{-18}

NOTE.—Bonferroni adjusted *P* values were computed for the same data included in table 5 using Student's *t*-test (*) and *P* values for combined data were computed using ANOVA (**).

sequences as the proxy for neutral evolution. As expected, considering many previous reports on strong, positive correlations between Ka and Ks, we observed that a substantial fraction of the analyzed genes was subject to significant purifying selection at synonymous sites (Akashi 1994; Mouchiroud, Gautier, and Bernardi 1995; Makalowski and Boguski 1998). By contrast, the finding that ~12% of the genes seemed to experience substantial positive selection at synonymous sites was surprising. A comparison of the distributions of Ks and Ki for the negative and positive gene sets (fig. 3) seems to rule out the possibility that the apparent positive selection in synonymous sites is actually due to purifying selection affecting the respective intronic sequences.

The fraction of rodent genes with apparent positive selection in synonymous codons is considerably greater than the fraction of mammalian genes that have been reported to exhibit positive selection on the level of the amino acid sequence, as reflected in Ka/Ks > 1 for the entire coding sequence. The specific numbers of mammalian genes that are subject to strong positive selection differ between studies, but a recent careful analysis of ~8,000 orthologous genes from humans and chimpanzees revealed only 35 genes with statistically significant positive selection manifest at the level of the entire protein sequence (Nielsen et al. 2005). Of the 1,562 genes analyzed here, only 2 had Ka/Ks >1 at the statistically significant level, indicating positive selection on the level of complete protein sequences (data not shown). Thus, it seems that, in comparable tests, 1 to 2 orders of magnitude more mammalian genes exhibit positive selection in synonymous than in nonsynonymous sites. This excess of positive selection in synonymous sites seems to reflect different balances of evolutionary forces that act at positions coding for amino acids and at synonymous positions. Protein sequences are almost universally subject to purifying selection of varying strength (Ka/Ks << 1 for most genes), which is likely to obscure more subtle effects of positive selection acting at some positions; indeed, site-specific positive selection detected by multiple alignment analysis appears to be common (Yang and Bielawski 2000; Zhang, Nielsen, and Yang 2005). By contrast, as shown here, readily detectable purifying selection on synonymous sites might affect only about one-quarter of the mammalian protein-coding genes such that positive selection is readily detectable against the, largely, neutral background of synonymous sites. Additionally, Ki and, especially, Ki-pseudo appear to be better neutral baselines than Ks such that positive selection at synonymous sites could be easier to detect than positive selection at nonsynonymous sites for which Ks is used as the baseline. This being said, a comparison of the distributions of Ka and Ks in rodent genes (fig. 1) indicates that Ka/Ks remains a reasonable measure of the strength of selection in proteins, given that protein sequences are subject to much more pronounced constraints than noncoding sites.

Finding the biological basis or at least a strong functional correlate of positive selection in synonymous sites turned out to be a challenge. The lack of significant correlation between Ks and Ka in the positive set indicates that positive selection in synonymous sites is decoupled from the evolution of the respective proteins. This conclusion is compatible also with the lack of any significant excess of a particular class of biological functions among the genes in the positive set. By contrast, analysis of links between selection in synonymous sites and gene expression and mRNA stability revealed nontrivial connections. Although there were no significant differences in the overall expression level or breadth between positively selected, negatively selected and neutral genes, the dependences between evolution rate and expression was remarkably different. In particular, and in contrast to negatively selected genes, among genes with positive selection in synonymous sites, those that are highly and widely expressed appear to

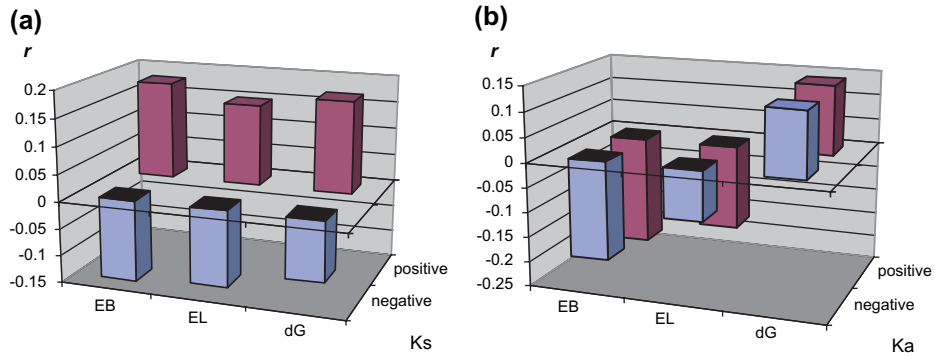

FIG. 6.—The structure of correlations between Ks (A) and Ka (B) and expression breadth (EB), expression level (EL), and predicted mRNA stability (dG) for the positive (purple) and negative (light blue) gene sets.

evolve faster (i.e., under a stronger selective pressure) than lowly expressed ones. The second clear and statistically highly significant correlation was with predicted stability of the mRNA secondary structure: the transcripts in the set of positively selected genes are predicted to be considerably less stable than those in the negative and neutral sets. The correlations between Ks, on the one hand, and expression and mRNA stability, on the other hand, were all of the same sign within the positive and negative sets but of the opposite signs between the sets (fig. 6A). This is compatible with a causal relationship between mRNA stability and expression levels measured in microarray experiments, with the link probably actualized through regulation of mRNA degradation. Therefore, although we technically cannot ascertain the direction of evolution without using a third species as an outgroup, we suspect that mRNA destabilization could be an important factor that, through its effect on mRNA stability and possibly also translation rates, drives positive selection in synonymous sites. Additionally or alternatively, it is conceivable that positive selection in synonymous positions is driven by the necessity to maintain interactions with other RNA species (Mattick and Makunin 2006).

We cannot be confident that the correlates of selection in synonymous site detected here, indeed, reflect the principal underlying selective forces. However, it is our hope that the demonstration of the wide spread of positive selection in synonymous sites in mammalian genes stimulates further theoretical and experimental studies aimed at the deeper characterization of the causes of this phenomenon.

Supplementary Material

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

Acknowledgments

We thank King Jordan, Alexei Kondrashov, Yuri Wolf, and John Wootton for helpful discussions. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS.

Literature Cited

Adams WT, Skopek TR. 1987. Statistical test for the comparison of samples from mutational spectra. J Mol Biol. 194:391–396.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics. 136:927–935.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688–693.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Avedisov SN, Rogozin IB, Koonin EV, Thomas BJ. 2001. Rapid evolution of a cyclin A inhibitor gene, roughex, in *Drosophila*. Mol Biol Evol. 18:2110–2118.

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

Chamary JV, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol Biol Evol. 21:1014–1023.

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

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 7:98–108.

Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J Mol Evol. 57:694–701.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol. 13:864–872.

Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. Nature. 397:344–347.

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.

Hoffman MM, Birney E. 2007. Estimating the neutral rate of nucleotide substitution using introns. Mol Biol Evol. 24:522–531.

Hughes AL, Yeager M. 1997. Comparative evolutionary rates of introns and exons in murine rodents. J Mol Evol. 45:125–130.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18:486.

Iida K, Akashi H. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene. 261:93–105.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. Gene. 345:119–126.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. J Theor Biol. 240:616–626.

Kouprina N, Mullokandov M, Rogozin IB, Collins NK, Solomon G, Otstot J, Risinger JI, Koonin EV, Barrett JC, Larionov V. 2004. The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. Proc Natl Acad Sci USA. 101:3077–3082.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.

Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet. 14:2221–2229.

Lercher MJ, Chamary JV, Hurst LD. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res. 14:1002–1013.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36:96–99.
Li WH. 1997. Molecular Evolution. Sunderland, MA: Sinauer.
Lin J. 1991. Divergence measures based on the Shannon entropy. IEEET Inform Theory. 37:145–151.
Lipman DJ, Wilbur WJ. 1984. Interaction of silent and replacement changes in eukaryotic coding sequences. J Mol Evol. 21:161–167.
Louie E, Ott J, Majewski J. 2003. Nucleotide frequency variation across human genes. Genome Res. 13:2594–2601.
Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. 12: 1827–1836.
Makalowski W, Boguski MS. 1998. Synonymous and non-synonymous substitution distances are correlated in mouse and rat genes. J Mol Evol. 47:119–121.
Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr Biol. 9:786–791.
Mattick JS, Makunin IV. 2006. Non-coding RNA. Hum Mol Genet 15 Spec No. 1:R17–29.
Mouchiroud D, Gautier C, Bernardi G. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J Mol Evol. 40:107–113.
Nielsen R, Bustamante C, Clark AG, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.
Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS. 2002. OWEN: aligning long collinear regions of genomes. Bioinformatics. 18:1703–1704.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927–931.
Pamilo P, Bianchi NO. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol. 10:271–281.
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.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res. 34:2428–2437.
Sharp PM, Li WH. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Smith NG, Hurst LD. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics. 153:1395–1402.
Su AI, Wiltshire T, Batalov S, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA. 101:6062–6067.
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.
Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. Hum Mol Genet 13 Spec No. 2:R245–254.
Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol. 23:1383–1390.
Waterston R, Lindblad-Toh HK, Birney E, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.
Wheeler DL, Barrett T, Benson DA, et al. 2006. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 34:D173–180.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc Biol Sci. 273: 1507–1515.
Wolfe KH, Sharp PM. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol. 37:441–456.
Wright F. 1990. The 'effective number of codons' used in a gene. Gene. 87:23–29.
Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci USA. 102:13526–13531.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution. 15:496–503.
Yeo G, Hoon S, Venkatesh B, Burge CB. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci USA. 101:15700–15705.
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.

William Martin, Associate Editor

Accepted May 17, 2007