A Simple Method for Estimating the Intensity of Purifying Selection in Protein-Coding Genes

Ron Ophir, Takeshi Itoh,† Dan Graur,* and Takashi Gojobori†*

*Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel; and †Center for Information Biology, National Institute of Genetics, Mishima, Japan

We propose a method by which the intensity of purifying selection on a functional protein-coding gene is estimated by using three aligned homologous sequences: a processed pseudogene (ψ), a functional paralog from the same species (g), and a functional ortholog from a different species (o). For each such trio, we calculate the numbers of nucleotide substitutions along the branches leading to ψ and g, i.e., K_ψ and K_g. If we assume that the mutation rates are the same in the genes and the pseudogenes and that mutations occurring in a pseudogene do not affect the fitness of the organism, we can show that the fraction of mutations that are selectively neutral, f_g, is equal to the ratio K_g/K_ψ. Since advantageous mutations occur only very rarely, such that they do not contribute significantly to the rate of molecular evolution, the fraction of deleterious mutations that are subject to purifying selection is $1 - f_g$. Therefore, the K_g/K_ψ ratio can be used directly to estimate the intensity of purifying selection, thereby isolating its effects on the rate of evolution from those of mutation. We compared the selection intensities of 12 orthologous protein-coding pairs from humans and murids. As expected, the fraction of mutations that are subject to purifying selection is strongest in the second codon position and weakest in the third. Interestingly, the mean fractions of effectively neutral mutations in the third codon position were only 41% and 42% for murids and humans, respectively, indicating that many synonymous mutations are subject to selective constraint. In several orthologous genes, we found that the intensity of purifying selection is very different between murid and human orthologous genes. There was no statistically significant difference in overall intensity of purifying selection between humans and murids. Thus, purifying selection does not seem to be an important factor contributing to the observed differences in the rates of evolution between these two taxa.

Introduction

Functional or selective constraint is defined as the degree of intolerance characteristic of a site or a locus toward nucleotide substitution, i.e., the proportion of deleterious mutations out of all mutations occurring at a site (Li and Graur 1991). In protein-coding genes, the functional constraint is usually estimated by the ratio of nonsynonymous (K_n) to synonymous (K_s) substitutions. For this method, it is assumed that synonymous mutations are always neutral, so that the rate of synonymous substitution reflects the rate of mutation, whereas the rate of nonsynonymous substitution is assumed to be determined by the joint effects of mutation (the rate of which is assumed to be equal to that at synonymous sites) and purifying selection against deleterious mutations. Since advantageous mutations are known to occur only very rarely (Endo, Ikeo, and Gojobori 1996), such that they do not contribute significantly to the rate of molecular evolution, the fraction of deleterious mutations that are subject to purifying selection (i.e., the functional constraint) is calculated as 1 minus the fraction of neutral mutations. The K_n/K_s ratio, however, may yield erroneous estimates if synonymous mutations are not always strictly neutral, i.e., if some synonymous mutations are selected against. In this case, K_s (the denominator) underestimates the number of effectively neutral mutations, and the functional constraint is consequently overestimated.

Key words: protein-coding genes, pseudogenes, purifying selection.

Address for correspondence and reprints: Dan Graur, Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel. E-mail: graur@post.tau.ac.il.

Mol. Biol. Evol. 16(1):49–53. 1999

In this study, we propose a simple analytical method to estimate the functional constraints of functional genes by comparing their rates of evolution with those of their conspecific paralogous pseudogenes. Specifically, we estimate the fraction of mutations that are selectively neutral in protein-coding regions from which processed pseudogenes have been derived and, by extension, the fraction of mutations that are deleterious. Such estimates can then be used to compute mean selection intensities, thus allowing us to compare the intensity of purifying selection among different evolutionary lineages. By using this method, we estimated the mean fractions of mutations that are selectively neutral for 12 orthologous protein-coding genes from humans and murids (rats and mice).

The Method

In order to estimate the intensity of purifying selection on a functional protein-coding gene, we need three aligned homologous sequences: a processed pseudogene (ψ), a functional paralog from the same species (g), and a functional ortholog from a different species (o). We use processed pseudogenes because they are mostly ''dead on arrival'' sequences; i.e., from the moment they were incorporated within the genome, they were devoid of function. The model tree for analyzing each data set is shown in figure 1. With three such sequences, it is possible to estimate the numbers of nucleotide substitutions along the branches leading to the pseudogene (K_ψ) and the functional gene (K_g) as

$$K_\psi = \frac{K_{\psi g} + K_{\psi o} - K_{go}}{2} \quad (1)$$

$$K_g = \frac{K_{\psi g} + K_{go} - K_{\psi o}}{2} \quad (2)$$

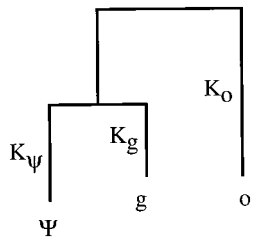

FIG. 1.—Model phylogenetic tree for a pseudogene (ψ), a functional paralog (*g*), and an outgroup ortholog of the functional gene (*o*). K_ψ, K_g, and K_o are the branch lengths.

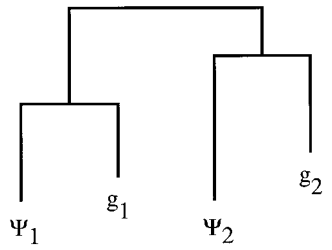

FIG. 2.—Model phylogenetic tree for comparing the intensity of purifying selection between two taxa (subscripts 1 and 2). ψ, = pseudogene; *g*, = functional paralog.

(Sarich and Wilson 1973), where the values of *K* are the numbers of substitutions per site between the sequences in the subscripts.

Following Kimura's (1983) model, the numbers of nucleotide substitutions along the two branches leading to ψ and *g* are given by

$$K_\psi = v_\psi f_\psi t \quad (3)$$

and

$$K_g = v_g f_g t, \quad (4)$$

where *v* is the total mutation rate per unit time, *f* is the fraction of mutations that are selectively neutral or nearly so, *t* is the time of divergence, and subscripts identify the branches. We note that advantageous mutations are assumed to occur only very rarely, such that they do not contribute significantly to the rate of molecular evolution.

We assume that the mutation rate is the same in the gene and the pseudogene, i.e., $v_g = v_\psi$. This assumption is reasonable, since mutations should occur at comparable rates regardless of whether they occur in a functional or a functionless sequence. We further assume that mutations occurring in a pseudogene do not affect the fitness of the organism, i.e., $f_\psi = 1$. The reason behind this assumption is that mutations occurring in a processed pseudogene, which is a ''dead on arrival'' stretch of DNA, are most probably not subject to selection.

Under these two assumptions, we obtain

$$\frac{K_g}{K_\psi} = f_g. \quad (5)$$

By definition, the fraction of deleterious mutations that are subject to purifying selection is $1 - f_g$. We can, therefore, use the K_g/K_ψ ratio to estimate the intensity of purifying selection, thereby isolating its effects on the rate of evolution from those of mutation. We note that a rigorous definition of f_g should include not only strictly neutral mutations, but also slightly deletrious mutations that have been fixed in the populations due to their finite sizes. Therefore, f_g is the fraction of effectively neutral mutations out of all mutations arising at a locus.

Following the suggestion of Wu and Li (1985), we may calculate the variances of K_ψ and K_g as

$$\mathrm{Var}(K_\psi) = \frac{1}{2}[\mathrm{Var}(K_{\psi g}) + \mathrm{Var}(K_{\psi o}) - \mathrm{Var}(K_{go})] \quad (6)$$

$$\mathrm{Var}(K_g) = \frac{1}{2}[\mathrm{Var}(K_{\psi g}) + \mathrm{Var}(K_{go}) - \mathrm{Var}(K_{\psi o})] \quad (7)$$

Therefore, the variance of f_g is given by

$$V(f_g) = \left(\frac{\mathrm{Exp}(K_g)}{\mathrm{Exp}(K_\psi)}\right)^2 \left(\frac{\mathrm{Var}(K_g)}{\mathrm{Exp}(K_g)^2} + \frac{\mathrm{Var}(K_\psi)}{\mathrm{Exp}(K_\psi)^2}\right). \quad (8)$$

Since the variance of f_g differs among genes, the standard error of the mean f_g ($\bar{f}_g$) along a branch was computed by

$$\mathrm{SE}(\bar{f}_g) = \sqrt{\frac{\sum_{i=1}^{n} \mathrm{Var}(f_{g,i})}{n^2}}, \quad (9)$$

where *n* is the number of genes and $f_{g,i}$ is the value of f_g for gene *i*.

For comparative purposes, we need four aligned homologous sequences: a processed pseudogene from species 1 (ψ_1) and its functional conspecific paralog (g_1), and a processed pseudogene from species 2 (ψ_2) and its functional conspecific paralog (g_2). For estimating K_ψ and K_g in species 1, we may use g_2 as outgroup. Similarly, we may use g_1 as outgroup for estimating K_ψ and K_g in species 2 (fig. 2).

It is well known that the fraction of neutral mutations varies greatly from gene to gene due to different functional constraints. Therefore, if we wish to determine whether different evolutionary lineages differ from one another in a systematic manner as far as the intensity of purifying selection is concerned (e.g., for testing the slightly deleterious model of molecular evolution, which predicts that mean selection intensities are stronger in species with large population sizes than in species with small population sizes), it is advisable to use the same genes in both lineages.

Data

All the processed pseudogenes from humans and murids were collected from the EMBL data bank. Using the FASTA program (Pearson 1990), we compiled a database comprising sets of four aligned sequences. Each quartet consisted of a human retropseudogene, a human functional paralog, a murid functional ortholog of the human functional gene, and its paralogous retropseu-

Table 1
Fractions of Acceptable Mutations (f_g) and Nonsynonymous to Synonymous Ratios (K_n/K_s) for 12 Orthologus Protein-Coding Genes in Murids and Humans

Gene	Murid ψ Accession No.	Human ψ Accession No.	Murid Gene Accession No.	Human Gene Accession No.	Murid f_g	Human f_g	K_n/K_s
Na/K-ATPase β-subunit	AF005897	AF005898	U59761	AF005896	0.175 ± 0.002	0.076 ± 0.003	0.431 ± 0.001
FAU 1	L33741	U02523	X65922	X65921	0.037 ± 0.002	0.266 ± 0.007	0.031 ± 0.000
Ferritin heavy chain	M73677	J04755	M24509	X03487	0.367 ± 0.027	0.075 ± 0.003	0.081 ± 0.000
UDP-glucuronosyltran	D38068	U39551	M34007	U89507	0.959 ± 0.005	0.675 ± 0.005	0.326 ± 0.001
Glutathione *S*-transferase	M14364	X65032	X02904	U21689	0.497 ± 0.006	0.287 ± 0.005	0.161 ± 0.000
Cytochrome *c*	K03239	D00266	K00750	M22877	0.177 ± 0.004	0.089 ± 0.006	0.123 ± 0.001
Glutathione peroxidase[a]	X91864	M93083	X03920	X13710	0.263 ± 0.002	0.272 ± 0.004	0.216 ± 0.001
Fatty acid-binding protein	U02885	U72237	U02883	Y10255	0	0.270 ± 0.004	0.133 ± 0.001
Histone H2b	L26163	K01888	U62673	X57985	0.192 ± 0.003	0.186 ± 0.004	0.037 ± 0.000
Phosphoglycerate kinase	M23961	K03201	M15668	V00572	0.053 ± 0.000	0.145 ± 0.001	0.038 ± 0.000
Ribosomal protein L7	M20743	Z68274	M14689	X52138	0.107 ± 0.002	0.152 ± 0.001	0.009 ± 0.000
γ-actin	M10142	M55082	M21495	X04098	0.491 ± 0.007	0.223 ± 0.001	0.004 ± 0.000
Mean ± SE					0.276 ± 0.038	0.225 ± 0.025	0.132 ± 0.003

[a] In the glutathione peroxidase gene, there is an opal stop codon at amino acid position 47 that encodes selenocysteine (Chambers et al. 1986; Mullenbach et al. 1987). This codon was excluded from the calculations.

dogene. To ensure independence of the data, in cases of multiple pseudogenes derived from the same functional gene, we only used one pseudogene for each gene. The databank was found to contain 12 suitable sets.

Computational Steps

Alignment was performed with the CLUSTAL W program (Thompson, Higgins, and Gibson 1994). Pairwise distances between sequences were calculated according to Kimura's (1980) two-parameter model. Gaps, as well as truncated regions in the pseudogenes, were excluded from the analysis. Numbers of synonymous and nonsynoymous substitutions, as well as associated standard errors, were computed according to the method of Comeron (1995). The ratio of nonsynonymous to synonymous substitution was calculated for the comparison between the human and murid functional genes. f_g values were computed separately for the three codon positions, as well as for all the positions together. Due to the finite length of the aligned sequences and the stochastic nature of the mutational process, f_g values may sometimes come out as slightly smaller than 0 or larger than 1. In such cases, we arbitrarily set them to 0 or 1, respectively. Mean f_g values for murids and humans were calculated as Σ(murid K_g + human K_g)/Σ(murid K_ψ + human K_ι).

Results

Pseudogenes evolve on average 3–4 times faster than their functional paralogs. f_g values for the 12 orthologous gene pairs are shown in table 1. The mean fraction of neutral mutations was 0.276 ± 0.038 and 0.225 ± 0.025 for murid and human orthologous genes, respectively. There was no statistically significant difference in mean f_g values between murids and humans (Wilcoxon's matched-pair test; $P = 0.388$). For some genes (e.g., gluthatione peroxidase, histone H2b), the intensity of purifying selection was similar in murids and humans. In most cases, however, selection intensity was found to vary widely between the two taxa. The most extreme example is the ferritin heavy chain, for which the intensity of purifying selection in humans is five times higher than that in murids. On average, the ratio K_n/K_s is about half of f_g. As expected, by using K_n/K_s, the intensity of selection is consistently overestimated. Notwithstanding, the values of K_n/K_s are positively correlated with the mean f_g values for murids and humans at the first and second codon positions (Spearman's rank-order test; $P = 0.030$ and $P = 0.028$, respectively).

A partition of f_g values by codon position is presented in table 2. As expected, the least constrained codon position is the third, followed by the first and second positions. On average, only about 41% and 42% of all the mutations occurring in the third codon position of protein-coding genes are acceptable in murids and humans, respectively. Given that about 70% of all possible mutations in this position are synonymous and assuming that some nonsynonymous mutations may be acceptable, these results indicate that a significant fraction of synonymous mutations are selected against. There is no statistically significant difference in the intensity of selection at the third codon position between murids and humans ($P = 0.875$). About 23% and 21% of all mutations in the first codon position are acceptable in murids and humans, respectively. Again, we found no statistically significant difference between the two taxa ($P = 0.959$). In the second codon position, in which all possible mutations are nonsynonymous, we found that in murids, approximately 11% of all mutations are acceptable, whereas the comparable value for humans is 7%. Although the difference is not statistically significant ($P = 0.091$), most probably because of the small sample size, there are tentative indications that nonsynonymous sites in human protein-coding genes may be subject to stronger selective constraint than are those in murid orthologs. A qualitative illustration of this difference can be seen from the fact that in 10 of the 12 human genes, all mutations occurring in the second po-

Table 2
Fractions of Acceptable Mutations (f_g) Partitioned by Codon Position for 12 Orthologus Protein-Coding Genes in Murids and Humans

Gene	Murid f_g First	Murid f_g Second	Murid f_g Third	Human f_g First	Human f_g Second	Human f_g Third
Na/K-ATPase β-subunit	0.282 ± 0.010	0.075 ± 0.006	0.225 ± 0.009	0	0	0.319 ± 0.009
FAU 1 gene	0	0.021 ± 0.002	0.132 ± 0.005	0.326 ± 0.030	0	0.266 ± 0.014
Ferritin heavy chain	1.000 ± 0.155	0	0	0	0	0.079 ± 0.007
UDP-glucuronosyltran	0.869 ± 0.129	0.817 ± 0.011	1.000 ± 0.023	0.736 ± 0.013	0.545 ± 0.012	0.757 ± 0.075
Glutathione *S*-transferase	0.324 ± 0.254	0.253 ± 0.014	0.813 ± 0.018	0.303 ± 0.265	0.303 ± 0.018	0.214 ± 0.011
Cytochrome *c*	0	0	0.374 ± 0.019	0	0	0.255 ± 0.023
Glutathione peroxidase	0.169 ± 0.016	0.098 ± 0.003	0.403 ± 0.006	0.396 ± 0.002	0	0.415 ± 0.012
Fatty acid-binding protein	0	0	0	0.687 ± 0.038	0	0.365 ± 0.011
Histone H2b	0.060 ± 0.029	0.041 ± 0.003	0.473 ± 0.012	0	0	0.318 ± 0.015
Phosphoglycerate kinase	0.002 ± 0.002	0.006 ± 0.000	0.121 ± 0.001	0.032 ± 0.004	0	0.378 ± 0.006
Ribosomal protein L7	0	0	0.389 ± 0.011	0.028 ± 0.002	0	0.796 ± 0.011
γ-actin	0	0	1.000 ± 0.050	0	0	0.834 ± 0.008
Mean ± SE	0.226 ± 0.089	0.109 ± 0.030	0.410 ± 0.181	0.209 ± 0.045	0.071 ± 0.027	0.416 ± 0.112

sition are deleterious, whereas the same phenomenon is seen in murids for only 5 genes.

Discussion

Our estimates of the intensity of purifying selection are consistently higher than those obtained with the K_n/K_s ratios. This is understandable, since synonymous mutations seem to be subject to some degree of purifying selection. Therefore, the rate of synonymous substitution must be lower than the rate of synonymous mutation. Notwithstanding, K_n/K_s ratios remain a very useful tool in evolutionary studies, particularly for identifying instances of suspected positive selection (e.g., Endo, Ikeo, and Gojobori 1996; Messier and Stewart 1997).

A possible weakness of our method may be in dealing with instances in which the rate of mutation differs between the functional gene and the derived retropseudogene. This may occur, for instance, if the restropseudogene is located in a genomic region that differs in GC content from the one in which the functional gene is located. We have preliminary data indicating that retropseudogenes tend to incorporate into genomic regions of GC content similar to that of the ones in which the paralogous functional gene is located (unpublished data).

Rates of gene evolution are higher in rodents than in humans (Laird, McConaughy, and McCarthy 1969; Kohne 1970; Wu and Li 1985; Gu and Li 1992; but see Easteal, Collet, and Betty 1995). This phenomenon is usually explained by mutational effects due to differences in generation times, efficiencies of DNA repair, GC contents, or metabolic rates (Li, Wu, and Luo 1984, 1985; Wu and Li 1985; Britten 1986; Filipski 1988; Martin and Palumbi 1993). In principle, however, the rate of evolution in protein-coding genes may also be determined by postmutational factors affecting the probability of fixation of newly emergent mutations. That is, the rate of gene evolution may vary among lineages because of differences in selection intensities. In this study, we show that the overall mean intensity of purifying selection against mutations in humans is not different from that in rodents. Therefore, the rate of evolution is most probably determined by a disparity in the rates of mutation between humans and murids. Selection does not seem to play an important role in determining the difference in the rates of evolution between humans and murids.

Acknowledgments

We thank Bernd Blasius, Inna Gur-El, Amit Huppert, Giddy Landan, and Lewi Stone for their help. D.G. was supported by a short-term fellowship from the Japan Society for the Promotion of Science and a grant from the Da'at Consortium for Developing Generic Technologies (Israel Ministry of Industry and Trade and Silicon Graphics Biomedical, Ltd.).

LITERATURE CITED

BRITTEN, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science **231**:1393–1398.

CHAMBERS, I., J. FRAMPTON, P. GOLDFARB, N. AFFARA, W. MCBAIN, and P. R. HARRISON. 1986. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the ''termination'' codon, TGA. EMBO J. **5**:1221–1227.

COMERON, J. M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41**:1152–1159.

EASTEAL, S., C. COLLET, and D. BETTY. 1995. The mammalian molecular clock. Landes, Austin.

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.

FILIPSKI, J. 1988. Sequence-based phylogeny in eukaryotic genomes. Nature **334**:571–572.

GU, X., and W.-H. LI. 1992. Higher rates of amino acid substitution in rodents than in humans. Mol. Phylogenet. Evol. **1**:211–214.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotides sequences. J. Mol. Evol. **16**:111–120.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KOHNE, D. E. 1970. Evolution of higher-organism DNA. Q. Rev. Biophys. **3**:327–375.

LAIRD, C. D., B. L. MCCONAUGHY, and B. J. MCCARTHY. 1969. Rate of fixation of nucleotide substitutions in evolution. Nature **224**:149–154.

LI, W.-H., and D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58–71.

———. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MARTIN, A. P., and S. R. PALUMBI. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proc. Natl. Acad. Sci. USA **90**:4087–4091.

MESSIER, W., and C. B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

MULLENBACH, G. T., A. TABRIZI, B. D. IRVINE, G. I. BELL, and R. A. HALLEWELL. 1987. Sequence of a cDNA coding for human glutathione peroxidase confirms TGA encodes active site selenocysteine. Nucleic Acids Res. **15**:5484.

PEARSON, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183**:63–98.

SARICH, V. M., and A. C. WILSON. 1973. Generation time and genomic evolution in primates. Science **179**:1144–1147.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4690.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

MARCY K. UYENOYAMA, reviewing editor

Accepted September 3, 1998