Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate

Laurent Duret and Dominique Mouchiroud

Laboratoire de Biométrie, Génétique et Biologie des Populations, Unité Mixte de Recherche Centre National de la Recherche Scientifique 5558, Université Claude Bernard, Villeurbanne, France

To determine whether gene expression patterns affect mutation rates and/or selection intensity in mammalian genes, we studied the relationships between substitution rates and tissue distribution of gene expression. For this purpose, we analyzed 2,400 human/rodent and 834 mouse/rat orthologous genes, and we measured (using expressed sequence tag data) their expression patterns in 19 tissues from three development states. We show that substitution rates at nonsynonymous sites are strongly negatively correlated with tissue distribution breadth: almost threefold lower in ubiquitous than in tissue-specific genes. Nonsynonymous substitution rates also vary considerably according to the tissues: the average rate is twofold lower in brain-, muscle-, retina- and neuron-specific genes than in lymphocyte-, lung-, and liver-specific genes. Interestingly, 5′ and 3′ untranslated regions (UTRs) show exactly the same trend. These results demonstrate that the expression pattern is an essential factor in determining the selective pressure on functional sites in both coding and noncoding regions. Conversely, silent substitution rates do not vary with expression pattern, even in ubiquitously expressed genes. This latter result thus suggests that synonymous codon usage is not constrained by selection in mammals. Furthermore, this result also indicates that there is no reduction of mutation rates in genes expressed in the germ line, contrary to what had been hypothesized based on the fact that transcribed DNA is more efficiently repaired than nontranscribed DNA.

Introduction

The process of base substitution during gene evolution can be split into two fundamental steps. First, there is a mutation, i.e., an alteration in DNA that has not been corrected by the repair systems. Second, there are selective forces and random genetic drift effects that will determine whether the new allele will become fixed in the population. The mutation rate reflects both the sensibility to mutagens, the fidelity of DNA polymerases, and the efficiency of DNA repair systems, whereas the rate of fixation of new mutations depends on their impact on fitness and on the effective population size. For selectively neutral mutations, the rate of substitutions is equal to the rate of mutation (Kimura 1983). Thus, measuring the substitution rates at sites that are not constrained by selection directly provides information on the mutation pattern. Conversely, deviation from neutral mutation rates is a good indicator of selective pressure.

The first studies on protein evolution (Dickerson 1971) revealed that the rate of amino acid substitution varies considerably among proteins (Li and Graur 1991; Bernardi, Mouchiroud, and Gautier 1993; Wolfe and Sharp 1993). This variation is thought to reflect mainly differences in functional constraints, i.e., in the proportion of the sequence that is critical to the function of the protein. Recently, analyses of a few vertebrate gene families have shown (1) that the degree of sequence conservation varies according to the tissue in which proteins are expressed (Kuma, Iwabe, and Miyata 1995; Hughes 1997) and (2) that broadly expressed proteins tend to be more conserved than tissue-specific ones (Hastings 1996). Both observations were interpreted as resulting from stronger functional constraints on proteins expressed in more diverse cellular environments.

In mammals, the rate of synonymous substitution also varies significantly among genes (Bernardi, Mouchiroud, and Gautier 1993; Wolfe and Sharp 1993; Mouchiroud, Gautier, and Bernardi 1995). It is, however, not yet clear whether this variation reflects variability in mutation rates along genomes or differences in selective pressure on silent sites. Many authors consider that silent sites are neutral because average substitution rates at synonymous sites are very close to substitution rates in pseudogenes or in the genome as a whole (Li and Graur 1991; Wolfe and Sharp 1993). However, there is evidence for selection on codon usage in mouse histone genes (Debry and Marzluff 1994), and comparisons of synonymous and nonsynonymous substitution rates suggest that silent positions may be to some extent under selective constraints (Mouchiroud, Gautier, and Bernardi 1995; Ohta and Ina 1995; Alvarez-Valin, Jabbari, and Bernardi 1998). Selection on synonymous codon usage has been demonstrated in many species, not only in bacteria but also in eukaryotes (including some invertebrates and plants; for a review, see Sharp et al. 1995). In all cases, the intensity of selection is positively correlated with gene expression level (Gouy and Gautier 1982; Sharp and Li 1986; Duret and Mouchiroud 1999). Thus, if such selection operates in mammals, one should also expect a correlation between synonymous substitution rate and gene expression level. It has been also proposed that the mutation rate might vary with gene expression pattern (Sullivan 1995). Indeed, it has been shown that nucleotide excision repair, one of the major

Abbreviations: aa, amino acid; CDS, protein-coding sequence; EST, expressed sequence tag; K_a, number of nonsynonymous substitutions per site; K_s, number of synonymous substitutions per site.

Key words: mammals, gene expression, substitution rate, noncoding regions, codon usage, DNA repair.

Address for correspondence and reprints: Laurent Duret, Laboratoire de Biométrie, Génétique et Biologie des Populations, Unité Mixte de Recherche Centre National de la Recherche Scientifique 5558, Université Claude Bernard, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France. E-mail: duret@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 17(1):68–74. 2000

Table 1
Distribution of Human Genes According to the Number of Tissues in Which they are Found to be Expressed (Based on EST Sequence Data)

No. of Genes	No. of Tissues
371 (16%)	0
824 (34%)	1–3
496 (21%)	4–6
280 (12%)	7–9
202 (8%)	10–12
126 (5%)	13–15
101 (4%)	16–19

DNA repair systems, is more efficient in transcribed DNA than in nontranscribed DNA (reviewed in Sullivan 1995). Thus, genes expressed in the germ line should be more efficiently repaired and hence evolve more slowly than others (Sullivan 1995).

In this paper, we studied the relationships between substitution rates and tissue distribution of gene expression to try to determine whether gene expression patterns affect mutation rates and/or selection intensity at different sites: synonymous and nonsynonymous codon positions and noncoding regions. We analyzed a large data set of 2,400 human/rodent orthologous genes and 834 pairs of mouse/rat orthologs. The tissue distribution of human genes was estimated by comparing their protein-coding sequences (CDSs) to a database of expressed sequence tags (ESTs) representing 19 tissues from three development states. These 19 tissues are expected to be representative of the whole organism. Hereafter, genes that are expressed in at least 16 tissues will be considered ubiquitous, whereas those that are detected in 0–3 tissues will be considered tissue-specific. Ubiquitous and tissue-specific genes make up, respectively, 4% and 50% of the data set (table 1). Our analysis provided no evidence for variation of the mutation rate according to gene expression pattern and no evidence for selection on synonymous sites but revealed a remarkable relationship between selective pressure on functional sites (in both coding and noncoding DNA) and tissue distribution of gene expression.

Materials and Methods

Sequence Data

Homologous protein-coding sequences (CDSs) common to humans (*Homo sapiens*) and rats (*Rattus norvegicus*) and/or mice (*Mus musculus*) were selected from the HOVERGEN database (Duret, Mouchiroud, and Gouy 1994) (release 27, November 1997) using the ACNUC retrieval system (Gouy et al. 1985). HOVERGEN phylogenetic trees were manually inspected to select orthologs and exclude paralogs. Protein sequences were aligned with CLUSTAL W (Thompson, Higgins, and Gibson 1994). CDSs were aligned using the protein alignment as a template. After alignment, CDSs of less than 150 homologous synonymous sites were excluded to reduce the influence of stochastic variations in synonymous rates in small sequences. The numbers of substitutions per site at synonymous sites (K_s) and at nonsynonymous sites (K_a) were calculated using Li's (1993) method. Numbers of substitutions per site in untranslated regions (UTRs) of human/rodent orthologous genes were taken from Makalowski and Boguski (1998) (http://www.ncbi.nlm.nih.gov/Makalowski/PNAS/index.html). We retained only values computed with UTRs longer than 150 nt (240 and 854 genes, respectively for 5′ UTRs and 3′ UTRs).

Expression Profiles

We selected from GenBank (release 110, December 1998; Benson et al. 1998) 679,286 human ESTs from 19 tissues: placenta, liver (fetal, adult), fetal heart, lung (fetal, adult), brain (fetal, infant, adult), breast, colon, testis, retina, uterus, lymphocyte, muscle, prostate, pancreas, and neuron. cDNA libraries from cell culture, tumors, pooled organs, or unidentified tissues were excluded. To limit stochastic variations in expression measures, we retained only cDNA libraries that had been sampled with at least 10,000 ESTs. Expression profiles of human CDSs were determined by counting the numbers of tissues in which they were represented by at least one EST. CDSs were first filtered with the XBLAST program (Claverie and States 1993) to mask repetitive elements (Alu, L1, MIR, microsatellites, etc.). CDSs were then compared with the EST data set using BLASTN2 (Altschul et al. 1997). BLASTN2 alignments showing at least 95% identity over 100 nt or more were counted as sequence matches. This criterion was chosen to be low enough to allow the detection of most ESTs despite sequencing error (the average sequence accuracy of ESTs is about 97%) (Hillier et al. 1996) but stringent enough to distinguish—in most cases—different members of highly conserved gene families (e.g., for β- and γ-actins, proteins are 98% identical and CDSs are 91% identical; for cardiac and skeletal α-actins, proteins are 99% identical and CDSs are 85% identical; for histones H3.3A and H3.3B, proteins are 100% identical and CDSs are 79% identical). The list of selected genes and their expression patterns is available at http://pbil.univ-lyon1.fr/datasets/Duret_Mouchiroud_1999/data .html.

Results

We measured the number of substitutions per site at synonymous sites (K_s) and at nonsynonymous sites (K_a) in 2,400 human/rodent and 834 mouse/rat orthologs. All the comparisons of K_a and K_s values that will be discussed below have been made between orthologous genes resulting from a same speciation event and thus having the same divergence date. Thus, variations in K_a or K_s directly reflect variations in substitution rates (number of substitutions per site per year). For the sake of simplicity, K_s and K_a will hereafter be directly taken as measures of substitution rate.

Substitution Rate in Coding Regions and Tissue-Distribution Breadth

Analysis of K_a values in human/rodent orthologs according to gene expression patterns revealed a sharp negative correlation between K_a and tissue distribution breadth (fig. 1). On average, tissue-specific proteins

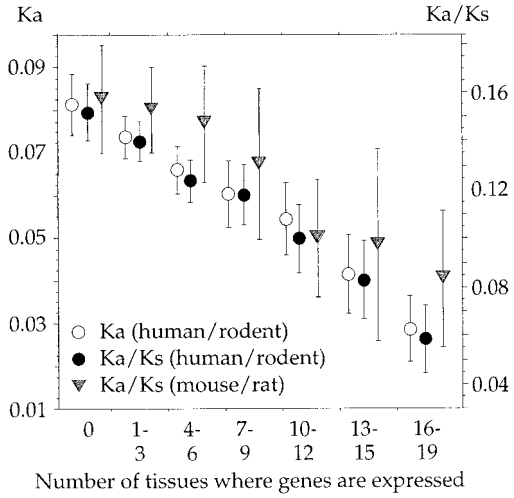

FIG. 1.—Relationship between the gene expression pattern and the nonsynonymous substitution rate (K_a) or the ratio of nonsynonymous/synonymous substitution rates (K_a/K_s). Human/rodent: N = 2,400. Mouse/rat: N = 834. Error bars indicate the 95% confidence interval.

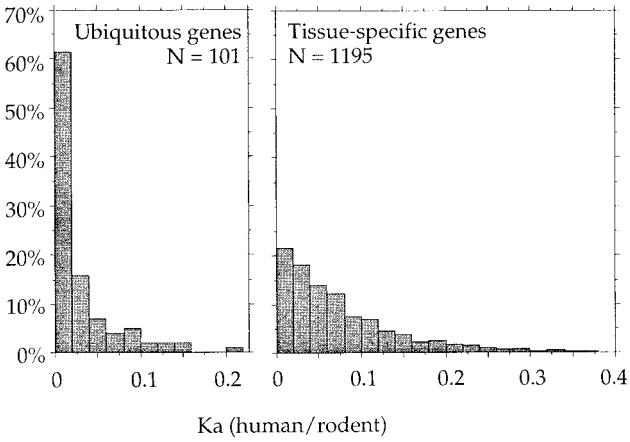

FIG. 2.—Frequency distribution of K_a values in tissue-specific genes (0–3 tissues) and ubiquitous genes (≥16 tissues).

evolve almost three times as fast as ubiquitous ones (table 2). If this variation is due to differences in mutation rate, K_s should vary accordingly. However, the K_a/K_s ratio shows exactly the same variation as K_a (fig. 1). Thus, the decrease in K_a demonstrates an increase in selective pressure on the amino acid sequence. The analysis of mouse/rat orthologs revealed exactly the same trend (fig. 1). There are, of course, some slowly evolving tissue-specific proteins. However, analysis of the distribution of K_a values clearly shows an overall shift toward high values in tissue-specific genes compared with ubiquitous ones (fig. 2).

We also analyzed K_s values to search for possible variations in silent substitution rates according to gene expression patterns. Analysis of mouse/rat orthologs showed no variation of K_s with tissue distribution breadth (fig. 3 and table 2). In human/rodent comparisons, we found a slight decrease in K_s in ubiquitous compared with tissue-specific genes (table 2). However, this trend is weak (see fig. 3), and this variation may be attributed to the correlation that we observed between K_s and K_a ($R = 0.55$, $P < 10^{-4}$), as noticed previously by others (Wolfe and Sharp 1993; Mouchiroud, Gautier, and Bernardi 1995; Makalowski and Boguski 1998). It has been suggested that this correlation might be due to neighboring effects (Wolfe and Sharp 1993). Indeed, it is known that the rate of mutation at a given base is influenced by the nature of its 5′- and 3′- flanking bases (Bains 1992; Hess, Blake, and Blake 1994). Hence, substitution rates at nonsynonymous sites may indirectly affect silent substitution rates. To test this hypothesis, we recalculated K_s ignoring all codons (or codon pairs) in which doublet substitutions (in positions 1–2, 2–3, and 3–1 of codons) occurred. Removal of those codons (which account for 3% of all silent sites and 19% of silent substitutions) abolishes the correlation between K_s and K_a ($R < 10^{-2}$, $P = 0.9$), which confirms that this correlation is due to neighboring effects. Moreover, we now find no decrease in K_s with increasing tissue distribution (table 2), which shows that the slight decrease of K_s noted above in ubiquitous genes is an indirect consequence of the higher selective pressure on nonsynonymous sites. Therefore, there is no evidence for variation of silent substitution rate with the expression pattern.

Variation in Nonsynonymous Substitution Rate in Tissue-Specific Genes According to the Tissue

The large data set of human/rodent orthologs analyzed here shows that there is significant variation in average K_a values of tissue-specific genes according to the tissues (fig. 4). The slowest evolutionary rates are

Table 2
Comparison of Substitution Rates in Ubiquitous and Tissue-Specific Genes

		Ubiq.		TS		TS/Ubiq.	Comparison of TS vs. Ubiq.	
		N	Rate	*N*	Rate	Ratio	*t*-test	*P*
Human/rodent.	K_a	101	0.029	1,195	0.076	2.62	6.57	<0.01%
	K_s	101	0.466	1,195	0.494	1.06	1.9	5.81%
	K_s-nd	101	0.425	1,195	0.403	0.95	−2.09	3.72%
	K 5′UTR	5	0.231	143	0.415	1.80	1.86	6.43%
	K 3′UTR	32	0.310	422	0.437	1.41	3.75	0.02%
Mouse/rat	K_a	45	0.014	435	0.028	2.00	3.034	0.25%
	K_s	45	0.174	435	0.175	1.01	0.11	91.14%
	K_s-nd	45	0.164	435	0.157	0.96	−0.95	34.11%

NOTE.—Ubiq. = ubiquitous genes; TS = tissue-specific genes; K_s-nd = K_s computed after doublet substitutions were removed (see text).

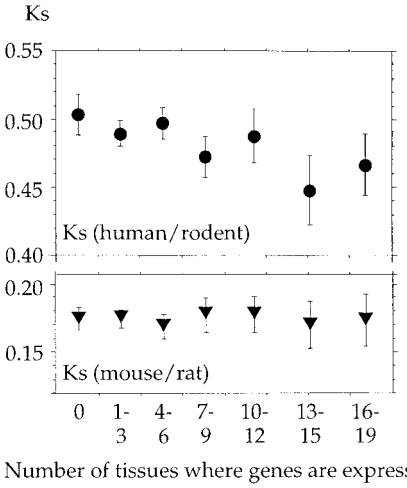

FIG. 3.—Relationship between the gene expression pattern and the synonymous substitution rate (K_s). Error bars indicate the 95% confidence interval.

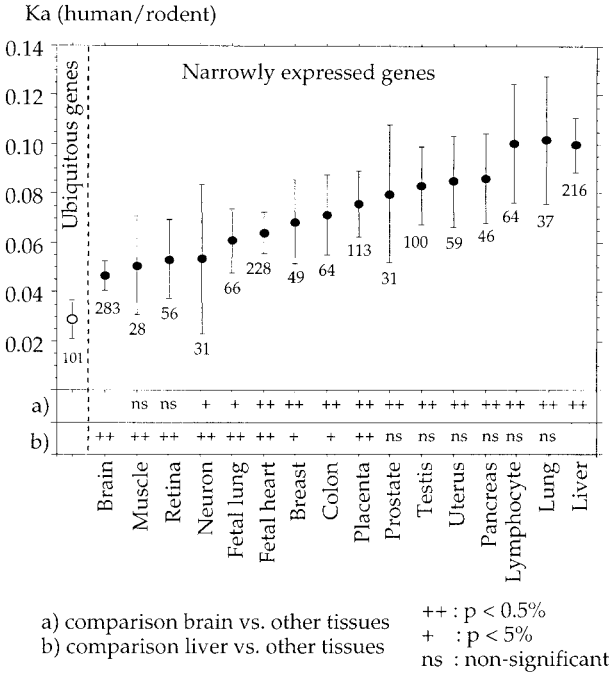

FIG. 4.—Average nonsynonymous substitution rates (K_a) in narrowly expressed genes (1–3 tissues) (black points). Average K_a values in ubiquitous genes (16–19 tissues) are shown for comparison (white point). Error bars indicate the 95% confidence interval. The number of genes is indicated for each sample. The sum of all samples is more than 824 (see table 1) because some genes are expressed in more than one tissue. Average K_a values of brain-specific and liver-specific genes were compared with other narrowly expressed genes (after genes common to compared samples were removed). The significance of Student's *t*-test is indicated.

found in brain-, muscle-, retina- and neuron-specific proteins, and the highest are found in lymphocyte-, lung-, and liver-specific proteins (more than twice as fast) (fig. 4 and table 3). The differences in average K_a values between the two extremes (brain and liver) and the other tissues are highly significant (fig. 4).

There is no significant variation of K_s (computed after doublet substitutions were removed) between tissue-specific genes from these different tissues (table 3). This confirms that the differences in K_a between narrowly expressed genes from different tissues reflect differences in selective pressure but not in mutation rate. Comparison of mouse/rat orthologs gives the same result: average K_a values for liver-specific genes are 2.7 times as high as those for brain (table 3).

Substitution Rates in 3′ UTRs and 5′ UTRs

Analyses of human/rodent orthologs have shown that substitution rates in coding and 5′ and 3′ noncoding regions are correlated (Ogata, Fujibuchi, and Kanehisa 1996; Makalowski and Boguski 1998). Obviously, these correlations cannot be attributed to the neighboring effects responsible for the K_a/K_s correlation. Interestingly, in human/rodent orthologs, the substitution rate within 3′ UTR (*K* 3′UTR) shows exactly the same relationship with the expression pattern as K_a: (1) *K* 3′UTR decreases steadily with increasing expression breadth (fig. 5 and table 2), and (2) liver-specific genes have significantly higher *K* 3′UTR values than brain-specific genes (table 3). The same trend is observed with 5′ UTRs (tables 2 and 3). However, the differences are not statistically significant, probably because of the small sample size. This finding confirms that 5′ and 3′ UTRs do not evolve as selectively neutral sequences but, instead, are functionally constrained (Duret, Dorkeld, and Gautier 1993) and

Table 3
Comparison of Substitution Rates in Brain-Specific and Liver-Specific Genes

		Brain		Liver		Liver/Brain	Comparison of Liver vs. Brain	
		N	Rate	*N*	Rate	Ratio	*t*-test	*P*
Human/rodent.	K_a	247	0.043	180	0.104	2.42	9.61	<0.01%
	K_s	247	0.454	180	0.531	1.17	5.83	<0.01%
	K_s-nd	247	0.399	180	0.409	1.03	1.04	29.71%
	K 5′UTR	30	0.339	23	0.434	1.28	1.68	9.92%
	K 3′UTR	75	0.378	64	0.458	1.21	2.62	0.97%
Mouse/rat	K_a	86	0.015	70	0.040	2.67	5.23	<0.01%
	*K*s	86	0.167	70	0.190	1.14	2.82	0.54%
	K_s-nd	86	0.157	70	0.166	1.06	1.2	23.30%

NOTE.—K_s-nd = K_s computed after doublet substitutions were removed (see text). Brain-specific genes (and, respectively, liver-specific genes) correspond to narrowly expressed genes (one to three tissues) not expressed in liver (respectively, brain).

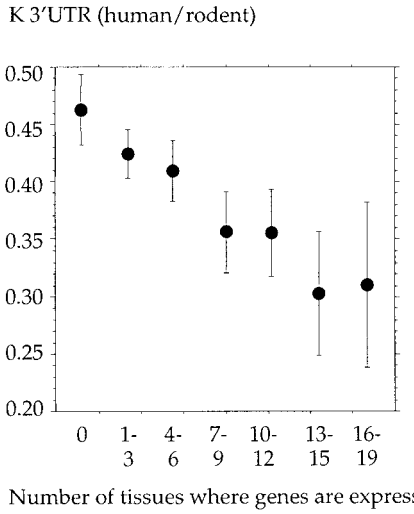

FIG. 5.—Relationship between tissue-distribution breadth and substitution rate within 3′ untranslated regions (K 3′UTR) ($N = 854$).

suggests that, as for coding sites, the selective pressure on UTRs is dependent on tissue distribution.

Discussion

Gene Expression and Mutation Rate

Early studies have shown that transcribed DNA sequences are preferentially subject to nucleotide excision repair (Bohr et al. 1985; Mellon et al. 1986). This connection between transcription and excision repair has been confirmed by many different works (reviewed in Sullivan 1995) and, notably, by the finding that the human basal transcription factor 2 (TFIIH) is also a component of the nucleotide excision repairsome (Schaeffer et al. 1993; Svejstrup et al. 1995). Several authors have suggested that this association between repair and transcription might result in differences in evolutionary rates in transcribed versus nontranscribed DNA (Boulikas 1992; Turker, Cooper, and Bishop 1993; Sullivan 1995). It is worth noting that in multicellular eukaryotes, the mutations that contribute to gene evolution are those that occur in the germ line. Our analyses show no direct relationship between gene expression pattern and silent substitution rates. Although we do not have data on germ line transcription, it is likely that the large majority of the tissue-specific genes of our data set are not expressed in the germ line, whereas most ubiquitous genes are, at least at some stages. Thus, if the above hypothesis were correct, we would have expected a significant decrease in K_s in ubiquitous genes compared with tissue-specific genes, a prediction that is not confirmed by our results. Thus, our results indicate that the variation in efficiency of DNA repair as related to the DNA transcriptional status does not significantly contribute to the variation of the mutation rate in the germ line.

Gene Expression and Selection on Silent Sites

As mentioned in the introduction, in all species in which selection affects synonymous codon usage, the intensity of selection is positively correlated with the gene expression level (Gouy and Gautier 1982; Sharp and Li 1986; Duret and Mouchiroud 1999). As a consequence of this stronger purifying selection, lower K_s values are expected in highly expressed genes compared with weakly expressed genes. Indeed, it has been shown both in bacteria and in drosophila that synonymous substitution rates are lower in genes with a strong codon usage bias (highly expressed) than in other genes (Sharp and Li 1987, 1989; Shields et al. 1988; Powell and Moriyama 1997). The fact that we did not observe any correlation between K_s and gene expression pattern in our data set thus suggests that silent sites are not constrained by selection in mammals. Indeed, we did not find any relationship between synonymous codon usage and gene expression among the 2,400 human genes in our data set (data not shown).

Gene Expression and Intensity of Selection

We found a remarkable negative correlation between K_a (and K_a/K_s) and tissue distribution breadth in both human/rodent and mouse/rat orthologs (fig. 1). This indicates that the selective pressure on nonsynonymous sites depends on the number of tissues in which genes are expressed. Since gene-specific nonsynonymous substitution rates are highly conserved in different mammalian lineages (Mouchiroud, Gautier, and Bernardi 1995), it is likely that this observation stands for all mammals. Indeed, we observed exactly the same effect in 482 human/bovine orthologous genes (data not shown). A similar trend has already been reported for vertebrates by Hastings (1996), who compared the amino acid substitution rates of tissue-specific and broadly expressed protein isoforms. Hastings (1996) proposed that the increase in selective pressure might result from the more diverse biochemical environments to which broadly expressed proteins are exposed. Broadly expressed proteins may interact with a greater variety of molecules and may have to function under a wider range of physical/chemical conditions (e.g., pH) than narrowly expressed proteins. Hence, more sites would be constrained by protein function.

Although this model probably explains a part of the variability in K_a, we do not think that variations in biochemical environments between different tissues are sufficient to account for the threefold decrease in K_a in ubiquitous versus tissue-specific genes. We propose an additional explanation to account for that observation. To simplify, let us consider two protein isoforms that have exactly the same function in the cell, X1, which is broadly expressed, and X2, which has a restricted tissue distribution. Assume that the biochemical environment is constant in all tissues and, finally, consider a mutation that reduces the activity of that protein. This mutation is likely to have a greater phenotypic effect (and hence a stronger impact on the fitness of the organism) in X1 than in X2 simply because it will affect more tissues or development stages. Thus, a slightly or mildly deleterious mutation is more likely to be counterselected when it occurs in a broadly expressed gene than when it occurs in a tissue-specific gene. Of course, sequences of several tissue-specific genes that are crucial for the organism are highly constrained. However, on average,

genes contain many sites at which mutations are not highly deleterious, and for all of those sites, the efficiency of selection will depend on the number of tissues in which genes are expressed. It is likely that this effect accounts for at least a part of the steady decrease in K_a with increasing tissue distribution breadth.

This effect should affect not only protein-coding sites, but also all other elements required for gene function. We have previously shown that many mammalian genes contain long regulatory elements within their 3′ UTRs, most of which are probably involved in posttranscriptional regulation of gene expression (Duret, Dorkeld, and Gautier 1993). Interestingly, we noted that such elements are 2.5-fold more frequent in widely expressed genes than in tissue-specific genes (Duret, Dorkeld, and Gautier 1993). Indeed, as for K_a, there is a negative correlation between substitution rate within 3′ UTR (K 3′UTR) and tissue distribution breadth (fig. 5). Thus, it seems that, as for coding sites, the efficiency of selection on regulatory elements increases with increasing tissue distribution.

Our results also show that the substitution rates of tissue-specific proteins vary considerably according to the tissue (fig. 4) and confirm the strong selective pressure on brain-specific proteins (Kuma, Iwabe, and Miyata 1995; Hughes 1997). Again, this variation in K_a had been interpreted in terms of functional constraints on protein sequence. It had been proposed that the stronger selective pressure in brain-specific proteins is a consequence of a higher complexity of biochemical networks in the brain compared with those of other tissues (Kuma, Iwabe, and Miyata 1995). Conversely, many of the lymphocyte-specific proteins are involved in the immunity response. Thus, the higher average K_a values in those proteins might reflect in part the positive selection for sequence diversity in response to environmental changes (Hughes 1997).

However, the differences in K 3′UTR (table 3) can obviously not be explained by such factors. One could argue that brain-specific genes contain more 3′ UTR regulatory elements than liver- or lymphocyte-specific genes. Indeed, it is possible that posttranscriptional regulation plays a more important role in tuning the expression level of brain-specific genes than in tuning those of liver- or lymphocyte-specific genes. However, the correlation between K_a and K 3′UTR suggests that both observations result from a same factor. Seemingly, mutations in coding regions or in regulatory elements both have, on average, higher impacts on fitness in genes expressed in brain than in liver-specific genes. This observation probably reflects the central role of the brain compared to peripheral organs.

In summary, the phenotypic impact of a mutation in a gene functional element (protein-coding, regulatory region, etc.) depends not only on its direct effect on the biochemical activity of this gene (or its product), but also on the number and the nature of tissues in which this gene is expressed.

Acknowledgments

We thank Manolo Gouy for many helpful comments. Data on UTR substitution rates were kindly provided by Wojciech Makalowski. This work is supported by the Centre National de la Recherche Scientifique.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. H. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ALVAREZ-VALIN, F., K. JABBARI, and G. BERNARDI. 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlation. J. Mol. Evol. **46**:37–44.

BAINS, W. 1992. Local sequence dependence of rate of base replacement in mammals. Mutat. Res. **267**:43–54.

BENSON, D. A., M. S. BOGUSKI, D. J. LIPMAN, J. OSTELL, and B. F. F. OUELLETTE. 1998. GenBank. Nucleic Acids Res. **26**:1–7.

BERNARDI, G., D. MOUCHIROUD, and C. GAUTIER. 1993. Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. **37**:583–589.

BOHR, V. A., C. A. SMITH, D. S. OKUMOTO, and P. C. HANAWALT. 1985. DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. Cell **40**:359–369.

BOULIKAS, T. 1992. Evolutionary consequences of nonrandom damage and repair of chromatin domains. J. Mol. Evol. **35**: 156–180.

CLAVERIE, J.-M., and D. J. STATES. 1993. Information enhancement methods for large scale sequence analysis. Comput. Chem. **17**:191–201.

DEBRY, R. W., and W. F. MARZLUFF. 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics **138**: 191–202.

DICKERSON, R. E. 1971. The structure of cytochrome c and the rates of molecular evolution. J. Mol. Evol. **1**:26–45.

DURET, L., F. DORKELD, and C. GAUTIER. 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. Nucleic Acids Res. **21**:2315–2322.

DURET, L., and D. MOUCHIROUD. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila,* and *Arabidopsis.* Proc. Natl. Acad. Sci. USA **96**:4482–4487.

DURET, L., D. MOUCHIROUD, and M. GOUY. 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. **22**:2360–2365.

GOUY, M., and C. GAUTIER. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**: 7055–7074.

GOUY, M., C. GAUTIER, M. ATTIMONELLI, C. LANAVE, and G. DI-PAOLA. 1985. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. **1**:167–172.

HASTINGS, K. E. M. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J. Mol. Evol. **42**: 631–640.

HESS, S. T., J. D. BLAKE, and R. D. BLAKE. 1994. Wide variations in neighbor-dependent substitution rates. J. Mol. Biol. **236**:1022–1033.

HILLIER, L., G. LENNON, M. BECKER et al. (26 co-authors). 1996. Generation and analysis of 280,000 human expressed sequence tags. Genome Res. **06**:807–828.

HUGHES, A. L. 1997. Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol. Biol. Evol. **14**:1–5.

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

KUMA, K., N. IWABE, and T. MIYATA. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. Mol. Biol. Evol. **12**:123–130.

LI, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. J. Mol. Evol. **36**: 96–99.

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

MAKALOWSKI, W., and M. S. BOGUSKI. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95**:9407–9412.

MELLON, I., V. A. BOHR, C. A. SMITH, and P. C. HANAWALT. 1986. Preferential DNA repair of an active gene in human cells. Proc. Natl. Acad. Sci. USA **83**:8878–8882.

MOUCHIROUD, D., C. GAUTIER, and G. BERNARDI. 1995. Frequencies of synonymous substitution in mammals are gene-specific and correlated with frequencies of non-synonymous substitutions. J. Mol. Evol. **40**:107–113.

OGATA, H., W. FUJIBUCHI, and M. KANEHISA. 1996. The size differences among mammalian introns are due to the accumulation of small deletions. FEBS Lett. **390**:99–103.

OHTA, T., and Y. INA. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. **41**:717–720.

POWELL, J. R., and E. N. MORIYAMA. 1997. Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94**: 7784–7790.

SCHAEFFER, L., R. ROY, S. HUMBERT, V. MONCOLLIN, W. VERMEULEN, J. H. HOEIJMAKERS, P. CHAMBON, and J. M. EGLY. 1993. DNA repair helicase: a component of BTF2 (TFIIH) basic transcription factor. Science **260**:58–63.

SHARP, P. M., M. AVEROF, A. T. LLOYD, G. MATASSI, and J. F. PEDEN. 1995. DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349**:241–247.

SHARP, P. M., and W. H. LI. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24**:28–38.

———. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4**:222–230.

———. 1989. On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28**:398–402.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. ''Silent'' sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

SULLIVAN, D. T. 1995. DNA excision repair and transcription: implications for genome evolution. Curr. Opin. Genet. Dev. **5**:786–791.

SVEJSTRUP, J. Q., Z. WANG, W. J. FEAVER, X. WU, D. A. BUSHNELL, T. F. DONAHUE, E. C. FRIEDBERG, and R. D. KORNBERG. 1995. Different forms of TFIIH for transcription and DNA repair: holo-TFIIH and a nucleotide excision repairosome. Cell **80**:21–28.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

TURKER, M. S., G. E. COOPER, and P. L. BISHOP. 1993. Region-specific rates of molecular evolution: a fourfold reduction in the rate of accumulation of ''silent'' mutations in transcribed versus nontranscribed regions of homologous DNA fragments derived from two closely related mouse species. J. Mol. Evol. **36**:31–40.

WOLFE, K. H., and P. M. SHARP. 1993. Mammalian gene evolution—nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37**:441–456.

DAVID IRWIN, reviewing editor

Accepted September 9, 1999