J Mol Evol (1997) 45:216–226

JOURNAL OF MOLECULAR EVOLUTION

Mutation Pattern Variation Among Regions of the Primate Genome

D. Casane,* S. Boissinot, B.H.-J. Chang, L.C. Shimmin, W.-H. Li**

Human Genetics Center, University of Texas Health Science Center, P.O. Box 20334, Houston, TX 77225, USA

Received: 10 February 1997 / Accepted: 21 April 1997

Abstract. We sequenced three argininosuccinate-synthetase-processed pseudogenes (ΨAS-A1, ΨAS-A3, ΨAS-3) and their noncoding flanking sequences in human, orangutan, baboon, and colobus. Our data showed that these pseudogenes were incorporated into the genome of the Old World monkeys after the divergence of the Old World and New World monkey lineages. These pseudogene flanking regions show variable mutation rates and patterns. The variation in the G/C to A/T mutation rate (u) can account for the unequal GC contents at equilibrium: 34.9, 36.9, and 41.7% in the pseudogene ΨAS-A1, ΨAS-A3, and ΨAS-3 flanking regions, respectively. The A/T to G/C mutation rate (v) seems stable and the u/v ratios equal 1.9, 1.7, and 1.4 in the flanking regions of ΨAS-A1, ΨAS-A3, and ΨAS-3, respectively. These ''regional'' variations of the mutation rate affect the evolution of the pseudogenes, too. The ratio u/v being greater than 1.0 in each case, the overall mutation rate in the GC-rich pseudogenes is, as expected, higher than in their GC-poor flanking regions. Moreover, a ''sequence effect'' has been found. In the three cases examined u and v are higher (at least 20%) in the pseudogene than in its flanking region—i.e., the pseudogene appears as mutation ''hot'' spots embedded in ''cold'' regions. This observation could be partly linked to the fact that the pseudogene flanking regions are long-standing unconstrained DNA sequences, whereas the pseudogenes were relieved of selection on their coding functions only around 30–40 million years ago. We suspect that relatively more mutable sites maintained unchanged during the evolution of the argininosuccinate gene are able to change in the pseudogenes, such sites being eliminated or rare in the flanking regions which have been void of strong selective constraints over a much longer period. Our results shed light on (1) the multiplicity of factors that tune the spontaneous mutation rate and (2) the impact of the genomic position of a sequence on its evolution.

Key words: Mutation pattern — Mutation rate — Regional effect — Sequence effect — Pseudogenes

Introduction

Because of its importance in molecular evolution, the pattern of point mutation has long been a subject of interest to molecular evolutionists (Vogel and Rörhborn 1966; Fitch 1967; Zuckerkandl et al. 1971; Gojobori et al. 1982; Li et al. 1984). Early on it was noted that point mutation is not random and a nucleotide does not mutate to the three others with equal probabilities. Indeed, the transition rate is higher than the transversion rate, though there are two types of transversion and only one type of transition, and the G/C to A/T mutation rate (u) is higher than the A/T to G/C mutation rate (v).

However, it is still not clear whether the mutation pattern is constant over the entire genome of an organism or is variable among regions of the genome. This is of particular interest with regard to the origin and maintenance of the compositional isochores in the mammalian

* *Present address:* Laboratoire de biologie du developpement, Anatomie comparee, Case 7077, Universite Paris 7, 2 Place Jussieu, 75251 Paris Cedex 05, France

** *Present address:* Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030, USA

Correspondence to: W.-H. Li

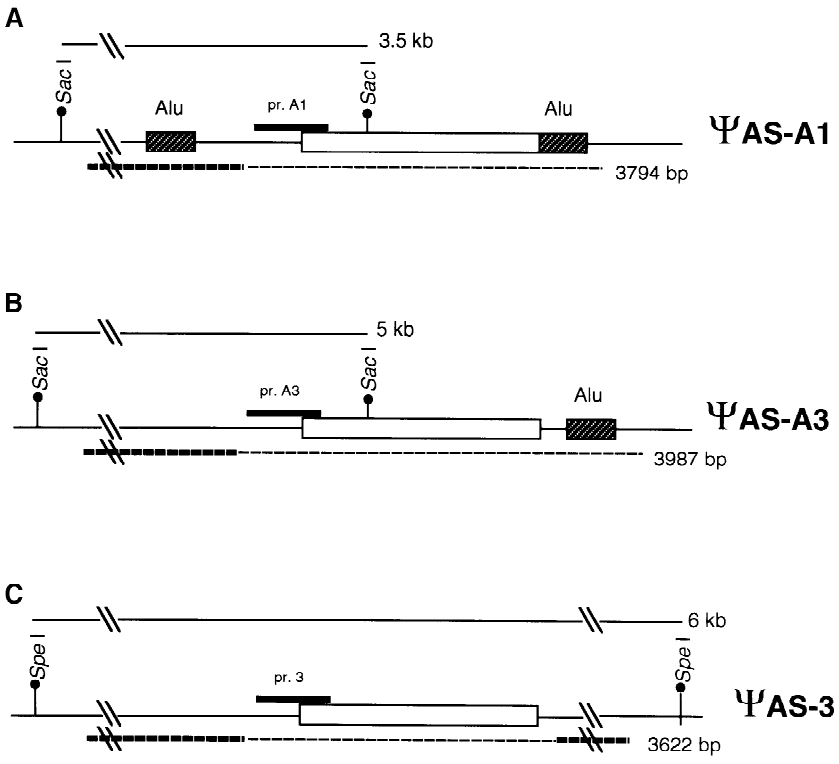

Fig. 1. Cloned and sequenced orangutan DNA regions. *Hatched box:* Alu sequence; *white box:* pseudogene. *Fine dashed line:* known human sequence. **A** ΨAS-A1 pseudogene and flanking regions. The 3.5-kb *Sac* I DNA fragment was identified using the pr. A1 probe (the *black bar* in the figure) and then cloned in λ ZAP II vector. Partial sequencing of a clone allowed the identification of an Alu sequence and unknown flanking sequence (*thick dashed line*). **B** ΨAS-A3 pseudogene and flanking regions. The 5-kb *Sac* I DNA fragment was identified using the pr. A3 probe (the *black bar*) and then cloned in λ ZAP II vector. Partial sequencing of a clone allowed the identification of the unknown flanking sequence (*thick dashed line*). **C** ΨAS-3 pseudogene and flanking regions. The 6-kb *Spe* I DNA fragment was identified using the pr. 3 probe (the *black bar*) and then cloned in λ ZAP II vector. Partial sequencing of a clone allowed the identification of the unknown flanking sequence (*thick dashed line*).

genome, which are long (>300 kb) DNA segments homogeneous in base composition (Bernardi et al. 1985, 1988; Bernardi 1995). Bernardi et al. (1985, 1988) proposed that GC-rich isochores arose because of functional (i.e., selective) advantages. This ''selectionist hypothesis'' is based on some puzzling observations: (1) The gene density is much higher in GC-rich isochores than in GC-poor isochores, and (2) GC-rich isochores exist only in the genome of warm-blooded vertebrates (mammals and birds) but not in the genome of cold-blooded vertebrates. The opposing view is that GC-rich isochores arose because of variation in mutational pressure over regions (Filipski 1987; Sueoka 1988, 1992; Wolfe et al. 1989; Holmquist and Filipski 1994). In particular, the mutationist hypothesis proposed by Wolfe et al. stipulates that isochores arose from mutational biases because of the changes in the composition of the nucleotide pool during replication of germ-line DNA.

Wolfe et al. (1989) showed that the rate of silent substitution varies among genes, depending on the base composition of the gene and its flanking DNA. An inverted-V-shaped distribution of mutation rates with respect to GC content was found and a model for explaining the observation was developed by Gu and Li (1994); see also Wolfe (1991) and Eyre-Walker (1992). However, analyses of more extensive data did not support this relation (Bulmer et al. 1991; Bernardi et al. 1993; Wolfe and Sharp 1993). Indeed, although the rate of synonymous substitution appeared to peak at approximately 60% GC, the GC content variation did not seem to be sufficient to explain the substitution rate variation.

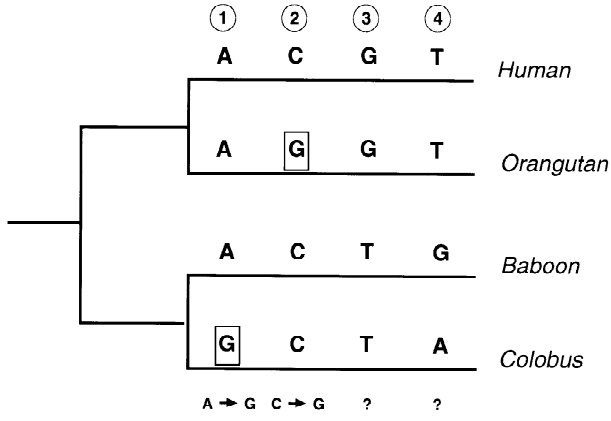

Fig. 2. Mutation pattern inference. For each polymorphic site the direction of the mutation and the ancestral nucleotide state were inferred, assuming that the ancestral nucleotide is the one that requires the minimum number of substitutions to account for the nucleotide differences at the site among the four sequences. (*1*) The most parsimonious explanation is a mutation from A to G (or from T to C on the other strand) in the *colobus* branch. (*2*) The most parsimonious explanation is a mutation from C to G (or from G to C on the other strand) in the *orangutan* branch. (*3*) and (*4*) The ancestral states cannot be uniquely inferred and such sites were excluded from the analysis. We also excluded deletions and insertions.

The above studies were based on comparison of homologous protein-coding genes between a single pair of species, but there are some drawbacks to using such data. First, they only allow the estimation of the variation of the overall mutation rate and give no information on the variation of the mutation pattern. Second, the spontaneous mutation rate is estimated under the assumption that

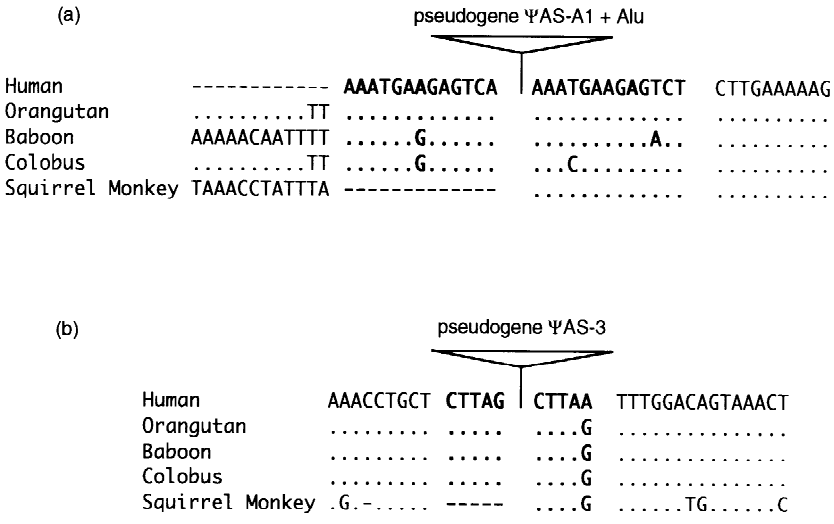

Fig. 3. Pseudogene insertion sites. **A** pseudogene ΨAS-A1 + Alu insertion site; **B** pseudogene ΨAS-3 insertion site. The pseudogenes are each flanked by direct repeats (*boldfaced sequences*). No pseudogene or Alu sequence and only one of the two flanking repeats were found in the squirrel monkey. A *dot (.)* indicates identity to the human sequence; *dashes (-)* indicate insertion or deletion required for the alignment.

it is equal to the substitution rate at synonymous sites. This assumes that synonymous sites are completely neutral. This assumption may not hold because there may be bias in codon usage and there may be compositional patterns of synonymous positions in homologous mammalian genes (Cacciò et al. 1995; Zoubak et al. 1995). Third, the GC content of the DNA region in which the gene is embedded is estimated using the GC content at the silent sites. However, this estimate may not accurately reflect the GC content of the flanking regions.

To avoid these problems we sequenced three processed pseudogenes with the same origin (the argininosuccinate synthetase gene) and their flanking regions (Fig. 1) in human, orangutan, baboon, and colobus. Using these data, we have addressed the issues discussed above.

Materials and Methods

Sources of Samples. Genomic DNA was isolated from 250 mg of baboon (*Papio cynocephalus*), orangutan (*Pongo pygmaeus*), and squirrel monkey (*Saimiri boliviensis*, a New World monkey) liver and human blood by the method of Ellsworth et al. (1993). The genomic DNA of *Colobus angolensis* (an Old World monkey) was a gift from Dr. Carol-Beth Stewart, the State University of New York at Albany.

Southern Blot Analysis. Five micrograms of genomic DNA was digested with 20 units of each restriction enzyme and separated by electrophoresis through 0.9% agarose gels. Gels were processed and DNA was transferred to a nylon membrane (Hybond N+, Amersham) according to the Southern method. The DNA was immobilized by baking at 80°C for 2 h. Membranes were prehybridized at 62°C for 2 h in hybridization buffer (6 × SSC, 5 × Denhardt, 0.1% SDS). The probes were generated by PCR and labeled by random priming with (α-^{32}P)dCTP. Hybridization was conducted overnight at 62°C. The membranes were washed twice at 62°C with 2 × SSC/0.1% SDS and twice with 0.5 × SSC/0.1% SDS.

Partial Library Construction and Screening. Five micrograms of orangutan genomic DNA was digested with 20 units of restriction enzyme and separated by electrophoresis through a 0.9% agarose gel. Restriction fragments with a size close to those of the fragments which hybridized with the probe were excised from the gel and purified. A partial library in λ ZAP II (Stratagene) was constructed according to manufacturer's specifications. The library was screened with the probe prepared in the previous step.

PCR Amplification. Overlapping sequences corresponding to the pseudogenes and flanking regions were amplified in each species. The sequences were amplified from 1 μg of genomic DNA by a regime of 94°C (1 min), 55°C (1 min), and 72°C (2 min) for a total of 35 cycles on a Perkin-Elmer/Cetus DNA thermal cycler in a reaction mix containing 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 1.5 mM $MgCl_2$, 0.1% Triton X-100, 200 mM dNTPs, 1 unit of Taq DNA polymerase, and 1 mM of each appropriate primer.

Cloning. Purified PCR products were ligated into pBluescript II SK+ (Stratagene) digested with EcoRV and tailed with a dTTP. The ligation mixtures were used to transform competent *E. coli* XL1/blue cells and single recombinant colonies were isolated.

Sequencing. For each sequence, three clones, each derived from an independant PCR and cloning reaction, were sequenced enzymatically with Sequenase version 2.0 sequencing kit (United States Biochemical) on double-stranded templates purified with Wizard Miniprep kits (Promega). The sequences of the clones were determined on both strands. Internal sequencing primers were designed as sequence information accumulated.

Sequence Alignment and Analyses. Multiple sequence alignments were made with the Pileup program (Devereux et al. 1984). The gap weight equals 5 and the gap length weight equals 0.3. Distance estimation was performed by Kimura's (1980) two-parameter method, using the MEGA package (Kumar et al. 1993).

Nucleotide Substitution Pattern Inference. The well-known phylogenetic relationships among the four species allow the inference of the mutation pattern as proposed by Gojobori et al. (1982). We included in the analysis the sites at which two and only two states exist, one state being present in one species and the other in the three other species. We assumed that the ancestral state was the one shared by the three species and that the mutation occurred in the other species (the most parsimonious inference). We excluded from the comparison the sites at which we could not decide the direction of the change (Fig. 2).

Table 1. Sequence lengths and base compositions

	Flanking regions			Pseudogenes			Functional AS gene
	ΨAS-A1	ΨAS-A3	ΨAS-3	ΨAS-A1	ΨAS-A3	ΨAS-3	
Length (bp)	1,619	2,109	1,992	1,592	1,630	1,630	1,527
%A	34.0	31.1	27.5	25.9	25.3	25.0	24.0
%T	31.2	31.6	30.8	20.4	20.5	21.1	19.5
%C	19.7	20.0	19.5	27.7	27.2	27.2	28.2
%G	15.2	16.9	22.2	26.1	26.9	26.7	28.2
%GC	34.9	36.9	41.7	53.8	54.1	53.9	56.4

Table 2. Indel size distribution

	Indel size (bp)										
Sequence[a]	1	2	3	4	5	6	7	10	22	27	Total
ΨAS-A1	10	1	5	0	1				1		18
ΨAS-A3	7	4	3	1	0	1		1		1	18
ΨAS-3	6	1	0	1	0	2					10
Fl. ΨAS-A1	3	3	2	1	1						10
Fl. ΨAS-A3	10	2	3	2	2	3	1				23
Fl. ΨAS-3	7	2	2	1	2	1					15
Total	43	13	15	6	6	7	1	1	1	1	94

[a] ΨAS: argininosuccinate synthetase pseudogene; Fl. ΨAS: flanking region of an argininosuccinate synthetase pseudogene

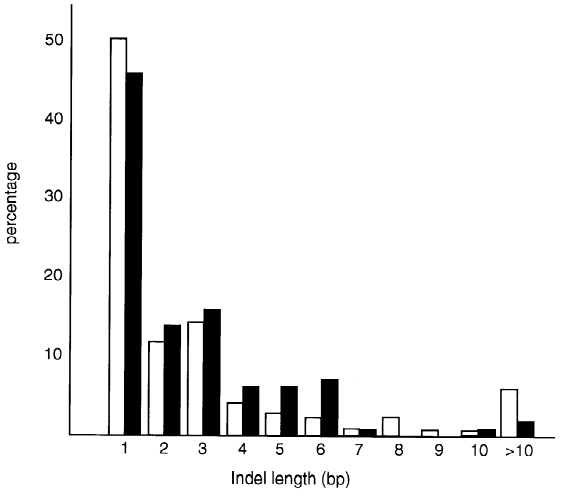

Fig. 4. Indel size distribution. Open bars: Indel size distributions based on 78 human processed pseudogenes (Gu and Li 1995). *Solid bars:* Indel size distribution from the three pseudogenes and their flanking regions (present study).

Results

Pseudogene and Flanking Region Sequences

In humans, in addition to the argininosuccinate synthetase (AS) gene, 14 AS pseudogenes are dispersed on 11 different chromosomes, including the X and Y chromosomes (Su et al. 1984). Seven AS pseudogenes and partial 5′ and 3′ flanking sequences had been sequenced (Freytag et al. 1984; Nomiyama et al. 1986). We constructed three probes (pr. A1, pr. A3, and pr. 3) by using PCR-amplified fragments of a 5′ flanking region and a short sequence of each of pseudogenes ΨAS-A1, ΨAS-A3, and ΨAS-3 (Fig. 1). Orangutan genomic DNA was digested by a set of restriction enzymes that cut in λ ZAP II vector, and the digested DNA was electrophoresed in an agarose gel, Southern blotted, and hybridized with probes pr. A1, pr. A3, and pr. 3. Three DNA fragments were identified (*Sac* I: 3.5 kb; *Sac* I: 5 kb *Spe* 1:6 kb), all of which contain long stretches of the pseudogene ΨAS-A1, ΨAS-A3, and ΨAS-3 flanking regions, respectively (Fig. 1). λ ZAP II partial libraries of these fragments were constructed and screened with probes pr. A1, pr. A3, and pr. 3, respectively. Phages bearing DNA fragments of interest were isolated and plasmid excision was performed in order to sequence about 2 kb of the 5′ flanking region of each pseudogene.

Human, orangutan, baboon, and colobus overlapping fragments of DNA covering a pseudogene and its flanking regions were PCR amplified and sequenced. In the 5′ flanking region of ΨAS-A1, an Alu sequence was identified. Other transposable sequences or long open reading frames (ORFs) were not observed.

Two DNA fragments in a squirrel monkey were amplified with primers flanking pseudogenes ΨAS-A1 and ΨAS-3, respectively, and sequenced. The pseudogenes were absent in each DNA fragment, and only a single copy of the two repeated sequences flanking a pseudogene in the other species were found. Moreover, no Alu sequence was found (Fig. 3). These results indicate that the insertions of the pseudogenes and the Alu sequences happened after the split of the New World and the Old World monkey lineages 35–40 million years (Myr) ago but before the emergence of the ape lineage 25–30 Myr ago.

The ΨAS-A1 pseudogene, the two Alu sequences, and the flanking regions sequenced constitute an uninterrupted DNA segment of 3,794 bp (the thick and thin dashed lines in Fig. 1A). The lengths of the ΨAS-A1 pseudogene and the flanking region are 1,592 bp and 1,619 bp, respectively (Table 1). The ΨAS-A3 pseudogene, the Alu sequence, and the flanking regions sequenced constitute an uninterrupted DNA segment of

Table 3. Mutation patterns in pseudogenes and their flanking regions[a]

	Flanking region								
	ΨAS-A1			ΨAS-A3			ΨAS-3		
G→A	17			18			18		
G→T	2			2			3		
G→C	2			4			4		
A→G	12			19			24		
A→T	8			4			4		
A→C	1			10			4		
T→G	4			6			9		
T→A	2			5			2		
T→C	21			14			15		
C→G	3			9			3		
C→A	6			4			3		
C→T	17			26			20		
Total	95			121			109		
Ts number	67			77			77		
% Ts (s.e.)	70 (4.6)			64 (4.4)			71 (4.3)		
	n	*l*	*p*	*n*	*l*	*p*	*n*	*l*	*p*
G/C→A/T	42	2,260	185	50	3,113	160	44	3,323	123
A/T→G/C	38	4,216	90	49	5,323	92	52	4,645	111
G/C→C/G	5	2,260	22	13	3,113	42	7	3,323	21
A/T→T/A	10	4,216	24	9	5,323	17	6	4,645	13
Total			321			311			277

[a] *n:* mutation number; *l:* number of G+C (or A+T) nucleotides screened for mutations i.e., sequence length × 4 × %GC / 100 (or sequence length × 4 × (1 − %GC) / 100); *p* (number of mutations normalized by the sequence length and the GC content): $n/l \times 10^4$; Ts: transition number; Values in parentheses were obtained by excluding all CG dinucleotides from comparisons.

3,987 bp (Fig. 1B). The lengths of the ΨAS-A3 pseudogene and the flanking regions are 1,630 bp and 2,109 bp, respectively (Table 1). The ΨAS-3 pseudogene (1,630 bp) and the flanking regions (1,992 bp) sequenced constitute a 3,622-bp uninterrupted DNA fragment (Fig. 1C).

Indel Distribution

In the alignment of the sequences from the four species studied, deletions and insertions (indels) of various sizes are scattered along the pseudogenes and their flanking regions (Table 2). The size distribution of indels (Fig. 4) is very similar to the one obtained from a large sample of pseudogenes (Gu and Li 1995)—that is, small indels occur much more frequently than large indels. The indel events are found in both the internal and the external branches of the phylogenetic trees, an observation that indicates the lack of function of the pseudogenes and their flanking regions.

Base Compositions in Pseudogenes and Flanking Regions

The pseudogenes show very similar base compositions (the average of four species): (%A: 25.9, 25.3, 25.0), (%T: 20.4, 20.5, 21.1), (%C: 27.7, 27.2, 27.2), and (%G: 26.1, 26.9, 26.7) in ΨAS-A1, ΨAS-A3, and ΨAS-3, respectively. The base composition in the AS functional gene is 24.0% A, 19.5% T, 28.2% C, and 28.2% G. The pseudogenes, like the functional genes, are relatively GC-rich (~54%). In contrast, the base compositions of their flanking regions are heterogeneous: (%A: 34.0, 31.1, 27.5), (%T: 31.2, 31.6, 30.8), (%C: 19.7, 20.0, 19.5), and (%G: 15.2, 16.9, 22.2) in the ΨAS-A1, ΨAS-A3, and ΨAS-3 flanking regions, respectively. The GC contents range from ~35 to ~42% but are considerably lower than the GC contents of the pseudogenes (Table 1).

Mutation Patterns in Flanking Regions

We excluded the Alu sequences in the analysis. In the flanking region of each of the three pseudogenes (ΨAS-A1, ΨAS-A3, and ΨAS-3), the number of mutations from G/C to A/T (42, 50, 44) is not statistically different from the number of mutations from A/T to G/C (38, 49, 52) (χ^2 test: $P = 0.65$, $P = 0.94$, $P = 0.41$) (Table 3). The GC contents in these flanking regions are thus at statistical equilibrium. If no selection is assumed (see later for a discussion of this assumption), the GC content

Table 3. Continued

Pseudogene									Total
ΨAS-A1			ΨAS-A3			ΨAS-3			
31 (22)			32 (29)			17 (11)			
6			3			6 (5)			
7			4			6 (4)			
20			15 (13)			20 (19)			
2			4			2			
5			5			3			
3			2			0			
10			3			2			
16 (15)			15 (14)			14 (13)			
7 (6)			4 (3)			6 (4)			
5 (4)			4			3			
25 (22)			35 (29)			33 (23)			
137			126			112			700
91			97			84			
66 (4.2)			77 (3.7)			75 (4.1)			
n	*l*	*p*	*n*	*l*	*p*	*n*	*l*	*p*	
67	3,426	196	74	3,527	210	59	3,514	168	336
(53)	(3,096)	(171)	(64)	(3,200)	(200)	(38)	(3,184)	(119)	
44	2,942	150	37	2,993	124	37	3,006	123	257
(43)		(146)	(34)		(113)	(35)		(116)	
14	3,426	40	8	3,527	23	12	3,514	34	59
	(3,096)	(45)		(3,200)	(25)		(3,184)	(38)	
12	2,942	41	7	2,993	23	4	3,006	13	48
		427			380			338	
		(403)			(361)			(286)	

at equilibrium (c) allows the estimation of the u/v ratio (Sueoka 1962):

$$u/v = (1 - c)/c$$

We obtain $u/v = 1.9$, 1.7, and 1.4 in the flanking regions of ΨAS-A1, ΨAS-A3, and ΨAS-3, respectively.

The substitution pattern in each pseudogene flanking region is reported (Table 3). After normalization by the sequence length and GC content, the G/C to A/T mutation number in the ΨAS-A1 flanking region is higher than those in the ΨAS-A3 and ΨAS-3 flanking regions (not statistically significant). The A/T to G/C mutation numbers are similar among the three pseudogene flanking regions (Fig. 5A).

Kimura's (1980) two-parameter distances in the flanking regions between the four species are reported in Table 4b, d, and f. All the other methods of distance estimation available with the software MEGA gave very similar estimates, as expected when distances are small (<10%). Under the assumption of rate constancy among lineages, the four distances *Hsa/Pcy*, *Hsa/Can*, *Ppy/Pcy*, and *Ppy/Can* must be equal (*Hsa*, human; *Ppy*, orangutan; *Pcy*, baboon; and *Can*, colobus). Indeed, in the flanking regions these distances are very close to each other. It is worth noting that the distance estimate depends on the DNA region studied—that is, it decreases with increasing GC content.

Mutation Patterns in Pseudogenes

The GC contents (54%) of the pseudogenes are higher than those observed in the flanking regions (Table 1). In contrast to the flanking regions, in each of the three pseudogenes (ΨAS-A1, ΨAS-A3, and ΨAS-3), the number of mutations from G/C to A/T (67, 74, 59) is statistically higher than the number of mutations from A/T to G/C (44, 37, 37) (χ^2 test: $P = 0.03$, $P < 0.001$, $P = 0.02$) (Table 3), implying that the GC contents of these pseudogenes are decreasing with time.

The substitution patterns in the peudogenes seem to follow the same trends as that observed in their flanking regions (Table 3 and Fig. 5A). However, a systematically higher normalized mutation number than in the corresponding flanking region is observed in each pseudogene for every type of mutation considered. The difference in the normalized G/C to A/T mutation number in pseudogene ΨAS-A1 and its flanking region (196 vs 185) is the only one that is not statistically significant.

Forty-one CG dinucleotide sites in the common ancestral sequence of the pseudogenes and the functional

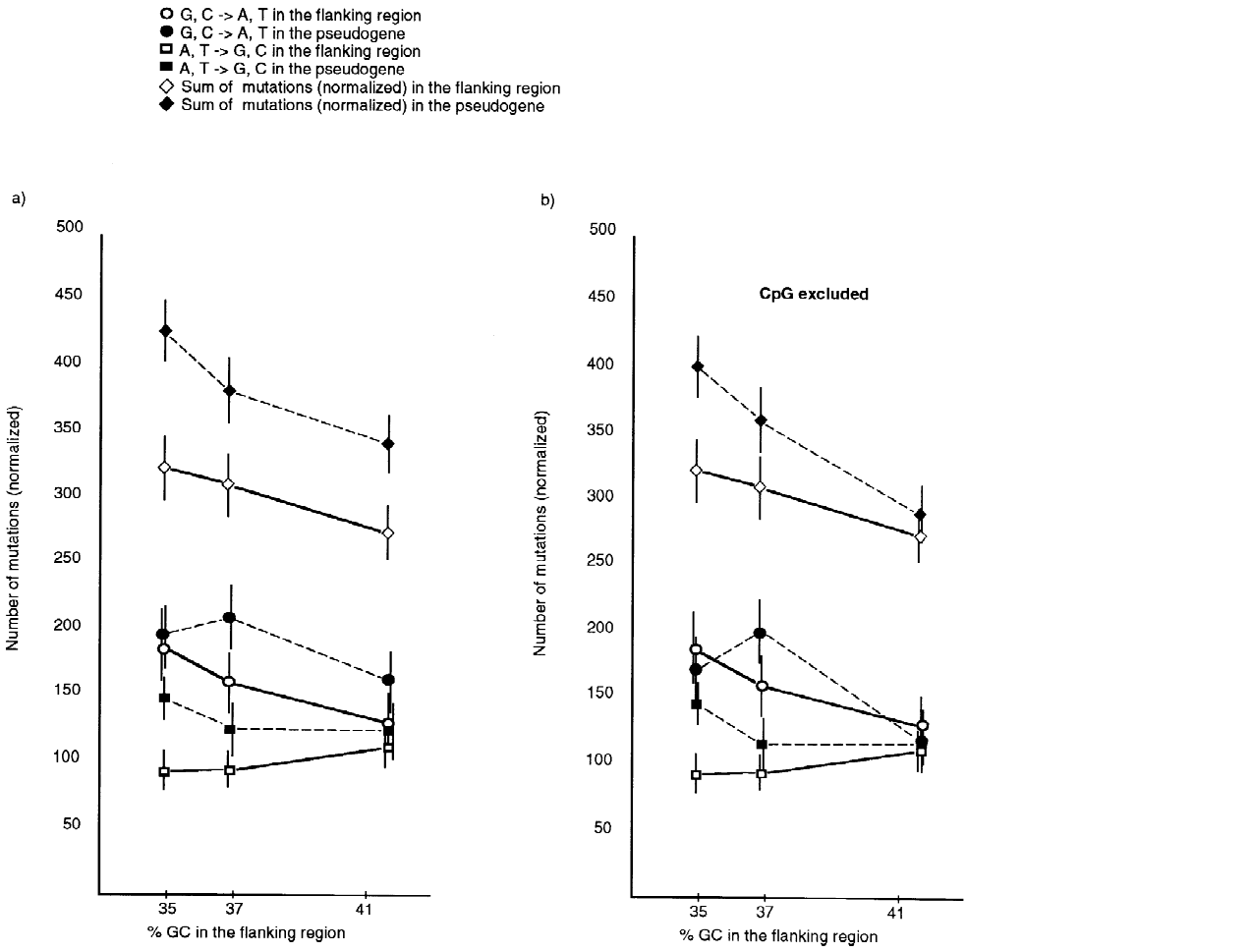

Fig. 5. Substitution rate variations. Plot of the number of substitutions (normalized) in pseudogenes and flanking regions versus flanking region GC content. See Table 3 for data normalization by sequence length and GC content. **A** all nucleotide sites included, **B** GC dinucleotide sites excluded.

gene were inferred. These sites were excluded and the mutation rates and patterns revised as in Li et al. (1984). The new estimates of G/C to A/T substitution rates are only slightly lower in ΨAS-A1 and ΨAS-A3 but substantially lower in ΨAS-3. The general trends are not changed by the exclusion of the CG dinucleotides (Table 3 and Fig. 5b).

Kimura's two-parameter distances between the four species are reported in Table 4a, c and e. The four distances *Hsa/Pcy*, *Hsa/Can*, *Ppy/Pcy*, and *Ppy/Can* are reported in Fig. 6b. The distances from the pseudogene data are longer than those estimated from the flanking regions. In addition, the ΨAS-A1 distances are longer than the ΨAS-3 distances, as is the case for the corresponding flanking region (Fig. 6A). The ΨAS-A3 distances are similar to those of the ΨAS-3 ones. Moreover, the four estimates for ΨAS-A3 are quite different, largely due to a particularly low number of mutations in the terminal branch leading to human and a particularly high number of mutations in the terminal branch leading to the baboon.

Discussion

An ideal experimental situation for studying the factors that affect the rate and pattern of spontaneous mutation would result from inserting identical neutral sequences in different regions of the genome and monitoring their evolution through time. Such an experiment is not feasible because evolution proceeds extremely slowly, but a molecular evolutionary process can mimic such a situation. Processed pseudogenes of a functional gene may be inserted into multiple sites in the genome. Because processed pseudogenes are likely to be devoid of function from the moment they are incorporated into the genome, they are useful for studying the rate and pattern of spontaneous mutation. In particular, 14 argininosuccinate synthetase pseudogenes were found on 11 different human chromosomes. We chose three of them (ΨAS-A1, ΨAS-A3, and ΨAS-3) that seem to have appeared during a short period of time and that might be found in all monkey genomes (Nomiyama et al. 1986). Four higher primate sequences of each pseudogene and their flanking

Table 4. Means (below diagonal) and standard errors (above diagonal) of the numbers of nucleotide substitutions per 100 sites between sequences

a)

Pseudogene ΨAS-A1 (53.8 %GC)

	Hsa	*Ppy*	*Pcy*	*Can*
Hsa		0.57	0.91	0.90
Ppy	4.6		0.94	0.92
Pcy	10.9	11.6		0.65
Can	10.7	11.2	6.1	

b)

ΨAS-A1 flanking region (34.9 %GC)

	Hsa	*Ppy*	*Pcy*	*Can*
Hsa		0.46	0.72	0.74
Ppy	3.1		0.70	0.73
Pcy	7.1	6.9		0.51
Can	7.7	7.5	3.8	

c)

Pseudogene ΨAS-A3 (54.1 %GC)

	Hsa	*Ppy*	*Pcy*	*Can*
Hsa		0.54	0.80	0.74
Ppy	4.3		0.86	0.81
Pcy	8.9	9.9		0.57
Can	7.7	9.0	4.8	

d)

ΨAS-A3 flanking region (36.9 %GC)

	Hsa	*Ppy*	*Pcy*	*Can*
Hsa		0.41	0.61	0.59
Ppy	3.2		0.60	0.59
Pcy	6.7	6.6		0.43
Can	6.4	6.3	3.5	

e)

Pseudogene ΨAS-3 (53.9 %GC)

	Hsa	*Ppy*	*Pcy*	*Can*
Hsa		0.56	0.81	0.85
Ppy	4.6		0.81	0.85
Pcy	9.0	9.1		0.52
Can	9.9	10.0	4.1	

f)

ΨAS-3 flanking region (41.7 %GC)

	Hsa	*Ppy*	*Pcy*	*Can*
Hsa		0.40	0.59	0.57
Ppy	2.9		0.57	0.56
Pcy	6.3	5.9		0.43
Can	5.8	5.7	3.4	

regions allowed us to analyze the pseudogene evolution in parallel with the flanking region evolution. Some ''regional'' and ''sequence'' effects on mutation rates have been revealed.

''Regional'' Effects on the Rate and Pattern of Mutation in the Flanking Regions

The flanking regions of the three pseudogenes have reached different GC content equilibria. In the absence of selection on the GC content, this means variation of the ratio u/v where u is the rate of mutation from G/C to A/T and v is the rate of mutation from A/T to G/C. v is rather stable in the three regions but u decreases with increasing GC content (Fig. 5A). These results suggest that the GC content variation among regions is mainly due to the variation of the mutation rate from G/C to A/T. If v is constant, then $u_3/u_{A1} = 1.4/1.9 = 0.74$, where u_3 and u_{A1} are the mutation rates from G/C to A/T in the flanking regions of ΨAS-3 and ΨAS-A1, respectively. This shows the importance of regional variation in mutation rate.

It is worth noting that the frequency of A is not equal to that of T in the flanking regions of ΨAS-A1 and ΨAS-3 and that the frequency of G is not equal to that of C in the flanking regions of all the pseudogenes. Equal frequencies of A and T on the one hand and G and C on the other hand are expected at equilibrium in noncoding sequences if no-strand mutation biases are involved (the type 2 parity rule) (Lobry 1995; Sueoka 1995). The violation of the type 2 parity rule is expected in coding sequences because there are asymmetrical constraints on the coding strand vs its complementary strand. Indeed, we found such a deviation in the pseudogenes, which reflects the coding function of the sequence from which they originated. The violation of the type 2 parity rule in noncoding sequences is more likely due to strand mutation biases. It could mean that not only the site of insertion of a sequence but also its position relative to the replication origin of the DNA segment in which it is embedded can affect its molecular evolution (Wu and Maeda 1987).

''Regional'' Effects on the Rate and Pattern of Mutation in the Pseudogenes

The comparison of the mutation patterns in different pseudogenes that were derived from the same functional gene is particularly interesting if their insertions into the genome occurred in a short period in the evolutionary

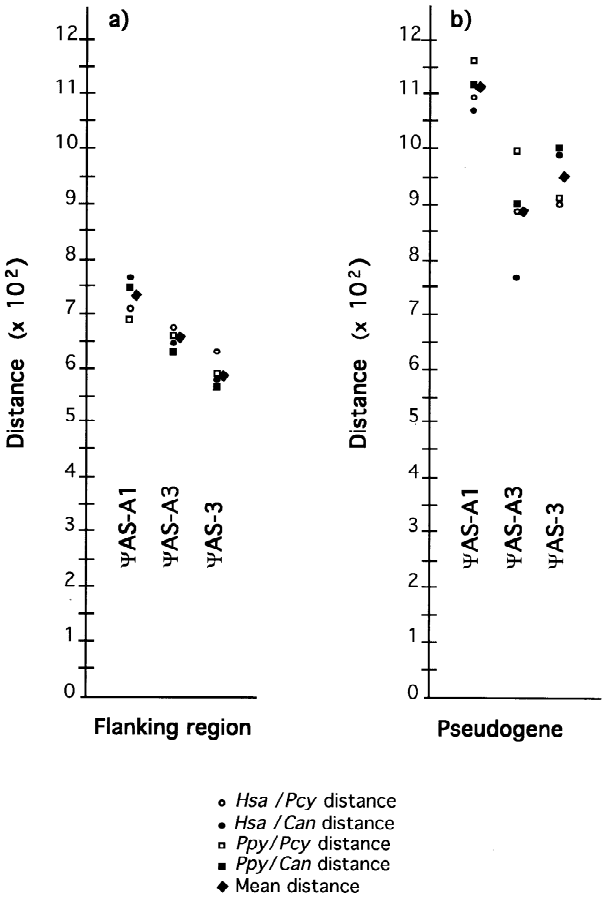

Fig. 6. Distance variations. Plot of Kimura's two-parameter distances between human and baboon (*Hsa/Pcy*), human and colobus (*Hsa/Can*), orangutan and baboon (*Ppy/Pcy*), orangutan and colobus (*Ppy/Can*) using three argininosuccinate synthetase pseudogene sequences or their flanking regions. Species abbreviations are: *Hsa, Homo sapiens; Ppy,* the orangutan *Pongo pygmaeus; Pcy,* the baboon *Papio cyanocephalus; Can,* the colobus *Colobus angolensis.*

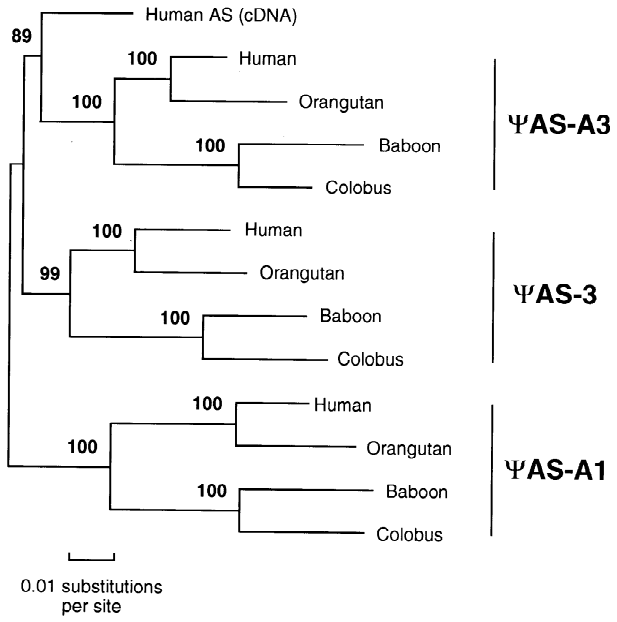

Fig. 7. Argininosuccinate synthetase pseudogene phylogeny. Human, orangutan, baboon, and colobus argininosuccinate synthetase pseudogene ΨAS-A1, ΨAS-A3, and ΨAS-3 sequences have been aligned with the argininosuccinate cDNA sequence and a phylogenetic tree inferred using the neighbor-joining method (Saitou and Nei 1987) and Kimura's two-parameter distances (the number of substitutions per site). Bootstrap values (based on 500 replications) are reported at each node. The highly mutable poly-A tails have been excluded from the analysis.

scale so that the pseudogenes allow the analysis of the evolution of the same sequence in different genomic environments. Our data show that the pseudogenes were incorporated into the genome between the separation of New World/Old World monkeys 35–40 Myr ago and the emergence of the ape lineage 25–30 Myr ago. This means a period of 10 million years at most between the first (ΨAS-A1) and the last insertion (ΨAS-A3) according to the phylogenic tree shown in Fig. 7. An examination of the phylogenic tree of the pseudogenes (Fig. 7) also suggests that the three insertions occurred at very similar times because the two internal branches that link the root of the tree to the node where the pseudogene ΨAS-A3 separates from the functional gene are very short. Note also that even now the nucleotide compositions of the pseudogenes are very similar (Table 1). Therefore, the set of three pseudogenes we chose is a good approximation of the ideal simultaneous insertion of three identical sequences in a common ancestor of the four species.

The pseudogene distances are longer than the distances in their flanking regions (Fig. 6). Estimates of $u/v > 1.0$ in the three regions could account for this observation. Indeed, if the mutation rates of G and C are higher than those of A and T, one can expect that the number of mutations is positively correlated with the GC content. However, ΨAS-A1 and ΨAS-3 have evolved at different rates, even though these sequences have the same GC content (Fig. 6B). As with their flanking regions the ΨAS-A1 distances are longer than the ΨAS-3 ones (Fig. 6A,B). This result confirms that the G and C mutation rates are lower in the ΨAS-3 region than in the ΨAS-A1 region. The ΨAS-A3 distances were thus expected to be somewhere between those in the ΨAS-A1 and ΨAS-3, as with the flanking region sequences. They are actually lower than expected (Fig. 6B), but this may be due to sampling effects (a short human terminal branch and a long colobus terminal branch).

''Sequence'' Effects on the Mutation Rate and Pattern

It is well known that u is greater than v (Li et al. 1984), so more mutations are expected in GC-rich sequences than in GC-poor sequences. Indeed, longer distances between sequences are obtained with GC-rich pseudogenes than with GC-poor flanking regions (Fig. 6A,B).

A more striking point is that estimates of the mutation rates u and v and the overall mutation rate are higher in

a pseudogene than in its flanking region (Fig. 5). This suggests that all nucleotide types in the pseudogenes mutate more rapidly than expected on the basis of the mutation rate pattern in their flanking regions.

The ratios of the sum of the normalized mutation numbers in ΨAS-A1, ΨAS-A3, and ΨAS-3 and in the flanking regions, respectively, are 427/321 (1.32), 380/311 (1.22), and 338/277 (1.22). This result shows a constant greater mutation rate (20% or higher) in the pseudogenes than in the flanking regions (Fig. 5A).

It is worth noting that the flanking regions have reached the GC content equilibrium, which is a slow process (Li and Graur 1991), because these regions have not been subject to significant selective pressures (no coding sequences) for a very long time. Conversely, selection has been strong on the nonsynonymous sites of the gene that gave rise to the pseudogenes. For about 35 million years the pseudogene sequences have been relieved of this constraint and the GC content has decreased slightly (from 56.4% to 54% GC) as expected under GC mutation pressure. Moreover, it is known that the mutation rate of a nucleotide depends on the neighboring nucleotides (Izuta et al. 1995; Mendelman et al. 1989). We propose that more mutable combinations of nucleotides have been conserved during evolution of the gene because of the selection pressure on the coding sequence. These more mutable sites may elevate the mutation rate in pseudogenes because pseudogenes are not subject to functional contraints. Of course, the higher mutation rates in these pseudogenes should be mainly due to their high GC content rather than due to some particular nucleotide combinations.

Because of the higher GC content and probably the existence of more mutable nucleotide combinations, these pseudogenes then appear as ''hot'' spots of mutation embedded in ''cold'' regions. When the GC content decreases and the highly mutable sites are consumed, the mutation rates will slow down to those observed in the flanking regions. A comparison of the substitution rates in older pseudogenes and their flanking regions may allow one to test this hypothesis.

No Selection on GC Content

The above discussion assumes that there is no selection on the GC content. Evidence for this assumption in the three regions studied can be derived from our data. First, if the different GC content equilibria in the flanking regions are maintained by selection (the mutation pattern being the same), this selection is very efficient in eliminating excess of A and T or G and C, because the observed numbers of mutations toward G or C and toward A or T are very similar. In other words, there is strong selection for maintaining the optimal GC percentage values. Then it is paradoxical to observe the maintenance of pseudogenes and Alu sequences, which have very different GC contents.

Second, a more quantitative argument can be found in our analyses. If we assume that the poor GC content in the flanking regions is controlled by selection, then we must observe an excess of substitutions of G and C toward A and T and a deficit of substitutions of A and T toward G and C in the GC-rich pseudogenes. An excess of mutation toward A and T is indeed observed, but there is also an excess of mutation toward G and C, which is incompatible with the selectionist hypothesis.

Third, the GC content decreases very slowly in the pseudogenes. The GC content is 56.4% in the human argininosuccinate mRNA (excluding the poly-A tail) and about 54% in the pseudogenes. This means only a 2.4% decrease after at least 30 million years; thus, it will take a very long time to reach the flanking-region GC-poor equilibrium state. Under such weak selection, it is quite inconceivable how the equilibrium state could be maintained in genomic regions with recurrent insertions. It is simpler to assume that the flanking regions of the pseudogenes are selectively neutral and have reached different GC-content equilibria under local mutation patterns.

Biological and Evolutionary Implications

The variations in the G and C spontaneous mutation rates will need further investigation. One explanation is linked to potential methylation of C residues. The transition from C to T can arise from conversion of methylated C residues to T residues upon deamination; about 90% of methylated C residues in vertebrate DNA occur at 5′-CG-3′ dinucleotides (Coulombre et al. 1978).

A higher C mutation rate is expected in CG dinucleotide-rich regions. When the CG dinucleotides are excluded from the analysis, the substitution rate in the pseudogene is somewhat less conspicuous (Fig. 5B). The maintenance of CG dinucleotides in the coding sequence can thus partially explain the higher substitution rate in the pseudogenes than in the flanking regions. Another explanation is suggested by the finding of a high level of organization of the genome in the nucleus and the association of replication and repair complexes with the nuclear matrix (Berezney and Jeon 1995). DNA sequence position in the nucleus might affect its replication timing and accuracy and repair efficiency (Holmquist and Filipski 1994).

Variations in the mutation pattern among regions may explain the poor fit of the model developed by Gu and Li (1994) to a large data set. This model could be true but only locally; a set of curves instead of only one could describe the relation between mutation rate and GC content for the whole genome. If all the data are analyzed together, no clear picture may thus appear.

As homologous genes can move from one region of the genome to another, it may be interesting to investigate whether the effects of the local mutation pattern on distance estimation can introduce bias that affects the accuracy of a phylogenetic method.

Acknowledgments. This study was supported by NIH grants and a French Government Fellowship. We thank a reviewer for valuable comments.

References

Berezney R, Jeon KW (eds) (1995) Structural and functional organization of the nuclear matrix. Academic Press, New York

Bernardi G (1995) The human genome: organization and evolutionary history. Annu Rev Genet 29:445–476

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228: 953–957

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Bernardi G, Mouchiroud D, Gautier C (1993) Silent substitutions in mammalian genomes and their evolutionary implications. J Mol Evol 37:583–589

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc Natl Acad Sci USA 88:5974–5978

Cacciò S, Zoubak S, D'Onofrio G, Bernardi G (1995) Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J Mol Evol 40:280–292

Coulombre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274: 775–780

Devereux J, Haberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12: 387–395

Ellsworth DL, Hewett-Emmett D, Li W-H (1993) Insulin-like growth factor II intron sequences support the hominoid rate-slowdown hypothesis. Mol Phylogenet Evol 2:315–321

Eyre-Walker A (1992) The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals. Genet Res 60:61–67

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217:184–186

Fitch WM (1967) Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. J Mol Biol 26:499–507

Freytag SO, Bock HG, Beaudet AL, O'Brien WE (1984) Molecular structures of human argininosuccinate synthetase pseudogenes. J Biol Chem 259:3160–3166

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Gu X, Li W-H (1994) A model for the correlation of mutation rate with GC content and the origin of GC-rich isochores. J Mol Evol 38: 468–475

Gu X, Li W-H (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J Mol Evol 40:464–473

Holmquist GP, Filipski J (1994) Organization of mutations along the genome: a prime determinant of genome evolution. TREE 9:65–69

Izuta S, Roberts JD, Kunkel TA (1995) Replication error rates for G.dGTP, T.dGTP, and A.dGTP mispairs and evidence for differential proofreading by leading and lagging strand DNA replication complexes in human cells. J Biol Chem 270:2595–2600

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

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetics analysis, version 1.01. The Pennsylvania State University, University Park, PA

Li W-H, Graur D (1991) Fundamentals of molecular evolution. Sinauer, Sunderland, MA

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutations reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326–330

Mendelman LV, Boosalis MS, Petruska J, Goodman MF (1989) Nearest neighbor influences on DNA polymerase insertion fidelity. J Biol Chem 264:14415–14423

Nomiyama H, Obaru KK, Jinno Y, Matsuda I, Shimada K, et al. (1986) Amplification of human argininosuccinate synthetase pseudogenes. J Mol Biol 192:221–233

Saitou S, Nei M (1987) The neighbor-joining method: a new method for constructing phylogenetic trees. Mol Biol Evol 4:406–425

Su T-S, Nussbaum RL, Airhart S, Ledbetter DH, Mohandas T, et al. (1984) Human chromosomal assignments for 14 argininosuccinate synthetase pseudogenes: cloned DNAs as reagents for cytogenetic analysis. Am J Hum Genet 36:954–964

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Vogel F, Rörhborn G (1966) Amino-acid substitution in haemoglobins and the mutation process. Nature 210:116–117

Wolfe KH (1991) Mammalian DNA replication: mutation biases and the mutation rate. J Theor Biol 149:441–451

Wolfe KH, Sharp PM (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441–456

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Wu C-I, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327:169–170

Zoubak S, D'Onofrio G, Cacciò S, Bernardi G, Bernardi G (1995) Specific compositional patterns of synonymous positions in homologous mammalian genes. J Mol Evol 40:293–307

Zuckerkandl E, Devancourt DJ, Vogel H (1971) Mutational trends and random process in the evolution of informational macromolecules. J Mol Biol 59:473–490