Copyright © 1998 by the Genetics Society of America

Isochore Evolution in Mammals: A Human-Like Ancestral Structure

Nicolas Galtier*,† and Dominique Mouchiroud*

Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Biométrie, Génétique et Biologie des Populations, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France and †Centre National de la Recherche Scientifique, Unité Propre de Recherche 9060, Génome et Populations, Université Montpellier 2, 34095 Montpellier Cedex, France

Manuscript received March 24, 1998
Accepted for publication September 8, 1998

ABSTRACT

Codon usage in mammals is mainly determined by the spatial arrangement of genomic G + C-content, *i.e.*, the isochore structure. Ancestral G + C-content at third codon positions of 27 nuclear protein-coding genes of eutherian mammals was estimated by maximum-likelihood analysis on the basis of a nonhomogeneous DNA substitution model, accounting for variable base compositions among present-day sequences. Data consistently supported a human-like ancestral pattern, *i.e.*, highly variable G + C-content among genes. The mouse genomic structure—more narrow G + C-content distribution—would be a derived state. The circumstances of isochore evolution are discussed with respect to this result. A possible relationship between G + C-content homogenization in murid genomes and high mutation rate is proposed, consistent with the negative selection hypothesis for isochore maintenance in mammals.

VERTEBRATE nuclear genomes are characterized by a peculiar structure regarding the spatial distribution of guanine and cytosine content (GC%): these genomes are mosaics of long, compositionally homogeneous DNA segments called isochores (Bernardi *et al.* 1985; Bernardi 1993). The isochore structure has been carefully studied in eutherian mammals, using both analytic ultracentrifugation experiments and DNA sequence analysis. Silent positions in protein-coding genes undergo no selection related to protein function and are therefore relevant markers of global evolutionary constraints applying at the isochore level (Clay *et al.* 1996). Actually, codon usage in mammals appears to be uniquely determined by local GC-richness (Sharp *et al.* 1995).

A major compositional change between human and murine genomes (rat, mouse) was found from comparative genome analysis: the variance of third codon position GC% (GC3) among protein-coding genes is higher in human than in rat and mouse (Mouchiroud *et al.* 1988; Mouchiroud and Gautier 1990); GC-rich isochores are richer in human, and GC-poor isochores are poorer. Rabbit, cow, and dog mainly show a human-like isochore structure: GC3 in protein-coding genes of these species is highly correlated to GC3 in human orthologous genes (Mouchiroud and Bernardi 1993). Robinson *et al.* (1997a) showed from relevant sampling of rodent species that the mouse/rat structure is common to the whole Muridae group. Consistent results were found from analytical ultracentrifugation (Salinas *et al.* 1986; Sabeur *et al.* 1993). Therefore, two major mammalian isochore patterns are known to date, which we call the Muridae pattern (with moderate interisochore GC% variability) and the nonrodent pattern (with high interisochore GC% variability). Taxon Muridae is a huge rodent family that includes two-thirds of rodent species and one-quarter of mammalian species (Wilson and Reeder 1993), among which are mouse, rat, and hamster.

Muridae are likely a monophyletic outgroup to the nonrodent eutherian orders whose isochore structure has been examined using DNA sequence data, namely Lagomorpha, Artiodactyla, Carnivora, and Primates, as supported by both mitochondrial and nuclear molecular data (Graur *et al.* 1991, 1996; d'Erchia *et al.* 1996; Cao *et al.* 1997; Janke *et al.* 1997; see Figure 1). Hence, the nature of the ancestral eutherian isochore structure is an open question; it may have been Muridae-like, nonrodent-like, intermediate, or different from any known present-day pattern. This question is a central one as far as mammalian molecular evolution is concerned because neither the circumstances of isochore evolution nor the evolutionary constraints maintaining this structure within mammalian genomes are well understood yet.

Recently, Galtier and Gouy (1998) devised a maximum-likelihood implementation of a new, nonhomogeneous model of DNA sequence evolution allowing diverging GC% among lineages. Given a phylogeny, a reasonable amount of information about past evolutionary modes can be extracted by this method from data sets of usual sizes. Especially, GC% at ancestral nodes

Corresponding author: D. Mouchiroud, Centre National de la Recherche Scientifique UMR 5558, Biométrie, Génétique et Biologie des Populations, Université Claude Bernard Lyon 1, 43, Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France.
E-mail: mouchi@biomserv.univ-lyon1.fr

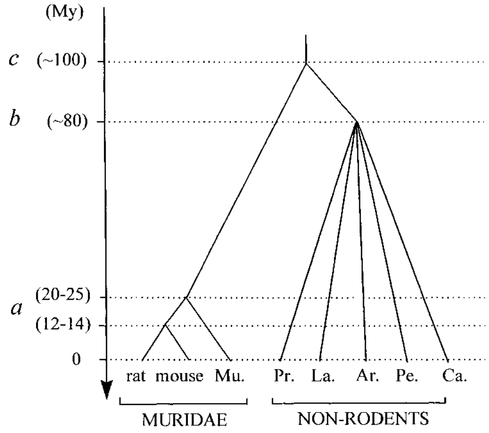

Figure 1.—Schematic phylogenetic tree of eutherian mammals. The radiation among non-Muridae eutherian lineages is considered unresolved. Mu., deeply branching Muridae; Pr., Primates; La., Lagomorpha; Ar., Artiodactyla; Pe., Perissodactyla; Ca., Carnivora. Under the molecular clock assumption, parameter ϕ is given by $(c - a)/[(c - a) + (c - b)]$, where a, b, and c are the dates of Muridae divergence, nonrodent divergence, and Muridae/nonrodent split, respectively.

is quite reliably recovered when data simulated under the model assumptions are used (Galtier and Gouy 1998). This tool appears suitable for studying mammalian isochore evolution because unequal transition and transversion and diverging GC% among lineages—*i.e.*, the very assumptions of Galtier and Gouy's model—are likely the major forces constraining synonymous substitution processes in mammalian genes.

In this article, we address the problems of recovering the ancestral eutherian isochore structure and locating the major genomic compositional changes in the mammalian phylogenetic tree by applying Galtier and Gouy's (1998) method to the third-codon positions of 27 nuclear protein-coding genes.

MATERIALS AND METHODS

Data: Sequences were extracted from the HOVERGEN database (Duret *et al.* 1994, release 25, July 1997). Orthologous genes were selected according to the number of represented species and the amount of GC3 change: at least six distinct eutherian species including at least two Muridae and two nonrodents were required, and the difference between human and mouse GC3 (both were available for all genes matching the above criteria) had to be >10% or <−10%, as suggested by Mouchiroud and Gautier (1990). These genes are likely to be reliable markers of the process of compositional change. For each gene family, orthology was checked by examining the phylogenetic tree provided by the HOVERGEN interface. Putative paralogous genes were discarded. Twenty-seven protein-coding genes were finally selected (Table 1). Non-Muridae rodent sequences were removed from data sets when available because both their phylogenetic location and isochore pattern are unclear (Nedbal *et al.* 1996; Robinson *et al.* 1997a,b). For each gene, amino acid sequences were automatically aligned using the CLUSTALW program (Thompson *et al.* 1994). Nucleotidic alignments were deduced from aligned amino acid sequences. Gap-containing sites were discarded. Phylogenetic trees were reconstructed by the neighbor-joining method (Saitou and Nei 1987) with the logdet distance (Steel 1993) applied to all three codon positions, using program PHYLO_WIN (Galtier *et al.* 1996). Tree topologies were rooted on the branch connecting Muridae to nonrodents. First- and second-codon positions were subsequently removed from the data sets. (GenBank accession numbers of the sequences used in this study are available upon request.)

Ancestral GC-content estimation: Galtier and Gouy's (1998) method was used to estimate the ancestral GC% at the third-codon position of each gene. This method relies on a new model of DNA substitution that allows varying evolutionary processes among lineages: two assumptions of usual models—namely homogeneity and stationarity—are relaxed to account for variable base compositions among present-day sequences. The assumed substitution process on any branch of a given rooted tree follows Tamura's (1992) model, with unequal equilibrium G + C contents (θ in Tamura 1992) among branches, so that GC% can vary with time and between lineages. The transition/transversion ratio is kept constant over the whole tree. The GC% at the root node (ω) and the precise location of the root on its branch (ϕ) are two additional parameters of the model. Parameter estimates are those values maximizing the likelihood of the model as defined by Felsenstein (1981). A reliable step-by-step optimization algorithm was designed to achieve this estimation task (Galtier and Gouy 1998). Accurate estimates of the ancestral GC% ω were recovered from data sets simulated under the model assumptions (Galtier and Gouy 1998), suggesting that a reasonable amount of information about past base compositions can be extracted from real data sets. For each gene, the above estimation algorithm was applied to third-codon position sequences, using the inferred neighbor-joining topology as a model tree.

To make inferences comparable among genes, ω estimates were calibrated taking into account Muridae and nonrodent mean GC3 values. A "nonrodent-likeness" index is defined,

$$\delta = \frac{\hat{\omega} - GC_M}{GC_{NR} - GC_M}, \qquad (1)$$

where $\hat{\omega}$ is the estimated ancestral GC3, GC_M is the mean GC3 among Muridae sequences, and GC_{NR} is the mean GC3 among nonrodent sequences. δ equals 0 if $\hat{\omega}$ equals GC_M or 1 if $\hat{\omega}$ equals GC_{NR}; δ is >1 if $\hat{\omega}$ is more extreme than GC_{NR} (*i.e.*, outside the $[GC_M, GC_{NR}]$ interval on the nonrodent side), and <0 if $\hat{\omega}$ is more median than GC_M (*i.e.*, outside the $[GC_M, GC_{NR}]$ interval on the Muridae side). GC_{NR} was computed taking phylogeny into account: equal weights were given to all nonrodent orders whatever the number of represented species in each order. A similar procedure was used to compute GC_M when the number of available Muridae species was >3.

RESULTS

Model/data adequacy: Twenty-seven genes were analyzed (Table 1). The total number of species varied from 6 to 16 among genes. The number of Muridae species was <4 excepting gene LCAT (Robinson *et al.* 1997b, 13 Muridae species). Nonrodent sampled orders were Primates, Lagomorpha, Carnivora, Artiodactyla, and Perissodactyla. GC3 differences between nonrodents and Muridae were positive in 17 genes out of 27; these genes are expected to be located in GC-rich isochores. Two major assumptions of Galtier and Gouy's (1998) model were empirically checked: A% = T% and C% =

TABLE 1

Ancestral GC% estimation for 27 mammalian genes

	Data					Estimates				
Gene[a]	Length[b]	n_M[c]	n_{NR}[d]	GC_M[e]	GC_{NR}[e]	$\hat{\omega}$[e]	δ[e]	$\hat{\omega}_{0.2}$[e]	$\hat{\omega}_{0.5}$[e]	$\hat{\omega}_{0.8}$[e]
alb[f]	417	2	7	53.5	43.6	45.3	0.84	45.6	45.2	44.7
alg[f]	141	2	7	68.4	86.2	90.7	1.25	86.3	90.5	90.7
amp[f]	953	2	4	67.8	81.6	85.1	1.25	80.3	84.8	85.0
ana[f]	314	2	6	72.2	82.8	94.3	2.09	93.3	94.0	94.3
apoa[f]	256	2	5	71.5	86.4	88.9	1.17	86.5	88.8	88.9
apoe[f]	305	2	5	76.9	90.9	95.3	1.31	93.5	95.2	95.3
atp[f]	302	2	4	67.1	57.2	61.0	0.61	61.0	60.8	59.7
b3a[f]	317	2	4	73.5	87.0	84.1	0.79	83.7	84.0	85.2
ckit[f]	950	2	4	60.7	50.7	52.5	0.83	53.5	52.2	51.1
cytp[f]	496	3	5	66.5	80.4	78.9	0.89	78.8	78.8	78.9
cytr[f]	666	3	3	75.2	89.1	89.3	1.01	84.3	88.6	89.1
dip[f]	409	2	4	67.8	84.4	83.8	0.97	78.9	83.3	83.8
inl5[f]	128	3	5	59.2	42.0	53.1	0.36	52.6	51.7	47.8
lcat[f]	260	13	3	59.7	76.2	72.0	0.75	67.5	70.3	72.1
prol[f]	223	3	6	52.2	64.1	57.6	0.45	57.5	57.7	59.7
pros[f]	181	2	5	74.3	86.1	91.5	1.46	87.0	91.2	91.4
rbp[f]	191	2	4	70.4	82.7	79.1	0.71	78.5	79.9	82.4
star[f]	284	3	3	64.4	75.7	73.1	0.77	73.9	73.1	72.3
thy[f]	138	2	6	55.8	45.0	47.2	0.80	46.9	47.3	47.3
inf[g]	151	3	8	61.1	49.1	40.7	1.69	51.8	44.2	41.2
lut[g]	139	2	7	69.8	79.9	92.8	2.28	82.8	91.2	92.4
timp[g]	203	2	7	62.1	75.9	78.2	1.13	71.7	77.7	78.8
angi[h]	359	4	4	64.9	49.5	62.8	0.14	62.1	59.3	53.4
inl2[h]	148	3	9	56.4	42.3	54.0	0.17	54.0	51.5	45.2
inl6[h]	193	2	12	46.3	56.1	47.4	0.08	48.1	47.2	56.9
ldha[h]	331	2	4	63.8	51.4	62.2	0.13	62.5	59.6	53.6
pit1[h]	283	2	6	53.5	40.3	44.2	0.70	49.1	44.3	42.2

[a] Gene names: *alb, serum albumin*; *alg, adult alpha globin*; *amp, aminopeptidase n/CD13*; *ana, anaphylatowin receptor C5*; *apoa, apolipoprotein A-1*; *apoe, apolipoprotein E*; *atp, Na+/K+ ATPase beta*; *b3a, beta 3 adregenic receptor*; *ckit, c-kit proto-oncogene*; *cytp, cytochrome P450c11*; *cytr, cytochrome P450 reductase*; *dip, dipeptidase*; *inl5,interleukin 5*; *lcat, lecithin-cholesterol acyltransferase*; *prol, prolactine*; *pros, prostaglandin D synthase*; *rbp, retinol binding protein*; *star, steroidogenic acute regulatory protein*; *thy, beta thyrotropin*; *inf, gamma interferon*; *lut, beta luteinizing hormone*; *timp, tissu inhibitor of metalloproteinase*; *angi, angiotensin II receptor type I*; *inl2, interleukin 2*; *inl6, interleukin 6*; *ldha, lactate dehydrogenase A*; *pit1, transcription factor pit-1*.

[b] Number of analyzed third-codon positions.

[c] Number of Muridae representative species.

[d] Number of nonrodent representative species.

[e] See definition in materials and methods.

[f] Root-insensitive genes.

[g] Root-sensitive genes unambiguously supporting a nonrodent-like ancestral pattern.

[h] Root-sensitive genes supporting either a nonrodent-like or Muridae-like ancestral pattern depending on the location of the root.

G% equalities, and constant substitution rate among sites.

In Tamura's (1992) substitution model (and consequently in Galtier and Gouy's model as well), sequence base composition is described by a single parameter, namely G + C content, so that A% = T% and C% = G% are the underlying assumptions. The AT (respectively GC)-skewness of the analyzed sequences was computed:

$$\text{AT-skewness} = \frac{A\% - T\%}{A\% + T\%}$$

$$\text{GC-skewness} = \frac{G\% - C\%}{G\% + C\%}. \qquad (2)$$

The distributions of both statistics over 219 compared genes are shown (Figure 2). Values >0.4 or <−0.4 are infrequent.

The amount of variation of substitution rates among sites was investigated by fitting to the data a substitution model involving gamma-distributed rates over sites (Yang 1994): the estimated shape parameter of the assumed gamma distribution (α in Yang 1994) gives insight about how variable rates are among sites. The

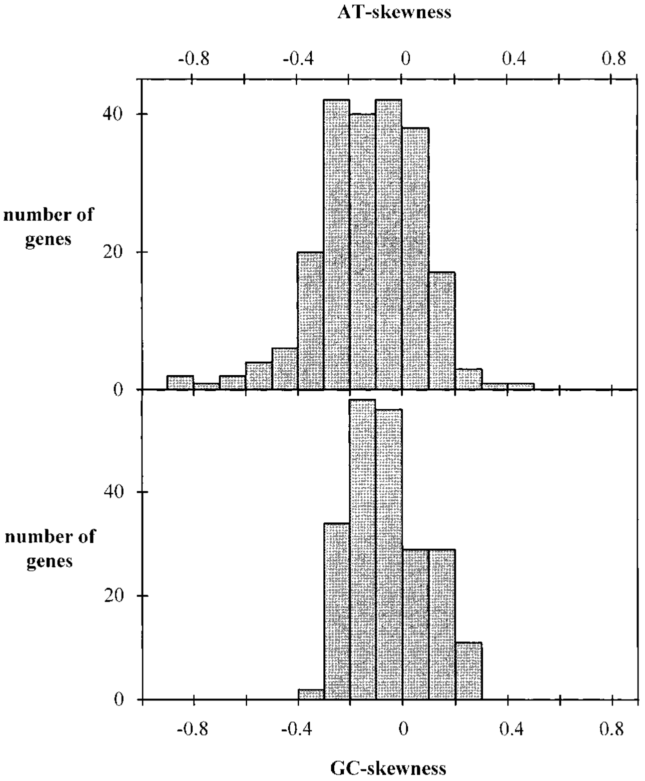

Figure 2.—Distribution of AT-skewness (top) and GC-skewness (bottom) among 219 mammalian third-codon-position sequences used in the present study.

mean estimated α value over 27 data sets was 2.58. α was >1 for 26 genes out of 27. Such values are characteristic of bell-shaped, low variance, distributions.

Simulations: Preliminary analyses above suggest that the major part of the data essentially meets the model assumptions. Nevertheless, because the estimation method has been assessed only under conditions matching exactly the model assumptions, one may wonder about the reliability of inferences when these assumptions are (slightly) violated. We conducted a simulation study to assess the robustness of the above method to departures from the A% = T%, G% = C%, and equal-rates-among-sites assumptions. The tree of Figure 1 was used as a model tree topology, with branch lengths proportional to those of Figure 1. Two distinct processes were performed, each one aiming to simulate the evolution of the third-codon positions of a G + C-rich gene. In the first procedure ("human-like ancestor"), a 300-nucleotide-long ancestral sequence was randomly drawn with 80% average G + C-content. The diverging evolution of eight sequences was simulated according to the "HKY + gamma" model (Hasegawa *et al.* 1985; Yang *et al.* 1994). This is a generalization of Tamura's model allowing unequal A- and T- (respectively G- and C-) pressures—resulting in sequences with nonzero expected AT (GC)-skewness—and unequal evolutionary rates among sites. Equilibrium G + C-content (*i.e.*, GC-pressure) was set to 80% in the nonrodent lineage, leaving the expected base composition unchanged, and to 25% in the Muridae lineage, so that the G + C-content in present-day Muridae sequences was around 65%. In the second simulation process ("mouse-like ancestor"), the ancestral G + C-content was set to 65%. It was left unchanged in the Muridae lineage but was increased toward 80% in the nonrodent lineage. The transition/transversion ratio was set to 4.

In both processes, AT- and GC-skewness varied from 0 (*i.e.*, A%/T% = 1) to 0.8 (*i.e.*, A%/T% = 9). Equal AT- and GC-skewness was assumed. The effect of among-site rate variation was also examined: equal rates, moderately variable rates (shape parameter of the gamma distribution set to 2.5), and highly variable rates (shape parameter set to 0.5) were used. Ancestral G + C-contents were estimated by applying the above-described maximum-likelihood method to the simulated data sets. Ten replicates were performed for each combination of AT (GC)-skewness and among-site heterogeneity. The method was considered successful when the estimated ancestral G + C-content was closer to the true value than to the mean G + C-content in present-day sequences with diverged base composition (namely Muridae sequences in process 1, nonrodent sequences in process 2). The number of successes out of 10 replicates is shown (Table 2), together with the mean estimated ancestral G + C-content.

The method appeared to be biased when AT (GC)-skewness was higher than 0.6. No sensitivity to among-site rate variation was found. Therefore, no significant bias is expected in the present work because most of the analyzed sequences have observed AT (GC)-skewness <0.4.

Ancestral GC% estimation: The main results are shown in Table 1. Ancestral GC3s estimated according to Galtier and Gouy's (1998) algorithm ($\hat{\omega}$) are given, as well as their calibrated version δ. The results clearly support the hypothesis of a nonrodent-like ancestral isochore pattern. $\hat{\omega}$ is closer to GC_{NR} than to GC_{M} ($\delta >$ 0.5) for 21 genes out of 27. The mean δ value over 27 genes is 0.85. Additional analyses were conducted to check the robustness of this general result.

Sensitivity to the location of the root: The actual location of the root on its branch ϕ (*i.e.*, the fraction of the root branch length lying on a given side of the root, say the Muridae side) was not reliably estimated when simulated data sets were used in Galtier and Gouy (1998). This has little effect on the estimation of ancestral GC% ω if the root branch is short. However, if a significant amount of the evolutionary change occurs along the root branch, inaccurate estimates of ϕ may mislead ω estimation. Because the branch connecting Muridae to nonrodents was generally long for the present data sets, we checked the sensitivity of the ω estimate to the location of the root: three additional estimation procedures were performed assuming fixed values for

TABLE 2

Sensitivity of the estimation method to AT(GC)- skewness and among-site rate variation assessed by simulations

	Human-like ancestor[a]			Mouse-like ancestor[b]		
AT- and GC-skewness	∞[c]	2.5	0.5	∞	2.5	0.5
0	80.0/10	80.7/10	78.8/10	63.9/10	67.3/9	65.7/10
0.2	80.2/10	81.7/10	80.9/10	66.0/9	69.5/10	69.0/8
0.4	79.1/10	78.9/10	80.0/10	64.0/9	69.8/6	68.8/7
0.6	76.7/8	73.9/7	79.3/9	73.0/5	72.8/4	72.7/3
0.8	76.4/8	77.2/8	78.9/10	72.3/6	74.6/3	73.9/2

[a] True ancestral G + C-content, 80%.

[b] True ancestral G + C-content, 65%.

[c] Shape parameters of the assumed gamma distribution among sites. Infinite means constant rates. Values given are mean estimated ancestral G + C-content (%)/number of successes (out of 10).

parameter ϕ, namely 0.2, 0.5, and 0.8. A 0.8 ϕ value means that the branch connecting the root to the ancestral Muridae node is four times longer than the branch connecting the root to the ancestral nonrodent node (Figure 1). The resulting ω estimates are called $\hat{\omega}_{0.2}$, $\hat{\omega}_{0.5}$, and $\hat{\omega}_{0.8}$, respectively (Table 1). ϕ-Dependent nonrodent-likeness $\delta_{0.2}$, $\delta_{0.5}$, and $\delta_{0.8}$ can be defined by replacing $\hat{\omega}$ by $\hat{\omega}_{0.2}$ (respectively $\hat{\omega}_{0.5}$, $\hat{\omega}_{0.8}$) in Equation 1.

Genes were classified into three groups depending on the sensitivity of the ω estimate to the ϕ value. For 19 genes out of 27, the highest difference between $\hat{\omega}_{0.2}$, $\hat{\omega}_{0.5}$, and $\hat{\omega}_{0.8}$ values was $<5\%$. These genes are called root insensitive (Table 1, top). The remaining 8 genes were again split into two groups. For 3 of them, the estimated ancestral GC3 was closer to GC_{NR} than to GC_{M} whatever the assumed ϕ value: $\delta_{0.2}$, $\delta_{0.5}$, and $\delta_{0.8}$ are >0.5 (Table 1, middle). For the last 5 genes, the ω estimate appears highly sensitive to the location of the root: $\hat{\omega}$ was either Muridae-like or nonrodent-like depending on ϕ (Table 1, bottom). Seventeen root-insensitive genes out of 19 supported the nonrodent ancestral pattern hypothesis; the mean δ value over 19 root-insensitive genes was 0.96. Inversely, genes supporting the Muridae ancestral pattern hypothesis generally belong to the root-sensitive group (4 out of 6): inferences vary depending on ϕ. When ϕ is set to 0.8, the estimated ancestral GC3 is closer to GC_{NR} than to GC_{M} for all 27 genes. These results provide additional support for the nonrodent ancestral isochore pattern hypothesis because genes supporting an alternative scheme appear less reliable regarding ω estimation and may have been misled by wrong ϕ estimates.

Obtaining *a priori* estimates for ϕ would help in interpreting root-sensitive results. This is not an easy task because it depends on species sampling, speciation times, and evolutionary rates. However, let us consider those genes in the data set (18 out of 27) where Muridae are represented by mouse and rat. If constant evolutionary rates among lineages are assumed (the molecular clock hypothesis), accepting 13, 80, and 100 mya as rough date estimates for, respectively, mouse/rat divergence (Jacobs and Downs 1994), nonrodent radiation, and murid/nonrodent divergence would result in $\phi = 0.81$ (Figure 1). Using the estimated date of Muridae radiation (20–25 mya, Hugueney and Mein 1993) rather than mouse/rat split gives $\phi = 0.79$–0.80. Accounting for the higher substitution rate in Muridae (Li *et al.* 1996) would increase these ϕ values. These are imprecise estimates. However, they suggest that ϕ values >0.5 are more likely than low ϕ values. This reinforces the nonrodent ancestral pattern hypothesis since ω estimates for root-sensitive genes are closer to GC_{NR} when ϕ is high.

Sensitivity to species sampling: The above results could be a consequence of unbalanced species sampling: nonrodent species are generally more numerous than Muridae in our data set, which might bias ω estimation toward the GC_{NR} value. Three genes, namely *albumin* (9 represented species), *interleukin 6* (14 species), and LCAT (16 species), were carefully studied to check this putative methodological artifact. Incomplete data sets were set up by removing species, and the above analyses were performed again. Results are shown in Table 3. Both $\hat{\omega}$ value and root-sensitivity remain almost unchanged when the number of species decreases, suggesting that asymmetric species sampling is unlikely to have misled the above findings.

Sensitivity to phylogeny: ω estimates may also be sensitive to the assumed phylogenetic tree. For three genes, we reconducted the above analyses after modifying the assumed tree topology. In each case, five distinct trees were randomly drawn by modifying those branching orders not considered firmly established. For genes *albumin* and *interleukin 6*, branching orders among nonrodent orders were randomly shifted. For gene *LCAT*, internal branches supported by bootstrap percentages >80 when all three codon positions are analyzed by the neighbor-joining method (logdet distance) were kept, but less supported branching orders were randomly shifted. Results are given in Table 4: the first line in

TABLE 3

Ancestral GC% estimation with variable number of species

Data					Estimates			
Gene	n_{m}	n_{NR}	$\mathrm{GC_M}$	$\mathrm{GC_{NR}}$	$\hat{\omega}$	$\hat{\omega}_{0.2}$	$\hat{\omega}_{0.5}$	$\hat{\omega}_{0.8}$
alb	2	7	53.5	43.6	45.3	45.6	45.2	44.7
alb	2	4	53.5	43.1	46.5	45.7	45.5	45.2
inl6	2	12	46.3	56.1	47.4	48.1	47.2	56.9
inl6	2	8	46.3	60.1	46.0	48.9	49.3	56.2
inl6	2	4	46.3	58.7	47.3	50.6	51.3	54.8
lcat	13	3	59.7	76.2	72.0	67.5	70.3	72.1
lcat	8	3	60.7	76.2	72.9	68.5	71.4	73.0
lcat	3	3	61.2	76.2	74.3	66.6	71.8	74.3

each part of the table recalls the initial ω estimate, and the next five lines are for modified trees. Again, ω estimates are remarkably stable when the assumed phylogenetic tree varies.

DISCUSSION

Locating the compositional change in the mammalian tree: When applied to 27 protein-coding genes showing variable GC3 in Muridae and nonrodent mammals, Galtier and Gouy's (1998) method unambiguously recovers a nonrodent-like ancestral pattern. Data are highly self-consistent with respect to this result: all 27 examined genes support the nonrodent-like ancestor hypothesis when a plausible ϕ value is assumed, which is quite unlikely to occur by chance. This result is not sensitive to species sampling, nor to assumed phylogenetic trees. Our method accurately recovered ancestral GC-contents from data sets simulated under the underlying model (Galtier and Gouy 1998). It also performed well when the model assumptions were slightly violated (this study). Therefore, we think that the above inferences deserve reasonable confidence.

TABLE 4

Ancestral GC% estimation with variable phylogeny

Data			Estimate:
Gene	$\mathrm{GC_M}$	$\mathrm{GC_{NR}}$	$\hat{\omega}$
alb	53.5	43.7	45.3
alb-1	53.5	43.7	47.8
alb-2	53.5	43.7	46.2
alb-3	53.5	43.7	47.1
alb-4	53.5	43.7	45.5
alb-5	53.5	43.7	47.2
il6	46.6	56.1	47.4
il6-1	46.6	56.1	47.2
il6-2	46.6	56.1	46.1
il6-3	46.6	56.1	48.7
il6-4	46.6	56.1	46.0
il6-5	46.6	56.1	46.0
lcat	59.7	76.2	72.0
lcat-1	59.7	76.2	72.1
lcat-2	59.7	76.2	72.0
lcat-3	59.7	76.2	72.5
lcat-4	59.7	76.2	72.4
lcat-5	59.7	76.2	72.6

Within-genome GC-content heterogeneity is far lower in fishes and amphibians than in mammals or birds (Bernardi and Bernardi 1990; Bernardi *et al.* 1997). Actually, the cold-blooded *vs.* warm-blooded discrepancy is the most striking pattern of isochore structure variability among vertebrates. This result, together with the deep phylogenetic location of Muridae among eutherian mammals, raised an attractive hypothesis: Muridae may have kept an "intermediate" state between poorly structured cold-blooded and highly structured nonrodent mammals (for example, Saccone *et al.* 1997). The present results suggest that this hypothesis should be dismissed. We propose that a highly structured genomic GC% compartmentation is ancestral in eutherian mammals. This structure has remained unchanged in many eutherian orders, while Muridae have been (are) undergoing a "reversal" toward a less-structured state, *i.e.*, GC-homogenization. A brief look at the eutherian phylogenetic tree (Figure 1) shows that the latter hypothesis, although less parsimonious than the alternative, is not less likely on general grounds: assuming a Muridae-like ancestral pattern would require that the nonrodent-like pattern either evolved during a short period of time—which is unlikely as far as the whole genomic structure is concerned—or evolved independently in several lineages—which is less parsimonious than the human-like ancestor hypothesis.

Several peculiar characteristics of the genomic evolution of rodents (and especially Muridae) have been reported, including a high rate of chromosomic rearrangements (Viegas-Pequinot *et al.* 1986), more numerous and longer microsatellites (Duret *et al.* 1995), and erosion of CpG islands (Matsuo *et al.* 1993; Cross *et al.* 1997). Most importantly, relative rate tests showed that both synonymous and nonsynonymous substitution rates in nuclear genes are significantly (possibly 10 times) higher in the rat / mouse lineage than in primates (Li *et al.* 1996). This discrepancy is likely a consequence of higher mutation rate; rodents show less efficient repair of DNA lesions (Hart and Setlow 1974) and lower generation time than most mammals. Therefore, the GC-homogenization we report occurred in the eutherian lineage undergoing the highest mutation rate. Both phenomena may be related, as we discuss below.

Isochore evolution in mammals: Debates rage on about the questions of isochore evolution and maintenance in mammalian genomes. Two main hypotheses compete. Bernardi has consistently argued that a negative selection pressure was acting to maintain this structure (see Bernardi 1993; Zoubak *et al.* 1995;

Alvarez-Valin *et al.* 1998); a possible relationship between isochore evolution and endothermy has been suggested. Sharp and colleagues criticized this view and proposed that the isochore structure is a consequence of varying mutational pattern within genomes (see Sharp *et al.* 1995). Several neutral, putatively isochore-forming molecular mechanisms were put forward, including varying free deoxyribonucleotide concentration through the cell cycle (Wolfe *et al.* 1989; but see Eyre-Walker 1992) and unequal repair pattern across chromosomes (Holmquist and Filipski 1994). Locating the compositional change on the Muridae branch in the mammalian tree provides a new argument. If isochores were the consequence of varying neutral mutation pattern across genomes, increasing mutation rate should result in increasing GC-heterogeneity. The opposite pattern is found: the isochore structure gets weaker in fast-evolving Muridae. This result, however, appears consistent with the alternative hypothesis. If negative selection against slightly deleterious variants is contributing to isochore maintenance, less structure is expected when DNA repair is less efficient because more deleterious mutations per generation accumulate.

The GC-content at third-codon positions of eutherian nuclear protein-coding genes is highly variable within genomes (23 to 98% in human) and highly correlated between genomes: the isochore structure appears well conserved among eutherian orders (Mouchiroud and Bernardi 1993). In one lineage, incidentally the ones evolving the fastest, GC% homogenized. GC3 homogenization in Muridae may be due to the loss of evolutionary constraints maintaining isochores in mammals, either mutational or selective. However, we still need to understand why such GC-pressure applying to most eutherian orders would slow down in murids. A less *ad hoc* hypothesis involves a relationship between GC-homogenization and high mutation rate, two peculiar features of Muridae genomic evolution. This scheme is consistent with selected *vs.* strictly neutral isochore maintenance in mammalian genomes.

The selective hypothesis leaves two points unexplained. First, we do not know which selective advantage may arise from highly structured isochores. Especially, is isochore evolution related to endothermy? Genome analysis of additional cold-blooded vertebrate species, *e.g.*, crocodilians, should help address this question. Second, the way selection may act within mammalian populations is unclear. Sharp *et al.* (1995) pointed out that a major role of natural selection is unlikely because the selective advantage of point mutations must be very low and effective population sizes small, so that random genetic drift should overcome selection. This criticism remains unanswered. Thus, the mechanisms of isochore evolution are far from being elucidated. This question is a challenging one for both molecular biologists and population geneticists.

LITERATURE CITED

Alvarez-Valin, F., K. Jabbari and G. Bernardi, 1998 Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. **46:** 37–53.

Bernardi, G., 1993 The vertebrate genome: isochores and evolution. Mol. Biol. Evol. **10:** 186–204.

Bernardi, G., and G. Bernardi, 1990 Compositional patterns in the nuclear genome of cold-blooded vertebrates. J. Mol. Evol. **31:** 265–281.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas *et al.*, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

Bernardi, G., S. Hughes and D. Mouchiroud, 1997 The major compositional transition in the vertebrate genome. J. Mol. Evol. **44:** 44–51.

Cao, Y., N. Okada and M. Hasegawa, 1997 Phylogenetic position of Guinea pigs revisited. Mol. Biol. Evol. **14:** 461–464.

Clay, O., S. Caccio, S. Zoubak, D. Mouchiroud and G. Bernardi, 1996 Human coding and non-coding DNA: compositional correlations. Mol. Phylogenet. Evol. **5:** 2–12.

Cross, S. H., M. Lee, V. H. Clark, J. M. Craig, A. P. Bird *et al.*, 1997 The chromosomal distribution of CpG islands in the mouse: evidence for genome scrambling in the rodent lineage. Genomics **40:** 454–461.

D'Erchia, A. M., C. Gissi, G. Pesole, C. Saccone and U. Arnason, 1996 The guinea-pig is not a rodent. Nature **381:** 597–600.

Duret, L., D. Mouchiroud and M. Gouy, 1994 HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. **22:** 2360–2365.

Duret, L., D. Mouchiroud and C. Gautier, 1995 Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40:** 308–317.

Eyre-Walker, A., 1992 Evidence that both G + C rich and G + C poor isochores are replicated early and late in the cell cycle. Nucleic Acids Res. **20:** 1497–1501.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Galtier, N., and M. Gouy, 1998 Inferring pattern and process: maximum-likelihood implementation of a new, non-homogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15:** 871–879.

Galtier, N., M. Gouy and C. Gautier, 1996 SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comp. Appl. Biosci. **12:** 543–548.

Graur, D., W. A. Hide and W.-H. Li, 1991 Is the guinea-pig a rodent? Nature **351:** 649–651.

Graur, D., L. Duret and M. Gouy, 1996 Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). Nature **379:** 333–335.

Hart, R. W., and R. B. Setlow, 1974 Correlation between deoxyribonucleic acid excision-repair and life-span in a number of mammalian species. Science **71:** 2169–2173.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Holmquist, G. P., and J. Filipski, 1994 Organization of mutations along the genome: a prime determinant of genome evolution. TREE **9:** 65–68.

Hugueney, M., and P. Mein, 1993 A comment on the earliest Spalacinae (Rodentia, Muroidea). J. Mammal. Evol. **1:** 215–223.

Jacobs, L. L., and W. R. Downs, 1994 The evolution of murine rodents in Asia, pp. 149–156 *in Rodent and Lagomorph Families of Asian Origin and Diversification*, edited by Y. Tomida, C. Li and T. Setoguchi. National Science Museum, Tokyo.

Janke, A., X. Xu and U. Arnason, 1997 The complete mitochondrial genome of Wallaroo (Macropus robustus) and the phylogenetic relationship among Monotrema, Marsupialia and Eutheria. Proc. Natl. Acad. Sci. USA **94:** 1276–1281.

Li, W.-H., D. L. Ellsworth, J. Krushkal, B. J. H. Chang and D. Hewett-Emett, 1996 Rates of nucleotide substitution in primates and rodents and the generation time hypothesis. Mol. Phylogenet. Evol. **5:** 182–187.

Matsuo, K., O. Clay, T. Takahashi, J. Silke and W. Schaffner, 1993 Evidence for erosion of mouse CpG island during mammalian evolution. Somatic Cell. Mol. Genet. **19:** 543–555.

Mouchiroud, D., and G. Bernardi, 1993 Compositional properties of coding sequences and Mammalian phylogeny. J. Mol. Evol. **37:** 109–116.

Mouchiroud, D., and C. Gautier, 1990 Codon usage changes and sequence dissimilarity between human and rat. J. Mol. Evol. **31:** 81–91.

Mouchiroud, D., C. Gautier and G. Bernardi, 1988 The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol. Evol. **27:** 311–320.

Nedbal, M. A., R. L. Honeycutt and D. A. Schlitter, 1996 Higher-level systematics of rodents (Mammalia:Rodentia): evidence from the mitochondrial 12S rRNA gene. J. Mammal. Evol. **3:** 201–237.

Robinson, M., C. Gautier and D. Mouchiroud, 1997a Evolution of isochores in rodents. Mol. Biol. Evol. **14:** 823–828.

Robinson, M., F. Catzeflis, J. Briolay and D. Mouchiroud, 1997b Molecular phylogeny of rodents, with special emphasis on murids: evidence from nuclear gene LCAT. Mol. Phylogenet. Evol. **8:** 423–434.

Sabeur, G., G. Macaya, F. Kadi and G. Bernardi, 1993 The isochore pattern of mammalian genomes and their phylogenetic implications. J. Mol. Evol. **37:** 93–108.

Saccone, C., S. Caccio, P. Perani, L. Andreozzi, A. Rapisarda *et al.*, 1997 Compositional mapping of mouse chromosomes and the identification of the gene-rich regions. Chromosome Res. **5:** 293–300.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Salinas, J., M. Zerial, J. Filipski and G. Bernardi, 1986 Gene distribution and nucleotide sequence organization in the mouse genome. Eur. J. Biochem. **160:** 469–478.

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. Ser. B Biol. Sci. **349:** 241–247.

Steel, M. A., 1993 Recovering a tree from the leaf colorations it generates under a Markov model. Appl. Math. Lett. **7:** 19–23.

Tamura, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol. Biol. Evol. **9:** 678–687.

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:** 4673–4680.

Viegas-Pequinot, E., D. Petit, T. Benazzou, M. Prod'Homme, M. Lombard *et al.*, 1986 Phylogénie chromosomique chez les Sciuridés, Gerbillidae et Muridae, et étude d'espèces appartenant à d'autres familles de Rongeurs. Mammalia **50:** 164–202.

Wilson, D. E., and D. M. Reeder, 1993 *Mammal Species of the World. A Taxonomic and Geographic Reference.* Smithsonian Institution Press, Washington, DC and London.

Wolfe, K. W., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Yang, Z., 1994 Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 105–111.

Yang, Z., N. Goldman and A. Friday, 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **12:** 451–458.

Zoubak, S., G. d'Onofrio, S. Caccio, G. Bernardi and G. Bernardi, 1995 Specific compositional pattern of synonymous positions in homologous mammalian genes. J. Mol. Evol. **40:** 293–307.

Communicating editor: G. A. Churchill