

Letters to the Editor

Expected Relationship Between the Silent Substitution Rate and the GC Content: Implications for the Evolution of Isochores

Gwenaël Piganeau, Dominique Mouchiroud, Laurent Duret, Christian Gautier

Laboratoire de Biométrie et de Biologie Evolutive, UMR-CNRS 5558, Université Claude Bernard, 69622 Villeurbanne Cedex, France

Received: 28 March 2001 / Accepted: 16 May 2001

Abstract. The relationship between the silent substitution rate (K_s) and the GC content along the genome is a focal point of the debate about the origin of the isochore structure in vertebrates. Recent estimation of the silent substitution rate showed a positive correlation between K_s and GC content, in contradiction with the predictions of both the regional mutation bias model and the selection or biased gene conversion model. The aim of this paper is to help resolve this contradiction between theoretical studies and data. We analyzed the relationship between K_s and GC content under (1) uniform mutation bias, (2) a regional mutation bias, and (3) mutation bias and selection. We report that an increase in K_s with GC content is expected under mutation bias because of either nonequilibrium of the isochore structure or an increasing mutation rate from AT toward GC nucleotides in GC-rich isochores. We show by simulations that CpG deamination tends to increase the mutation rate with GC content in a regional mutation bias model. We also demonstrate that the relationship between K_s and GC under the selectionist or biased gene conversion model is positive under weak selection if the mutation selection equilibrium GC frequency is less than 0.5.

Key words: Substitution rate — Evolution of isochores — Mutation bias — CpG deamination

Introduction

The relationship between the silent substitution rate (K_s) and the GC content of the genome is a focal point of the debate about the origin of the isochore structure in vertebrates (Bielawski et al. 2000; Hurst and Williams 2000). Isochores (for review see Bernardi 2000) were defined as regions with a homogeneous GC content and thought to be a common feature of the genomes of amniotes (Hughes et al. 1999). Analysis of the sequence of the human genome has ruled out a strict notion of isochores as compositionally homogeneous as defined by Bernardi (2000) but the genome clearly contains large regions of a distinctive GC content [International Human Genome Sequencing Consortium (IHGSC) 2001]. We retain the word “isochore” to designate these “GC content domains” (IHGSC 2001). Currently, three hypotheses compete to account for the emergence of the isochore structure: regional mutation biases along the genome (Sueoka 1988; Wolfe et al. 1989; Holmquist and Filipiski 1994; Francino and Ochman 1999), biased gene conversion (Holmquist 1992; Eyre-Walker 1993), and selection over GC nucleotides (Bernardi et al. 1985; Eyre Walker 1999). The molecular processes able to generate regional mutation biases along the genome are not identified yet. One hypothesis discussed considers that a regional mutation bias could occur because of a variation in the relative concentration of GC versus AT nucleotides during replication (Wolfe et al. 1989; Gu and Li 1994). This model predicts an inversed “u” or “v” relationship between K_s and the equilibrium GC content of

the sequence (Gu and Li 1994). But this model relies on the assumption that equally GC-rich isochores replicate at the same moment, which has not been confirmed by experimental observations (Eyre Walker 1992). On the other hand, if there is selection pressure favoring a high GC content, K_s is expected to decrease with GC content for most of the parameter values of mutation and selection coefficients (Hurst and Williams 2000). This prediction also holds for the biased gene conversion model of the isochore structure, as biased gene conversion and selectionist models lead to equivalent modeling assuming independence between selected sites (Nagylaki 1983).

Recent estimation of K_s by a maximum-likelihood method (Goldman and Yang 1994; Yang and Nielsen 1998) showed a positive correlation between K_s and GC content (Bielawski et al. 2000; Hurst and Williams 2000), in contradiction to the predictions of both the regional mutation bias model of Gu and Li (1994) and the selectionist or biased gene conversion model. Here we report the relationship between K_s and GC content under (1) a uniform mutation bias model, (2) a regional mutation bias model, and (3) a selectionist or biased gene conversion model and discuss the rationale of the parameter space leading to an increase in K_s and GC content.

Uniform Mutation Bias Model

The directional mutation pressure theory was introduced by Sueoka (1962) to explain the homogeneity within genomes of bacterial species and the heterogeneity between bacterial species. This theory was also proposed to explain the isochore structure in the genome of vertebrates (Sueoka 1988). Let us call v the rate of mutation from GC to AT nucleotides and u the mutation rate from AT to GC nucleotides. Theoretically, u and v depend on base frequencies, however, if strand symmetry is assumed (that is, $A = T$ and $G = C$), this dependency vanishes. Supporting this hypothesis, strand symmetry has been observed in primate sequences (Francino and Ochmann 2000). So if we call f the frequency of GC in a sequence, and ignoring the other types of mutations (transversions between A and T and between G and C), an assumption supported by data analysis [they represent less than 6% of the total number of substitutions in repeated elements (IHGSC 2001)], at any generation, the number of mutations m is

$$m = vf + u(1 - f) = u + f(v - u) \quad (1)$$

The equilibrium GC frequency f^* is given by (Sueoka 1962)

$$f^* = u/(u + v) \quad (2)$$

If there is no selection, the substitution rate is expected to be equal to the neutral mutation rate if $Nm \ll 1$, where N is the effective population size (Kimura 1968).

From Eq. (1) it follows that K_s increases with f if $v > u$, that is, $f^* < 0.5$, and decreases with f if $v < u$ ($f^* > 0.5$). So if we assume that there is only one constant biased mutation rate for the genome so that $f^* < 0.5$ [the average GC content of the human genome is 0.41 (IHGSC 2001)], K_s is expected to increase with the GC content along the genome. But in this case, the GC-rich isochores are expected to disappear.

The isochore structure is known to have evolved differently among the vertebrates. For example, the isochore structure of the rodent genome is attenuated compared to the isochore structure of the human genome: the GC-richer regions are less rich and the AT-richer regions are GC richer (Mouchiroud et al. 1988). It is therefore possible that the GC content of isochores keeps on changing in mammalian species. A previous analysis of substitution in three retropseudogenes concluded that the GC content of the analyzed sequences was at equilibrium (Casane et al. 1997). But a wider study on the pattern of nucleotide substitution in repeated elements (IHGSC 2001) provides strong support for a nonequilibrium isochore structure on a larger sample. Assuming equilibrium base composition, under any model of evolution, the number of substitutions toward AT should equal the number of substitutions toward GC. The analyses reveal that there is an overall substitution bias from GC to AT substitutions, increasing in the GC-poorer regions. The rate of substitution leads to $f^* < 0.5$ for each class of GC content (IHGSC 2001), so that a positive relationship between K_s and GC content is expected. Abandoning the equilibrium hypothesis has important consequences, as most of the estimations of the substitution rates between sequences assume that the base content of the sequence is at equilibrium (e.g., Yang 1994). Furthermore, the rejection of the regional mutation bias hypothesis to explain the isochore structure (Eyre Walker 1999) no longer holds if an equilibrium GC content is not achieved (Eyre Walker 1997): assuming equilibrium, the number of polymorphic sites segregating toward AT should equal the number of polymorphic sites segregating toward GC, but nonequilibrium could well explain the biased polymorphism pattern.

Regional Mutation Bias Model

Now if one assumes that the isochores result from regional mutation patterns along the genome (Sueoka 1988; Wolfe et al. 1989; Holmquist and Filipinski 1994), then the mutation rate in each region at equilibrium GC content, m_i^* , is given by

$$K_{s_i}^* = m_i^* = 2v_i u_i / (u_i + v_i) = 2v_i f_i^* = 2u_i (1 - f_i^*) \quad (3)$$

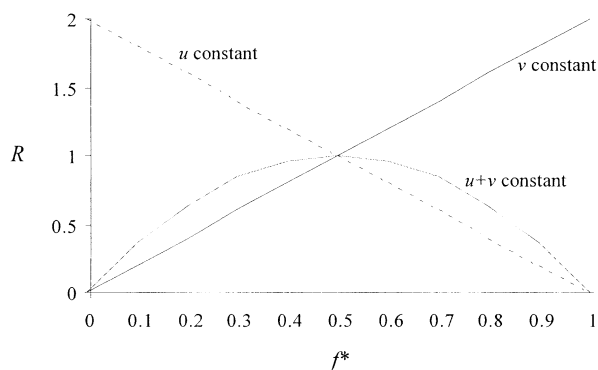


Fig. 1. Evolution of the relative substitution rate R with the biased mutation equilibrium GC content f^* . GC mutates toward AT at rate v and AT mutates toward GC at rate u . R is the substitution rate divided by the substitution rate at $f^* = 0.5$ [from Eqs. (3) and (4); see text].

The mutation rate thus depends on two parameters, u_i and v_i , and there are infinite possibilities for the evolution of m_i^* with f_i^* . Let us consider three very simple cases. Let us assume that u_i is constant over the genome and equals u and that the difference in GC content arises through the variation of v_i . From Eq. (3) this gives that the substitution rate K_S^* is a decreasing function of the equilibrium GC content. The reciprocal case is to assume that v is constant along the genome but that u_i varies. Then the substitution rate is an increasing function of the equilibrium GC frequency (Fig. 1). A last special case is to assume that u_i and v_i vary along the genome but that their sum is constant, $u_i + v_i = c$. In this case the substitution rate is a second-order polynomial of f^* reaching a maximum when $f = 0.5$.

$$K_{S_i}^* = 2cf_i^*(1 - f_i^*) \quad (4)$$

We are not aware of any molecular mechanism suggesting such variations of u_i or v_i . Under a model where v is constant, one expects fewer mutations in AT-rich regions; this is supported by the positive (though weak) correlation between nucleotide diversity and GC content (International SNP Map Working Group 2001). At this point, we have demonstrated that the regional mutation bias model should not be restricted to an inverted-“ u ” relationship between K_S and GC content (Gu and Li 1994).

CpG Deamination and Mutation Rate

Hurst and Williams (2000) proposed that the increase of K_S and GC content could be due to the CpG deamination process. We tested this assumption by assuming regional mutation rates u_i and v_i along the genome and an additional mutation rate for CpG dinucleotides: k . CpG deaminations are expected to occur at a 10-fold higher rate than other mutations (Gianelli et al. 1999). Under such a mutation model the equilibrium GC frequency,

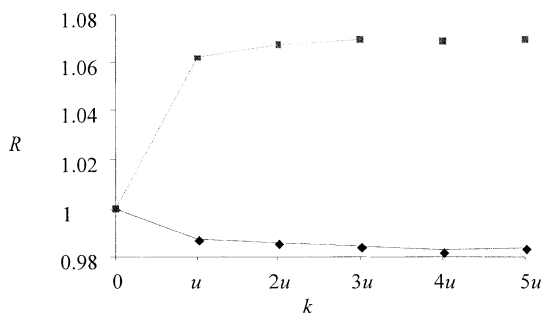


Fig. 2. Evolution of the relative mutation rate R with the CpG deamination rate k at equilibrium base frequencies. Squares: $u/(u + v) = 0.6$. Diamonds: $u/(u + v) = 0.4$. R is the relative mutation rate compared to the mutation rate for $k = 0$ (for $k = 0$ the mutation rate from a CpG dinucleotide to a TpG dinucleotide is equal to the mutation rate from C to T).

g^* , cannot be estimated analytically (Duret and Galtier 2000) and we checked by simulation the overall mutation rate at equilibrium base frequencies. We found that for $u_i < v_i$ ($f^* < 0.5$) the mutation rate decreases with increasing k (Fig. 2). This is paradoxical, as one may expect an increase in the overall mutation rate when a new mutation rate is added. This is because for u_i and v_i so that $f^* < 0.5$, the mutation rate increases with the GC frequency f [Eq. (1)]. CpG deamination decreases the equilibrium GC frequency $g^* < f^*$ (g^* is obtained from $AT \leftrightarrow GC$ and $CpG \rightarrow TpG$ or CpA mutations). At base frequency equilibrium, the expected CpG in the sequence is very low so that the increased mutation of CpG dinucleotides is not sufficient to compensate for the decreased mutation due to the lower GC content. On the contrary, when u_i and v_i are such that $f^* > 0.5$, the mutation rate increases with decreasing GC content [Eq. (1)]. The CpG mutations decrease f so that the increase in mutation due to the decrease in GC content, and the CpG mutations both contribute to an increase in the mutation rate relative to the mutation rate without special CpG mutations. In conclusion, CpG deamination tends to decrease the mutation rate in GC-poor regions and to increase the mutation rate in GC-rich regions, so that it could well explain an increase in the mutation rate with GC content, assuming that u and v are such that without deamination, the mutation rates are similar in different GC regions [if the mutation rate decreases with the GC content (Fig. 1), CpG deamination will reduce the increase but not necessarily reverse it into a positive relationship].

Selection or BGC

The rate of substitution S is equal to the product of the number of mutations by generation by the probability of fixation of each mutation. Let us call p the probability of fixation of a mutation from AT to GC and q the probability of fixation of a mutation from GC to AT. Then the

average rate of substitution is (Eyre Walker and Bulmer 1995)

$$S = u(1 - f)p + vfq = u + f(vq - up) \quad (5)$$

so that the relationship between the synonymous substitution rate and the equilibrium GC content depends on the sign of $vq - up$. The rate of substitution from AT toward GC is up and the rate of substitution from GC toward AT is vp . The selection model becomes equivalent to the mutation bias model by replacing u by up and v by vq . The equilibrium GC frequency at mutation selection drift equilibrium f^* in a diploid population of effective population size N , assuming that $Nu \ll 1$ and $Nv \ll 1$, is (Li 1987)

$$f^* = \frac{up}{vq + up} = \frac{e^{4Ns}}{v/u + e^{4Ns}} \quad (6)$$

s is the selection coefficient for GC nucleotides; for the biased gene conversion model s is replaced by c , the rate of biased gene conversion from AT toward GC (Nagylaki 1983). If $vq - up > 0$, S increases with f^* ; otherwise, S decreases with f^* . Now if one assumes that $v = u$ (no mutation bias), as $q < p$, S decreases with increasing f^* (Eyre-Walker and Bulmer 1995). Interestingly, when there is a bias in the mutation rates in opposition to the effect of selection, S increases with f^* as long as $v/u > p/q$, that is, the mutation bias is greater than the ratio of the probability of fixation of a GC compared over the probability of fixation of an AT mutation.

Following Eyre-Walker and Bulmer (1995), we assume that the mutation rates $u \ll s$ and $v \ll s$ and

$$p = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \quad (7)$$

$$q = \frac{1 - e^{2s}}{1 - e^{4Ns}} \quad (8)$$

Substituting for q and p in Eq. (5) enables us to define the range of parameters s and v/u for which there is an increase in K_S with GC content f of a sequence (Fig. 3). Comparison with the corresponding values of f^* shows that a positive correlation between K_S and f^* occurs only if the selection mutation equilibrium frequency f^* is lower than 0.5, as the line $vq = up$ corresponds to an equilibrium value of $f^* = 0.5$. If one assumes that all the regions of the genome are at mutation selection equilibrium, then the relationship between K_S and GC is an inverted “u”, the maximum mutation rate being reached for $f^* = 0.5$ (for fixed v/u the relationship between K_S and f^* can be inferred by following a horizontal line in Fig. 3).

So if we want to explain the observed relationship between K_S and GC content (Bielawski et al. 2000; Hurst and Williams 2000), we have to assume that the selection

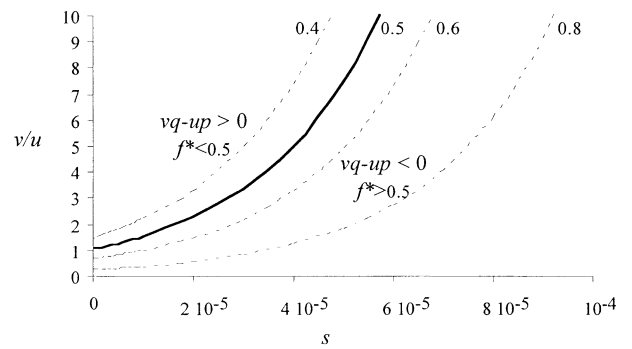


Fig. 3. Evolution of the sign of $vq - up$ in the parameter space s and v/u ($N = 10^4$). $vq - up > 0$ for a positive relationship between K_S and f (see text). Dashed lines: Isoclines for which the equilibrium GC frequency at selection mutation equilibrium is $f^* = 0.4, 0.5, 0.6,$ and 0.8 . Bold line: $vq = up$ corresponds to $f^* = 0.5$.

mutation equilibrium GC frequency is lower than 0.5, again raising the question about how GC-rich isochores have emerged. Dunn et al. (2001) suggested that the lack of correlation between K_S and codon usage in *Drosophila* could be due to a release of selection on codon usage in *Drosophila simulans*. So, again, the positive correlation can be explained by a nonequilibrium structure of the isochores.

Conclusion

We have shown that a regional mutation bias model explain the observed positive correlation between the rate of silent substitution and the GC content if the equilibrium GC content is due to the variation of the mutation from AT to GC. CpG deamination tends to increase the overall mutation rate in GC-rich regions and decrease the mutation rate in GC-poor regions. Thus, if the other mutation rates responsible for the variation of GC content along the genome do not produce a decrease in the mutation rate with GC content, CpG deamination will produce a positive correlation between K_S and GC content. The selectionist (or biased gene conversion) model can explain the positive correlation only if the mutation selection GC frequency is lower than 0.5. We suggest that there are many arguments favoring the more simple hypothesis that the isochore structure is not at equilibrium. The potential nonequilibrium of the isochore structure is a challenging hypothesis for future work, as the equilibrium property is the starting assumption for most parameter estimations (K_S) and the test of the evolutionary models of isochore evolution (Eyre Walker 1999). Nonequilibrium of isochore structure would even raise the difficulty of studying the processes by which the isochore structure originated.

Acknowledgments. The first author would like to thank Marie Semon for access to unpublished results and Laurent Gueguen and Adam Eyre

Walker for useful discussions. This work is supported by the French Bioinformatic program.

References

- Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3–17
- Bernardi G, Olofson B, Filipski J, *et al.* (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958
- Bielawski JP, Dunn KA, Yang Z (2000) Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* 156:1299–1308
- Casane D, Boissinot S, Chang BHJ, Shimmin LC, Li WH (1997) Mutation pattern variation among regions of the primate genome. *J Mol Evol* 45:216–226
- Dunn KA, Bielawski JP, Yang Z (2001) Substitution rates in *Drosophila* nuclear genes: Implications for translational selection. *Genetics* 157:295–305
- 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 Acid Res* 20:1497–1501
- Eyre Walker A (1993) Recombination and mammalian genome evolution. *Proc R Soc Lond B* 252:237–243
- Eyre Walker A (1997) Differentiating between selection and mutation bias. *Genetics* 147:1983–1987
- Eyre Walker A (1999) Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* 152:675–683
- Eyre Walker A, Bulmer M (1995) Synonymous substitution rates in enterobacteria. *Genetics* 140:1407–1412
- Filipski J (1988) Why the rate of silent codon substitutions is variable within a vertebrate's genome. *J Theor Biol* 134:159–164
- Francino MP, Ochman H (1999) Isochores result from mutation, not selection. *Nature* 400:30–31
- Francino MP, Ochman H (2000) Strand symmetry around the β globin origin of replication in primates. *Mol Biol Evol* 17:416–422
- Fryxell KJ, Zuckerkandl E (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol* 17:1371–1383
- Goldman N, Yang Z (1994) A codon based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725–736
- Gianelli F, Anagnostopoulos T, Green PM (1999) Mutation rates in human. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from *Hemophila B*. *Am J Hum Genet* 65:1580–1587
- Gu X, Li WH (1994) A model for the correlation of mutation rate with GC content & the origin of GC rich isochores. *J Mol Evol* 38:468–475
- Holmquist GP (1992) Chromosome bands, their chromatinic flavours, and their functional features. *Am J Hum Genet* 51:17–37
- Holmquist GP, Filipski J (1994) Organization of mutations along the genome: A prime determinant of genome evolution. *Trends Ecol Evol* 9:65–69
- Hughes S, Zelus D, Mouchiroud D (1999) Warm-blooded isochore structure in Nile crocodile and turtle. *Mol Biol Evol* 16:1521–1527
- Hurst LD, Williams EJB (2000) Covariation of GC content and the silent site substitution rate in rodents: Implications for methodology and for the evolution of isochores. *Gene* 261:107–114
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933
- Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217:624–626
- Li WH (1987) Models of nearly neutral mutations with particular implications for non random usage of synonymous codons. *J Mol Evol* 24:337–345
- Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distance of coding sequences and DNA molecules in human and murids. *J Mol Evol* 21:311–320
- Nagylaki (1983) Evolution of a finite population under gene conversion. *Genetics* 80:6278–6281
- Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 85:2653–2657
- Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci USA* 85:2653–2657
- Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285
- Yang Z (1994) Estimating the pattern of nucleotide substitution. *J Mol Evol* 39:105–111
- Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 46:409–418