Mutation rates differ among regions of the mammalian genome

Kenneth H. Wolfe, Paul M. Sharp* & Wen-Hsiung Li†

Department of Genetics, Trinity College, Dublin 2, Ireland
† Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, Texas 77225, USA

In the traditional view of molecular evolution, the rate of point mutation is uniform over the genome of an organism and variation in the rate of nucleotide substitution among DNA regions reflects differential selective constraints[1,2]. Here we provide evidence for significant variation in mutation rate among regions in the mammalian genome. We show first that substitutions at silent (degenerate) sites in protein-coding genes in mammals seem to be effectively neutral (or nearly so) as they do not occur significantly less frequently than substitutions in pseudogenes. We then show that the rate of silent substitution varies among genes and is correlated with the base composition of genes and their flanking DNA. This implies that the variation in both silent substitution rate and base composition[3] can be attributed to systematic differences in the rate and pattern of mutation over regions of the genome. We propose that the differences arise because mutation patterns vary with the timing of replication of different chromosomal regions in the germline. This hypothesis can account for both the origin of isochores in mammalian genomes[4] and the observation[5] that silent nucleotide substitutions in different mammalian genes do not have the same molecular clock.

Early estimates of the substitution rates in pseudogenes and at silent sites in protein-coding sequences suggested that pseudogenes evolve more rapidly which implies that silent sites are under some selective constraint[6,7]. However, these studies used very limited data and rather simple methods of estimating rates. Here we use two approaches to compare the rate of substitution in functionless DNA (that is, in pseudogenes and intergenic regions) with that at silent sites in genes; the latter is measured in terms of the number of nucleotide substitutions per fourfold degenerate site (designated K_4). First, we compare the divergences in pseudogenes and in functional genes across the same species pair. If silent sites in functional genes are selectively constrained, then they should have accumulated fewer substitutions.

By far the most substantial data set available is for DNA sequences from humans and Old World monkeys (Table 1). The mean degree of divergence at silent sites is clearly not lower than that in pseudogenes. There are rate differences among genes, but for only one gene (β-globin) out of 13 is the silent rate significantly lower than the pseudogene rate, suggesting that the β-globin rate is an extreme random deviate. Nevertheless, perhaps a better method is to compare a recently generated pseudogene with its functional homologue from the same species, using the functional gene from a second species as a reference. This method can be applied to both processed and unprocessed pseudogenes, but is more accurate in the former case because processed pseudogenes can be assumed to become functionless as soon as they are formed. The three LDH-A processed pseudogenes in mouse[8] provide excellent data for this approach because they are reasonably large and because they were produced after the mouse-rat split, so the rat LDH-A gene can be used as a reference. The ratios of the substitution rates in the three LDH-A pseudogenes to the rate at silent sites in the functional gene are 1.1, 0.8 and 0.7. Again, there is little, if any, evidence for a significantly higher rate of substitution in pseudogenes than at silent sites. We therefore conclude that most silent (fourfold degenerate) sites are effectively neutral, and that their substitution rate is a good measure of the mutation rate, because under selective neutrality the substitution rate is expected to be equal to the mutation rate[1].

* To whom correspondence should be addressed.

Table 1 Rates of nucleotide substitution at silent sites and in functionless DNA for human versus Old World monkey sequences

Gene	$K_4 \pm$ s.e.m.	GC_4*	L_4†
α_1-Antitrypsin	0.077 ± 0.022	73.6	184
Apolipoprotein A-I	0.061 ± 0.025	88.4	126
Apolipoprotein E	0.062 ± 0.021	89.1	175
Chorionic gonadotrophin-β	0.065 ± 0.027	81.3	99
Erythropoietin	0.110 ± 0.035	71.7	110
β-Globin	0.039 ± 0.028	68.2	54
Insulin	0.130 ± 0.054	82.3	62
Metallothionein I	0.125 ± 0.113	85.5	21
Metallothionein II	0.096 ± 0.072	87.0	23
β-Myosin heavy chain	0.179 ± 0.056	86.5	97
Pepsinogen A	0.108 ± 0.026	73.7	194
Transforming growth factor-β	0.064 ± 0.019	83.5	203
Triose phosphate isomerase	0.112 ± 0.033	59.8	128
Total for silent sites	0.089 ± 0.009	78.7	1,474
η-Globin pseudogene locus	0.074 ± 0.006	42.9	2,071
η-δ Globin intergenic DNA	0.076 ± 0.006	38.7	2,771
Total for functionless DNA	0.075 ± 0.004	40.5	4,842

K_4 (the corrected number of substitutions per fourfold degenerate site) was calculated as described in Fig. 1. References to the original DNA sequence data are available on request.
* G+C content at fourfold degenerate sites (%).
† Number of fourfold degenerate sites compared.

Next we focus on the variation among genes in the substitution rate at silent sites (K_4) and show that it is correlated with their G+C content (GC_4). We compare DNA sequences between mouse and rat, first concentrating on the larger genes (those with over 175 silent sites) (Fig. 1*a*). For the 17 genes with $GC_4 \geq 50\%$, K_4 clearly decreases as GC_4 increases (correlation coefficient $r = -0.73$, $P < 10^{-4}$); the highest K_4 is nearly twice the lowest value. On the other hand, there is a significant positive correlation between substitution rate and GC_4 in the A+T-rich genes, though the number of points is small ($r = +0.82$; $P < 0.05$). This latter observation suggests that the low substitution rates in genes with a high G+C content cannot be simply explained by assuming that G and C nucleotides mutate less frequently than A and T nucleotides. On reflection, it is difficult to think of any other reasonable relationship between K_4 and GC_4, although this has not previously been reported. In fact, others have suggested either no systematic variation (in silent substitution rate) among genes[9], or a simple linear relationship between substitution rate and G+C content[10,11]. When smaller genes are included (making a total of 88 genes), the same peaked relationship as that shown in Fig. 1*a* is seen, but the correlation coefficients are decreased because of the stochastic effects (particularly on K_4) of short genes. The above comparison requires no knowledge of the date of speciation, the data set used is large, and the K_4 values are fairly large but far from saturation. The mouse-rat comparison is preferable to a human-rodent comparison because of the shorter timescale involved; some human genes differ considerably in base composition[12] or even in chromosomal map position[13] from their rodent homologues.

The variance of GC_4 in the 23 large rodent genes (Fig. 1*a*) is more than ten times that expected by chance (under a binomial distribution with mean 58.6%), suggesting that the factors determining G+C content are not uniform for all genes. Of course, one might argue that the GC_4 value in any gene is at a particular optimum and that the observed relationship between K_4 and GC_4 is a consequence of selection maintaining that GC_4 value in the face of a mutation rate that is uniform across the genome. But this would require strong selective constraints, whereas we have shown here that silent sites seem to be effectively neutral. Furthermore, the direction of selection on G+C-rich genes

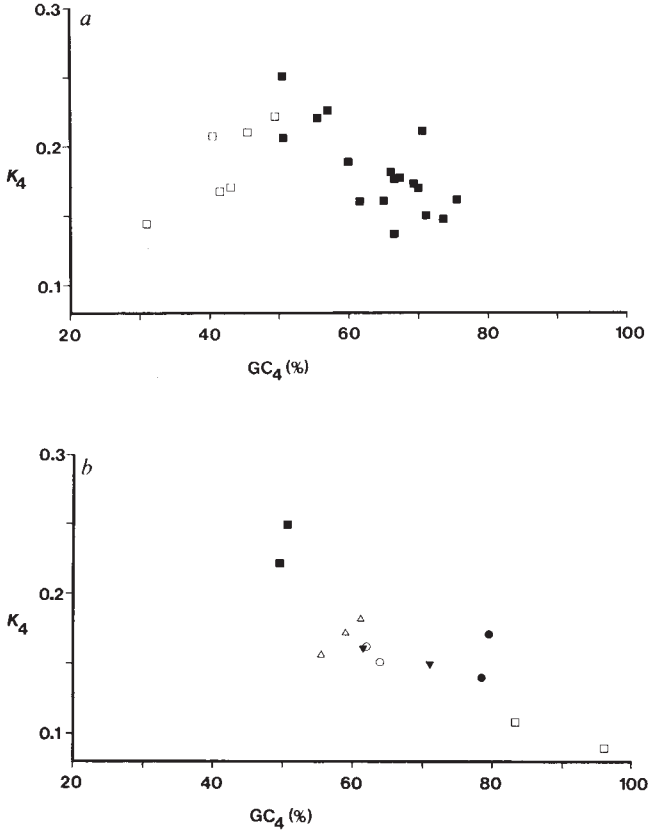

Fig. 1 *a*, Plot of silent substitution rate (K_4) versus G+C content at silent sites (GC_4) for 23 genes compared between mouse and rat. Only large genes (with ≥175 fourfold degenerate sites) are compared. ■, 17 genes for which $GC_4 \geq 50\%$ (encoding alcohol dehydrogenase-1, α-fetoprotein, CD4 antigen, c-*mos* and c-*myc* oncogenes, creatine kinase M, cytochromes $P_1$450 and $P_3$450, β-glucuronidase, malic enzyme, mitochondrial aspartate aminotransferase and malate dehydrogenase, neural cell adhesion molecule NCAM-140, neurofilament M, oestrogen receptor, protein disulphide isomerase and renin); □, 6 genes with $GC_4 < 50\%$ (encoding albumin, α_2-amylase, glucocorticoid receptor, leukocyte common antigen (Ly-5, T200), ornithine decarboxylase and UDP-glucuronosyltransferase). Full references for the sequences are available from the authors on request; all are taken from GenBank (release 50) or the recent literature. K_4 was calculated by the method of Tajima and Nei[30], which requires no assumptions about the base composition of the sequences compared, other than that they be at equilibrium. Any codons at which mouse and rat differed by more than one nucleotide substitution were ignored. This avoided (1) the need to apply weights to the different mutational pathways possible between multiply-substituted codons, and (2) any regions within a gene in which the protein sequences of the two species are considerably different and might not share a common origin (for example as a result of insertions, deletions or alternative splicing). If, as we assume, the sequences being compared have diverged only through a process of random point mutation, ignoring these codons should not alter the synonymous substitution rate calculated. *b*, Relationship between K_4 and GC_4 for several groups of physically linked genes: □, metallothioneins I and II (5 kb apart); ■, albumin and α-fetoprotein (14 kb); ▼, cytochromes $P_1$450 and $P_3$450 (<25 kb); ●, γ_A and γ_D crystallins (35 kb); ○, Ly-2 and Ly-3 antigens (36 kb); △, immunoglobulin heavy chains C_δ $C_{\gamma 3}$ and C_ε (50 kb and 90 kb).

would have to be opposite to that on A+T-rich genes. Thus our finding that the substitution rate and the base composition of silent sites vary together in a systematic way is most simply explained by supposing that the pattern of mutation is different for different genes. Most germline mutations are thought to arise from misincorporation errors made by the DNA replication apparatus[14,15]. It has been demonstrated that different genes replicate at different stages of the cell cycle in differentiated cells[16], and this is presumably also true in the germline. The number and type of replication errors are likely to vary during the cell cycle if the chemical environment in the nucleus changes. In fact, the abundances (both relative and absolute) of free dNTPs in the nucleus change with time[17], as do the activities of the DNA polymerase enzymes and their accessory proteins[18]. We have examined a theoretical model (to be detailed elsewhere) of the relationship between the mutation rate and the G+C content of both the nucleotide precursor (dNTP) pools and the template DNA. This model is based on a simple kinetic description of DNA replication[19], as applied to the replication of long double-stranded DNA molecules. It can explain the occurrence of the maximum mutation rate at an intermediate value of GC_4, provided that GC_4 is positively correlated with the proportion of dGTP plus dCTP in the dNTP pools. This assumption seems plausible as we are considering only effectively neutral sites in DNA. There is also some experimental evidence to support this: as S-phase progresses, the G+C content of both the dNTP pools and the replicating DNA decreases[16,17,20]. As K_4 is a consequence of the mutation rate, under this model the relationship seen in Fig. 1*a* can be generated in the absence of selection.

Replication origins in mammalian DNA are located 50–300 kilobases (kb) apart, and clusters of 12–100 adjacent origins tend to replicate simultaneously[16]. Thus, regions of DNA of 1,000 kb or more could form temporal units of replication, and all DNA within a unit would be expected to show the same pattern and rate of mutation. Thus, under our model, genes which are physically close to each other in the genome should have similar values of K_4 and GC_4. This is indeed the case for those genes for which data are available (Fig. 1*b*): several pairs of genes which are the results of ancient tandem duplications show paired results for K_4 and GC_4, as do genes of the IgH constant-region locus. These IgH genes span 140 kb on mouse chromosome 12, and are known to be part of a single replicon[21].

These replication units seem to correspond to the 'isochores' (large blocks of DNA with homogeneous base compositions) found experimentally in mammalian nuclear DNA[4], and probably also to high-resolution Giemsa chromosomal bands[16]. We therefore propose that isochores arise as a result of the synchronous replication of megabase stretches of DNA under varying dNTP pool conditions. Although our model based on variation in the dNTP precursor pools can provide a simple explanation for the observed variation of mutation rates and patterns around the genome, it is of course not the only possible explanation. For example, Filipski[22] has proposed that isochores are formed as a consequence of the repair of DNA in different types of chromatin by different DNA polymerase enzymes, and there is now some experimental evidence for between-gene differences in efficiency of DNA repair[23]. Our observations could also be explained if DNA is replicated by several distinct DNA polymerase holoenzymes with different error propensities. But recent data indicate that there is only one replicative polymerase complex for mammalian nuclear DNA[18,24].

We conclude that much of the intragenomic variation in silent substitution rate and base composition in mammals results from variation in the process of mutation, rather than from natural selection[25,26]. The implication that silent sites in mammalian genes are effectively neutral contrasts with prokaryotic genes, in which differences in translational efficiency lead to selective differences between codons[27] and thus to lower silent substitution rates in genes that are highly expressed[2]. Our assertion is supported by several arguments, however. First, the base composition of silent sites in mammalian genes is strongly correlated with that of introns and flanking sequences of the same gene[3],

suggesting that it is determined largely by local mutation pressures. In addition, it has recently been suggested on the basis of analysis of dinucleotide frequencies that codon usage in human genes is determined by context-dependent mutational patterns[28]. Second, any suggestion of tissue-specific or expression level-related patterns in mammalian codon usage must be set against the observation that the human α- and β-globin genes have very different codon usage. Third, as the effective population sizes of mammalian species are quite small[29], it would require large selective differences for selection to be effective, but it is unlikely that mammals have sufficient reproductive excess to allow such selection at a very large number of codons. In summary, we propose that, in the mammalian genome, variation in the silent substitution rate and variation in base composition are two facets of the same phenomenon.

We thank D. W. Nebert for cytochrome P_{450} mapping data. This study was supported in part by the European Community BAP (to P.M.S.) and the NIH (to W.H.L).

Received 5 July; accepted 7 December 1988.

1. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
2. Sharp, P. M. & Li, W.-H. *Molec. biol. Evol.* **4**, 222–230 (1987).
3. Aota, S. & Ikemura, T. *Nucleic Acids Res.* **14**, 6345–6355 (1986).
4. Bernardi, G. *et al. Science* **228**, 953–958 (1985).
5. Li, W.-H., Tanimura, M. & Sharp, P. M. *J. molec. Evol.* **25**, 330–342 (1987).
6. Li, W.-H., Gojobori, T. & Nei, M. *Nature* **292**, 237–239 (1981).
7. Miyata, T. & Hayashida, H. *Proc. natn. Acad. Sci. U.S.A.* **78**, 5739–5743 (1981).
8. Fukasawa, K. M. *et al. Genetics* **115**, 177–184 (1987).
9. Miyata, T. *et al. J. molec. Evol.* **19**, 28–35 (1982).
10. Filipski, J. *J. theor. Biol.* **134**, 159–164 (1988).
11. Smithies, O., Engels, W. R., Devereux, J. R., Slightom, J. L. & Shen, S-h. *Cell* **26**, 345–353 (1981).
12. Mouchiroud, D. & Gautier, C. *Molec. biol. Evol.* **5**, 192–194 (1988).
13. Nadeau, J. H. & Taylor, B. A. *Proc. natn. Acad. Sci. U.S.A.* **81**, 814–818 (1984).
14. Friedberg, E. C. *DNA Repair* (Freeman, New York, 1985).
15. Topal, M. D. & Fresco, J. R. *Nature* **263**, 285–289 (1976).
16. Holmquist, G. P. *Am. J. hum. Genet.* **40**, 151–173 (1987).
17. Leeds, J. M., Slabaugh, M. B. & Mathews, C. K. *Molec. cell Biol.* **5**, 3443–3450 (1985).
18. Kelly, T. & Stillman, B. (eds) *Cancer Cells 6: Eukaryotic DNA Replication* (Cold Spring Harbor Laboratory, New York, 1988).
19. Fersht, A. R. & Knill-Jones, J. W. *Proc. natn. Acad. Sci. U.S.A.* **78**, 4251–4255 (1981).
20. Holmquist, G., Gray, M., Porter, T. & Jordan, J. *Cell* **31**, 121–129 (1982).
21. Brown, E. H. *et al. Molec. cell. Biol.* **7**, 450–457 (1987).
22. Filipski, J. *FEBS Lett.* **217**, 184–186 (1987).
23. Bohr, V. A., Phillips, D. H. & Hanawalt, P. C. *Cancer Res.* **47**, 6426–6436 (1987).
24. Prelich, G. & Stillman, B. *Cell* **53**, 117–126 (1988).
25. Bernardi, G. & Bernardi, G. *J. molec. Evol.* **24**, 1–11 (1986).
26. Gillespie, J. H. *Genetics* **113**, 1077–1091 (1986).
27. Ikemura, T. *Molec. biol. Evol.* **2**, 13–34 (1985).
28. Hanai, R. & Wada, A. *J. molec. Evol.* **27**, 321–325 (1988).
29. Nei, M. & Graur, D. *Evol. Biol.* **17**, 73–118 (1984).
30. Tajima, F. & Nei, M. *Molec. biol. Evol.* **1**, 269–285 (1984).

Nonmutagenic carcinogens induce intrachromosomal recombination in yeast

Robert H. Schiestl

Department of Biology, University of Rochester, Rochester, New York 14627, USA
Prairie Biological Research Ltd, 10515-36A Avenue, Edmonton, Alberta T6J 2H7, Canada
GeneBioMed, Inc., PO Box 18121, 12 Corners Station, Rochester, New York 14618, USA

To identify environmental carcinogens there is a need for inexpensive and reliable short-term tests[1], but certain human or animal carcinogens are persistently undetectable as mutagents with the Ames assay[2–5] or with other short-term tests currently in use[6,7]. Thus there is a need for short-term tests which detect carcinogens missed by the Ames assay[1]. Because of the association of carcinogenesis with genome rearrangement[8–12], a system screening for intrachromosomal recombination resulting in genome rearrangement has been constructed in *Saccharomyces cerevisiae*[13]. Evaluation of this system shows inducibility by a variety of carcinogens[3–7] not detectable by the Ames assay or various other short-term tests. In the light of these results it is tempting to speculate that 'nongenotoxic carcinogens' are in fact genotoxic but, in the past, the tools to measure the genetic alterations they induce have been inappropriate.

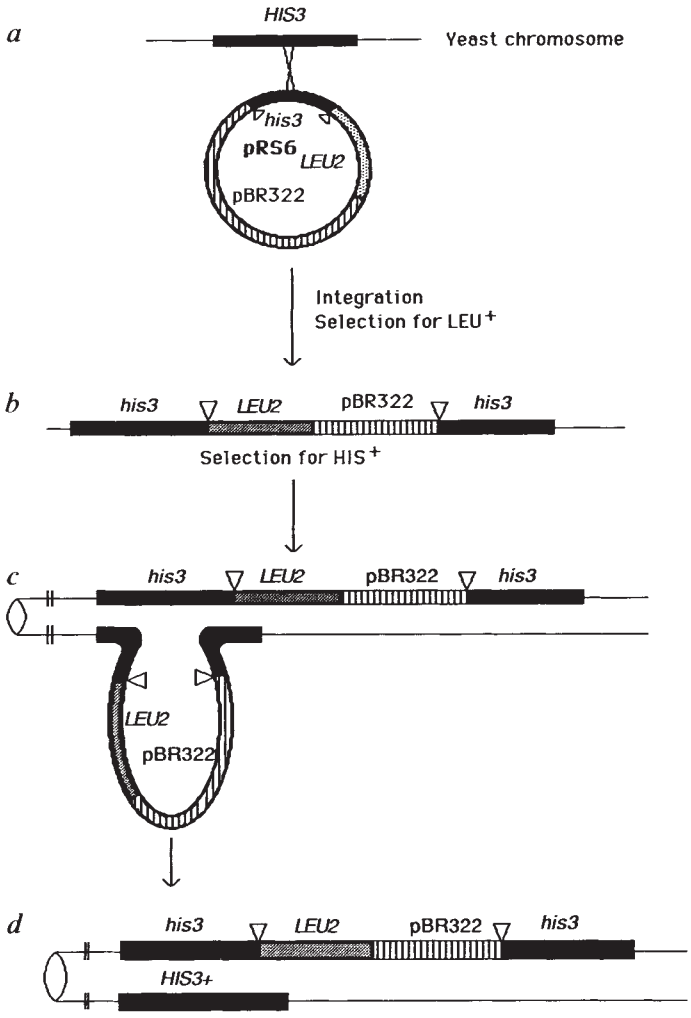

Fig. 1 Plasmid pRS6 which contains an internal fragment of the *HIS3* gene was cut within the internal *his3* fragment and integrated into the geome at the *HIS3* locus *a*. This creates a duplication of the *his3* gene in which one allele is deleted for its 3′ end and the other for its 5′ end *b*, The two alleles share about 400 base pairs of homology and thus can recombine with each other to revert to the $HIS3^+$ allele. As shown previously $HIS3^+$ recombinants do not arise by plasmid excision or by unequal sister chromatid exchange[13]. In these studies the frequency of plasmid excision was determined by putting a yeast origin of replication sequence onto the integrating plasmid. Excised plasmids could be recovered but were 100-fold less frequent than the frequency of HIS^+ formation which suggests that plasmid excision does not occur in the majority of recombinants. Reciprocal products expected to result from sister chromatid exchange were analysed by Southern blotting. The pattern characteristic for sister chromatid exchange was not found in any of 25 events examined. Thus, as a likely alternative mechanism conversion between sister chromatids (*c*) is suggested which result in deletion of the integrated plasmid on one chromatid (*d*). After segregation of the two chromatids the HIS^+ recombinant should show a HIS^+ leu^- phenotype which is found in 99% of all HIS^+ recombinants[13]. Thus, it is proposed, as a possible model, that a double strand break initiates the recombination event which is extended by an exonuclease to a gap. This gap is repaired by gene conversion from the sister chromatid as donor[13].