J Mol Evol (2006) 62:803–806
DOI: 10.1007/s00239-005-0228-7

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

The GC Content of Primates and Rodents Genomes Is Not at Equilibrium: A Reply to Antezana

Laurent Duret

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS, Univ. Lyon 1 Villeurbanne Cedex, France

Received: 29 September 2005 / Accepted: 4 January 2006

To study the patterns of neutral substitutions in the human genome, we have recently analyzed a large data set of alignments of orthologous noncoding DNA sequences from human, chimpanzee, and baboon (Meunier and Duret 2004). We observed that the base composition of the human genome is not at equilibrium: substitutions from G or C to A or T (hereafter referred to as GC $\rightarrow$ AT substitutions) are more numerous than AT $\rightarrow$ GC substitutions. Antezana (2005) has re-analyzed the genomic alignment data that we had compiled. In contradiction to our results, he found that the GC-content of the human genome is close to the equilibrium. The explanation he proposed for this discrepancy is that Meunier and Duret "used a malfunctioning dinucleotide-level simulation procedure out of concern for context-dependent mutation effects." I show here that in fact, Antezana (2005) used an erroneous procedure to count substitutions that ignored the hypermutability of CpG dinucleotides, and therefore led to systematically overestimating the number of AT $\rightarrow$ GC substitutions.

Antezana (2005) used parsimony to count substitutions in alignments of orthologous human, chimpanzee, and baboon nongenic DNA sequences: substitutions to human or chimpanzee were retrieved from sites at which the baboon base and the base in one of the two nonbaboon sequences were identical but different from the base in the other nonbaboon sequence. It is well established that because of multiple substitutions, parsimony may be erroneous when patterns of substitutions are biased (Eyre-Walker 1998). It is also well known that in mammals, CpG dinucleotides are mutational hot spots: the rate of transition (C $\rightarrow$ T or G $\rightarrow$ A) at CpG sites is about 10 times higher than at non-CpG sites (Giannelli et al. 1999). Thus, although the average rate of divergence (excluding indels) between human and chimpanzee is 1.2%, the divergence at CpG sites is about 15.2% (CSAC 2005). Hence, as mentioned in our article (Meunier and Duret 2004), there is an important frequency of homoplasy at CpG sites, and therefore parsimony must be used with caution.

To illustrate this problem of homoplasy at CpG sites, let us take a simple example, very similar to the real situation in our human/chimp/baboon alignments: two species (species1 and species2) and an outgroup, such that the evolutionary distance at non-CpG sites is 0.01 substitutions/site between species1 and species2 and 0.05 substitutions/site between the outgroup and the two other species (Fig. 1a), and the rate of substitution at CpG sites is 10 times higher than at non-CpG sites. Let us consider a site that corresponds to a T in species1, a C in species2, a T in the outgroup and that is followed by a C, conserved in the three species (Fig. 1b, c). The scenario proposed by the simple parsimony method predicts that the ancestral sequence was TC and that a single T $\rightarrow$ C substitution occurred in the species2 lineage. The probability of that scenario is 5×10^{-3} (Fig. 1b). The second most likely scenario involves two independent substitutions, and is 40 times less likely than the first one (Fig. 1c). Thus, in that situation, the parsimony approach can be considered as reliable. Now consider

Correspondence to: Laurent Duret; *email:* duret@biomserv.univ

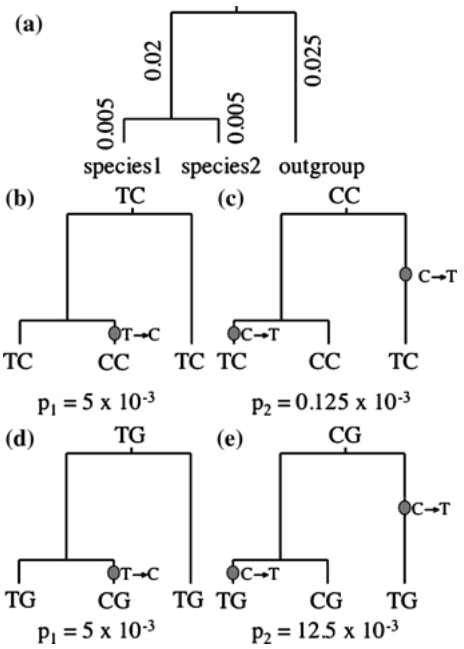

Fig. 1. Illustration of the artifact of the maximum parsimony method to count substitutions at CpG sites. The phylogeny of the three species used to infer substitutions is shown in (**a**). Branch lengths indicate the rate of substitution per site at non-CpG positions. Substitution rates at CpG sites are considered 10 times higher than at non-CpG sites. The first alignment (TC/CC/TC) corresponds to a situation where the parsimony method is reliable: the most parsimonious scenario (one single substitution) (**b**) is 40 times more likely than the first alternative scenario (**c**). The second alignment (TG/CG/TG) corresponds to a situation where the parsimony method is not reliable: the most parsimonious scenario (one single substitution) (**d**) is 2.5 times less likely than the alternative scenario that involves two independent substitution at CpG sites (**e**).

a site that—as in the first example—corresponds to a T in species1, a C in species2, a T in the outgroup but that is followed by a G, conserved in the three species (Fig. 1d, e). As in the previous example, the scenario proposed by the simple parsimony method predicts that a single T $\rightarrow$ C substitution occurred in the species2 lineage, and the probability of that scenario is 5×10^3 (Fig. 1d). The alternative scenario predicts that the ancestral sequence was CG (i.e., a CpG site), and that two independent C $\rightarrow$ T substitutions occurred in the species1 lineage and in the outgroup. Because the rate of substitution is 10 times higher at CpG sites, this scenario (that involves two C $\rightarrow$ T substitutions) is 2.5 times more likely than the one predicted by the simple parsimony approach (that predicts a single T $\rightarrow$ C substitution). In other words, the parsimony approach used by Antezana (2005) systematically overestimates the number of AT $\rightarrow$ GC substitutions, because of homoplasy at CpG sites, and this of course leads to overestimation of the equilibrium GC content.

This artifact of the parsimony method is a major problem even for very closely related species. Indeed, substitutions at CpG sites constitute 25% of all substitutions observed between human and chimpanzee (CSAC 2005). This is the reason why, as clearly mentioned in our article, we took care to analyze separately CpG and non-CpG sites and to exclude those sites for which the ancestral state was unsure (Meunier and Duret 2004). This analysis showed that GC $\rightarrow$ AT substitutions clearly outnumber AT $\rightarrow$ GC substitutions, even if only non-CpG sites are considered (Table 1 in Meunier and Duret 2004). The excess of GC $\rightarrow$ AT over AT $\rightarrow$ GC substitutions is more pronounced in GC-rich isochores than in GC-poor isochores. These observations have led to the conclusion that there is an overall decrease of the GC-content of GC-rich isochores in the human genome, which we have called the ''erosion'' of GC-rich isochores.

Antezana (2005) also analyzed with the same parsimony method the pattern of substitutions in homologous coding regions from human, mouse, and rat. Given the evolutionary distance between primates and rodents, even non-CpG sites are affected by homoplasy (the average synonymous substitution rate between primate and rodents is about 0.6 substitutions per site Waterston et al. 2002). Hence, the numbers of substitutions inferred by Antezana (2005) in the rodent lineages are clearly unreliable.

There is another problem in the article by Antezana (2005): the method he used to compute the equilibrium GC-content (GC*) assumes that all sites evolve independently (i.e., that the probability of substitution at a given base does not depend on the nature of flanking bases). It is well established that in reality this assumption is not correct, and that the strongest neighboring effect is by far that of CpGs (Hess et al. 1994). Indeed, the frequency of CpGs in the human genome is only about 23% of what would be expected if all sites were evolving independently (Bird 1980). If sequences were at equilibrium, then the procedure used by Antezana would have given the correct estimate of GC*. However, when sequences are not at equilibrium, then it is necessary to use more realistic models DNA sequence evolution with neighbor-dependent mutations, such as the one proposed by Arndt et al. (2003a). This is clearly shown in a recent paper by Arndt and Hwa (2005), where they investigated the impact of taking into account of neighbor-dependent nucleotide substitution processes on the estimate of substitution rates and of GC*.

It should be stressed that the erosion of GC-rich isochores in the genomes of primates and rodents had been previously demonstrated by many independent works. This erosion was first observed by analyzing patterns of substitutions in transposable elements in the human genome (Lander et al. 2001; Arndt et al. 2003b, 2005). It might be argued that the pattern of substitution in repeated sequence does not perfectly

reflect the evolution of unique DNA. Indeed, it has been shown in mammals that the rate of substitution at CpG sites is higher in repeated sequences than in unique DNA, most probably because of a higher level of methylation (Kricker et al. 1992; Meunier et al. 2005). However, in those repeated sequences, even non-CpG sites show an excess of GC $\rightarrow$ AT over AT $\rightarrow$ GC substitutions (Fig. 1 in Meunier et al. 2005). Moreover, it has been shown that this pattern is not restricted to repetitive DNA, but is also observed at synonymous sites of exons, not only in humans (Duret et al. 2002), but also in rodents (Duret et al. 2002; Smith and Eyre-Walker 2002) and cetartiodactyls (Duret et al. 2002). The latter result was criticized because the cetartiodactyls species that we had analyzed were too distantly related and, therefore, the parsimony approach that we had used was not reliable (Alvarez-Valin et al. 2004). Indeed, the analysis of synonymous substitutions by a maximum likelihood approach confirmed the erosion of GC-rich isochores in primates and in rodents, but not in cetartiodactyles (Belle et al. 2004). The erosion of GC-rich isochores in primates was again confirmed by the analysis of substitutions in introns and intergenic regions (Webster et al. 2003; Meunier and Duret 2004). This erosion of GC-rich isochores has also been noted in carnivores, but not in lagomorphs or perissodactyls (Belle et al. 2004).

In conclusion, there is ample evidence for an erosion of GC-rich isochores in rodents and primates. The assertion made by Antezana (2005) that the GC content of their genomes is close to equilibrium is based on an erroneous count of substitutions and an inappropriate method to estimate the equilibrium GC content. This paper illustrates again the fact that even with very closely related species, parsimony should be used with caution and that it is essential to take into account neighbor-dependent mutations if we want to understand the evolution of genomes.

Acknowledgments. I thank Peter Arndt for his helpful comments. This work was supported by the Centre National de la Recherche Scientifique.

References

Alvarez-Valin F, Clay O, Cruveiller S, Bernardi G (2004) Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect. Mol Phylogenet Evol 31:788–793

Antezana MA (2005) Mammalian GC content is very close to mutational equilibrium. J Mol Evol 61:834–836

Arndt PF, Burge CB, Hwa T (2003a) DNA sequence evolution with neighbor-dependent mutation. J Comput Biol 10:313–322

Arndt PF, Hwa T (2005) Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21:2322–2328

Arndt PF, Hwa T, Petrov DA (2005) Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J Mol Evol 60:748–763

Arndt PF, Petrov DA, Hwa T (2003b) Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol Biol Evol 20:1887–1896

Belle E, Duret L, Galtier N, Eyre-Walker A (2004) The decline of isochores in mammals: An assessment of the GC content variation along the mammalian phylogeny. J Mol Evol 58:653–660

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N (2002) Vanishing GC-rich isochores in mammalian genomes. Genetics 162:1837–1847

Eyre-Walker A (1998) Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686–690

Giannelli F, Anagnostopoulos T, Green PM (1999) Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. Am J Hum Genet 65:1580–1587

Hess ST, Blake JD, Blake RD (1994) Wide variations in neighbor-dependent substitution rates. J Mol Biol 236:1022–1033

Kricker MC, Drake JW, Radman M (1992) Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes. Proc Natl Acad Sci U S A 89:1075–1079

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol 21:984–990

Meunier J, Khelifi A, Navratil V, Duret L (2005) Homology-dependent methylation in primate repetitive DNA. Proc Natl Acad Sci USA 102:5471–5476

Smith NG, Eyre-walker A (2002) The compositional evolution of the murid genome. J Mol Evol 55:197–201

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri

J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Webster MT, Smith NG, Ellegren H (2003) Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol Biol Evol 20:278–286