Available online at www.sciencedirect.com
ScienceDirect
Gene 385 (2006) 50–56
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Are GC-rich isochores vanishing in mammals?

Jianying Gu [a,b], Wen-Hsiung Li [b,*]

[a] Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, United States
[b] Department of Biology, College of Staten Island, City University of New York, Staten Island, NY 10314, United States

Received 30 November 2005; accepted 30 March 2006
Available online 5 August 2006

Abstract

Several studies of nucleotide substitution patterns in mammalian species suggested that GC-rich isochores might be vanishing in mammalian genomes. However, the number of genes and the number of genomes included in these studies might not have given a reliable broad view of the trend in GC change in mammals. It is therefore worth exploiting this issue with a broader coverage of mammalian genomes using a reliable approach, the maximum likelihood approach. We have applied two maximum likelihood methods to infer the ancestral GC contents of 176 mammalian genes from representative eutherian species and at least one marsupial species. Except for a large GC decrease in marsupial genes, we found no general decreasing trend in GC content in GC-rich genes or in other genes among eutherian mammals; indeed, the GC content of GC-rich genes appears to have increased in recent times in some genomes, e.g., the rabbit. For the large GC decrease in marsupials, it could be mainly due to the great reduction in chromosome number, which could lead to a large reduction in recombination rate and thus also a large reduction in the rate of gene conversion. Since many eutherian mammals still maintain a fairly large number of chromosomes, it is unlikely that GC-rich isochores are vanishing in these mammals.

Keywords: GC content; GC-rich isochores; Maximum likelihood; Mammalian genomes

1. Introduction

Compositional isochores are long DNA segments with relatively homogeneous base composition and GC-rich isochores have been found in warm-blooded vertebrates such as mammals and avians and also in some cold-blooded reptiles (Bernardi et al., 1985; Hughes et al., 1999). The GC content of isochores varies across genomes and seems to be correlated with various genomic aspects, including gene density and repetitive element density (Pavlicek et al., 2001). The GC content does not appear to be stationary in mammalian genomes. Duret et al. (2002) reported evidence for vanishing of GC-rich isochores in some closely related mammals. Such decreasing pattern has also been found in non-coding regions in primates (Smith et al., 2002; Webster et al., 2003) and in repetitive retroelements in mammals (Arndt et al., 2003). Recently, by analyzing 41 genes in mammals, Belle et al. (2004) showed that a significant decrease in GC content occurred during the early stage of mammalian evolution, as early as the divergence between the eutherian and marsupial lineages. Also, they demonstrated that in most mammalian orders, including primates, rodents, carnivores, cetartiodactyls and perrisodactyls, a weak but still significant decrease occurred in the GC-rich genes studied (Belle et al., 2004).

The maximum parsimony method was applied in some studies of GC content evolution in closely related species by Duret et al. (2002) and Smith et al. (2002), leading to suggestions of vanishing GC-rich isochores in primates, rodents and cetartiodactyls. However, if there is substantial variation in the rate of substitution among sites, parsimony can give biased results, even when the levels of divergence are quite modest (Eyre-Walker, 1998; Alvarez-Valin et al., 2004). Furthermore, the artifact of parsimony based ancestor reconstruction has been shown to be strong enough to explain the vanishing of GC-rich regions, at least in cetartiodactyls (Alvarez-Valin et al., 2004). Therefore, it is worth testing whether GC-rich isochores are actually vanishing from mammalian genomes, using the maximum likelihood approach. Maximum likelihood methods are more appropriate for relatively high levels of divergence between sequences than the maximum

[*] Corresponding author.
E-mail address: whli@uchicago.edu (W.-H. Li).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.03.026

parsimony method (Arndt et al., 2003, Belle et al., 2004). Belle et al. (2004) applied a maximum likelihood method based on a nonhomogeneous substitution model to a total of 41 orthologous genes in species ranging from eutherian mammals to marsupials, representing roughly 130 million years of evolutionary history. From the common ancestor of mammals to extant eutherian species, a significant GC content decrease across all genes has been identified, largely due to GC-rich genes. Further, the GC content change in different evolutionary phases suggested that the decrease in GC content was stronger at the beginning of mammalian evolution.

In this study, we first used a maximum likelihood method (Yang et al., 1995) to infer the ancestral nucleotide states and to estimate the ancestral GC content at each internal node in the mammalian phylogeny. Next, we used a second maximum likelihood method that implements a model of heterogeneous substitution rate among sites and nonstationary base composition across branches of a phylogeny (Galtier and Gouy, 1998). The GC content at each node of a tree can be estimated under this substitution model. This allows us to study isochore evolution in terms of GC content change along the mammalian tree.

The completion of a draft genomic sequence of a marsupial (*Monodelphis domestica*) and the accumulation of sequences in various mammalian species provide a good opportunity to readdress the question of whether GC-rich isochores are vanishing in eutherian mammals. We collected datasets including 176 orthologous mammalian genes with at least one representative sequence of the following eutherian taxa: primates, rodents, lagomorphs, and cetartiodactyls or perissodactyls or carnivores and also marsupial sequences as outgroups and performed the maximum likelihood analysis. Overall, no evidence was found to support a general decline in GC content in the evolution of GC-rich genes along eutherian lineages, although a moderate level of decrease in GC content in GC-rich genes was observed in the primate and rodent lineages. In fact, we observed an increasing trend of GC content in the Lagomorpha and artiodactyl lineages.

2. Materials and methods

2.1. Data

Mammalian genes with representative eutherian sequences were extracted from the HOVERGEN (Duret et al., 1994) database (release 42; Duret et al., 1994). All pairwise blastp searches were conducted among members of each gene family excluding partial sequences (length <80% full length of the human sequence). The orthologous sequences of a gene were defined as the best reciprocal blast hits in different mammalian species (Lee et al., 2002) and checked by phylogenetic analysis (see below). The protein sequences were aligned using ClustalW with the default parameters (Thompson et al., 1994) and adjusted manually if necessary. The length of the alignable region between marsupial and eutherian sequences is longer than 80% of the full length sequence and the sequence identity should be greater than 70%. Phylogenetic analysis was performed to confirm the orthology using the neighbor-joining method (MEGA2) with the Poisson distance between protein sequences. The known phylogenetic relationship of all eutherian groups was given based on the well-accepted consensus phylogenies (Murphy et al., 2001; Liu et al., 2001 for Eutheria; Colgan (1999) for metatherian; tree of life URL http://tolweb.org/tree/phylogeny.html). For the cases in which gene trees showed a striking disagreement with the known phylogeny, a blastp search against the non-redundant database at NCBI was applied to reduce potential paralogy problems. For each orthologous gene, at least one representative sequence from the following eutherian groups was used in our analysis: primates, rodents, lagomorphs, and cetartiodactyls or perissodactyls or carnivores. The orthologous genes of the metatherian outgroup (marsupial, *M. domestica*) were extracted from the newly available marsupial genome database (ENSEMBL, 2005 version 2.0, http://www.ensembl.org/index.html). For each gene, the orthologous eutherian and metatherian protein sequences were aligned and the corresponding nucleotide sequence alignments were deduced by back-translating the aligned protein sequences.

2.2. Inference of ancestral sequences and GC contents

A maximum likelihood method (http://abacus.gene.ucl.ac.uk/software/paml.html) was employed to reconstruct the ancestral nucleotide sequence at each interior node and at the root of a given phylogeny, based on the aligned extant nucleotide sequences, under the TN93 model, which allows for sites to have different rates of substitution (Yang et al., 1995). The program allows one to model the rate variation among sites according to a gamma distribution. The maximum likelihood estimates of parameters in the model (e.g., branch length) were used to compare the posterior probabilities of the ancestral nucleotide state at each site. In contrast to the maximum parsimony approach, which assigns a single ancestral state for each nucleotide site, this method (referred to as PAML in this paper) assigns the nucleotide state with the highest posterior probability to the inferred ancestral state. The corresponding GC content (including the GC content at position 3, denoted by GC_3, and the GC content at positions 1 and 2, denoted by GC_{12}) in each extant or ancestral taxon can be calculated based on the observed or derived nucleotide states at each site.

We also estimated the ancestral GC content at each node in the phylogeny using the maximum likelihood method developed by Galtier and Gouy (1998) (referred to as NHML in this paper). This method implements a model of nucleotide substitution that accounts for variable base compositions among extant sequences compared to traditional models. The substitution process on each branch follows Tamura's (1992) model with unequal equilibrium GC contents among branches, so that the GC level can vary with time and among lineages. Instead of directly inferring the ancestral base state, the ancestral level of GC contents of a phylogeny was parameterized in the model. The GC content at the root and the location of the root are two additional parameters of the model. The Newton–Raphson optimization algorithm was implemented to estimate the values