A New Framework for Studying the Isochore Evolution: Estimation of the Equilibrium GC Content Based on the Temporal Mutation Rate Model

Satoshi OOta*[,1], Kazuhiro Kawamura[2], Yosuke Kawai[3], and Naruya Saitou[4]

[1]Department of Biological Systems, RIKEN BioResource Center, Tsukuba, 305-0074 Japan

[2]Institute of Mathematics, University of Tsukuba, Tsukuba, 305-8571 Japan

[3]Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, Kusatsu, 525-8577 Japan

[4]Division of Population Genetics, National Institute of Genetics, Mishima, 411-8540 Japan

*Corresponding author: E-mail: oota@riken.jp.

Abstract

Isochore is the genome-wide mosaic structure in guanine-cytosine (GC) content. The origin of isochores is thought to have emerged in the ancestral amniote genome, and the GC-rich isochore is eroded in the mammalian lineages. However, there are many enigmas in the isochore evolution: 1) although all the mammalians, birds, and even reptiles, which are clearly polyphyletic, have isochore, opossum and platypus lack GC-rich and GC-poor isochore classes; 2) although the isochore is predicted to vanish according to a fairly robust theory, a completely opposite conclusion was led in some mammalian lineages; and 3) the major three hypotheses on the isochore evolution cannot explain observed evidences without flaws. So far compositional evolution has been studied under the assumption that per base pair rate of GC→AT (u) and AT→GC (v) mutations are temporally constant (the constant model). With this model alone, however, it is difficult to explain the isochore evolution. We propose a simple model for compositional evolution based on the temporal per base pair rate of mutations (the variable model). In this model, rates u and v vary depending on temporal GC contents. Mathematically, the variable model is an expansion of the constant model. By using high-density human single nucleotide polymorphism data, we compared the variable model with the constant model. Although the variable model gave consistent results with the constant model, it can potentially describe the complicated isochore evolution, which the constant model cannot explain. The versatile characteristics of the variable model may shed new light on the mysterious isochore evolution.

Key words: isochore, equilibrium, the human genome, the chimpanzee genome, SNP.

Introduction

One of the distinctive features of the amniote genome is the huge mosaic structure in guanine-cytosine (GC) content variation known as isochore (Bernardi et al. 1985). In mammals, GC-rich isochore has three families: that is, H1, H2, and H3 (Bernardi 1993). The large-scale variability of base composition is significantly related to various genomic features: densities of LINE and ALU elements, level of methylation, recombination rates, and gene densities (Mouchiroud et al. 1991; Eyre-Walker 1993; Duret et al. 1995; Jabbari and Bernardi 1998; Smit 1999; Fullerton et al. 2001; Lander et al. 2001).

Biological meaning of isochore is still an open question: why did isochore emerge and why was it maintained against frequent mutations in putatively evolutionarily neutral regions (e.g., intergenic regions)? There are mainly three hypotheses in regarding to the isochore evolution: 1) selection (Bernardi et al. 1985), 2) the mutation bias (Sueoka 1988; Wolfe et al. 1989), and 3) biased gene conversion (Holmquist 1992; Eyre-Walker 1993) hypotheses. So far no evidence was found to exclusively support one of them, suggesting that the isochore evolution harbors a complicated evolutionary mechanism.

When we consider phylogenies of various organisms whose genome data are recently available, the isochore evolution is further enigmatic. Figure 1 shows polyphyletic characteristics of isochore families. In this figure, the isochore
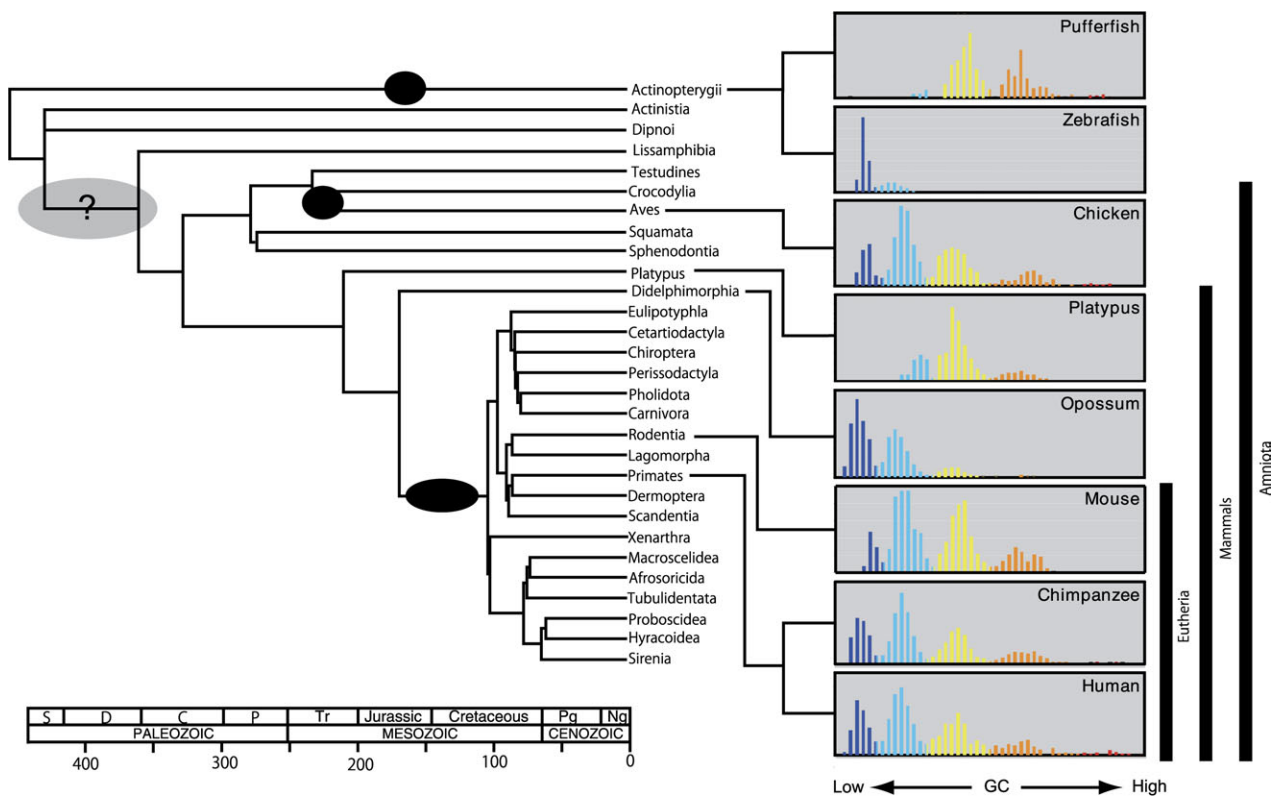
Fig. 1.—Polyphyletic characteristics of relative amounts of isochore families. Although primates, Rodentia, and Aves have clear isochore structures, Didelphimorphia and platypus lack GC-rich and GC-poor isochore families, respectively. Gray ellipse (indicated by a question symbol) represents a putative origin of GC-rich isochore. Black ellipses represent alternative (multiple) origins of GC-rich isochore. Scale at bottom represents the geological era (S, Silurian; D, Devonian; C, Carboniferous; P, Permian; Tr, Triassic; Pg, Paleogene; Ng, Neogene). Time unit is million years. Histograms represent weighted distributions of relative amounts of isochore families of various organisms (Costantini et al. 2006). Abscissa and ordinate are GC contents and relative amounts of sequences, respectively. Colors in the histograms represent five isochore families (blue, L1; cyan, L2; yellow, H1; orange, H2; red, H3) (for details, see Costantini et al. 2006). Phylogenetic tree and histograms are based on Hedges et al. (2006) and Costantini et al. (2006). Nomenclature of OTUs does not follow precise taxonomic hierarchies.

maps were used (Costantini et al. 2006, 2009) to represent isochore structures. Instead of using mere GC content distribution in the genome, this approach employs the plateau values reached by the standard deviation of GC levels of isochors belonging to different isochore families, that is, the results are expected to reflect spatial isochore structures in terms of the original definition. Duret et al. (2002) claimed that the origin of isochore is the ancestral amniote genome (gray ellipse in fig. 1), and the GC-rich isochore is eroded in the mammalian lineages. However, although primates (human and chimpanzee), Rodentia (mouse), and Aves (chicken) have clear isochore structures, Didelphimorphia (opossum) and platypus lack GC-rich and GC-poor isochore families, respectively. These suggest lineage-specific isochore evolution or multiple emergence of isochore (black ellipses in fig. 1), both of which are hard to be explained by the existing theories.

One of the difficulties underlying the isochore studies is that there were no sufficient data to analyze the isochore evolution with an appropriate evolutionary model. Major part of isochore resides in noncoding genomic regions, which are in general not alignable between distantly related species like human and mouse. Conventional evolutionary models are empirically based on substitution patterns in coding regions (in many cases, substitution patterns in synonymous sites and introns are employed to exclude selective pressures). In other words, we implicitly assume that evolution of alignable genic regions fairly reflect genome-level isochore evolution.

By using synonymous substitution patterns in coding sequences, Duret et al. (2002) and Belle et al. (2004) predicted that mammalian isochore is vanishing. Meanwhile, Alvarez-Valin et al. (2004) and Gu and Li (2006) claims that the vanishing isochore theory may be wrong due to misinference of ancestral states because the "parsimony-based" method may not be appropriate to analyze the long-term evolution. Gu and Li (2006) thus applied the maximum likelihood (ML) method, which is more robust than the parsimony-based method, using the nonhomogeneous model (Galtier and Gouy 1998), to reconstruct ancestral states. Surprisingly,

they led to an opposite conclusion to that of Duret et al (2002): the GC content of GC-rich genes appears to have increased in recent times at least in some lineages, for example, the rabbit genome. Galtier and Gouy (1998) claimed that the nonhomogeneous model is able to estimate accurate ancestral compositions under a given evolutionary model. As mentioned above, however, conventional evolutionary models (including the nonhomogeneous model) were originally devised to analyze genic regions not to analyze genomic regions. The definition of isochore is the genome-wide nonrandom spatial variation in GC content. Taking them into consideration, we should be cautious to analyze isochore evolution by using the conventional models.

Webster et al. (2003) examined the compositional evolution of noncoding DNA by using 1.8 Mb genomic alignments of human, chimpanzee, and baboon. According to mutation patterns inferred by parsimony in human single nucleotide polymorphisms (SNPs), the equilibrium GC content of each GC content class was not significantly different. This again suggests homogenization of the isochore structure (the vanishing isochore theory). Because it may be inappropriate to parsimoniously reconstruct ancestral GC levels among even closely related species, they carefully compared results by the conventional parsimony-based method with those by the ML method and found virtually no difference between them. Therefore, the model-dependent mis-inference is likely avoided in the genomic substitution patterns in the human and chimpanzee lineage.

Since most of the SNP sites in intergenic regions are approximately subject to neutral evolution, human SNP data and the human and chimpanzee genomes are ideal extant data. Today, large amounts of reliable human SNP data are available at the international HapMap project (International HapMap Consortium 2004; Couzin 2006). To infer reliable mutation patterns, we used the human genome and SNP data obtained from the HapMap database and dbSNP as well as the chimpanzee genome as outgroup (a reference). In this paper, we focus on mutation patterns, which can be estimated from the SNP data and the reference genome data, to study the isochore evolution. Our objectives are: 1) by using the high-density SNP data set to detect characteristic evolutionary patterns that were missed in the previous studies and 2) to develop a simple mathematical model of the compositional evolution.

Basic Concepts for the Model

To estimate the equilibrium GC content, in general, the per base pair rate of GC → AT (u) and AT → GC (v) mutations have been assumed to be temporally constant (Webster et al. 2003): that is, a time-homogeneous Markov process is used in the model. We call this model "the constant model." However, as Galtier and Gouy (1998) mentioned, this may be an unrealistic assumption. The constant model

implicitly assumes that a genomic region has position-specific mutation rates u and v. On the other hand, it is shown that both of rates u and v are positively correlated to GC contents (Smith et al. 2002; Webster et al. 2003). This raises a question: can rates u and v be temporally constant while GC content is changing?

According to the previous studies, inferred mutation patterns (in fact, synonymous substitution patterns in many cases) have a tendency to converge toward an equilibrium state. If it is exactly true, by using a sufficiently rich data set and a reliable evolutionary model, we can estimate a set of similar equilibrium values to each other independent from initial GC contents.

We propose a simple model that explains compositional evolution without a complicated parameter set. We call this model "the variable model": unlike the constant model, the per base pair rate of GC → AT (u) and AT → GC (v) mutations vary depending on time. It is virtually impossible to estimate functions u and v of time t by using currently available data. Instead, we propose a more specific model under the assumption that rates u and v are constant in each GC content class but vary in time. In the constant model, u and v are fixed depending on "observed" GC content. In the variable model, u and v are determined depending on "temporal" GC content: that is, they are functions of temporal GC content GC(t) (GC content at time t). Therefore, this is not a time-homogeneous Markov model and is determined by two functions $u(\text{GC}(t))$ and $v(\text{GC}(t))$. Once GC content at time t is given, the variable model allows us to estimate values of u and v at time t. Consequently, GC content at time $t+1$ is recursively estimated from the one at time t. Because time t is arbitrary, we use reconstructed ancestral GC content as an initial GC content at $t = 0$. The constant model can be defined as a special case of the variable model (($u(\text{GC}(t))$ and $v(\text{GC}(t))$ are constant functions here). In this paper, we formalize the variable model as an extension of the constant model and evaluate the model by using an actual rich data set: the human and chimpanzee genomes.

Materials and Methods

Analysis of Mutation Patterns on the Human Lineage

A total of 3,100,436 SNPs of the human genome were obtained from the HapMap database (The International HapMap Consortium 2003) and dbSNP (Sherry et al. 1999). By using Blast program (Altschul et al. 1990), we mapped the SNPs on the chimpanzee genome (National Center for Biotechnology Information [NCBI] PanTro1.1) by using 5′ and 3′ flanking regions of the SNPs as queries. To distinguish coding and transcribed regions from the other genomic regions, we used the Ensembl annotation (release41) of the human genome from the UCSC genome annotation database for the March 2006 GenBank freeze
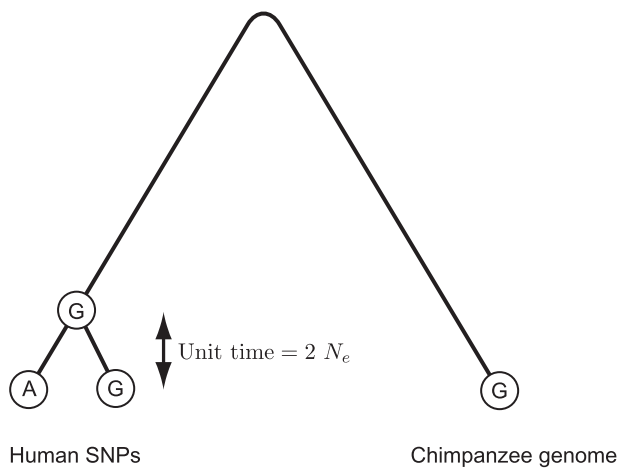
FIG. 2.—Schematic representation of inference of a mutation pattern by using human SNP and the chimpanzee genome as a reference. In this figure, the human genome has a polymorphism adenine (A) and guanine (G) in a site. A homolog site in the chimpanzee genome has G. Therefore, an ancestral state of the human genome is inferred to be G, according to the parsimonious method. Note that the coalescence time of the human SNP is regarded as a unit time in the following formalization. Expected length of the unit time is approximately $2N_eX20$ years because the average time to coalescence for two randomly chosen SNPs is $2N_e$ generations, where N_e is an expected effective population size of the human lineage.

assembled by NCBI (hg18, Build 36.1). We also excluded putative CpG islands and repetitive regions from the human genome, which were predetected by the UCSC genome database. The human genome was divided into 2 kb windows (nonoverlapped bins). Only biallelic SNPs, in which the chimpanzee root was the same as one of the two human alleles (fig. 2), and windows greater than or equal to 100 bp were used in this analysis.

The genome data are mixture of derived and ancestral states in polymorphic loci. To estimate accurate per base pair rate of GC→AT (u) and AT→GC (v) mutations (Sueoka 1988, 1993), we need to reconstruct nucleotide compositions of the ancestral genome of the human population. Ancestral and derived nucleotide compositions are generally nearly identical to each other in the human genome. However, especially when the number of SNP data is huge (i.e., SNP density is high) and/or an effective window (bin) size can happen to be small, an effect of the heterogeneity is not negligible. We reconstructed "ancestral" chromosomes of the human population by the parsimony-based method.

We basically followed Webster et al. (2003) to estimate the per base pair rate of mutations for various GC content classes. They estimated equilibrium GC contents by using five GC content classes ($GC_1 < 0.3$, $0.3 \leq GC_2 < 0.4$, $0.4 \leq GC_3 < 0.5$, $0.5 \leq GC_4 < 0.6$, and $GC_5 \leq 0.6$, where GC_i is the i th GC content class) with parsimony-inferred mutation patterns. Owning to the high-density SNP data,

we categorized the known mutations in human into 50 GC classes by 49 GC content values (0.02, 0.04, . . . , and 0.998). We performed bootstrap resampling from the effective bins, assuming that mutations independently occurred in different bins.

Estimation of the Equilibrium GC Contents by the Constant and Variable Models

We consider expected GC content transitions with discrete time $t = 0, 1, 2, . . .$ in the following description.

The Constant Model. The constant model is based on the following two assumptions:

Assumption 1. The mutation rates u and v are constant in each GC class and further.

Assumption 2. The mutation rates u and v are constant in time as well.

Let $\bar{N}_{GC}(t)$ be the expected number of G + C sites at time t. Under these assumptions, $\bar{N}_{GC}(t+1)$ and $\bar{N}_{GC}(t)$ satisfy the following formula

$$\bar{N}_{GC}(t + 1) = \bar{N}_{GC}(t) + \bar{N}_{AT \to GC}(t) - \bar{N}_{GC \to AT}(t), \quad (1)$$

where $\bar{N}_{AT \to GC}(t)$ and $\bar{N}_{GC \to AT}(t)$ are the expected numbers of transitions AT->GC and GC->AT at time t, respectively. In the constant model, $\bar{N}_{AT \to GC}(t)$ and $\bar{N}_{GC \to AT}(t)$ are written as $\bar{N}_{AT \to GC}(t) = v\bar{N}_{AT}(t)$ and $\bar{N}_{GC \to AT}(t) = u\bar{N}_{GC}(t)$, respectively.

Note that $\bar{N}_{AT}(t) + \bar{N}_{GC}(t) \equiv L$, the total number of sites. From these we obtain

$$\begin{aligned}\bar{N}_{GC}(t + 1) &= \bar{N}_{GC}(t) + v\bar{N}_{AT}(t) - u\bar{N}_{GC}(t) \\ &= \bar{N}_{GC}(t) + v(L - \bar{N}_{GC}(t)) - u\bar{N}_{GC}(t) \\ &= (1 - u - v)\bar{N}_{GC}(t) + vL.\end{aligned}$$

Dividing both sides by L and noticing that $GC(t) = \bar{N}_{GC}(t)/L$, we obtain

$$GC(t + 1) = (1 - u - v)GC(t) + v, \quad (2)$$

where $GC(t)$ is GC content at time t. Suppose that $GC(t)$ has reached an equilibrium f^* at time t_e. Then $f^* = GC(t_e) = GC(t_e + 1) = \cdots$. By equation (2), we obtain the equilibrium formula (Sueoka 1988, 1993) $f^* = \frac{v}{u+v}$.

The Variable Model. *The Derivation of the Variable Model by Generalizing the Constant Model.* Now we retain Assumption 1 and replace Assumption 2 by:

Assumption 3. The mutation rates u and v vary in time.

Furthermore, we make a stronger assumption:

Assumption 3a. The mutation rates u and v are functions $u = u(GC(t))$ and $v = vGC(t)$ of the GC content at time t, $GC(t)$.

We do not claim that Assumption 3a is theoretically justified. First of all, there is no a priori reason to assume that mutation rates u and v depend only on $GC(t)$. Secondly, the

mutation rates u and v, if they vary in time, should be regarded as stochastic processes and the distributions of $u(GC(t))$ and $v(GC(t))$ should play an important role. Assumption 3a states that u and v follow deterministic rules defined by the functions $u(GC(t))$ and $v(GC(t))$.

On the basis of Assumption 3a, the formula (2) is changed as follows

$$GC(t + 1) = GC(t) - GC(t)\{u(GC(t)) \\ + v(GC(t))\} + v(GC(t)). \quad (3)$$

When the GC content (at the site we are examining) has reached the equilibrium f_∞ at time t_e, then we obtain $f_\infty = GC(t_e) = GC(t_e + 1) = GC(t_e + 2) = \ldots$. From this with equation (3), we obtain $f_\infty = f_\infty - f_\infty\{u(GC(t_e)) + v(GC(t_e))\} + v(GC(t_e))$ and Sueoka's formula in this context takes an analogs form

$$f_\infty = \frac{v(GC(t_e))}{u(GC(t_e)) + v(GC(t_e))}. \quad (4)$$

We still do not know explicit formulas of $u(GC(t))$ and $v(GC(t))$. These functions are to be estimated from the actual mutation patterns (AT->GC and GC->AT, respectively) by using the 50 GC classes described above. Because formula 3 is concerned with the GC contents at time t and time $t + 1$, we choose the initial time $t = 0$ as an initial state or $t = 1$ as an observed state.

By the generalization, the model no longer follows the stationery Markov process because the transition matrices changes depending on time. Before proceeding further, we make a digression to an auxiliary consideration.

An Auxiliary Consideration on Potential Equilibria. We introduce an auxiliary function $f(x)$ defined by

$$f(x) = (1 - u(x) - v(x))x + v(x), \quad 0 \le x \le 1, \quad (5)$$

where $u(x)$ and $v(x)$ are smooth functions, such that $0 < u(x)$, $v(x) << 1$. The change of variable $x = GC(t)$ gives equation (3) and the function $f(x)$ stands for a transition rule of the GC content. Because $f(x)$ is given by the two mutation rates $u(x)$ and $v(x)$, a temporal GC content $GC(t)$ is also given by the two mutation rates in a recursive way. We have $f(0) = v(0) > 0$ and $f(1) = 1 - u(1) < 1$. These two inequalities imply that there is at least a value x_e such that $f(x_e) = x_e$. Such value is also called a "fixed point." There may be several fixed points, each of which corresponds to a potential equilibrium.

The behavior of iterations of the function f near a potential equilibrium x_e is described by the absolute value of the derivative, $|f'(x_e)|$ (most likely <1 or >1 and not exactly equal to 1 in our context). Let $f^{(2)}(x) = f(f(x))$, $f^{(3)}(x) = f(f(f(x)))$, and so on.

Class I: $0 < f'(x_e) < 1$: When we start with a value $x_0 < x_e$, which is "sufficiently close" to x_e, then we have $x_0 < f(x_0) <$
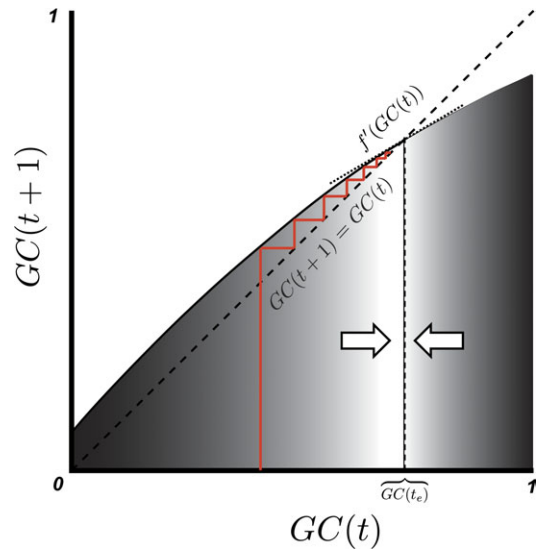

Fig. 3.—Schematic representations of a typical transition of GC content under the $f(GC(t))$ framework (red lines). The diagonal broken line represents $f(GC(t)) = GC(t+1) = GC(t)$, that is, a set of fixed points and, in this case, the equilibrium states. The intersection of $GC(t)$ and $GC(t + 1)$ represents an equilibrium. GC content transition (red lines) asymptotically approaches the equilibrium (GC_e). The variable model predicts asymptotical convergence of transitions on potential equilibrium $GC(t_e)$, where t_e is equilibrium time. The total derivative $f'(GC(t))$ satisfies condition $0 < f'(GC(t_e)) < 1$ (Class I).

$f^{(2)}(x_0) < \ldots < f^{(n)}(x_0) < \to x_e$ as $n \to \infty$. Similarly when we start with a value $x_0 > x_e$, which is sufficiently close to x_e, then we have $x_0 > f(x_0) > f^{(2)}(x_0) > \ldots > f^{(n)}(x_0) > \to x_e$ as $n \to \infty$ (see fig. 3).

Class II: $-1 < f'(x_e) < 0$: When we start with a value $x_0 < x_e$, which is sufficiently close to x_e, then we have an "oscillation" $x_0 <x_e> f(x_0)$ áxₑñ $> f^{(2)}(x_e) <x_e> f^{(3)}(x_e)< \ldots$ and $f^{(n)}(x_0) \to x_e$ as $n \to \infty$ (see fig. 4a).

Class III: $|f'(x_e)| > 1$: Here, the behavior of iterations of f exhibits a sharp contrast with Classes I and II. Even if we start with a value x arbitrarily close to x_e, the iterated values $f^{(n)}(x)$ are "repelled" from x_e, and it is not possible to predict their asymptotic behavior in general.

Although sometimes there may be periodic values x_p in the sense that after some iteration of the function f, x_p comes back to itself. In formula, this is expressed as: $x_p = f^{(n)}(x_p) := f(f(f \ldots (n \text{ times}) (x_p)))$.

If the derivative $(f^{(n)})'(x_p) = f'(f^{(n-1)}(x_p)) f'(f^{(n-1)}(x_p)) \ldots f'(x_p)$ falls into Class I or Class II, we may derive similar conclusions on the periodic value x_p (see fig. 4b).

In our application to the actual data set, $f(x)$ is considerably close to x. It is convenient to introduce a function $g(x) = x - f(x)$. Noticing that $g'(x) = 1 - f'(x)$, we have

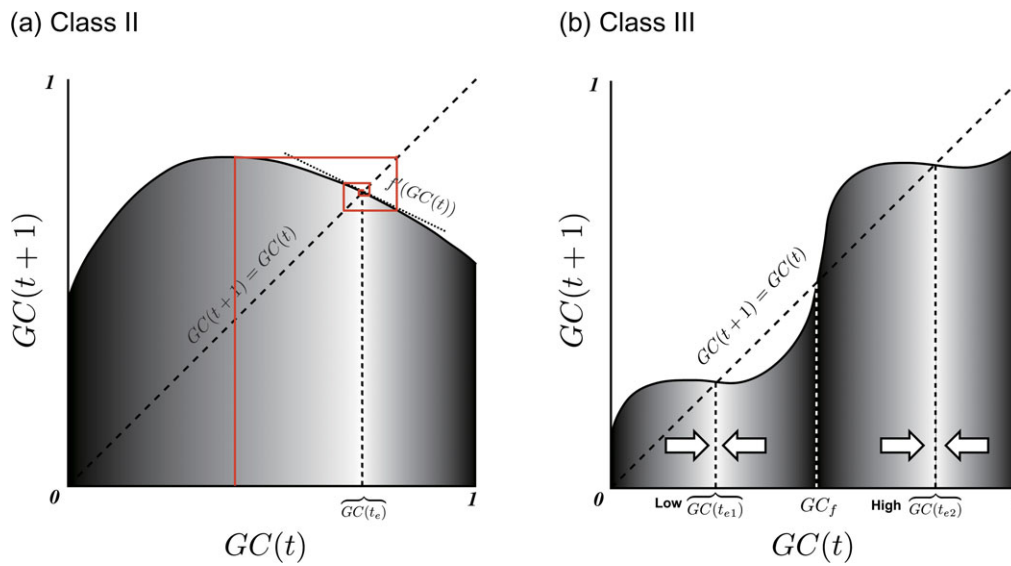$$\text{Class I}: 0 < f'(x_e) < 1 \quad \text{if and only if} \quad 0 < g'(x_e) < 1, \quad (6.1)$$

Fig. 4.—Schematic representations of "atypical" transitions of GC content under the $f(GC(t))$ framework (red lines). (a) Class II: oscillatory convergence of transitions on a potential equilibrium $GC(t_e)$. The total derivative $f'(GC(t))$ must satisfy condition $-1 < f'(GC(t_e)) < 0$. (b) Class III: divergence of transitions from fixed points.

$$\text{Class II}: \quad -1 < f'(x_e) < 0 \quad \text{if and only if} \quad 1 < g'(x_e) < 2,$$
(6.2)

$$\text{Class III}: \quad |f'(x_e)| > 1 \quad \text{if and only if} \quad g'(x_e) < 0 \text{ or } 2 < g'(x_e).$$
(6.3)

A Preliminary Assessment. To briefly assess characteristics of the variable model, we first performed three kinds of computations by using human chromosome 21 (HSA21): 1) quadratic fitting and numeric: u and v were given by nonlinear regression to fit the five sets of replicated data to quadratic models, and the equilibrium GC contents were estimated by giving a large value of time t (say, 2,000 unit times); 2) quadratic fitting and analytic: the equilibrium GC contents were analytically estimated under assumption that $GC(t_e - 1) \simeq GC(t_e) = GC(t_e + 1)$ holds, where t_e is time when the GC content reaches the equilibrium; 3) GC class: u and v are estimated according to discrete GC classes (as a step function), say, $0 \leq GC < 0.3$, $0.3 \leq GC < 0.4$, $0.4 \leq GC < 0.5$, $0.5 \leq GC < 0.6$, and $0.6 \leq GC < 1.0$. Six initial GC contents were used (0.1, 0.3, 0.4, 0.5, 0.6, and 0.8).

Note that the 2,000 unit times are 2,000 times as long as the average time to coalescence for two randomly chosen SNPs: approximately 800 My, assuming that the effective population size for the human lineage is 10,000 and the generation time for human is 20 years.

Results

Figure 5 shows transitions of GC content under the constant model with various initial GC content classes (0.2, 0.3, 0.4,

0.5, 0.6, 0.7, and 0.8). To examine overall characteristics of transitions of GC content, $u(GC(t))$ and $v(GC(t))$, are simplified by nonlinear regression: $\hat{u} = 0.002014 - 0.000735\hat{GC}^2$ and $\hat{v} = 0.003486 - 0.000812\hat{GC}^2$, respectively. The fitting was better in the quadratic models than the linear models: R-squared values of rates u and v are 0.623 and 0.547 in the linear models, respectively, whereas those in the quadratic models are 0.700 and 0.551, respectively. In the constant model, each $\hat{GC}$ was given as an initial GC content for each GC content class. Although it is known that GC contents under the constant model take extremely long time to reach the equilibria (Sueoka 1962), an estimated convergence time is around 350 My under the given conditions: 1) the intergenic regions we used are subject to neutral evolution, 2) the effective population size of the human lineages is 10,000, and 3) the generation time for human is 20 years. It is approximately equal to or rather shorter than 400 My, which corresponds to the divergence time of amniotes: a putative age of the origin of isochore (Duret et al. 2002) (fig. 1). Table 1 shows comparison of GC content equilibria of HSA21 between the constant and variable models. The equilibrium GC contents under the constant model were fairly similar. Meanwhile, owing to the high-density SNP data, the small differences appeared to be significant (fig. 6a). Although Webster et al. (2003) did not report any significant differences of the equilibrium among GC classes, our results were consistent with theirs in terms of those similar equilibria, suggesting homogenization or "vanishing" isochore under the constant model. Meanwhile, under the variable model, all the GC content categories converged on the exactly same equilibrium (table 1 and fig. 6b), which
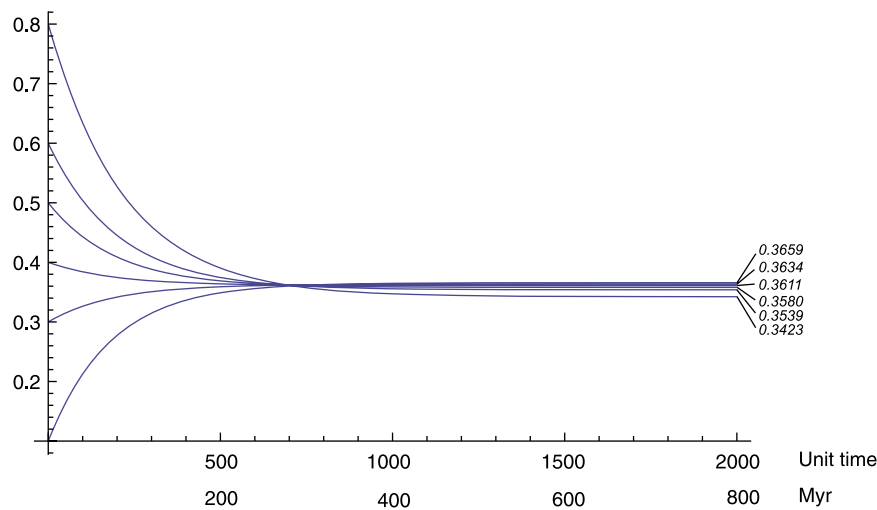
FIG. 5.—Transitions of CG content of chromosome 21 (HSA21) under the constant model. Italic bold figures represent initial GC content; Italic figures represent GC content values at $t = 2{,}000$ (putative unit times to reach the equilibria). A unit time represents approximately 400,000 years.

is still close to the equilibrium values under the constant model.

Next, we examined transition behaviors of GC content based on the $f(GC(t))$ framework, the one which describes $GC(t + 1)$ by a function $f(GC(t))$ of $GC(t)$. We used the functions $u(GC(t))$ and $v(GC(t))$ instead of the simplified quadratic models and the stepwise functions of GC content classes. Figure 7 shows the per base pair rate of $GC \rightarrow AT$ (u) and $AT \rightarrow GC$ (v) mutations corresponding to GC content in human chromosome 3 (HSA3). Note that figure 7a and b shows the relationships between GC content and rates u and v, respectively, instead of the net mutation rates. By using synonymous substitution patterns, Smith et al. (2002) showed that both u and v are positively correlated to GC content. However, detailed relationships estimated

from the rich SNP data were more complicated in the intergenic regions. Owing to the high-density SNP data, we could detect multimodal structures of relationships between GC and per base rates u and v. These multimodal structures of estimated $u(GC(t))$ and $v(GC(t))$ were observed in all chromosomes (data not shown). According to the re-sampled data, the structural relationships between GC content and the per base pair rates are moderately stable across the chromosome, suggesting that the relationships are universal at intrachromosomal level. Figure 8 shows relationships between $GC(t)$ and $g(GC(t)) = GC(t) - f(GC(t)) = GC(t) - GC(t - 1)$ under the variable model (HSA21). A root of $g(GC(t)) = 0$ is a fixed point (see section, An auxiliary consideration on potential equilibria) of $f(GC(t))$ and, in this case, a potential equilibrium value since the derivative at the fixed point satisfied the condition given by formulas (6.1) and (6.2) (note that $g(GC(t))$ is the function of $GC(t)$, not t, in our context). This value belongs to Class I in the transition behavior of GC content (fig. 3). As shown in figure 8, while $g(GC(t))$ varies depending on resampled SNP data from HSA21, the variation is moderate (for details, see figure legends).

We furthermore elaborated on characteristics of $g(GC(t))$ for each replicated data set. Figure 9 shows distribution of fixed points of $f(GC(t))$ and the derivatives at fixed points of $g(GC(t))$ in chromosome 1 (HSA1). There are mainly three clusters: low (~ 0.2), medium (~ 0.37), and high (~ 0.7). The derivatives at the fixed points have a tendency to be negative, positive (or nearly zero), and negative for the low, medium, and high clusters, respectively. Because many of the fixed points of the medium cluster satisfy the conditions given by formulas (6.1) and (6.2), they are suggested to be potential equilibria. However, there are a considerable number of fixed points that do not follow the tendency: for

Table 1

Comparison of GC Content Equilibria of HSA21 between the Constant and Variable Models

		Equilibrium		
		Variable		
		Quadratic Fitting		
Initial GC	Constant	Numeric	Analytic	GC Class
0.1	0.3659	0.3620		0.3683
0.3	0.3634	0.3620		0.3683
0.4	0.3611	0.3620		0.3683
0.5	0.3580	0.3620	0.3620	0.3683
0.6	0.3539	0.3620		0.3683
0.8	0.3423	0.3620		0.3683

NOTE.—Initial GC: GC content at time $t = 0$, which corresponds to GC content classes; Constant, the constant model; Variable, the variable models; Quadratic fitting, u and v are given as quadratic models of local GC content; Numeric, numeric estimation of the equilibrium GC content by giving a large value to time t; Analytic, analytical estimation of the equilibrium GC content; GC class, u and v are estimated according to discrete GC classes (as a step function). For detail, see text.
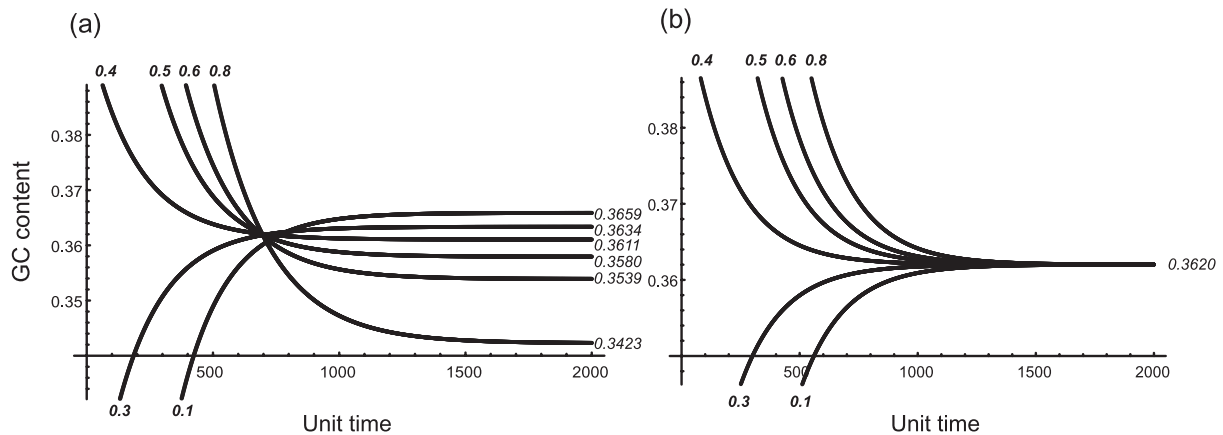
Fɪɢ. 6.—Close-ups of transitions of CG content of HSA21 under (*a*) the constant model and (*b*) the variable model. Although all the GC categories showed similar but different equilibria under the constant model, all the GC content categories converged on the same equilibrium under the variable model. For legend, see figure 5.

example, some of the fixed points at low and high GC content clusters satisfy the condition (6.3), many of which are thought to be "repulsive" fixed points leading to the adjacent potential equilibria. Figure 10 shows distribution of fixed points of $g(GC(t))$ in all chromosomes. CpG and non-CpG SNPs were distinguished. Overall, there are 2–4 clusters of fixed points for each of the data sets, suggesting that plural GC content equilibria and/or divergent behaviors under the variable model. GC content equilibria under the constant model (indicated by dashed and dotted lines for the all SNP and non-CpG SNP data, respectively) are well matched in mean GC values of one of the clusters, many of which are potential equilibria under the variable model. Surprisingly, except for these clusters, there were virtually no fixed points between them. This means that randomly sampled data from each chromosome share the characteristics, suggesting that the $f(GC(t))$ framework is considerably stable in the given data set.

Figure 11 shows the genome-level derivatives of $g(GC(t))$ at the fixed points appeared in figure 10. It is obvious that all chromosomes share similar characteristics in terms of distribution of the clusters. Note that we did not elaborate upon Class III except for repulsive fixed points because it was impossible to predict their asymptotic behavior in general.

Discussion

There are evolutionary enigmas in compositional evolution: 1) the opossum (Mikkelsen et al. 2007) and platypus (Warren et al. 2008) genomes lack GC-rich and GC-poor isochore families, respectively, whereas all the other mammalians, birds, and even reptiles, which are polyphyletic, have clear isochore structures (Duret et al. 2002); 2) rat and mouse show reduction of GC content variation (murine shift); 3) the ML-based inference can sometimes lead to a completely opposite conclusion against that by the
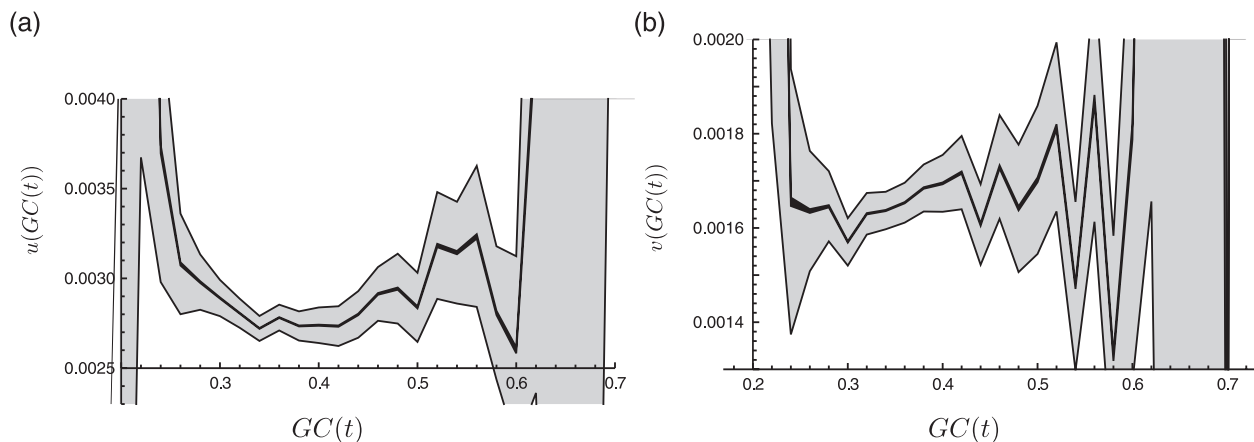

Fɪɢ. 7.—The per base pair rate of GC->AT (*u*) (*a*) and AT->GC (*v*) (*b*) mutations corresponding to GC content in chromosome 3 (HSA3). Thick (gray) and thin (black) regions indicate 95% confidence intervals based on a normal distribution fitted to 1,000 bootstrap replications of HSA3 data and for their means, respectively.
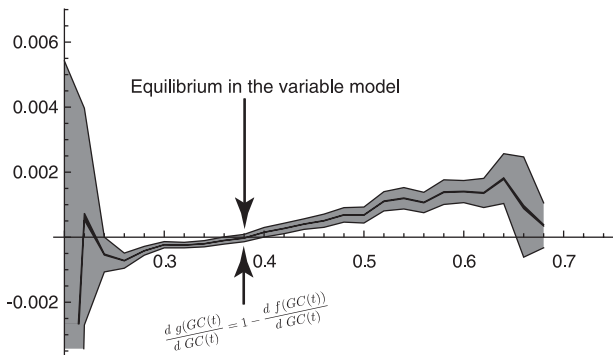
Fig. 8.—An example of relationships between GC(*t*) and *g*(GC(*t*)) = GC(*t*) − *f*(GC(*t*)) = GC(*t*) − GC(*t* + 1) in the variable model (HSA21). The upper and lower arrows indicate an equilibrium GC value and the slope of *g*(GC(*t*)) at the fixed point, respectively. For legend, see figure 7.

maximum parsimony-based inference (Gu and Li 2006); 4) the major three hypotheses, which are all sophisticated and plausible, cannot explain the isochore evolution without flaw (Eyre-Walker and Hurst 2001). These suggest that the existing compositional evolution frameworks need to be reconsidered.

The constant model (the time-homogeneous/stationary Markov model) is simple and intuitive. Many compositional evolution studies, explicitly or implicitly, have used this model: that is, the per base pair rate of mutations is assumed to be temporally constant. As a consequence, transitions of GC content are necessarily gradual and asymptotic to equilibrium states. However, with this model alone, it is difficult to elucidate the above four enigmatic phenomena.
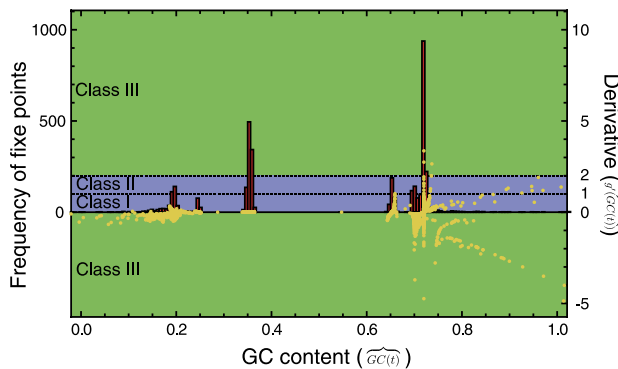

Fig. 9.—Distribution of fixed points of *f*(GC(*t*)) in chromosome 1 (HSA1). Left axis: frequencies of fixed points of *f*(GC(*t*)) in HSA1 (histogram); Right axis: the derivative at of fixed points of *g*(GC(*t*)) (yellow dots). Classes I and II: "attractive" fixed points or potential equilibria; Class III: "divergent" fixed points (green). Note that, in actual computation, the fixed points were given by roots of *g*(GC(*t*)) = 0. Therefore, the three classes in terms of behaviors of GC content transition were determined as follows: (*a*) 0 ≤ *g*′(GC(*t_e*)) < 2: convergence, suggesting an equilibrium (purple); (*b*) *g*′(GC(*t_e*)) > 2 or *g*′(GC(*t_e*)) < 0: divergence (green). The constant model shows equilibria around GC content ~0.4, which are consistent with those of the constant model with substitution patterns of GC₃.

Alvarez-Valin et al. (2004) and Gu and Li (2006) used the ML method with the nonhomogeneous model (i.e., the "time-nonhomogeneous" Markov model) (Galtier and Gouy 1998) to reconstruct ancestral states and obtained the opposite conclusion from the "vanishing isochore" hypothesis at least in certain lineages. We should note that their nonhomogeneous model is composed of different kinds of the time-homogeneous Markov models, which are branch specific in a phylogenetic tree of the compared organisms. Therefore, if we consider only one branch between ancestral and descend nodes, this model is identical to the time-homogeneous Markov model for estimating the ancestral base compositions.

The variable model is essentially different from the "nonhomogeneous" model. Our model is literally "time-heterogeneous" even on a branch. An a priori assumption that the variable model requires is that each GC content class has a moderately fixed per base pair rate as shown in section The Variable Model (Materials and Methods). This assumption may look like too strong. However, as long as GC content homogenization is predicted under the constant model, this assumption is implicitly required also in the constant model: if this assumption is violated, the same GC content class will have different equilibria, namely, the homogenization will not occur. Therefore, both the constant and variable models share the same assumption at this point.

One might be confused by the complicated characteristics of the variable model. Overall, the variable model has 2-fold characteristics: 1) if mutation patterns belong to Classes I and II, the variable model has a unique equilibrium regardless of initial GC content values and 2) if mutation patterns belong to Class III, the variable model can have plural equilibria. In the constant model, it is required that mutation patterns are temporally constant in a genomic region: that is, the constant model assumes that initial GC content-specific (or rather position-specific) mutation patterns last forever. This is a special case of the variable model when *u*(GC(*t*)) and *v*(GC(*t*)) are constant functions. Therefore, the variable and constant models are not antagonistic to each other, but the former is an extension of the latter.

In the net compositional evolution, the variable model provided consistent results with the constant model (fig. 10). Meanwhile, there exist plural clusters of fixed points, suggesting plural equilibria. For example, data belonging to Classes I and II in figure 9 indicate potential GC content equilibria, according to formulas (6.1) and (6.2). A question arising here is whether the isochore structure could reflect this multiplicity of equilibrium GC contents. The observed distribution of GC content across genomic regions in human is apparently unimodal. However, this does not contradict the multiplicity of equilibrium GC contents. Note that 1) the definition of isochore is the heterogeneity of spatial distribution of GC content not multimodality of the gross
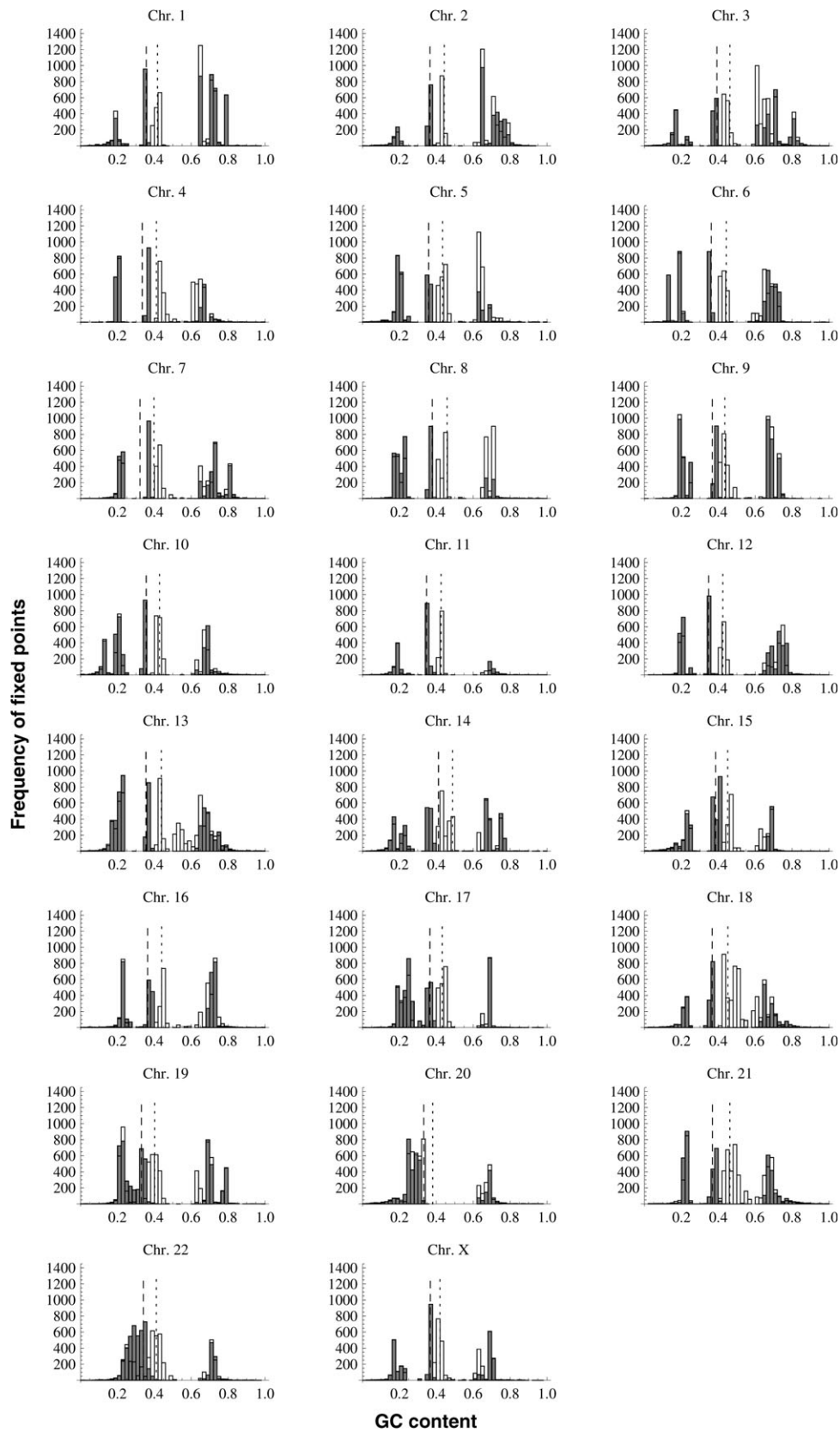
distribution (see the isochore maps in fig. 1) and that 2) we observe a snapshot of a GC content transition. As the previous studies (Webster et al. 2006; Duret and Arndt 2008) suggested, GC content is probably not at equilibrium in many mammalian lineages. Furthermore, unlike the constant model, the variable model allows a change of a mutation pattern at a genomic region. This makes it difficult to intuitively follow a behavior of GC content distribution. For example, if the transition is oscillatory (Class II), it is quite difficult to predict how the gross GC content distribution behaves. Besides that, it is not obvious that each chromosome shows a unimodal GC content distribution if we extract genomic regions putatively subject to the neutral evolution: for example, chromosomes 17 and 22 are apparently exceptional.

We assume that each chromosome has a pair of specific functions $u(GC(t))$ and $v(GC(t))$. The bootstrapped data share the similar characteristics in terms of distribution of fixed points overall (fig. 10; for the whole genome perspectives, see also supplementary figs. S1 and S2, Supplementary Material online). Although this assumption is moderate for the net compositional evolution, the functions $u(GC(t))$ and $v(GC(t))$ vary depending on positions in a chromosome. The nonlinearity of the variable model opens another possibility on the transitions of GC contents. As was shown in Li and Yorke (1975), a function $f(x)$ ($0 \leq x \leq 1$) satisfying a certain "global" condition (roughly illustrated in fig. 12; for detail, see Li and Yorke [1975]) exhibits a rather complicated behavior on its iterations. For example:

i) Arbitrarily small difference of initial conditions may cause a large deviation of long-term iterations;
ii) There may be many periodic points with almost arbitrary period.

This type of iteration behavior is generally called "chaotic," and many mathematical studies have been made on chaotic dynamical systems. Although no such transition function is clearly observed in our data, there could be a possibility that future studies find transition functions that exhibit chaotic behaviors. This might shed another light on the isochore evolution.

The variable model can explain both the gradual (and/or asymptotic) and the drastic (and/or chaotic) compositional transitions. By considering the present genomic compositions as a snapshot of dynamic transitions, the paradoxes of isochore are explained under the variable model: that is, a local genomic composition can occasionally converge on a certain GC content value and occasionally can diverge with tiny changes of mutation patterns. Under those conditions, GC content transitions are more rapid than those predicted by the constant model because tiny fluctuations of

$u(GC(t))$ and $v(GC(t))$ lead to drastic difference in GC content transitions.

How did isochore evolve? The previous study (Duret et al. 2002) suggests the origin of isochore is the ancestral amniote genome (fig. 1). Although their theory is plausible for many extant genome data, there are counterexamples: for example, the opossum (Mikkelsen et al. 2007) and platypus (Warren et al. 2008) genomes. As mentioned above, homogenization of isochore (vanishing isochore) is supported by the constant model. Because the homogenization takes extremely long time to reach the equilibrium (Sueoka 1962), we are tempted to consider that the observed isochore is just remnants of the original isochore of the ancestral amniote genome. However, under the constant model, the convergence time of vanishing isochore is around 350 My (see figs. 5 and 6). Because the estimated divergence time between the amniote and its counterparts is around 400 My (Hedges et al. 2006), it is hard to think that the extant isochore is mere remnants of the ancient isochore. Furthermore, we should note that the estimated convergence time (350 My) may be an overestimation according to the following reason. We set the unit time to $2 N_e \times 20 = 800,000$ years, assuming that the effective population size of the human lineage is 10,000. Because an effective population size is the harmonic mean of the actual numbers, the long-term effective population size tends to be dominated by the smallest temporal population size in a lineage: that is, the human effective population size (10,000) is putatively an overestimation as the long-term effective population size. Therefore, it is reasonable to think that the 800,000 years is an upper limit of the unit time and that the ancestral amniote genome would have reached the equilibrium by today under the constant model.

To solve this contradiction, we need to assume lineage-specific isochore evolution and/or multiple emergence of isochore. Those phenomena are hard to be explained under the constant model.

According to the variable model, tiny fluctuation of parameters ($u(GC(t))$ and $v(GC(t))$), which are easily observed in real data, can lead to drastic changes in the genome compositions under the variable model. On the other hand, if mutation patterns in a genome do not satisfy any conditions resulting oscillatory, plural, or chaotic transition behaviors (Classes II and III), the genome does not have clear isochore in our context. Therefore, the genome that does not have clear GC-rich isochore (e.g., the opossum genome) is predicted to harbor fixed points belonging to only Class I (we need to wait for high-density SNP data of the opossum genome to confirm this prediction as well as closely related species as a reference).

←

FIG. 10.—Distribution of fixed points of $g(GC(t))$ in human chromosomes (open bars: all SNP data; shaded bars: non-CpG SNP data). Subsets of the fixed points are potential equilibria. Abscissa and ordinate are GC content and frequency of fixed points, respectively. Dashed and dotted lines indicate estimated GC content equilibria in each chromosome under the constant model by using all SNP data and non-CpG SNP data, respectively.
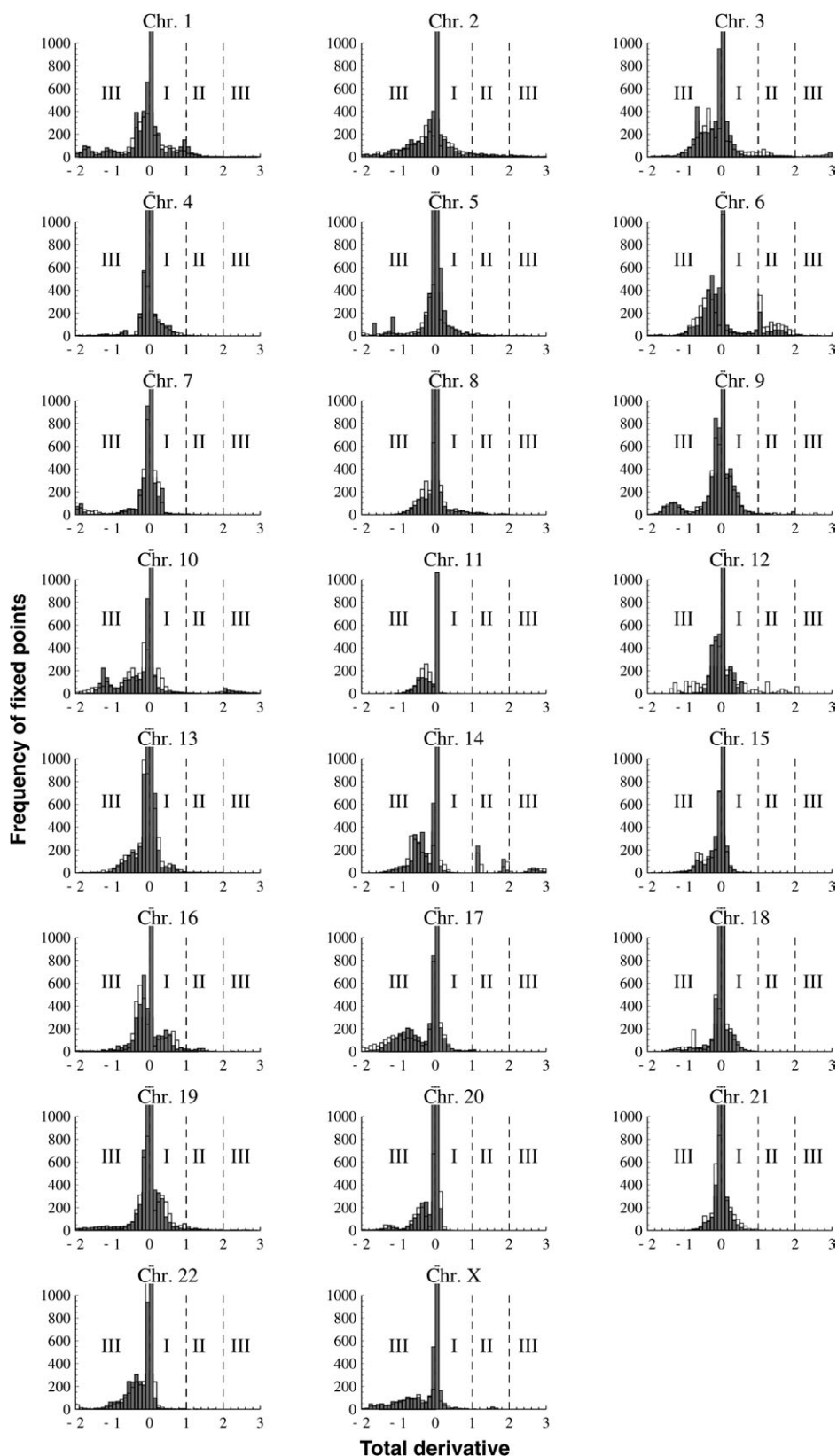
FIG. 11.—The genome-level derivatives of $g(GC(t))$ at the fixed points (open bars: all SNP data; shaded bars: non-CpG SNP data). The three classes are indicated by roman numerals: convergence or potential equilibrium (Classes I and II); divergence (Class III). Part of histograms are truncated for comparison. Note that each replicated data set has plural fixed points in general (the sum of fixed points are greater than 1,000).
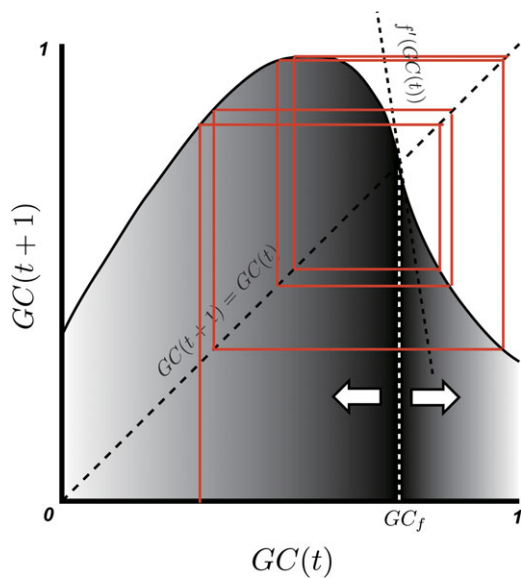
Fig. 12.—A schematic representation of a chaotic behavior of GC content transitions. Red lines represent the transitions of GC content. Diagonal broken lines represent GC(t + 1) = GC(t): that is, the condition of fixed points (potential equilibria).

Note that the variable model requires a pair of functions u(GC(t)) and v(GC(t)), which are derived only from mutation patterns of unfunctional intergenic regions. They are (putatively) selectively neutral. Although, the neutrality is not essential in our model, it is unlikely that vast amounts of the genomic regions that we used in the analyses are under strong selection. The variable model may be able to explain emergence and maintenance of isochore under the neutral theory.

The previous studies (Eyre-Walker 1999; Duret et al. 2002; Smith et al. 2002; Webster et al. 2003; Duret and Galtier 2009) suggested that the fixation process in mammals is biased toward G and C. The fixation biases are, however, primarily discussed in the context of evolutionary changes in synonymous sites and introns as well as intergenic regions. The data we used (SNP data only from unfunctional intergenic regions) showed a different view: for example, Webster et al. (2003) estimated the overall equilibrium GC content 0.41 based on the parsimony-inferred substitution data since human–chimpanzee and the common ancestor (95% confidence interval is 0.40–0.42). Those values are consistent with our equilibrium estimations both by the variable and constant models when the non-CpG SNP data were used (see fig. 10).

We categorized fixed points of f(GC(t)) derived from the replicated data into three classes: Classes I, II, and III. Each class theoretically shows a characteristic behavior in transitions of GC content: asymptotic convergence (Class I), oscillatory convergence (Class II), and divergence (Class III). Note that those transitions described here are local behaviors: that is, a single fixed point is considered in each

class. If plural fixed points are involved (and this is a common case), a global behavior of GC content transitions will be more complicated. Further theoretical studies of f(GC(t)) will provide insight to the compositional evolution under the variable model. We used high-density SNP data of the human population and closely related species, chimpanzee, as a reference. Recently, high-speed sequencing technologies are emerging. More abundant and reliable SNP data of various species will be available soon. At a point of view on polymorphism, the isochore evolution is still open to further discussion due to lack of the data. New rich data will make it possible to shed light on dynamics of the isochore evolution. With accumulation of reliable SNP data not only of human but also of the other animals (e.g., laboratory mouse, zebra fish, opossum, and platypus), studies on compositional evolution will reach a new era.

Supplementary Material

Supplementary figures S1–S2 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

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.

Belle EM, 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.

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

Bernardi G, et al. 1985. The mosaic genome of warm-blooded vertebrates. Science. 228:953–958.

Costantini M, Cammarano R, Bernardi G. 2009. The evolution of isochore patterns in vertebrate genomes. BMC Genomics. 10:146.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. Genome Res. 16:536–541.

Couzin J. 2006. GENOMICS: the HapMap gold rush: researchers mine a rich deposit. Science. 312:1131.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285–311.

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

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. 1993. Recombination and mammalian genome evolution. Proc Biol Sci. 252:237–243.

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, Hurst LD. 2001. The evolution of isochores. Nat Rev Genet. 2:549–555.

Fullerton SM, Bernardo Carvalho A, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 18:1139–1142.

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

Gu J, Li WH. 2006. Are GC-rich isochores vanishing in mammals? Gene. 385:50–56.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 22:2971–2972.

Holmquist GP. 1992. Chromosome bands, their chromatin flavors, and their functional features. Am J Hum Genet. 51:17–37.

The International HapMap Consortium. 2003. The International HapMap Project. Nature. 426:789–796.

International HapMap Consortium. 2004. Integrating ethics and science in the International HapMap Project. Nat Rev Genet. 5:467–475.

Jabbari K, Bernardi G. 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene. 224:123–127.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Li T-Y, Yorke JA. 1975. Period Three Implies Chaos. The American Mathematical Monthly. 82:985–992.

Mikkelsen TS, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 447:167–177.

Mouchiroud D, et al. 1991. The distribution of genes in the human genome. Gene. 100:181–187.

Sherry ST, Ward M, Sirotkin K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 9:677–679.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657–663.

Smith NG, Webster MT, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. Genome Res. 12:1350–1356.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A. 48:582–592.

Sueoka N. 1993. Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J Mol Evol. 37:137–153.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A. 85:2653–2657.

Warren WC, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 453:175–183.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. Mol Biol Evol. 23:1203–1216.

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

Wolfe KH, Sharp PM, Li W-H. 1989. Mutation rates differ among regions of the mammalian genome. Nature. 337:283–285.

Associate editor: Yoshihito Niimura