Analytical Expression of the Purine/Pyrimidine Autocorrelation Function after and before Random Mutations

DIDIER G. ARQUES*
Equipe de Biologie Théorique, Université de Franche-Comté, Laboratoire d'Informatique de Besançon, 16 Route de Gray, 25030 Besançon, France

AND

CHRISTIAN J. MICHEL
Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie de Belfort-Montbéliard, BP 527, 90016 Belfort, France

Received 5 May 1993; revised 17 November 1993

ABSTRACT

The mutation process is a classical evolutionary genetic process. The type of mutations studied here is the random substitutions of a purine base R (adenine or guanine) by a pyrimidine base Y (cytosine or thymine) and reciprocally (transversions). The analytical expressions derived allow us to analyze in genes the occurrence probabilities of motifs and d-motifs (two motifs separated by any d bases) on the R/Y alphabet under transversions. These motif probabilities can be obtained after transversions (in the evolutionary sense; from the past to the present) and, unexpectedly, also before transversions (after back transversions, in the inverse evolutionary sense, from the present to the past). This theoretical part in Section 2 is a first generalization of a particular formula recently derived. The application in Section 3 is based on the analytical expression giving the autocorrelation function (the d-motif probabilities) before transversions. It allows us to study primitive genes from actual genes. This approach solves a biological problem. The protein coding genes of chloroplasts and mitochondria have a preferential occurrence of the 6-motif $\mathrm{YRY}(N)_6\mathrm{YRY}$ (maximum of the autocorrelation function for $d = 6$, $N = \mathrm{R}$ or Y) with a periodicity modulo 3. The $\mathrm{YRY}(N)_6\mathrm{YRY}$ preferential occurrence without the periodicity modulo 3 is also observed in the RNA coding genes (ribosomal, transfer, and small nuclear RNA genes) and in the noncoding genes (introns and 5′ regions of eukaryotic nuclei). However, there are two exceptions to this $\mathrm{YRY}(N)_6\mathrm{YRY}$ rule: the protein coding genes of eukaryotic nuclei, and prokaryotes, where $\mathrm{YRY}(N)_6\mathrm{YRY}$ has the second highest value after $\mathrm{YRY}(N)_0\mathrm{YRY}$ (YRYYRY) with a periodicity modulo 3. When we go backward in time with the analytical expression, the protein coding genes of both eukaryotic nuclei and prokaryotes retrieve the $\mathrm{YRY}(N)_6\mathrm{YRY}$ preferential occurrence with a periodicity modulo 3 after 0.2 back transversions per

*To whom correspondence should be addressed.

655 Avenue of the Americas, New York, NY 10010 0025-5564/94/$7.00

base. In other words, the actual protein coding genes of chloroplasts and mitochondria are similar to the primitive protein coding genes of eukaryotic nuclei and prokaryotes. On the other hand, this application represents the first result concerning the mutation process in the model of DNA sequence evolution we recently proposed. According to this model, the actual genes on the R/Y alphabet derive from two successive evolutionary genetic processes: an independent mixing of a few nonrandom types of oligonucleotides leading to genes called primitive followed by a mutation process in these primitive genes. Indeed, the mutation process can simulate statistical properties identified in genes, e.g., the variations between $YRY(N)_0YRY$ and $YRY(N)_6YRY$, which could not have been so far simulated with the mixing process.

1. INTRODUCTION

The autocorrelation function defined in [1, 2] and given in a generalized form in Section 3.2 allows us to analyze in gene populations the mean occurrence probabilities of d-motifs, a d-motif being two motifs (a series of a few nucleotides, e.g., a trinucleotide) separated by any d bases. This autocorrelation function allows the identification of nonrandom statistical properties in genes on the purine (R)/pyrimidine (Y) alphabet (R = adenine or guanine, Y = cytosine or thymine) [1–5]: periodicities (modulo 2, 3, etc.) and subperiodicities, the preferential occurrence of the 6-motif $YRY(N)_6YRY$ (N = R or Y) in various genes (global maximum of the autocorrelation function for $d = 6$ in functional and taxonomic genes), local maxima, etc. These properties are important as they have a biological meaning, e.g., a periodicity modulo 3 reveals a protein (coding) gene, a periodicity modulo 2, large alternating R/Y stretches found in noncoding genes, etc. They are related to a specific nucleotide ordering in the genes which can be studied with the development of models simulating molecular evolution.

The mutation process is a classical evolutionary genetic process analyzed by different theories, e.g., the neutral theory [6, 7]. The type of mutations studied here is the random substitutions of R by Y and Y by R (transversions). Two analytical expressions solved in Section 2 allow us to analyze the d-motif probabilities (the autocorrelation function) after transversions (in the evolutionary sense, from the past to the present) and also before transversions (after back transversions, in the inverse evolutionary sense, from the present to the past). Different properties and a generalization to d-motifs with motifs of any base length are also derived from these formulas. The application in Section 3, based on the analytical expression giving the autocorrelation function after back transversions, solves a biological problem. Indeed, it demonstrates that the protein genes of both eukaryotic nuclei and prokaryotes which do not have the $YRY(N)_6YRY$ preferential occurrence found in the protein genes of chloroplasts and mitochondria, in the RNA coding

genes (ribosomal, transfer, and small nuclear RNA genes), and in the noncoding genes (introns and 5′ regions of eukaryotic nuclei), retrieve this property after 0.2 back transversions per base. These primitive protein genes of eukaryotic nuclei and prokaryotes have the $YRY(N)_6YRY$ preferential occurrence with a periodicity modulo 3 such as that of the actual protein genes of chloroplasts and mitochondria. In the Discussion (Section 4), we briefly recall the properties of the autocorrelation function. Then, we summarize the calculus methods of the autocorrelation function which are not analytical. We present the properties of the analytical expressions solved here, in particular those associated with the generalization and the inverse evolutionary sense. Then, the mutation process is replaced in the oligonucleotide mixing model which we have recently developed [3]. Finally, we discuss the results of other methods and the biological data which may support the result obtained with the application.

2. THEORY

2.1. *RECALL OF THE POISSON EVEN/ODD DISTRIBUTION ASSOCIATED WITH RANDOM SUBSTITUTIONS*

Let s be a sequence of base length $\ell(s)$ on the alphabet {R,Y} (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine). This sequence s is subjected to transversions, i.e., random substitutions of a base R (resp. Y) by the base Y (resp. R) at random sites in s. Let x be the number of transversions per base (per base site) in average, i.e., the total number of transversions in s divided by the length $\ell(s)$ of s. Note that in the following, "x transversions" always stand for "x transversions per base in average."

The transversion number of a base in a given site of a sequence subjected to random substitutions per base of mean x follows a Poisson law of parameter x (e.g., the classical proofs in [8, p. 447; 6, p. 69; 7, p. 40]).

We define $P_{R\to R}(x)$ (resp. $P_{R\to Y}(x)$, $P_{Y\to R}(x)$, $P_{Y\to Y}(x)$) as being the probability that a base R (resp. R, Y, Y) in the sequence s before the substitution process is R (resp. Y, R, Y) after x transversions. Then, $\mathcal{O}(x)=P_{R\to Y}(x)=P_{Y\to R}(x)=(1-e^{-2x})/2$ is the probability that a given site in s is subjected to an odd number of transversions and $\mathcal{E}(x)=P_{R\to R}(x)=P_{Y\to Y}(x)=1-\mathcal{O}(x)=(1+e^{-2x})/2$ is the probability that a given site in s is subjected to an even number of transversions (proof detailed in [5]). The two formulas $\mathcal{E}(x)$ and $\mathcal{O}(x)$ obtained on the alphabet {R,Y} are similar to those obtained on the alphabet {A,C,G,T} with the one-parameter model (a unique rate of substitutions; see [9]) and with the two-parameter model (a rate of transitions and a rate of transversions; see [10]).

2.2. *ANALYTICAL EXPRESSION OF THE* d*-MOTIF PROBABILITY AFTER RANDOM SUBSTITUTIONS*

Let the motif m be a trinucleotide (series of three bases) on the alphabet {R, Y}, i.e., $m \in \{\text{RRR},\ldots,\text{YYY}\}$. Let the d-motif $m_1(N)_d m_2$ be two motifs m_1 and m_2 separated by any d bases N (N = R or Y), d being a given constant, i.e., $m_1(N)_d m_2 \in \{\text{RRR}(N)_d\text{RRR}, \text{RRR}(N)_d\text{RRY},\ldots,\text{YYY}(N)_d\text{YYY}\}$. By convention, in the following the indexes i or $j \in [1, 64]$ represent the d-motifs $\text{RRR}(N)_d$ RRR, $\text{RRR}(N)_d\text{RRY},\ldots,\text{YYY}(N)_d\text{YYY}$ in the alphabetic order. Let $\mathcal{I}(j,i)$ be the number of identical bases in the same d-motif site between the d-motifs i and j, e.g., $\mathcal{I}(1,2) = 5$, 1 representing $\text{RRR}(N)_d\text{RRR}$ and 2, $\text{RRR}(N)_d\text{RRY}$. The d-motif probabilities after x transversions (at time t) can be obtained from the d-motif probabilities before the substitution process (at time 0) (see Figure 1), τ being the unknown number of transversions per base in average between the times 0 and today.

THEOREM 1

Let $[P_i(x)]_{1 \leqslant i \leqslant 64}$ *be the probabilities of the d-motifs* i, $i \in [1, 64]$, *in a given sequence after* x *transversions. Then,*

$$P_i(x) = \sum_{j=1}^{64} P_j(0)\mathcal{E}(x)^{\mathcal{I}(j,i)}\mathcal{O}(x)^{6-\mathcal{I}(j,i)} \tag{1}$$

$$P_i(x+y) = \sum_{j=1}^{64} P_j(x)\mathcal{E}(y)^{\mathcal{I}(j,i)}\mathcal{O}(y)^{6-\mathcal{I}(j,i)} \tag{2}$$

$$P_i(\tau) = \sum_{j=1}^{64} P_j(x)\mathcal{E}(\tau-x)^{\mathcal{I}(j,i)}\mathcal{O}(\tau-x)^{6-\mathcal{I}(j,i)}, \tag{3}$$

where $P_i(\tau)$ *represents the actual d-motif probabilities.*

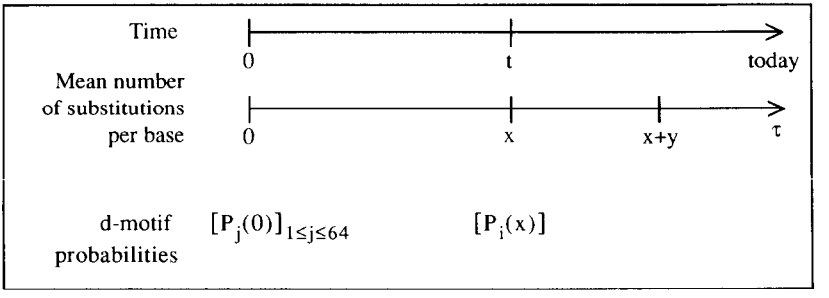


FIG. 1. d-motif probabilities after transversions.

Proof. (1) Formula deduced from the fact that the probability of the d-motif j, $j \in [1, 64]$, giving the d-motif i, $i \in [1, 64]$, after x transversions, is $\mathscr{E}(x)^{\mathscr{I}(j,i)}\mathscr{O}(x)^{6-\mathscr{I}(j,i)}$. Precisely, let M_i^x be the event "a d-motif, randomly chosen in a sequence after x transversions, is of type i." Then,

$$\begin{aligned} P_i(x) &= P(M_i^x) \\ &= \sum_{j=1}^{64} P\left(M_j^0\right) \times P\left(M_i^x \mid M_j^0\right) \\ &= \sum_{j=1}^{64} P_j(0) \times P(d\text{-motif } j \to d\text{-motif } i \text{ after } x \text{ random} \\ &\qquad\qquad \text{substitutions per base in average}) \\ &= \sum_{j=1}^{64} P_j(0)\mathscr{E}(x)^{\mathscr{I}(j,i)}\mathscr{O}(x)^{6-\mathscr{I}(j,i)}. \end{aligned}$$

(2)

$$\begin{aligned} &\sum_{k=1}^{64} P_k(x)\mathscr{E}(y)^{\mathscr{I}(k,i)}\mathscr{O}(y)^{6-\mathscr{I}(k,i)} \\ &\quad = \sum_{k=1}^{64} \left(\sum_{j=1}^{64} P_j(0)\mathscr{E}(x)^{\mathscr{I}(j,k)}\mathscr{O}(x)^{6-\mathscr{I}(j,k)} \right) \\ &\qquad \times \mathscr{E}(y)^{\mathscr{I}(k,i)}\mathscr{O}(y)^{6-\mathscr{I}(k,i)} \qquad \text{by (1)} \\ &\quad = \sum_{j=1}^{64} P_j(0) \left(\sum_{k=1}^{64} \mathscr{E}(x)^{\mathscr{I}(j,k)}\mathscr{O}(x)^{6-\mathscr{I}(j,k)}\mathscr{E}(y)^{\mathscr{I}(k,i)}\mathscr{O}(y)^{6-\mathscr{I}(k,i)} \right) \\ &\quad = \sum_{j=1}^{64} P_j(0)\mathscr{E}(x+y)^{\mathscr{I}(j,i)}\mathscr{O}(x+y)^{6-\mathscr{I}(j,i)} \\ &\qquad\qquad \text{(consequence of exponential properties)} \\ &\quad = P_i(x+y). \end{aligned}$$

(3) Particular case of the formula (2) with $y = \tau - x$. It gives the actual probabilities in function of the past probabilities.

Remarks. The formula $P_i(x)$ (1) is a particular case of the formula (2) with $x = 0$ and $y = x$. The formula $P_i(x)$ (1) converges as expected toward the random value $1/64 = 0.015625$ when the number x of transversions increases, whatever the d-motif i and whatever the d-motif probabilities $P_j(0)$ (consequence of negative exponentials).

The formula $P_i(x)$ (1) can be generalized to two motifs of base lengths λ_1 and λ_2:

$$P_i(x) = \sum_{j=1}^{2^{\lambda_1+\lambda_2}} P_j(0)\mathscr{E}(x)^{\mathscr{I}(j,i)}\mathscr{O}(x)^{\lambda_1+\lambda_2-\mathscr{I}(j,i)}.$$

2.3. *ANALYTICAL EXPRESSION OF THE* d*-MOTIF PROBABILITY BEFORE RANDOM SUBSTITUTIONS (AFTER RANDOM BACK SUBSTITUTIONS)*

The problem of Section 2.3 is the inverse problem of Section 2.2. Let τ (resp. x) be the number of transversions per base in average between the times 0 and today (resp. t) (see Figure 2). Let y be the number of transversions per base in average between the times t and today, i.e., $\tau = x + y$. In the previous problem of Section 2.2, the reference time is the time 0 (before the substitution process), while in the inverse problem, the reference time is today (after the substitution process). *Note*: In the following, "x (resp. y) transversions" always stands for "x (resp. y) transversions per base in average."

Therefore, the inverse problem consists in expressing $Q_i(y)$ $(= P_i(\tau - y))$ in function of $[Q_j(0) = P_j(\tau)]_{1 \leqslant j \leqslant 64}$ and more generally $Q_i(y + z)$ $(= P_i(\tau - y - z))$ in function of $[Q_j(z) = P_j(\tau - z)]_{1 \leqslant j \leqslant 64}$, i and $j \in [1, 64]$ representing the alphabetic order of d-motifs.

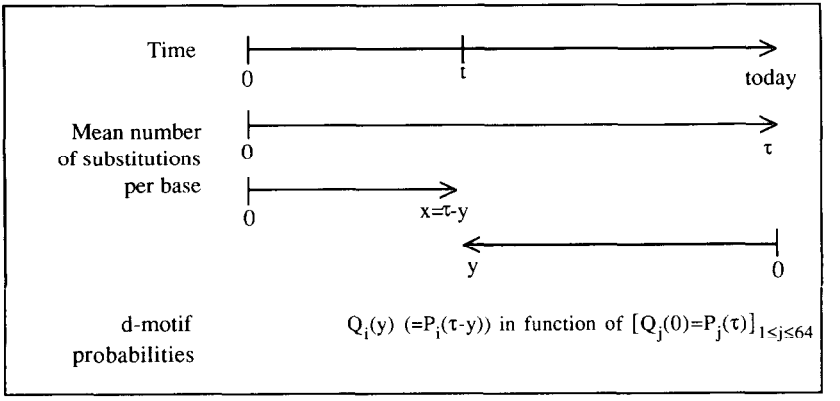


FIG. 2. d-motif probabilities before transversions.

PROPOSITION 2

$$Q_i(y+z) = \sum_{j=1}^{64} Q_j(z)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{6-\mathscr{I}(j,i)} \tag{4}$$

$$Q_i(y) = \sum_{j=1}^{64} P_j(\tau)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{6-\mathscr{I}(j,i)}, \tag{5}$$

with $\mathscr{E}(-y) = (1+e^{2y})/2$ *and* $\mathscr{O}(-y) = (1-e^{2y})/2$.

Proof. (4) The inverse matrix of $[\mathscr{E}(y)^{\mathscr{I}(j,i)}\mathscr{O}(y)^{6-\mathscr{I}(j,i)}]_{1\leqslant i,j\leqslant 64}$ associated with formula (2) in Theorem 1 is $[\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{6-\mathscr{I}(j,i)}]_{1\leqslant i,j\leqslant 64}$. Then, formula (2) implies that

$$P_i(x) = \sum_{j=1}^{64} P_j(x+y)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{6-\mathscr{I}(j,i)}.$$

Then,

$$\begin{aligned} Q_i(y+z) &= P_i(\tau-y-z) \\ &= \sum_{j=1}^{64} P_j(\tau-z)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{6-\mathscr{I}(j,i)} \\ &= \sum_{j=1}^{64} Q_j(z)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{6-\mathscr{I}(j,i)}. \end{aligned}$$

(5) Particular case of formula (4) with $z=0$.

Remarks. The formula $Q_i(y)$ (5) will be used in the application in Section 3 to determine the d-motif probabilities after back transversions in the protein coding genes, the actual d-motif probabilities $P_j(\tau) = Q_j(0)$ being computed from gene databases. Contrary to the formula $P_i(x)$ (1), the formula $Q_i(y)$ (5) does not converge when the number y of transversions increases (see Section 2.4).

The formula $Q_i(y)$ (5) can be generalized to two motifs of base lengths λ_1 and λ_2 in the same way as with $P_i(x)$ (1):

$$Q_i(y) = \sum_{j=1}^{2^{\lambda_1+\lambda_2}} P_j(\tau)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{\lambda_1+\lambda_2-\mathscr{I}(j,i)}.$$

2.4. *BIOLOGICAL MEANING OF THE PREVIOUS FORMULAS*

$$P_i(x) = Q_i(y) \qquad \text{if } x + y = \tau \text{ and } 0 \leqslant x, y \leqslant \tau$$

(see Figure 2). The formula $P_i(x)$ (1) gives the evolution of the d-motif probabilities when we go from the past to the present and when the number of transversions increases from 0 to τ (after transversions). $P_i(x)$ can be obtained either exactly by analytical expression or approximately by computer simulation (simulation of random substitutions in simulated sequences).

The formula $Q_i(y)$ (5) gives the inverse evolution of the d-motif probabilities, when we go from the present to the past and when the number of transversions decreases from τ to 0 (before transversions or after back transversions). Contrary to $P_i(x)$, $Q_i(y)$ does not converge when the number y of transversions increases. However, the vector $[Q_i(y)]_{1 \leqslant i \leqslant 64}$ must remain a probability vector, i.e., the 64 values $Q_i(y)$ must be bounded between 0 and 1 (and of sum 1). Therefore, the condition $0 \leqslant Q_i(y) \leqslant 1$ for i in [1, 64] implies a maximal number of transversions. Another difference with $P_i(x)$ lies in the fact that $Q_i(y)$ can only be obtained by analytical expression and not by computer simulation. Indeed, as the site and the order of previous substitutions are unknown, it is impossible to reproduce the effects of back substitutions in the exact nucleotide ordering of actual genes.

3. APPLICATION: RANDOM BACK SUBSTITUTIONS IN THE PROTEIN CODING GENES

3.1. *PRESENTATION OF THE PROBLEM*

Since 1987 the biological problem related to the YRY$(N)_6$YRY preferential occurrence in the protein (coding) genes has remained unsolved.

The 6-motif YRY$(N)_6$YRY studied with the autocorrelation function analyzing the occurrence probability of the d-motif YRY$(N)_d$YRY by varying d between 0 and 99 (100 points) has a preferential occurrence because it has the highest value among 100 points in almost all gene populations [2, 3]. It should be remembered that a curve with 100 different points can lead to 100! (10^{158}) possible curve shapes and that a maximal value common to n different gene populations has a probability equal to $1/100^n$ in the random case. Such a statistical evaluation with the gene populations available in 1987 leads to a YRY$(N)_6$YRY probability of order 10^{-12} if the nucleotide distribution in genes is random [2]. Furthermore, since 1987, the preferential occurrence of

YRY$(N)_6$YRY was observed in several new gene populations and subpopulations ([3]; data not shown).

The YRY$(N)_6$YRY preferential occurrence is found in genes according to three main types [3]: (1) the YRY$(N)_6$YRY preferential occurrence without periodicity observed in RNA coding genes (ribosomal, transfer, and small nuclear RNA genes), (2) the YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 2 identified in noncoding genes (introns and 5′ regions of eukaryotic nuclei) by deleting their large alternating R/Y stretches (not detailed here), and (3) the YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 3 for $d \equiv 0[3]$ (maximal values for $d = 0$, 3, 6, etc.) found in protein genes according to two subtypes which are analyzed and solved here with the following protein gene populations: chloroplasts (1002 genes, 812 kb), mitochondria (813 genes, 666 kb), eukaryotic nuclei (13,997 genes, 19,148 kb) and prokaryotes (5729 genes, 6866 kb). These gene populations are obtained from the release 32 of the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Data Library in the same way as described in previous studies (see, e.g., [3] for a description of data acquisitions).

In the protein genes of chloroplasts and mitochondria, YRY$(N)_6$YRY has the highest probability with a periodicity modulo 3 nonuniform (the top curve $d \equiv 0[3]$ and the bottom curve $d \equiv 1,2[3]$ are not horizontal) (Figures 3a and 3b).

In the protein genes of eukaryotic nuclei and prokaryotes, YRY$(N)_6$YRY does not have the highest probability. It occurs after YRY$(N)_0$YRY (YRYYRY) with a periodicity modulo 3 uniform (Figures 4a and 5a). According to the statistical analysis performed on the gene populations available, this statistical perturbation only exists in these two gene populations and their subpopulations, e.g., nuclear protein genes of primates and rodents (data not shown). Furthermore, because of the law of large numbers (in [3], p. 752, Section 2.3.3), nonrandom statistical properties identified with populations made of several hundreds of genes are stable from a statistical point of view. Since the 10th release of the EMBL gene database, the second highest value of YRY$(N)_6$YRY after YRY$(N)_0$YRY was observed in these two populations with each new release. In summary, this perturbation cannot be attributed to any statistical bias. However, no explanation has been proposed so far for this unexpected second highest value of YRY$(N)_6$YRY.

The formula $Q_i(y)$ (5), using the d-motif probabilities $P_j(\tau) = Q_j(0)$ of actual genes (see Figure 2 and also Figure 8 in Section 4) obtained from gene databases, allows us to determine by varying d the autocorrelation function after back transversions (Section 3.2.1). We will show

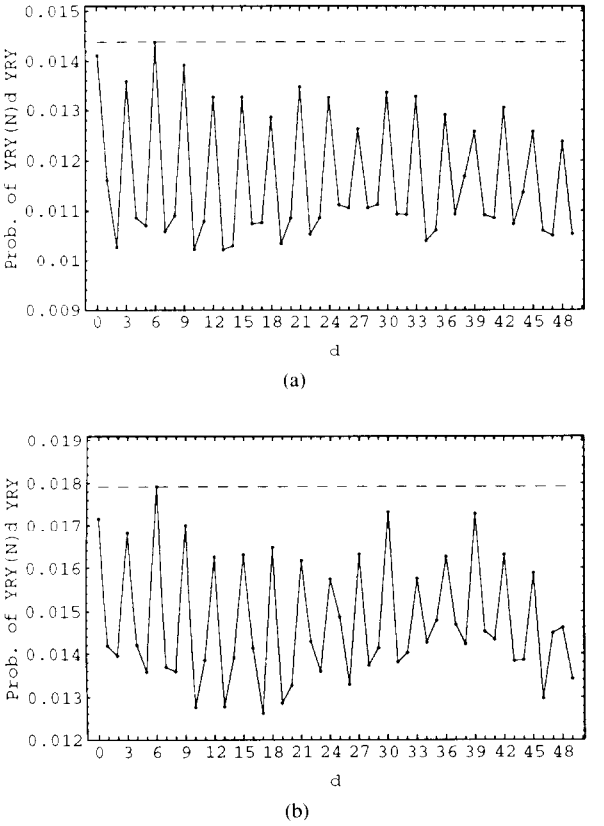


FIG. 3. YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 3 nonuniform in the actual protein coding genes of chloroplasts and mitochondria. Autocorrelation function analyzing the actual probability ($Q_{46}(0, d)$) of YRY$(N)_d$YRY in the protein coding genes of (a) chloroplasts; (b) mitochondria. The horizontal axis represents the number d of bases N between 2 YRY, $d \in [0,49]$, i.e., YRY$(N)_d$YRY. The vertical axis represents the YRY$(N)_d$YRY probability.

that the protein genes of both eukaryotic nuclei and prokaryotes have the YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 3 after 0.2 back transversions similar to the actual protein genes of chloroplasts and mitochondria.

3.2. *AUTOCORRELATION FUNCTION*

3.2.1. Autocorrelation function analyzing the probability of the d-motif YRY(N)$_d$YRY after random back substitutions. The inverse substitution

process of the 6-motif YRY$(N)_6$YRY, the 0-motif YRY$(N)_0$YRY, and the periodicity modulo 3 in the protein genes of eukaryotic nuclei and prokaryotes is studied by using the formula $Q_i(y)$ (5) for $i = 46$, i.e., $Q_{46}(y)$ (YRY$(N)_d$YRY is the 46th d-motif in the alphabetic order) and by varying d, noted $Q_{46}(y,d)$, in the following. $Q_{46}(y,d)$ is an autocor-

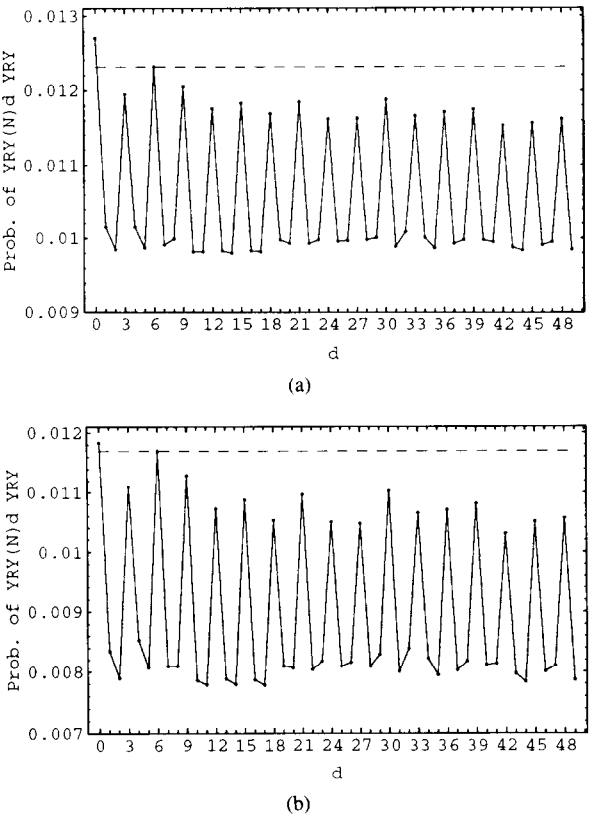
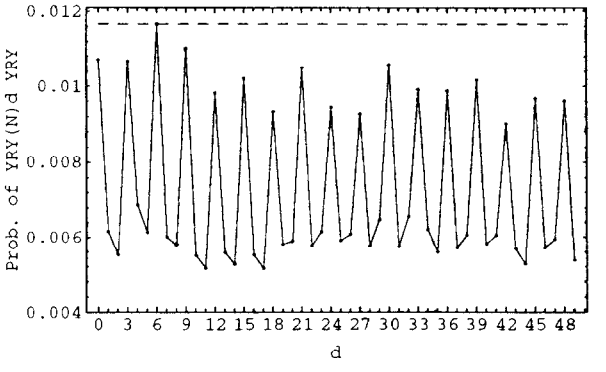


FIG. 4. YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 3 nonuniform before 0.2 transversions per base (after back transversions) in the protein coding genes of eukaryotic nuclei. Autocorrelation function analyzing the probability of YRY$(N)_d$YRY in the protein coding genes of eukaryotic nuclei: (a) actual ($Q_{46}(0,d)$): second highest probability of YRY$(N)_6$YRY after YRY$(N)_0$YRY with a periodicity modulo 3 uniform; (b) before $y = 0.1$ transversions per base ($Q_{46}(0.1,d)$): same probability for YRY$(N)_6$YRY and YRY$(N)_0$YRY; (c) before $y = 0.2$ transversions per base ($Q_{46}(0.2,d)$): highest probability of YRY$(N)_6$YRY with a periodicity modulo 3 nonuniform. The horizontal axis represents the number d of bases N between 2 YRY, $d \in [0,49]$, i.e., YRY$(N)_d$YRY. The vertical axis represents the YRY$(N)_d$YRY probability.



(c)

FIG. 4. (*Continued*)

relation function by definition

$$Q_{46}(y,d) = \sum_{j=1}^{64} P_j(\tau,d)\mathscr{E}(-y)^{\mathscr{I}(j,46)}\mathscr{O}(-y)^{6-\mathscr{I}(j,46)}, \qquad (6)$$

with $j \in [1, 64]$ representing the d-motifs in the alphabetic order, $d \in [0,49]$ being the number of any bases N between the two trinucleotides, y being the number of back transversions per base in average, and $P_j(\tau,d)$ being the previous probabilities $P_j(\tau)$ in which d varies.

Remarks. For $y = 0$ (before the inverse substitution process), it can be easily verified in the formula $Q_{46}(y,d)$ (6) that $Q_{46}(0,d) = P_{46}(\tau,d)$.

In order to compute the formula $Q_{46}(y,d)$ (6), it is easier to rewrite it as follows:

$$Q_{46}(y,d) = \frac{1}{64}\sum_{k=0}^{6}\left(\sum_{j/\mathscr{I}(j,46)=k} P_j(\tau,d)\right)(1+e^{2y})^k(1-e^{2y})^{6-k}.$$

In order to use the formula $Q_{46}(y,d)$, the actual d-motif probabilities $P_j(\tau,d) = Q_j(0,d)$ are computed by generalizing the previous autocorrelation function definition [2] to any trinucleotide on the R/Y alphabet. This method is described in what follows.

3.2.2. Autocorrelation function analyzing the actual probabilities of d-*motifs.* Let F be a gene population with $n(F)$ DNA sequences. Let s be a sequence in F with a length $\ell(s)$. Let j be the alphabetic order of the d-motifs $\mathrm{RRR}(N)_d\mathrm{RRR}$, $\mathrm{RRR}(N)_d\mathrm{RRY},\ldots,\mathrm{YYY}(N)_d\mathrm{YYY}$, $j \in$

[1, 64] and $d \in [0,49]$, i.e., these d-motifs are characterized by $j \in [1, 64]$, their number in the alphabetic order, and by $d \in [0,49]$, their number of any bases N between the two motifs m_1 and m_2 ($m_1, m_2 \in$ {RRR,..., YYY}). In the following, we call a d-motif j, the jth d-motif in the alphabetic order. For each s of F, the counter $c_{d,j}(s)$ counts the

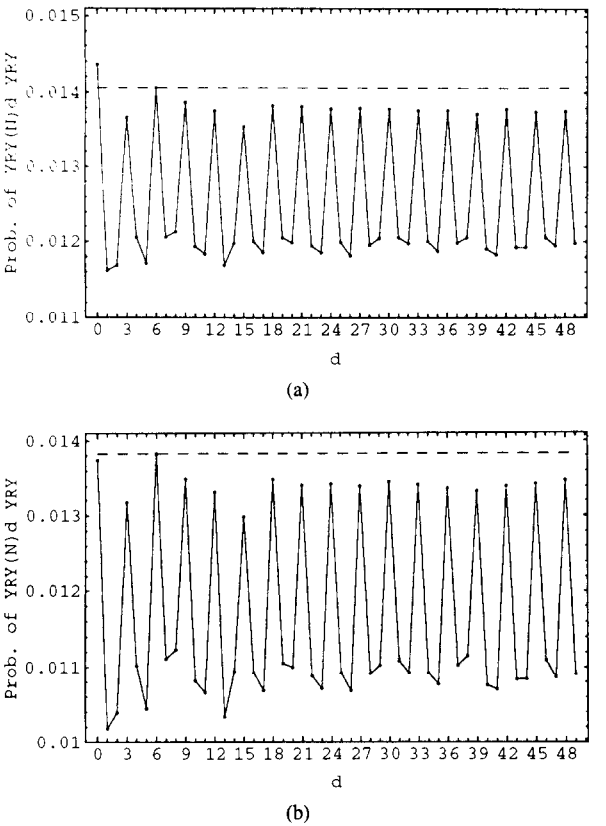
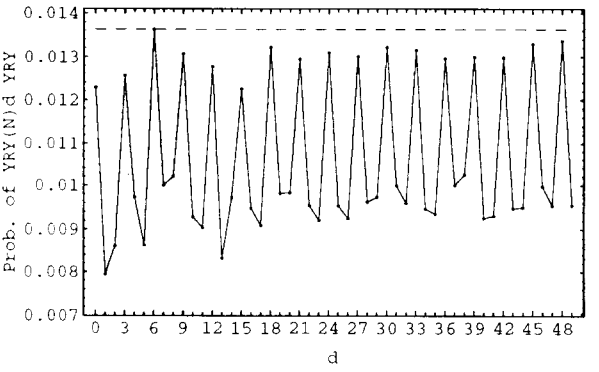


FIG. 5. YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 3 nonuniform before 0.2 transversions per base (after back transversions) in the protein coding genes of prokaryotes. Autocorrelation function analyzing the probability of YRY$(N)_d$YRY in the protein coding genes of prokaryotes: (a) actual ($Q_{46}(0,d)$): second highest probability of YRY$(N)_6$YRY after YRY$(N)_0$YRY with a periodicity modulo 3 uniform; (b) before $y = 0.1$ transversions per base ($Q_{46}(0.1,d)$): same probability for YRY$(N)_6$YRY and YRY$(N)_0$YRY; (c) before $y = 0.2$ transversions per base ($Q_{46}(0.2,d)$): highest probability of YRY$(N)_6$YRY with a periodicity modulo 3 nonuniform. The horizontal axis represents the number d of bases N between 2 YRY, $d \in [0,49]$, i.e., YRY$(N)_d$YRY. The vertical axis represents the YRY$(N)_d$YRY probability.



(c)

FIG. 5. (*Continued*)

occurrences of the d-motif j in s. In order to count the occurrences of the d-motif j in the same conditions for all d, only the first $\ell(s)-54$ $(=\ell(s)-(49+6)+1)$ bases of s are examined $(49+6$ is the maximal length of the d-motif $j)$. The occurrence probability $o_{d,j}(s)$ of the d-motif j for s is then equal to $c_{d,j}(s)/[\ell(s)-54]$, i.e., the ratio of the counter by the total number of current bases read. The actual probability $Q_j(0,d)=P_j(\tau,d)$ of the d-motif j for F is finally equal to $[\Sigma_{s\in F} o_{d,j}(s)]/n(F)$.

Remarks. The function $d \to Q_{46}(0,d)$ giving the mean occurrence probability of the d-motif YRY$(N)_d$YRY in a gene population is our classical definition of the autocorrelation function. In order to have a sufficient number of occurrences of the d-motif j for $d=49$, the probability $Q_j(0,d)$ is computed with sequences having a minimal length of 300 bases.

3.2.3. Graphical representation of the autocorrelation function analyzing the probability of the d-motif YRY(N)$_d$YRY. The autocorrelation function analyzing the probability of the d-motif YRY$(N)_d$YRY (Figures 3–5) is represented as follows: (1) the abscissa shows the number d of bases N between 2 YRY by varying d between 0 and 49; and (2) the ordinate gives the probability of YRY$(N)_d$YRY.

Figures 3a, 3b, 4a, and 5a show the actual probability $Q_{46}(0,d)$ of YRY$(N)_d$YRY in the protein genes of chloroplasts, mitochondria, eukaryotic nuclei, and prokaryotes, respectively.

Figures 4b and 5b show the probability $Q_{46}(0.1,d)$ of YRY$(N)_d$YRY before $y=0.1$ transversions (after 0.1 back transversions) in the protein genes of eukaryotic nuclei and prokaryotes, respectively.

Figures 4c and 5c show the probability $Q_{46}(0.2, d)$ of YRY$(N)_d$YRY before $y = 0.2$ transversions (after 0.2 back transversions) in the protein genes of eukaryotic nuclei and prokaryotes, respectively.

3.3. *EFFECTS OF BACK TRANSVERSIONS*

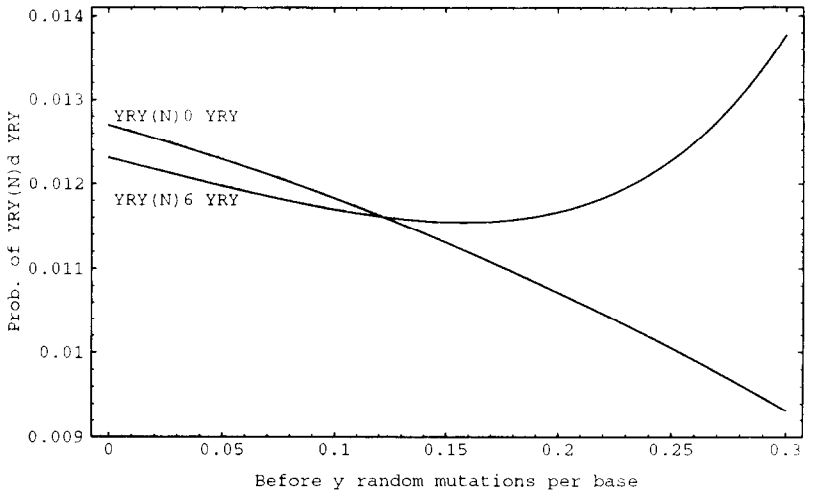
3.3.1. Effects of back transversions in the eukaryotic nuclear protein coding genes. Unexpectedly, after 0.1 back transversions in the eukaryotic nuclear protein genes, YRY$(N)_6$YRY and YRY$(N)_0$YRY have nearly the same probability (Figure 4b) (the point at 6 being slightly less than the point at 0). Surprisingly, after 0.2 back transversions, there is the YRY$(N)_6$YRY preferential occurrence (Figure 4c).

In order to analyze precisely the modifications of these two points in the inverse evolutionary sense, Figure 6a shows their probability variation continuously in the transversion range [0,0.3]. At the beginning of the inverse substitution process, the probabilities of YRY$(N)_0$YRY and YRY$(N)_6$YRY decrease with a higher slope for YRY$(N)_0$YRY which crosses YRY$(N)_6$YRY after 0.12 back transversions. Then, YRY$(N)_0$YRY decreases while YRY$(N)_6$YRY increases.

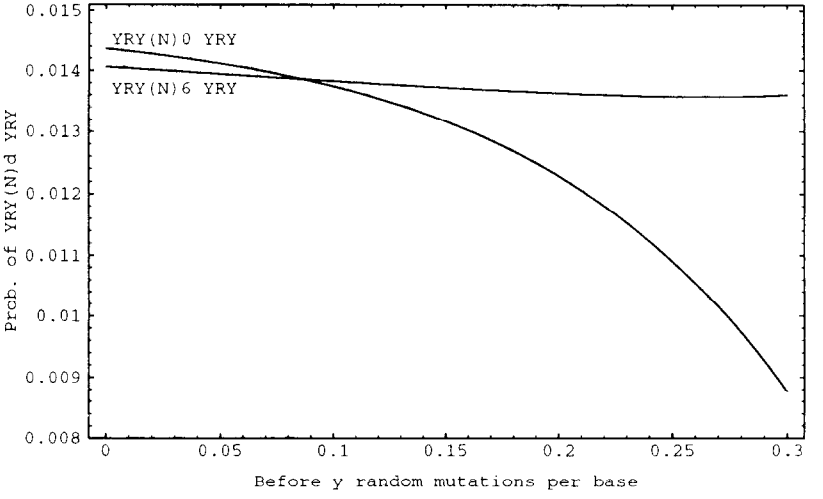
The YRY$(N)_6$YRY preferential occurrence obtained after 0.2 back transversions is associated with a periodicity modulo 3 nonuniform (Figure 4c) similar to the protein genes of chloroplasts and mitochondria. The probability of the bottom curve $d \equiv 1,2[3]$ gets closer to 0 (Figure 4c). After 0.3 back transversions, there are some points in the bottom curve with negative values (data not shown) leading to a maximal number of transversions equal to 0.3 in the eukaryotic nuclear protein genes.

3.3.2. Effects of back transversions in the prokaryotic protein coding genes. Surprisingly, the inverse substitution process in the prokaryotic protein genes is very similar to that of eukaryotic nuclei (Figures 5a–c). It also leads to the YRY$(N)_6$YRY preferential occurrence with a periodicity modulo 3 nonuniform after 0.2 back transversions (Figure 5c).

Figure 6b shows that the probabilities of YRY$(N)_0$YRY and YRY$(N)_6$YRY decrease with a higher slope for YRY$(N)_0$YRY which crosses YRY$(N)_6$YRY after 0.09 back transversions. It seems that YRY$(N)_6$YRY does not increase, which is in contrast to the case of eukaryotic nuclei. However, by analyzing the probability variation of YRY$(N)_6$YRY in the transversion range [0,0.8], Figure 7 reveals that YRY$(N)_6$YRY in prokaryotes also increases but with a delay of about 0.15 transversions compared to YRY$(N)_6$YRY in eukaryotic nuclei. The maximal number of transversions is also equal to 0.3 (data not shown), as in eukaryotic nuclei. It should be stressed that this maximal



(a)



(b)

FIG. 6. Probability variation of YRY$(N)_0$YRY ($Q_{46}(y,0)$) and YRY$(N)_6$YRY ($Q_{46}(y,6)$) with the inverse substitution process continuously in the transversion range [0,0.3] for the protein coding genes of (a) eukaryotic nuclei: decrease of the YRY$(N)_0$YRY probability and increase of the YRY$(N)_6$YRY probability; (b) prokaryotes: decrease of the YRY$(N)_0$YRY probability. The horizontal axis represents the number y of back transversions per base, $y \in [0,0.3]$. The vertical axis represents the probabilities of YRY$(N)_0$YRY and YRY$(N)_6$YRY.

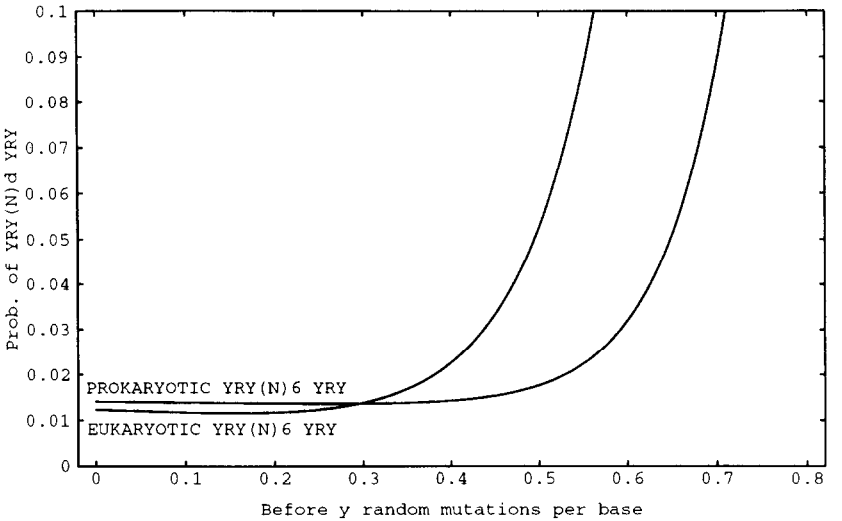


FIG. 7. Probability variation of YRY$(N)_6$YRY ($Q_{46}(y,6)$) with the inverse substitution process continuously in the transversion range [0,0.8] for the protein coding genes of eukaryotic nuclei and prokaryotes: increase of the YRY$(N)_6$YRY probability. The horizontal axis represents the number y of back transversions per base, $y \in [0,0.8]$. The vertical axis represents the probability of YRY$(N)_6$YRY.

value of 0.3 concerns a mean number of transversions, i.e., the case where all R/Y base sites are equiprobably substituted. However, as mentioned in [3], Section 3.3.2, some base sites can have a higher (and also lower) transversion rate compared to the average 0.3. This also explains that an analysis of the probability variation of some d-motifs in a transversion range exceeding [0,0.3], e.g., YRY$(N)_6$YRY in [0,0.8], is not irrelevant.

4. DISCUSSION

The autocorrelation function as defined in [1, 2] and Section 3.2 avoids the decrease of probabilities when the number d of bases between the two motifs increases. Indeed, the side effect induced by the end of the gene is corrected in order to have the same occurrence of d-motifs in the gene, whatever the number d. Therefore, this autocorrelation function is without bias; simple, as it is based on the frequency concept and with a graphical representation biologically interpretable; interesting, as it studies not only the frequency of two motifs but also the distance between them; general, as a motif is a particular case of a d-motif with two motifs separated by 0 base and as a gene is a particular

case of a population with one gene; and, finally, stable, as the d-motif probabilities are computed at the gene population level, i.e., populations made of several hundreds of genes (consequence of the law of large numbers; see [3, p. 752, Sect. 2.3.3]).

These reasons explain why several methods were developed to compute this autocorrelation function (detailed in [11]). For gene populations or single genes, the autocorrelation function is computed with an occurrence counting algorithm (see Section 3.2.2). For simulated genes created either by an independent mixing or a Markov mixing of oligonucleotides, the autocorrelation function is computed with an approximated simulation algorithm of linear complexity (function of d and of the number and the length of sequences created) and with an exact calculus algorithm of polynomial complexity (function of d and of the number and the length of oligonucleotides chosen). Both algorithms are necessary and complementary (see the complexity problems analyzed in [11]).

A new method based on analytical expressions allows us to study the autocorrelation function (the R/Y d-motif probabilities) after transversions (in the evolutionary sense; from the past to the present) and before transversions (after back transversions, in the inverse evolutionary sense; from the present to the past). Different properties and a generalization to d-motifs with motifs of any base length are also derived from these formulas. The formulas obtained here are simple and general enough to be applied in a series of situations. In particular, the generalization of $P_i(x)$ and $Q_i(y)$ (in the remarks of Sections 2.2 and 2.3) with $\lambda_1 = 1$ and $\lambda_2 = 2$ (or $\lambda_1 = 2$ and $\lambda_2 = 1$) and with $d = 0$ allows to study the R/Y codon probabilities under transversions in both evolutionary senses and to retrieve the particular formula derived in [5].

As the site and the order of previous substitutions are unknown, it is impossible to reproduce the effects of back substitutions in the nucleotide ordering of actual genes (unidirectional arrow in Figure 8). Unexpectedly, some statistical measures of the substitution process can be inverted, allowing to go backward in time. Indeed, if the substitution process is studied, not with the nucleotide ordering, but with probabilities (obtained from a good statistical function analyzing the nucleotide ordering, e.g., the autocorrelation function) then it can be inverted. Indeed, the formula $P_i(x)$ (1) giving the d-motif probabilities after x transversions can be inverted, and the inverse formula $Q_i(y)$ (5) gives the d-motif probabilities before y transversions. Therefore, the d-motif probabilities of primitive genes can be determined by applying the formula $Q_i(y)$ (5) with the d-motif probabilities $Q_j(0)$ of actual genes obtained from gene databases. Furthermore, the identification of nonrandom properties in the primitive genes, e.g., a d-motif with a maximal

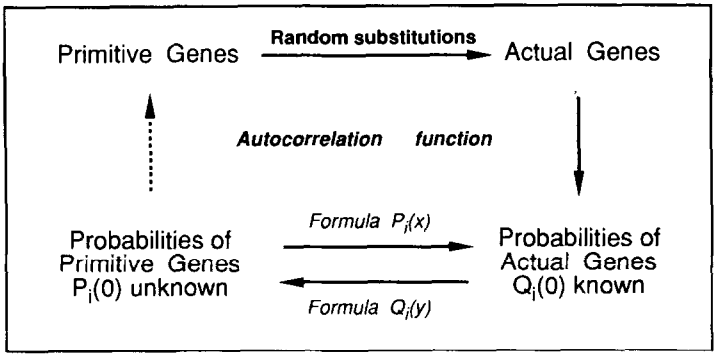


FIG. 8. Study of the substitution process in the evolutionary sense (from the past to the present) and in the inverse evolutionary sense (from the present to the past) with probabilities.

value or periodicities, implies some rules in the nucleotide ordering of these primitive genes (dashed arrow in Figure 8). Finally, this approach allows to determine the maximal number of transversions for a given gene population.

The study of the substitution process confirms and improves the model of DNA sequence evolution recently proposed and according to which actual genes on the R/Y alphabet are the result of two successive evolutionary genetic processes [3].

The first genetic process is the mixing of a few nonrandom types of oligonucleotides (series less than 10 bases) leading to genes (series of several hundreds of bases) called primitive genes. In this model, it was proved in particular that (1) the mixing is independent by using initially a Markov mixing; and (2) so far, five oligonucleotides identified are involved in this mixing: first YRYRYR, YRYYRY, YRY$(N)_6$ [3] and later also R^8, Y^8 [4]. The primitive genes resulting from this oligonucleotide mixing have the main (but not all) nonrandom statistical properties observed in actual genes on the R/Y alphabet, in particular the periodicities modulo 2 and 3 and the YRY$(N)_6$YRY preferential occurrence [3]. The modification of a base, a length or a probability in the mixing of one of these oligonucleotides leads to primitive genes without the properties observed in actual genes: there is no correlation between simulated genes and real genes. However, the mixing process is insufficient at least for two reasons: (1) a few nonrandom properties cannot be so far generated by a mixing process, in particular the perturbation with the second highest value of YRY$(N)_6$YRY after YRY$(N)_0$YRY with a periodicity modulo 3; and (2) the nonrandom

properties generated by a mixing process have higher probabilities compared to those in actual genes.

We had earlier proposed from a statistical point of view but without a proof, that a second genetic process related to the mutation process must be added after the mixing process (in [3, Sect. 1 and 3.3.2]). In this article we offer the first proof that a substitution process improves the simulation of the genetic reality. While the mixing process mainly acts on the relative values in the simulated curves (on the curve shape) [3], the substitution process mainly acts on the absolute values in the simulated curves. Indeed, the global effect of transversions decreases as expected the probability variations between the top curve $d \equiv 0[3]$ and the bottom curve $d \equiv 1,2[3]$ and also inside each curve. For example, by analyzing the transversions in the evolutionary sense (i.e., from the past [Figures 4c and 5c] to the present [Figures 4a and 5a]), the amplitude of the periodicity modulo 3 (mean probability difference between the top and bottom curves) decreases from 0.004 (Figures 4c and 5c) to 0.002 (Figures 4a and 5a), the probability curve shape of $YRY(N)_6YRY$ decreases (increases with back transversions [Figure 7]), etc. However, some local effects cannot be intuitively predicted as a single substitution in a gene modifies several d-motifs, e.g., the probability curve shape of $YRY(N)_0YRY$ which unexpectedly increases (decreases with back transversions [Figure 6]), etc.

The application proposed here shows that the protein genes of both eukaryotic nuclei and prokaryotes have the $YRY(N)_6YRY$ preferential occurrence with a periodicity modulo 3 after 0.2 back transversions such as that of the actual protein genes of chloroplasts and mitochondria. Therefore, more transversions have occurred in the protein genes of eukaryotic nuclei and prokaryotes than with chloroplasts and mitochondria. In the introns and 5′ regions of eukaryotic nuclei, the $YRY(N)_6YRY$ preferential occurrence is hidden by large alternating R/Y stretches, whereas in the protein genes of eukaryotic nuclei and prokaryotes, it is hidden by random substitutions.

The lower rate of transversions in protein genes of organelles (chloroplasts and mitochondria) compared to eukaryotic nuclei and prokaryotes may be related to a globally lower rate of substitutions in organelle genes. Contrary to our approach, in which substitutions studied by transversions are analyzed in gene populations (e.g., all the mitochondria), other methods, experimental (e.g., restriction-enzyme mapping) and statistical, also support the previous assumption. The method of Li et al. [12] and others, in which substitutions classified as nonsynonymous and synonymous are analyzed in a few genes, showed that [13, p. 86; 14, p. 142; 15, 16] (1) in plants, the synonymous rates in organelle genes are lower than those in nuclear genes (the synonymous rates of

plant mitochondrial, chloroplast, and plant nuclear genes are in the approximate ratio 1:3:12); (2) in plants, the nonsynonymous rates in organelle genes are lower than those in nuclear genes (the nonsynonymous rates of plant mitochondrial, chloroplast, and plant nuclear genes are in the approximate ratio $1:1:>2$); (3) the synonymous rates in plant organelle genes are lower than those in mammalian nuclear genes (plant mitochondrial, primate nuclear, and rodent nuclear genes are in the approximate ratio 1:2–5:10–20; chloroplast, primate nuclear, and rodent nuclear genes are in the approximate ratio 1:1:4); and (4) the synonymous rates in plant nuclear genes are similar to those in mammalian nuclear genes. However, mitochondrial genes of mammals, in contrast to those of plants, have a higher rate of substitutions than in mammalian nuclear genes but mainly related to transitions [17, 18]. Therefore, the transversion rate in mammalian mitochondrial genes could well be of the same order as that in plant mitochondrial genes. In addition, the mitochondrial gene population used here may have a lower rate of substitutions (and more certainly for transversions) than in (plant and/or mammalian) nuclear genes as the mitochondrial population contains mammalian as well as plant genes. In summary, previous studies may support the observation obtained here with a lower rate of transversions in protein genes of organelles than of nuclei.

The lower rate of substitutions in organelle genes may be explained by the existence of several space and functional constraints at the DNA sequence level. First, the number of protein genes in chloroplasts (about 80) and mitochondria (about 20) is very small compared to the one in eukaryotic nuclei and prokaryotes (estimated between 1000 and 3000 in *Escherichia coli*). Second, each organelle gene is often present in a single copy per genome. Third, the organelle genes code for proteins related to only a few functions: photosynthesis, respiration, and transcription and translation needed to express those genes. Fourth, the expression of these organelle genes is often optimized: (1) in animal mitochondria, the protein genes have neither 5′ nor 3′ regions: they simply begin directly with the initiator codon for protein synthesis and end with a partial terminator codon (T or TA); (2) several mitochondrial protein genes overlap, e.g., ATPase subunits 6 and 8, cytochrome oxidase subunit 2, and an adjacent ORF, etc. (reviewed in [19]). However, the reduced rate of substitutions at the DNA sequence level can be compensated by different molecular processes at the RNA level which are specific to organelles: (1) a genetic code with variations from the universal code in mammalian mitochondria [20]; (2) a particular translating code in chloroplasts [20]; (3) RNA editing by insertion and/or deletion of nucleotides in mitochondrial transcripts of trypanosomes [21] and of a slime mold [22]; (4) RNA editing by polyadeny-

lation generating UAA stop codons in transcripts of vertebrate mitochondria [23]; and (5) RNA editing by transition in plant mitochondrial transcripts [24, 25] and recently also observed in chloroplasts transcripts [26, 27], etc.

The numerical results of the application (the difference number 0.2, the maximal number 0.3) provide new information about gene mutation as they are obtained with an analytical expression analyzing back transversions in gene populations. These results have to be added to the list of other rates analyzing gene mutation [6, 7, 13, 14]: transition, synonymous and nonsynonymous rates, rates expressed in function of the time, etc.

The formulas $P_i(x)$ and $Q_i(y)$ are simple enough to be directly used. We are nevertheless currently implementing these analytical expressions and their generalization in the Analysis of Gene Evolution (AGE) software [11].

We thank Dr. Nouchine Soltanifar and the referees for their advice. This work was supported by CNRS Grant GDR 1029 and by an INSERM grant (contrat de recherche externe No. 930101).

REFERENCES

1 D. G. Arquès and C. J. Michel, Study of a perturbation in the coding periodicity, *Math. Biosci.* 86:1–14 (1987).

2 D. G. Arquès and C. J. Michel, A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups, *J. Theor. Biol.* 128:457–461 (1987).

3 D. G. Arquès and C. J. Michel, A model of DNA sequence evolution: 1. Statistical features and classification of gene populations. 2. Simulation model. 3. Return of the model to the reality, *Bull. Math. Biol.* 52:741–772 (1990).

4 D. G. Arquès, C. J. Michel, and K. Orieux, Identification and simulation of new non-random statistical properties common to different populations of eukaryotic non-coding genes, *J. Theor. Biol.* 161:329–342 (1993).

5 D. G. Arquès and C. J. Michel, Analytical expression of the purine/pyrimidine codon probability after and before random mutations, *Bull. Math. Biol.* 55:1025–1038 (1993).

6 M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, 1987.

7 M. Nei, *Molecular Evolutionary Genetics*, Columbia University Press, New York, 1987.

8 W. Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1968.

9 T. H. Jukes and C. R. Cantor, Evolution of protein molecules, In H. N. Munro, ed., *Mammalian Protein Metabolism*, Academic Press, New York, 1969, pp. 21–132.

10 M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotides sequences, *J. Mol. Evol.* 16:111–120 (1980).
11 D. G. Arquès, C. J. Michel, and K. Orieux, Analysis of gene evolution: The software AGE, *Comput. Appl. Biosci.* 8:5–14 (1992).
12 W.-H. Li, C.-I. Wu, and C.-C. Luo, A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes, *Mol. Biol. Evol.* 2:150–174 (1985).
13 W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution*, Sinauer Associates, MA, 1991.
14 R. K. Selander, A. G. Clark, and T. S. Whittam, *Evolution at the Molecular Level*, Sinauer Associates, MA, 1991.
15 K. H. Wolfe, W.-H. Li, and P. M. Sharp, Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs, *Proc. Natl. Acad. Sci. USA* 84:9054–9058 (1987).
16 K. H. Wolfe, P. M. Sharp, and W.-H. Li, Rates of synonymous substitution in plant nuclear genes, *J. Mol. Evol.* 29:208–211 (1989).
17 W. M. Brown, E. M. Prager, A. Wang, and A. C. Wilson, Mitochondrial DNA sequences of primates: Tempo and mode of evolution, *J. Mol. Evol.* 18:225–239 (1982).
18 G. G. Brown and M. V. Simpson, Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes, *Proc. Natl. Acad. Sci. USA* 79:3246–3250 (1982).
19 M. W. Gray, The endosymbiont hypothesis revisited, *Int. Rev. Cytol.* 141:233–357 (1992).
20 T. H. Jukes and S. Osawa, The genetic code in mitochondria and chloroplasts, *Experientia* 46:1117–1126 (1990).
21 R. Benne, J. Van Den Burg, J. P. J. Brakenhoff, P. Sloof, J. H. Van Boom, and M. C. Tromp, Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA, *Cell* 46:819–826 (1986).
22 R. Mahendran, M. R. Spottswood, and D. L. Miller, RNA editing by cytidine insertion in mitochondria of *Physarum polycephalum*, *Nature* 349:434–438 (1991).
23 D. Ojala, J. Montoya, and G. Attardi, tRNA punctuation model of RNA processing in human mitochondria, *Nature* 290:470–474 (1981).
24 P. S. Covello and M. W. Gray, RNA editing in plant mitochondria, *Nature* 341:662–666 (1989).
25 W. Schuster, R. Hiesel, B. Wissinger, and A. Brennicke, RNA editing in the cytochrome b locus of the higher plant *Oenothera berteriana* includes a U-to-C transition, *Mol. Cell. Biol.* 10:2428–2431 (1990).
26 B. Hoch, R. M. Maier, K. Appel, G. L. Igloi, and H. Kössel, Editing of a chloroplast mRNA by creation of an initiation codon, *Nature* 353:178–180 (1991).
27 R. M. Maier, B. Hoch, P. Zeltz, and H. Kössel, Internal editing of the maize chloroplast ndhA transcript restores codons for conserved amino acids, *Plant Cell* 4:609–616 (1992).