Bulletin of Mathematical Biology Vol. 55, No. 6, pp. 1025–1038, 1993
Printed in Great Britain

0092–8240/93$6.00+0.00
Pergamon Press Ltd

ANALYTICAL EXPRESSION OF THE PURINE/PYRIMIDINE CODON PROBABILITY AFTER AND BEFORE RANDOM MUTATIONS

■ DIDIER G. ARQUÈS
Equipe de Biologie Théorique,
Université de Franche-Comté,
Laboratoire d'Informatique de Besançon,
16 route de Gray,
25030 Besançon, France
■ CHRISTIAN J. MICHEL*
Equipe de Biologie Théorique,
Université de Franche-Comté,
Institut Universitaire de Technologie de Belfort-Montbéliard,
BP 527,
90016 Belfort, France

Recently, we proposed a new model of DNA sequence evolution (Arquès and Michel. 1990b. *Bull. math. Biol.* **52**, 741–772) according to which actual genes on the purine/pyrimidine (R/Y) alphabet (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine) are the result of two successive evolutionary genetic processes: (i) a mixing (independent) process of non-random oligonucleotides (words of base length less than 10: YRY$(N)_6$, YRYRYR and YRYYRY are so far identified; N = R or Y) leading to primitive genes (words of several hundreds of base length) and followed by (ii) a random mutation process, i.e. transformations of a base R (respectively Y) into the base Y (respectively R) at random sites in these primitive genes. Following this model the problem investigated here is the study of the variation of the 8 R/Y codon probabilities RRR, . . . , YYY under random mutations. Two analytical expressions solved here allow analysis of this variation in the classical evolutionary sense (from the past to the present, i.e. after random mutations), but also in the inverted evolutionary sense (from the present to the past, i.e. before random mutations). Different properties are also derived from these formulae. Finally, a few applications of these formulae are presented. They prove the proposition in Arquès and Michel (1990b. *Bull. math. Biol.* **52**, 741–772), Section 3.3.2, with the existence of a maximal mean number of random mutations per base of the order 0.3 in the protein coding genes. They also confirm the mixing process of oligonucleotides by excluding the purine/pyrimidine contiguous and alternating tracts from the formation process of primitive genes.

1. Introduction. Several models of purine/pyrimidine (R/Y) code structures were proposed (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine). Barrell and Clark (1974) showed that the anticodon in transfer RNA is bordered by an R in the 3′ location and by two Y in the 5′ location. Based on this

* Author to whom correspondence should be addressed.

anticodon region structure, and on the translation model developed by Woese (1970), Crick *et al.* (1976) proposed that the primitive coding genes are constituted of codons of the type RRY, with RNY (N = R or Y) as a less likely alternative. These concepts were again considered by Eigen and Schuster (1978) who favoured codons of the type RNY, giving more symmetry between R and Y and between the two strands. A gene statistical analysis by Shepherd (1981) and Smith *et al.* (1983) supports the RNY codon model. R/Y bases are also involved in the geometry of the DNA double helix, which is implied in the regulation of genes. For example, R–Y steps open preferentially their base planes towards the major groove of the B-DNA double helix, while Y–R steps towards the minor groove (Dickerson and Drew, 1981). The R/Y contiguous and alternating tracts which will be used as applications of the formulae can strongly modify the B-DNA double helix (see the Discussion).

Recently, we proposed a new model of DNA sequence evolution according to which actual genes on the R/Y alphabet are the result of two successive evolutionary genetic processes (Arquès and Michel, 1990b).

The first genetic process is the mixing of non-random oligonucleotides (words of base length less than 10) leading to genes (words of base length of several hundreds) called primitive genes. It was proved in particular that: (i) The mixing is independent by using initially a Markov mixing. (ii) Three oligonucleotides, YRY$(N)_6$, YRYRYR and YRYYRY (so far identified), are involved in this mixing. (iii) The primitive genes resulting from this oligonucleotide mixing have the main non-random statistical properties observed in the actual genes on the R/Y alphabet, in particular the periodicities modulo 2 and 3 (Arquès and Michel, 1990a) and the preferential occurrence of the motif YRY$(N)_6$YRY (Arquès and Michel, 1987a, 1987b). The modification of a specified base (R or Y), a length or a probability in the mixing of one of these three oligonucleotides leads to primitive genes without the genetic properties mentioned above. Note that the RNY codon model is a particular case of the oligonucleotide mixing model (Arquès and Michel, 1990b, p. 763, Section 3.3.3). However, the non-random properties in the primitive genes occur with a higher probability compared to the actual genes. This is the reason why a second genetic process must be added after the mixing process (Arquès and Michel, 1990b, Sections 1 and 3.3.2).

The second genetic process is related to random processes, random insertions and deletions of nucleotides (Arquès and Michel, 1990b, Section 1, point 2; 1992), but mainly random mutations, i.e. transformations of a base R (respectively Y) into the base Y (respectively R) at random sites in the primitive genes (Arquès and Michel, 1990b, Section 1, point 2, Section 3.3.2). Indeed, the mixing process acts on the relative values in the simulated curves (on the curve shape), i.e. it leads to non-random properties. However, the random mutation process acts on the absolute values in the simulated curves (no effect on the curve shape: random mutations are a noise in terms of signal processing), i.e. it cannot lead to non-random properties. The mutation process is a classical evolutionary genetic process and well accepted as it is retrieved in several molecular theories of evolution (e.g. Kimura, 1987; Nei, 1987). Nevertheless, we had proposed in Arquès and Michel (1990b, Section 3.3.2) the existence of a maximal random mutation rate of the order 1/2 per specified base (R or Y), however, without proof.

The problem investigated here is the study of the variation of the eight R/Y codon probabilities RRR, . . . , YYY under random mutations. Two analytical expressions solved here allow the analysis of this variation in the classical evolutionary sense (from the past to the present, i.e. after random mutations), but also in the inverted evolutionary sense (from the present to the past, i.e. before random mutations). Different properties are also derived from these formulae. Finally, a few applications of these formulae are presented. They prove the proposition in Arquès and Michel (1990b, Section 3.3.2) with the existence of a maximal mean number of random mutations per base of the order 0.3 in the protein coding genes. They also confirm the mixing process of oligonucleotides by excluding the R/Y contiguous and alternating tracts from the formation process of primitive genes.

2. Method and Results.

2.1. *Recall of the Poisson even/odd distribution associated with random mutations.* Let S be a sequence on the alphabet $\{R, Y\}$ (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine). This sequence S is subjected to random mutations, i.e. transformations of a base R (respectively Y) into the base Y (respectively R) at random sites s in S. Let x be the mean number of random mutations per base site (per base) between times 0 and t (see also Scheme 1 below), i.e. $x = \lambda t$, where λ is the mean number of random mutations per base site per unit of time.

Under classical assumptions (Feller, 1968, p. 447; Kimura, 1987, p. 69; Nei, 1987, p. 40; Haldane, 1927, p. 839; Haldane, 1990, appendix) the counting process $\{N(t), t \geqslant 0\}$, giving the number of random mutations per base site in the time interval $[0, t]$, is a Poisson process with rate $\lambda > 0$, i.e. the probability $P_n(t)$ of a base site to be subjected to n random mutations in the time interval $[0, t]$ is:

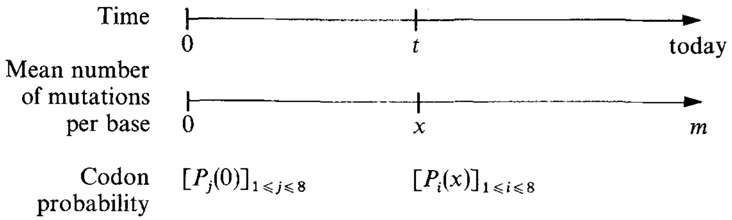
$$P_n(t) = P(N(t) = n) = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} = \mathrm{e}^{-x} \frac{x^n}{n!}.$$

Therefore, the probability of a base site to have changed, i.e. R→Y or Y→R (respectively, not to have changed, i.e. R→R or Y→Y) after x random mutations per base site on average in the time interval $[0, t]$ is the probability of this base site being subjected to an odd (respectively even) number of random mutations, i.e.:

$$\mathcal{O}(x)=\sum_{k\geqslant 0} P_{2k+1}(t)=\frac{1-\mathrm{e}^{-2x}}{2}, \text{ respectively } \mathcal{E}(x)=\sum_{k\geqslant 0} P_{2k}(t)=\frac{1+\mathrm{e}^{-2x}}{2}.$$

These two equations 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 mutations; Jukes and Cantor, 1969) and with the two-parameter model (a rate of transitions and a rate of transversions; Kimura, 1980).

2.2. *Analytical expression of the codon probability after random mutations.* The codon probability after x random mutations per base (at time t) can be obtained from the codon probability before the mutation process (at time 0) (see Scheme 1), m being the unknown number of random mutations per base between the times 0 and today:


Time
0
t
today
Mean number of mutations per base
0
x
m
Codon probability
$[P_j(0)]_{1\leqslant j\leqslant 8}$
$[P_i(x)]_{1\leqslant i\leqslant 8}$

Scheme 1.

By convention the index $j\in[1, 8]$ represents the codons RRR, . . . , YYY in alphabetical order. Let $P_j(0), j\in[1, 8]$, be the probability of the codon j being in a given sequence before the mutation process. Let $P_i(x)$, $i\in[1, 8]$, be the probability of the codon i being in a given sequence after x random mutations per base on average; depending on the context, we note indifferently $P_1(x)$ or $P_{\mathrm{RRR}}(x)$, etc. Let $\mathcal{I}(j, i)$ be the number of identical bases in the same codon site between the codons i and j. Then $P_i(x)$ can be expressed as a function of $[P_j(0)]_{1\leqslant j\leqslant 8}$. Indeed:

$$\begin{aligned} P_i(x) &= \sum_{j=1}^{8} P_j(0)\times P(\text{codon } j\to\text{codon } i \text{ after } x \text{ random mutations per base})\\ &= \sum_{j=1}^{8} P_j(0)\mathcal{E}(x)^{\mathcal{I}(j,i)}\mathcal{O}(x)^{3-\mathcal{I}(j,i)} \qquad (1)\\ &= \tfrac{1}{8}\sum_{j=1}^{8} P_j(0)\,(1+\mathrm{e}^{-2x})^{\mathcal{I}(j,i)}(1-\mathrm{e}^{-2x})^{3-\mathcal{I}(j,i)}\\ &= \tfrac{1}{8}\sum_{k=0}^{3}\Bigg(\sum_{j/\mathcal{I}(j,i)=k} P_j(0)\Bigg)(1+\mathrm{e}^{-2x})^{k}(1-\mathrm{e}^{-2x})^{3-k}. \end{aligned}$$

2.2.1. Properties of $P_i(x)$. Property 1: The formula giving $P_i(x)$ is true for a codon frequency $P_i(0)$, computed either in a modulo 3 frame (obvious) or without frame. Indeed, the codon frequency computed without frame is the mean of the codon frequencies computed in the 3 modulo 3 frames. Therefore, the formula $P_i(x)$ remains true in the "without frame" case by linear combination of the three formulae $P_i(x)$, which are true in all three frames. Note: According to the definitions in Section 2.2 a codon is any trinucleotide (the term codon is used for convenience). In biology, this definition is too general and the term codon is restricted to a trinucleotide in the open reading frame.

Property 2: The codon frequency computed without frame leads to the following property with the codons RRY, YRR, RYY and YYR: $P_{\mathrm{RRY}}(x)=P_{\mathrm{YRR}}(x)$ and $P_{\mathrm{RYY}}(x)=P_{\mathrm{YYR}}(x)$, whatever the mean number x of random mutations per base; in particular $P_{\mathrm{RRY}}(0)=P_{\mathrm{YRR}}(0)$ and $P_{\mathrm{RYY}}(0)=P_{\mathrm{YYR}}(0)$. Indeed, in an infinite sequence the codon YRR (respectively RYY) preceding a series of R (respectively Y) can be bijectively associated with the codon RRY (respectively YYR) following this series of R (respectively Y) (see the scheme below for the case of a series of R).

$$\ldots\boxed{\mathrm{YRR}}\,\mathrm{RRR}\ldots\mathrm{RRR}\,\boxed{\mathrm{RRY}}\ldots$$

Property 3: If x and z are mean numbers of random mutations per base, then $(P_i(x))(z)=P_i(x+z)$.

2.2.2. Two applications of the formula $P_i(x)$ in tracts. The formula $P_i(x)$ is applied in the purine/pyrimidine contiguous and alternating tracts as they are currently studied experimentally. Indeed, these two basic tracts are associated with important biological functions described in the Discussion. The results of these applications are presented below and commented on in the Discussion.

2.2.2.1. Application 1: Purine/pyrimidine codon probability in the purine contiguous tract after random mutations. If the sequence S is the purine contiguous tract before the mutation process, then $P_{\mathrm{RRR}}(0)=1$ and $P_j(0)=0$ if $j\neq 1$ (all codons are of type RRR). Then:

$$P_i(x)=\tfrac{1}{8}(1+\mathrm{e}^{-2x})^{\mathscr{J}(1,i)}(1-\mathrm{e}^{-2x})^{3-\mathscr{J}(1,i)}$$

leading to:

$$P_{\mathrm{RRR}}(x)=\tfrac{1}{8}(1+\mathrm{e}^{-2x})^3 \qquad \text{(curve C1 in Fig. 1a)}$$

$$\begin{aligned}P_{\mathrm{RRY}}(x)&=P_{\mathrm{RYR}}(x)=P_{\mathrm{YRR}}(x)\\&=\tfrac{1}{8}(1+\mathrm{e}^{-2x})^2(1-\mathrm{e}^{-2x}) \qquad \text{(curve C2 in Fig. 1a)}\end{aligned}$$

$$P_{\mathrm{RYY}}(x)=P_{\mathrm{YRY}}(x)=P_{\mathrm{YYR}}(x)$$
$$=\tfrac{1}{8}(1+\mathrm{e}^{-2x})(1-\mathrm{e}^{-2x})^2 \quad \text{(curve C3 in Fig. 1a)}$$
$$P_{\mathrm{YYY}}(x)=\tfrac{1}{8}(1-\mathrm{e}^{-2x})^3 \quad \text{(curve C4 in Fig. 1a).}$$

As before the mutation process the codon probability in the purine contiguous tract [i.e. $P_{\mathrm{RRR}}(0)=1$] is identical whatever the frame, the eight codon probabilities $P_i(x)$ after random mutations are also identical whatever the frame [because they are obtained from the same formula (1) in Section 2.2] and equal to the probabilities $P_i(x)$ computed without frame. The equalities mentioned above, $P_{\mathrm{RRY}}(x)=P_{\mathrm{YRR}}(x)$ and $P_{\mathrm{RYY}}(x)=P_{\mathrm{YYR}}(x)$, are the consequence of property 2 in Section 2.2.1 (verified: curve C2 and curve C3 in Fig. 1a).

Remark. By exchanging the complementary bases R and Y the formulae in the pyrimidine contiguous tract remain the same as the ones in the purine contiguous tract.

2.2.2.2. Application 2: Purine/pyrimidine codon probability in the purine/pyrimidine alternating tract after random mutations. If the sequence S is the purine/pyrimidine alternating tract before the mutation process then $P_{\mathrm{RYR}}(0)=P_{\mathrm{YRY}}(0)=1/2$ and $P_j(0)=0$ otherwise (all codons are of type RYR and YRY) and $\mathscr{I}(6,i)=3-\mathscr{I}(3,i)$ for all i (3 representing RYR and 6, YRY). Then:

$$P_i(x)=\tfrac{1}{16}[(1+\mathrm{e}^{-2x})^{\mathscr{I}(3,i)}(1-\mathrm{e}^{-2x})^{3-\mathscr{I}(3,i)}+(1+\mathrm{e}^{-2x})^{3-\mathscr{I}(3,i)}(1-\mathrm{e}^{-2x})^{\mathscr{I}(3,i)}]$$

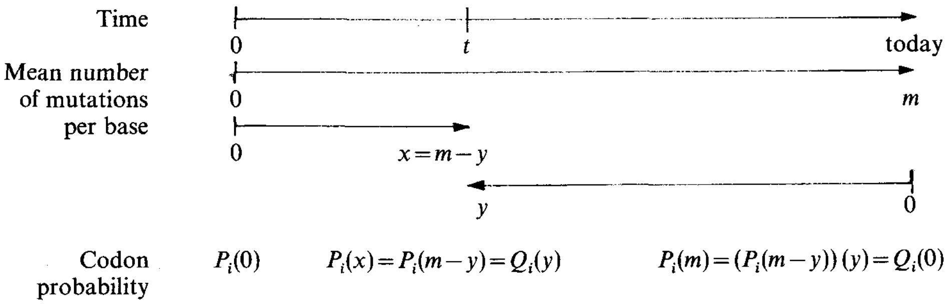
leading to:

$$P_{\mathrm{RYR}}(x)=P_{\mathrm{YRY}}(x)=\tfrac{1}{8}(1+3\mathrm{e}^{-4x}) \quad \text{(curve C1 in Fig. 1b)}$$
$$P_{\mathrm{RRR}}(x)=P_{\mathrm{RRY}}(x)=P_{\mathrm{RYY}}(x)=P_{\mathrm{YRR}}(x)=P_{\mathrm{YYR}}(x)=P_{\mathrm{YYY}}(x)=\tfrac{1}{8}(1-\mathrm{e}^{-4x})$$
$$\text{(curve C2 in Fig. 1b).}$$

As before the mutation process the codon probabilities in the purine/pyrimidine alternating tract [i.e. $P_{\mathrm{RYR}}(0)=P_{\mathrm{YRY}}(0)=1/2$] are identical whatever the frame, the eight codon probabilities $P_i(x)$ after random mutations are also identical whatever the frame and equal to the probabilities $P_i(x)$ computed without frame. The equalities mentioned above, $P_{\mathrm{RRY}}(x)=P_{\mathrm{YRR}}(x)$ and $P_{\mathrm{RYY}}(x)=P_{\mathrm{YYR}}(x)$, are the consequence of property 2 in Section 2.2.1 (verified: curve C2 in Fig. 1b).

2.3. *Analytical expression of the codon probability before random mutations.* The problem of Section 2.3 is the inverted problem of Section 2.2. Let m (respectively x) be the mean number of random mutations per base between the times 0 and today (respectively t) (see Scheme 2). Let y be the mean number of random mutations per base between times t and today, i.e. $m=x+y$.

In the previous problem of Section 2.2 the reference time is the time 0 (before the mutation process), while in the inverted problem the reference time is today (after the mutation process).



Scheme 2.

Therefore, the inverted problem consists of inverting the equation in Section 2.2, giving $P_i(x)$ as a function of $P_i(0)$, and to express $Q_i(y)=P_i(m-y)$ as a function of $Q_i(0)=P_i(m)$, $i\in[1, 8]$ representing the alphabetical order of codons RRR, . . . , YYY [depending on the context, we note indifferently $Q_1(y)$ or $Q_{\mathrm{RRR}}(y)$, etc.]. The solution obtained is:

$$Q_i(y)=\sum_{j=1}^{8} P_j(m)\mathscr{E}(-y)^{\mathscr{I}(j,i)}\mathscr{O}(-y)^{3-\mathscr{I}(j,i)}$$

with:

$$\mathscr{E}(-y)=\frac{1+\mathrm{e}^{2y}}{2}$$

$$\mathscr{O}(-y)=\frac{1-\mathrm{e}^{2y}}{2}$$

leading to:

$$Q_i(y)=\tfrac{1}{8}\sum_{k=0}^{3}\left(\sum_{j/\mathscr{I}(j,i)=k} P_j(m)\right)(1+\mathrm{e}^{2y})^k(1-\mathrm{e}^{2y})^{3-k}.$$

Proof. The formula $Q_i(y)$ can be proved in two ways:

(i) by inversion of the matrix $(\mathscr{E}(x)^{\mathscr{I}(j,i)}\mathscr{O}(x)^{3-\mathscr{I}(j,i)})_{1\leqslant i,\, j\leqslant 8}$ associated with formula (1) in Section 2.2;

(ii) by generalization of property 3 in Section 2.2.1 to negative numbers, i.e. by replacing x by m and z by $-y$. ■

2.3.1. Properties of $Q_i(y)$. The properties of $Q_i(y)$ are similar to those of $P_i(x)$.

(1) The formula $Q_i(y)$ is true for a codon frequency $P_i(m)$ computed either in a modulo 3 frame or without frame.

(2) The codon frequency computed without frame leads to the following property with the codons RRY, YRR, RYY and YYR: $Q_{\mathrm{RRY}}(y)=Q_{\mathrm{YRR}}(y)$ and $Q_{\mathrm{RYY}}(y)=Q_{\mathrm{YYR}}(y)$, whatever the mean number y of random mutations per base.

(3) If y and z are mean numbers of random mutations per base, then $(Q_i(y))(z)=Q_i(y+z)$.

2.3.2. Applications of the formula $Q_i(y)$ in eukaryotic protein coding genes. The formula $Q_i(y)$ is particularly interesting as the actual codon frequencies $P_i(m)$ are known and can be computed with gene databases. Contrary to the formula $P_i(x)$, which converges as expected towards the random value $1/8=0.125$ when x increases (a consequence of the negative exponentials and see e.g. Figs 1a–b), the formula $Q_i(y)$ does not converge when y increases (see e.g. Fig. 2). However, the vector $(Q_i(y))_{1 \leqslant i \leqslant 8}$ must remain a vector of probability, i.e. the eight values $Q_i(y)$ must be bounded between 0 and 1 (and of sum 1). In summary, the formula $Q_i(y)$ gives the codon probability before y random mutations per base and the condition $0 \leqslant Q_i(y) \leqslant 1$ for i in [1, 8] implies a maximal mean number of random mutations per base.

The mean codon frequencies in eukaryotic protein coding genes computed in the open reading frame (modulo 3 frame) of a population from the EMBL database containing 13,169 available genes (6718 kb), lead to the following $P_i(m)$ values: $P_{\mathrm{RRR}}(m)=0.1772$; $P_{\mathrm{RRY}}(m)=0.1593$; $P_{\mathrm{RYR}}(m)=0.1067$; $P_{\mathrm{RYY}}(m)=0.1561$; $P_{\mathrm{YRR}}(m)=0.0732$; $P_{\mathrm{YRY}}(m)=0.0923$; $P_{\mathrm{YYR}}(m)=0.1028$; $P_{\mathrm{YYY}}(m)=0.1324$. The formula $Q_i(y)$ applied with these $P_i(m)$ values (Fig. 2) leads to a maximal mean number of random mutations per base equal to 0.27. Indeed, the decrease of the probability of the codon YRR reaches 0 when $y=0.27$. This result is commented on in the Discussion.

2.3.3. Properties of $P_i(x)$ and $Q_i(y)$. $P_i(x)=Q_i(y)$ if $x+y=m$ and $0 \leqslant x, y \leqslant m$ (see Scheme 2). The formula $P_i(x)$ gives the evolution of the codon probabilities when we go from the past to the present and when the number of random mutations increases from 0 to m. $P_i(x)$ can be obtained either exactly by analytical expression or approximately by computer simulation (simulation of random mutations in simulated sequences). The formula $Q_i(y)$ gives the inverted evolution of the codon probabilities when we go back in time and when the number of random mutations decreases from m to 0. The main 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: it is not possible to simulate

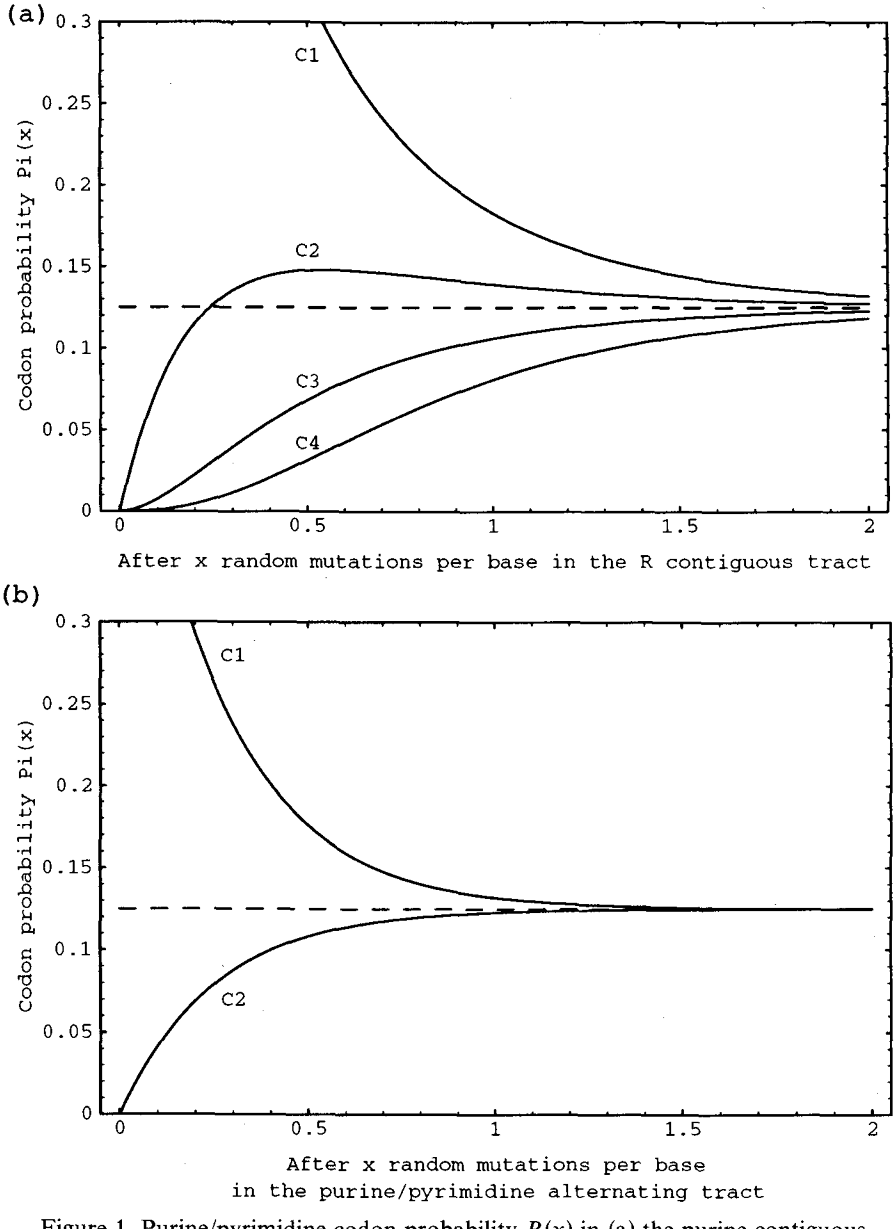


Figure 1. Purine/pyrimidine codon probability $P_i(x)$ in (a) the purine contiguous tract and (b) in the purine/pyrimidine alternating tract after x random mutations per base, $x \in [0, 2]$. (a) Curve C1 represents the codon RRR; curve C2 the codons RRY, RYR and YRR; curve C3 the codons RYY, YRY and YYR; curve C4 the codon YYY. Curve C2 first increases up to the maximum $2^2/3^3 = 0.1481$ (value greater than the random one) reached at $x = (\text{Log } 3)/2 = 0.5493$ random mutations per base, then decreases. (b) Curve C1 represents the codons RYR and YRY; curve C2 the codons RRR, RRY, RYY, YRR, YYR and YYY. The curves converge as expected towards the random value 0.125 (horizontal dash line).

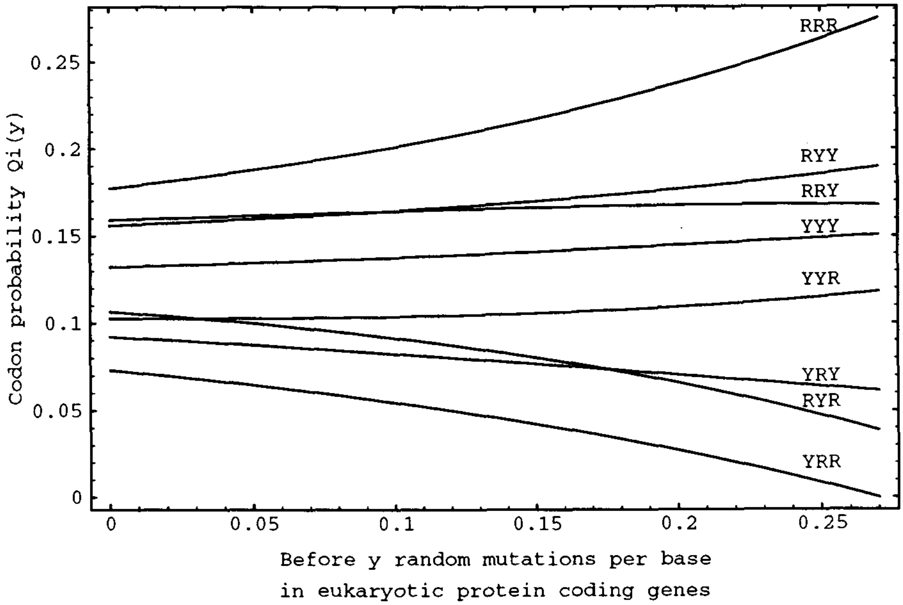


Figure 2. Purine/pyrimidine codon probability $Q_i(y)$ in eukaryotic protein coding genes (computed in open reading frame) before y random mutations per base, $y \in [0, 0.27]$. The probability $Q_{\mathrm{YRR}}(y)$ is equal to 0 when $y = 0.27$; therefore, the maximal mean number of random mutations per base in eukaryotic protein coding genes is 0.27.

inverted mutations as the site and the order of the previous mutations are not known.

3. Discussion. The analytical expression of the purine/pyrimidine codon probability after and also before random mutations (inverted evolutionary process, i.e. from the present to the past) is solved in this paper. Different properties are also derived from these formulae. The formulae obtained are simple and general enough to be applied in a series of situations. In particular, the formulae are independent of the nucleotide ordering in the sequence, e.g. they remain true for a codon frequency computed in a modulo 3 frame or not (see properties 1 in Sections 2.2.1 and 2.3.1). Finally, a few applications of these formulae were presented and are commented below.

Two applications of the formula giving $P_i(x)$ deal with the purine/pyrimidine (R/Y) contiguous and alternating tracts. These tracts were chosen as an increasing number of experimental studies (the use of the experimental method called PCR: Polymerase Chain Reaction; Mullis and Faloona, 1987; Erlich *et al.*, 1991; Weber and May, 1989) show that these two basic tracts (also called microsatellites) are associated with important biological functions. Mainly four biological functions of the R/Y contiguous tracts were identified, depending on their location in gene subregions. In intergenic regions these tracts play a structural role in chromosome or nucleosomal organization (Beasty and Behe, 1988), in exons, a coding function, in introns, an RNA splicing function (Keller and Noon, 1984) and in promoter regions, a regulatory function of gene expression (Larsen and Weintraub, 1982). The R/Y alternating tracts have the potential for forming Z-DNA (Wang *et al.*, 1979; Konopka *et al.*, 1985), which may play a role in gene regulation (Hamada *et al.*, 1984), in genetic recombination (Treco and Arnheim, 1986) and in condensing and decondensing DNA (Stallings *et al.*, 1991). After the identification of a function for a tract, a classical experimental approach consists of mutating the tract in order to quantify its function. For example mutations in the R contiguous tracts of promoter regions allow to study the sensitivity to single strand nucleases and the unwound state of the DNA structure which is involved in gene expression (e.g. Evans and Efstratiadis, 1986; Hanvey *et al.*, 1988). Mutations in the Y contiguous tracts located in the 3′ splice site in mammalian introns enable to analyse their splicing efficiency (e.g. Wieringa *et al.*, 1984; Ruskin and Green, 1985). Mutations in the R/Y alternating tracts allow the characterization of the Z-DNA form (e.g. Ellison *et al.*, 1985). On the other hand, evolutionary mutations (in contrast to the experimental mutations) are an important cause of the tract polymorphism (length, number of repeats and intensity of mutations; Hamada *et al.*, 1984; Tautz and Renz, 1984; Jeffreys *et al.*, 1988; Weber, 1990). However, the quantification of this tract polymorphism remains open (e.g. Weber, 1990). The random mutation model developed here may find solutions to these tract mutation problems. The formula $P_i(x)$ shows that only the probability of RRR decreases in the R contiguous tract after random mutations (or YYY in the Y contiguous tract; application 1 in Section 2.2.2 and Fig. 1a). Therefore, the lowering of the biological function in mutated R contiguous tracts might be related to their content in RRR (or in more general motifs containing RRR), decreasing according to $(1+e^{-2x})^3/8$ (x being the mean number of random mutations per base). The lowering of the biological function in mutated R/Y alternating tracts would be related to their content in RYR and YRY decreasing according to $(1+3e^{-4x})/8$ (application 2 in Section 2.2.2 and Fig. 1b). On the other hand, the formula $P_i(x)$ may also be used to precisely date the mutated tracts and to construct phylogenetic trees. Finally, it should be noted that even with these two simple sequences the probability curve form for some codons cannot be intuitively predicted, in particular the surprising form of the curve C2 in Fig. 1a.

The application of the formula $Q_i(y)$ in eukaryotic protein coding genes shows that the maximal mean number of random mutations per base is equal to 0.27 (Section 2.3.2 and Fig. 2). The formula $Q_i(y)$ applied in protein coding genes of prokaryotes and viruses leads for both taxonomic groups to a maximal value equal to 0.25 (data not shown). Even if the frequency for some R/Y codons is very different from eukaryotes to prokaryotes or viruses, protein coding genes have a maximal mean number of random mutations per base of the order 0.3. This result proves the proposition in Arquès and Michel's (1990b) Section 3.3.2 stating the existence of a maximal random mutation rate. The order of 1/2 given is close to 0.3. Two additional reasons concerning the stability of this result are briefly discussed. First, the R/Y codon frequencies are computed in a population of several thousands of protein coding genes. Due to the law of large numbers (Arquès and Michel, 1990b, Section 2.3.3) these mean frequencies are stable from a statistical point of view: the codon frequency difference between two (or even three) successive EMBL releases is less than 0.1% (data not shown). Second, this codon frequency difference has been tested in the case of eukaryotic protein coding genes. For the eight R/Y codons this frequency difference is stable (less than 1%) up to 0.6 back random mutations per base, then it exponentially increases (data not shown). Therefore, the formula $Q_i(y)$ is stable for a number of random mutations per base, which greatly exceeds the maximal value 0.3. The maximal value of 0.3 concerns a mean number of random mutations per base, i.e. the case where all R/Y bases in the primitive genes are equiprobably mutated. However, as mentioned in Arquès and Michel (1990b), Section 3.3.2, some R/Y sites can have a higher (and also lower) mutation rate compared to the average 0.3. Figure 2 also shows that the codon frequencies were not random in the past as they diverge from the random value 0.125. The codon YRR occurred with the lowest frequency (also verified with the protein coding genes of prokaryotes and viruses: data not shown), suggesting that primitive genes would have contained stop codons only rarely (YRR codes in particular for the three stop codons).

The R/Y codon frequencies are completely different whether they are computed in the open reading frame (ORF), in frame 1 (ORF shifted from one base) or in frame 2 (ORF shifted from two bases). For example, with the eukaryotic protein coding genes the frequency of the codon YRR in the ORF is the lowest, 0.0732, but its frequency in frame 1 is equal to 0.1311 and in frame 2 to 0.1689; the frequency of the codon YYR in the ORF is equal to 0.1028, but its frequency in frame 1 is equal to 0.1647 and in frame 2 to 0.0949 (see Section 2.3.2; the data with frames 1 and 2 are not shown), etc. This codon frequency difference in frame is retrieved with all codons and with protein coding genes of other taxonomic groups: prokaryotes, viruses, chloroplasts and mitochondria (data not shown). This classical observation leads to an important result about primitive genes (genes before the mutation process in our model of gene evolution). Indeed, as random mutations in R/Y contiguous and alternating tracts lead to codon frequencies that are identical whatever the frame (see Sections 2.2.2.1 and 2.2.2.2), these tracts could not have been the formation process of primitive genes. Therefore, this random mutation model, excluding the R/Y contiguous and alternating tracts from the formation process of primitive genes, is in agreement with the mixing model of oligonucleotides being at the origin of primitive genes (the classical proof consisted of verifying that the R/Y contiguous and alternating tracts cannot lead to the non-random statistical properties observed in the actual genes and mentioned in the Introduction).

The formulae $P_i(x)$ and $Q_i(y)$ can be easily generalized to motifs different from the codon, i.e. to motifs of any base length. Otherwise, these formulae are simple enough to be directly used. However, we are currently implementing these analytical expressions and their generalization in the software AGE (Analysis of Gene Evolution) (Arquès *et al.*, 1992).

Note Added in Proof. The analytical expressions of the probability of motifs of any base length on the genetic alphabets {R, Y} and {A, C, G, T} and on the protein alphabet after and before random mutations have been derived (Arquès and Michel, submitted).

We thank Dr Nouchine Soltanifar and a Referee for their advice.

LITERATURE

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

Arquès, D. G. and C. J. Michel. 1987b. A purine–pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. *J. theor. Biol.* **128**, 457–461.

Arquès, D. G. and C. J. Michel. 1990a. Periodicities in coding and noncoding regions of the genes. *J. theor. Biol.* **143**, 307–318.

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

Arquès, D. G. and C. J. Michel. 1992. A simulation of the genetic periodicities modulo 2 and 3 with processes of nucleotide insertions and deletions. *J. theor. Biol.* **156**, 113–127.

Arquès, D. G., C. J. Michel and K. Orieux. 1992. Analysis of Gene Evolution: the software AGE. *Comp. appl. Biosci.* **8**, 5–14.

Barrell, B. G. and B. F. C. Clark. 1974. *Handbook of Nucleic Acid Sequences.* Oxford: Joynson-Bruvvers.

Beasty, A. M. and M. J. Behe. 1988. An oligopurine sequence bias occurs in eukaryotic viruses. *Nucl. Acids Res.* **16**, 1517–1528.

Crick, F. H. C., S. Brenner, A. Klug and G. Pieczenik. 1976. A speculation on the origin of protein synthesis. *Orig. Life* **7**, 389–397.

Dickerson, R. E. and H. R. Drew. 1981. Kinematic model for B-DNA. *Proc. natn Acad. Sci. U.S.A.* **78**, 7318–7322.

Eigen, M. and P. Schuster. 1978. The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften* **65**, 341–369.

Ellison, M. J., R. J. Kelleher III, A. H.-J. Wang, J. F. Habener and A. Rich. 1985. Sequence-dependent energetics of the B-Z transition in supercoiled DNA containing nonalternating purine–pyrimidine sequences. *Proc. natn Acad. Sci. U.S.A.* **82**, 8320–8324.

Erlich, H. A., D. Gelfand and J. J. Sninsky. 1991. Recent advances in the polymerase chain reaction. *Science* **252**, 1643–1651.

Evans, T. and A. Efstratiadis. 1986. Sequence-dependent S1 nuclease hypersensitivity of a heteronomous DNA duplex. *J. biol. Chem.* **261**, 14771–14780.

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

Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Camb. Phil. Soc.* **27**, 838–844.

Haldane, J. B. S. 1990. *The Causes of Evolution.* Princeton, NJ: Princeton University Press.

Hamada, H., M. Seidman, B. H. Howard and C. M. Gorman. 1984. Enhanced gene expression by the poly(dT-dG)·poly(dC-dA) sequence. *Mol. cell. Biol.* **4**, 2622–2630.

Hanvey, J. C., J. Klysik and R. D. Wells. 1988. Influence of DNA sequence on the formation of non-B right-handed helices in oligopurine·oligopyrimidine inserts in plasmids. *J. biol. Chem.* **263**, 7386–7396.

Jeffreys, A. J., N. J. Royle, V. Wilson and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**, 278–281.

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

Keller, E. B. and W. A. Noon. 1984. Intron splicing: a conserved internal signal in the introns of animal pre-mRNAs. *Proc. natn Acad. Sci. U.S.A.* **81**, 7417–7420.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotides sequences. *J. molec. Evol.* **16**, 111–120.

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

Konopka, A. K., J. Reiter, M. Jung, D. Zarling and T. M. Jovin. 1985. Concordance of experimentally mapped or predicted Z-DNA sites with positions of selected alternating purine-pyrimidine tracts. *Nucl. Acids Res.* **13**, 1683–1701.

Larsen, A. and H. Weintraub. 1982. An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. *Cell* **29**, 609–622.

Mullis, K. B. and F. A. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase catalysed chain reaction. In *Methods in Enzymology.* R. Wu (Ed.), Vol. 155. San Diego: Academic Press.

Nei, M. 1987. *Molecular Evolutionary Genetics.* Washington, DC: Columbia University Press.

Ruskin, B. and M. R. Green. 1985. The role of the 3′ splice site concensus sequence in mammalian pre-mRNA splicing. *Nature* **317**, 732–734.

Shepherd, J. C. W. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. natn Acad. Sci. U.S.A.* **78**, 1596–1600.

Smith, T. F., M. S. Waterman and J. R. Sadler. 1983. Statistical characterization of nucleic acid sequence functional domains. *Nucl. Acids Res.* **11**, 2205–2220.

Stallings, R. L., A. F. Ford, D. Nelson, D. C. Torney, C. E. Hildebrand and R. K. Moyzis. 1991. Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. *Genomics* **10**, 807–815.

Tautz, D. and M. Renz. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucl. Acids Res.* **17**, 6463–6471.

Treco, D. and N. Arnheim. 1986. The evolutionary conserved repetitive sequence $d(TG \cdot AC)_n$ promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. *Molec. Cell. Biol.* **6**, 3934–3947.

Wang, A. H.-J., G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. Van Boom, G. Van der Marel and A. Rich. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680–686.

Weber, J. L. 1990. Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. *Genomics* **7**, 524–530.

Weber, J. L. and P. E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396.

Wieringa, B., E. Hofer and C. Weissmann. 1984. A minimal intron length but no specific internal sequence is required for splicing the large rabbit b-globin intron. *Cell* **37**, 915–925.

Woese, C. R. 1970. Molecular mechanics of translation: a reciprocating ratchet mechanism. *Nature* **226**, 817–820.

Received and revised 18 January 1993