Analytical Solutions of the Dinucleotide Probability after and before Random Mutations

Didier G. Arquès† and Christian J. Michel‡§

†*Equipe de Biologie Théorique, Université de Franche-Comté, Laboratoire d'Informatique de Besançon,* 16 *route de Gray,* 25030 *Besançon, France and* ‡*Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie de Belfort-Montbéliard, BP* 527, 90016 *Belfort, France*

(*Received on* 22 *July* 1994, *Accepted in revised form on* 3 *February* 1995)

The mutation process is a classical evolutionary genetic process mainly based on the (random) substitutions of one base (A = Adenine, C = Cytosine, G = Guanine, T = Thymine) for another. Two analytical solutions derived here allow us to analyse in genes the occurrence probabilities of motifs (e.g. dinucleotides) after substitutions (in the evolutionary sense: from the past to the present) and, unexpectedly, also before substitutions (after back substitutions, in the inverse evolutionary sense: from the present to the past). We generalize on the alphabet {A, C, G, T} of the analytical solutions and of the properties derived on the alphabet {R, Y} (R = purine = A or G, Y = pyrimidine = C or T). Application of the theory is based on the analytical solution giving the probabilities of the 16 dinucleotides AA, . . . , TT in the protein (coding) genes of (nuclear) eukaryotes, viruses and prokaryotes and in (eukaryotic) introns after back substitutions (called primitive genes). After back substitutions, four of 16 dinucleotides—CG, TA, GT and AC—occur with low probabilities in each of these four primitive gene populations, except for CG in the primitive prokaryotic protein genes. In the primitive eukaryotic protein genes, the dinucleotide AT has also a significant low probability.

We present the properties of the two analytical solutions, and the functions which may have these five dinucleotides in primitive genes are described in terms of biological signals.

1. Introduction

The mutation process is a classical evolutionary genetic process that has been analysed by different theories (e.g. the neutral theory: Kimura, 1987; Nei, 1987). There are several types of mutation: substitutions, transitions, transversions, missense mutations, nonsense mutations, silent mutations, insertions, deletions, etc. The type of mutations studied here are the (random) substitutions of one base (A = Adenine, C = Cytosine, G = Guanine, T = Thymine) for another. Substitutions are expressed as the mean number of substitutions per base site. The problem investigated here is the occurrence probability variation of motifs (e.g. dinucleotides) under substitutions.

Two analytical solutions derived in Section 2 allow us to analyse in genes the occurrence probabilities of dinucleotides after substitutions (in the evolutionary sense: from the past to the present) and unexpectedly, also before substitutions (after back substitutions, in the inverse evolutionary sense: from the present to the past). Different properties and a generalization to motifs of any base length are also derived from these formulas. This theoretical part is the generalization on the alphabet {A, C, G, T} of the analytical solutions and of the properties recently derived on the alphabet {R, Y} (R = purine = A or G, Y = pyrimidine = C or T) (Arquès & Michel, 1993, 1994).

The application in Section 3 studies the inverse substitution process in four gene populations: the

§ Author to whom correspondence should be addressed.

†Present address: Equipe de Biologie Théorique Université de Marne-la-Vallée, Institut Gospard Monge, 2 rue de la butte verte, 93160 Noisy le Grand, France.

protein (coding) genes of (nuclear) eukaryotes, viruses and prokaryotes and the (eukaryotic) introns with the analytical solution giving the probabilities of the 16 dinucleotides AA, . . . , TT before substitutions. Genes before (after back) substitutions are called primitive genes. Four dinucleotides—CG, TA, GT and AC—occur with low probabilities in each of these four primitive gene populations, except for CG in the primitive prokaryotic protein genes. On the other hand, the dinucleotide AT has a low probability in the primitive eukaryotic protein genes.

In the Discussion, the properties of the two analytical solutions allow us to understand in particular why the substitution process associated with probabilities can be analysed in both evolutionary senses and also why the substitution process in the inverse evolutionary sense can only be studied by analytical solution (exactly), while the substitution process in the evolutionary sense can be studied both by analytical solution (exactly) and computer simulation (approximately). The biological functions which may have these five dinucleotides (CG, TA, GT, AC and AT) in primitive genes are described. As the probabilities of these five dinucleotides in the inverse evolutionary sense are lower than their actual probabilities, they can be rare enough in primitive genes in order to be related to biological signals. The type of biological signals associated with each dinucleotide is deduced from the conserved location and the function of each dinucleotide in the actual genes. Therefore, in primitive genes, CG would be a gene activity signal of eukaryotes; TA, a stop signal of protein genes; GT or its complementary dinucleotide AC, a start signal of introns; and AT, a start signal of eukaryotic protein genes.

2. Theory

2.1. RECALL

Let s be a DNA sequence of base length $l(s)$. This sequence s is subjected to (random) substitutions, i.e. random transformations of a base B into a base B', $B, B' \in \{A, C, G, T\}$ and $B \neq B'$, at random sites in s. Let x be the number of substitutions per base site (in short per base) in average, i.e. the total number of substitutions in s divided by the length $l(s)$ of s. (Note: in the following, "x substitutions" always stand for "x substitutions per base site in average").

The substitition number of a base in a given site of a sequence subjected to substitutions per base site of mean x, follows a Poisson law of parameter x (see e.g. the classical proofs in Feller, 1968: 447; Kimura, 1987: 69; Nei, 1987: 40).

We define $F(x)$ (respectively $G(x)$) as being the probability that a base B, $B \in \{A, C, G, T\}$, in the sequence s before the substitution process is B (respectively B', $B' \in \{A, C, G, T\}$ and $B' \neq B$) after x substitutions. Then, it can be proved that (similar to the formulas obtained with the one-parameter model of Jukes & Cantor, 1969).

$$F(x) = (1 + 3e^{-4x/3})/4$$

and

$$G(x) = (1 - F(x))/3 = (1 - e^{-4x/3})/4.$$

Proof

Let B, $B \in \{A, C, G, T\}$, be a base before the substitution process. At the substitution step k, $k = 0, \ldots$, the base B can generate β_k bases B and β'_k bases B', $B' \in \{A, C, G, T\}$ and $B' \neq B$. The numbers β_k and β'_k are first expressed in function of k. At the substitution step k, there is the obvious relation $\beta_k + \beta'_k = 3^k$ with the initial conditions for k, $k = 0$: $\beta_0 = 1$ and $\beta'_0 = 0$, and $k = 1$: $\beta_1 = 0$ and $\beta'_1 = 3$. At the substitution step $k+1$, $\beta_{k+1} = \beta'_k$ (substitution of B' into B) and $\beta'_{k+1} = 3\beta_k + 2\beta'_k$ (substitution of B into B' and substitution of B' into B'', $B'' \in \{A, C, G, T\}$ and $B'' \neq B' \neq B$). Note that $\beta_{k+1} + \beta'_{k+1} = 3^{k+1}$. Therefore, the two following recurrence relations are obtained:

$$\begin{cases} \beta_{k+1} = \beta'_k = 3\beta_{k-1} + 2\beta'_{k-1} = 3\beta_{k-1} + 2\beta_k \\ \beta'_{k+1} = 3\beta_k + 2\beta'_k = 3\beta'_{k-1} + 2\beta'_k \end{cases}$$

These two relations can be solved in function of k with the initial conditions for $k = 0$ and $k = 1$:

$$\begin{cases} \beta_k = \frac{1}{4}(3(-1)^k + 3^k) \\ \beta'_k = \frac{3}{4}(-1(-1)^k + 3^k) \end{cases}$$

Then, the probability to have a base B (respectively B') at the substitution step k is equal to $\beta_k/3^k$ (respectively $\beta'_k/3^k$).

Let x be the number of substitutions per base site in the sequence and let $F(x)$ be the probability that a base B in the sequence before the substitution process is B after x substitutions. Then

$$F(x) = \sum_{k \geqslant 0} e^{-x} \frac{x^k}{k!} \frac{\beta_k}{3^k}$$

(sum for all substitution steps k of the probability of a base site to be subjected to k substitutions (see also

Arquès & Michel, 1993) times the probability of having the base B at the substitution step k).

$$=\tfrac{1}{4}\,\mathrm{e}^{-x}\sum_{k\geqslant 0}\frac{x^k}{k!}\frac{(3(-1)^k+3^k)}{3^k}$$

$$=\tfrac{1}{4}\,\mathrm{e}^{-x}\sum_{k\geqslant 0}\frac{x^k}{k!}+\tfrac{3}{4}\,\mathrm{e}^{-x}\sum_{k\geqslant 0}\frac{\left(\dfrac{-1}{3}\,x\right)^k}{k!}$$

$$=\tfrac{1}{4}+\tfrac{3}{4}\,\mathrm{e}^{-4x/3}.$$

Let $G(x)$ be the probability that a base B in the sequence before the substitution process is B' after x substitutions. Then

$$G(x)=\tfrac{1}{3}\,(1-F(x))=\tfrac{1}{4}-\tfrac{1}{4}\,\mathrm{e}^{-4x/3}.$$

2.2. ANALYTICAL SOLUTION OF THE DINUCLEOTIDE PROBABILITY AFTER RANDOM SUBSTITUTIONS

Let the motif m be a dinucleotide, i.e. $m\in\{AA,\ldots,TT\}$. By convention, in the following the indexes i or $j\in[1, 16]$ represent the dinucleotides AA, . . . , TT in the alphabetical order. Let $Id(j, i)$ be the number of identical bases in the same dinucleotide site between the dinucleotide i and j, e.g. $Id(1, 2)=1$, 1 representing AA and 2, AC.

The dinucleotide probabilities $[P_i(x)]_{1\leqslant i\leqslant 16}$ after x substitutions (at time t) can be obtained from the dinucleotide probabilities $[P_j(0)]_{1\leqslant j\leqslant 16}$ before the substitution process (at time 0) (Scheme 1), τ being the unknown number of substitutions per base site on average between times 0 and the present.

Theorem 1

Let $[P_i(x)]_{1\leqslant i\leqslant 16}$ be the probabilities of the dinucleotide i, $i\in[1, 16]$, in a given sequence after x substitutions. Then

$$P_i(x)=\sum_{j=1}^{16}P_j(0)F(x)^{Id(j,i)}G(x)^{2-Id(j,i)} \qquad (1)$$

$$P_i(x+y)=\sum_{j=1}^{16}P_j(x)F(y)^{Id(j,i)}G(y)^{2-Id(j,i)} \qquad (2)$$

$$P_i(\tau)=\sum_{j=1}^{16}P_j(x)F(\tau-x)^{Id(j,i)}G(\tau-x)^{2-Id(j,i)} \qquad (3)$$

where $P_i(\tau)$ represents the actual dinucleotide probabilities.

Proof. (1) The formula deduced from the fact that the probability of the dinucleotide j, $j\in[1, 16]$, giving the dinucleotide i, $i\in[1, 16]$, after x substitutions, is

$$F(x)^{Id(j,i)}G(x)^{2-Id(j,i)}.$$

Let M_i^x be the event "a dinucleotide randomly chosen in a sequence is after x substitutions of type i". Then

$$P_i(x)=P(M_i^x)$$

$$=\sum_{j=1}^{16}P(M_j^0)\times P(M_i^x\,|\,M_j^0)$$

$$=\sum_{j=1}^{16}P_j(0)\times P(\text{dinucleotide } j\rightarrow\text{dinucleotide } i \text{ after } x \text{ substitutions per base site on average})$$

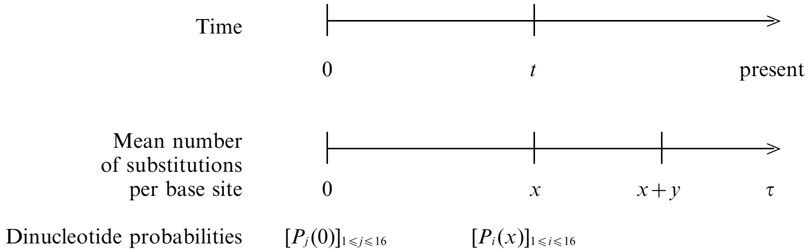
$$=\sum_{j=1}^{16}P_j(0)F(x)^{Id(j,i)}G(x)^{2-Id(j,i)}$$

(2)

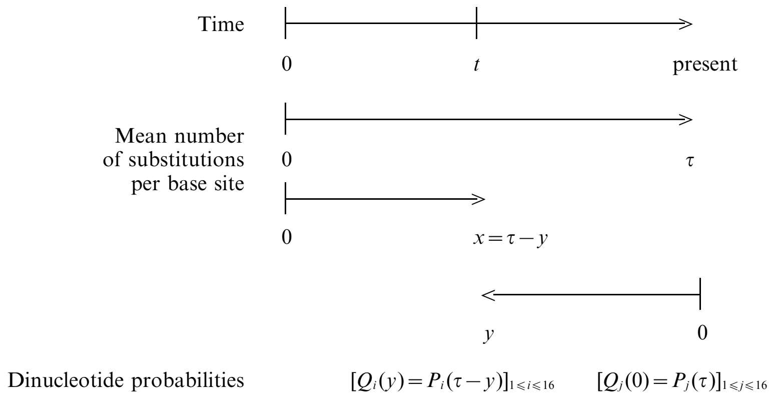
$$\sum_{k=1}^{16}P_k(x)F(y)^{Id(k,i)}G(y)^{2-Id(k,i)}$$

$$=\sum_{k=1}^{16}\left(\sum_{j=1}^{16}P_j(0)F(x)^{Id(j,k)}G(x)^{2-Id(j,k)}\right)$$

$$\times F(y)^{Id(k,i)}G(y)^{2-Id(k,i)} \quad \text{by (1)}$$


Time
0
t
present
Mean number of substitutions per base site
0
x
x+y
τ
Dinucleotide probabilities
$[P_j(0)]_{1\leqslant j\leqslant 16}$
$[P_i(x)]_{1\leqslant i\leqslant 16}$

SCHEME 1. Dinucleotide probability after random substitutions.


Time
0
t
present
Mean number of substitutions per base site
0
τ
0
$x=\tau-y$
y
0
Dinucleotide probabilities
$[Q_i(y)=P_i(\tau-y)]_{1\leqslant i\leqslant 16}$
$[Q_j(0)=P_j(\tau)]_{1\leqslant j\leqslant 16}$

SCHEME 2. Dinucleotide probability before random substitutions (after random back substitutions).

$$=\sum_{j=1}^{16}P_j(0)\left(\sum_{k=1}^{16}F(x)^{Id(j,\,k)}G(x)^{2-Id(j,\,k)}\right.$$

$$\left.\times F(y)^{Id(k,\,i)}G(y)^{2-Id(k,\,i)}\right)$$

$$=\sum_{j=1}^{16}P_j(0)F(x+y)^{Id(j,\,i)}G(x+y)^{2-Id(j,\,i)}$$

(consequence of exponential properties)

$$=P_i(x+y)$$

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

Remarks

The formula (1) is a particular case of formula (2) with $x=0$ and $y=x$.

Formula (1) converges as expected towards the random value $1/16=0.0625$ when the number of substitutions, x, increases, whatever the dinucleotide i and whatever the dinucleotide probabilities $P_j(0)$ (consequence of negative exponentials).

Formula (1) can be generalized to a motif of a base length λ:

$$P_i(x)=\sum_{j=1}^{4^\lambda}P_j(0)F(x)^{Id(j,\,i)}G(x)^{\lambda-Id(j,\,i)}$$

This generalization with $\lambda=3$ allows the codon probabilities after substitution to be studied.

2.3. ANALYTICAL SOLUTION OF THE DINUCLEOTIDE PROBABILITY BEFORE RANDOM SUBSTITUTIONS (AFTER RANDOM BACK SUBSTITUTIONS)

The problem here is the opposite of that in the previous section. Let τ (respectively x) be the number of substitutions per base site in average between times 0 and the present (respectively t) (Scheme 2). Let y be the number of substitutions per base site on average between times t and the present, i.e. $\tau=x+y$. In Section 2.2, the reference time is time 0 (before the substitution process) while in the inverse problem, the reference time is the present (after the substitution process).

In the following, "x (respectively y) substitutions" always stand for "x (respectively y) substitutions per base site in average".

Therefore, the inverse problem consists in expressing

$$[Q_i(y)=P_i(\tau-y)]_{1\leqslant i\leqslant 16}$$

in the function

$$[Q_j(0)=P_j(\tau)]_{1\leqslant j\leqslant 16}$$

and more generally

$$[Q_i(y+z)=P_i(\tau-y-z)]_{1\leqslant i\leqslant 16}$$

in the function

$$[Q_j(z)=P_j(\tau-z)]_{1\leqslant j\leqslant 16},$$

i and $j\in[1,16]$ representing the alphabetical order of dinucleotides.

Proposition 2

$$Q_i(y+z)=\sum_{j=1}^{16}Q_j(z)F(-y)^{Id(j,\,i)}G(-y)^{2-Id(j,\,i)} \qquad (4)$$

$$Q_i(y)=\sum_{j=1}^{16}P_j(\tau)F(-y)^{Id(j,\,i)}G(-y)^{2-Id(j,\,i)} \qquad (5)$$

with $F(-y)=(1+3\,\mathrm{e}^{4y/3})/4$ and $G(-y)=(1-\mathrm{e}^{4y/3})/4$

Proof. (4) The inverse matrix of

$$[F(y)^{Id(j,\,i)}G(y)^{2-Id(j,\,i)}]_{1\leqslant i,j\leqslant 16}$$

associated with formula (2) in Theorem 1 is

$$[F(-y)^{Id(j,\,i)}G(-y)^{2-Id(j,\,i)}]_{1\leqslant i,j\leqslant 16}\,.$$

Then, formula (2) implies:

$$P_i(x)=\sum_{j=1}^{16} P_j(x+y)F(-y)^{Id(j,\,i)}G(-y)^{2-Id(j,\,i)}.$$

Then,

$$\begin{aligned}Q_i(y+z)&=P_i(\tau-y-z)\\&=\sum_{j=1}^{16} P_j(\tau-z)F(-y)^{Id(j,\,i)}G(-y)^{2-Id(j,\,i)}\\&=\sum_{j=1}^{16} Q_j(z)F(-y)^{Id(j,\,i)}G(-y)^{2-Id(j,\,i)}\end{aligned}$$

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

Remarks

Formula (5) will be used in the application in Section 3 to determine the dinucleotide probabilities after back substitutions in the protein genes of eukaryotes, viruses and prokaryotes and in the eukaryotic introns, the actual dinucleotide probabilities $P_j(\tau)=Q_j(0)$ being computed from gene databases.

In contrast to formula (1), formula (5) does not converge when the number of substitutions, y, increases (see Section 2.4).

Formula (5) can be generalized to a motif of a base length λ in the same way as for formula (1):

$$Q_i(y)=\sum_{j=1}^{4^{\lambda}} P_j(\tau)F(-y)^{Id(j,\,i)}G(-y)^{\lambda-Id(j,\,i)}$$

This generalization with $\lambda=3$ allows the codon probabilities before substitutions to be studied.

2.4. BIOLOGICAL MEANING OF THE PREVIOUS FORMULAS

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

(see Scheme 2)

The formula $P_i(x)$ gives the evolution of the dinucleotide probabilities when we go from the past to the present and when the number of substitutions increases from 0 to τ (after substitutions). $P_i(x)$ can be obtained either exactly by analytical solution or approximately by computer simulation (simulation of random substitutions in simulated sequences).

The formula $Q_i(y)$ gives the inverse evolution of the dinucleotide probabilities when we go from the present to the past and when the number of substitutions decreases from τ to 0 (before substitutions or after back substitutions). In contrast to $P_i(x)$, $Q_i(y)$ does not converge when the number of back substitutions, y,

TABLE 1

	Protein genes of eukaryotes			Protein genes of viruses			Protein genes of prokaryotes			Introns of eukaryotes		
Dinucleotide	$Q_i(0)$ (%)	$Q_i(0.34)$ (%)	$Q_i(0)$–$Q_i(0.34)$	$Q_i(0)$ (%)	$Q_i(0.44)$ (%)	$Q_i(0)$–$Q_i(0.44)$	$Q_i(0)$ (%)	$Q_i(0.50)$ (%)	$Q_i(0)$–$Q_i(0.50)$	$Q_i(0)$ (%)	$Q_i(0.15)$ (%)	$Q_i(0)$–$Q_i(0.15)$
AA	7.80	9.23	−1.43	8.69	11.42	−2.72	8.13	12.75	−4.63	8.42	9.20	−0.78
AC	5.89	4.97	0.92	6.20	5.48	0.72	5.53	3.37	2.16	4.90	4.31	0.59
AG	7.33	8.40	−1.07	6.88	7.24	−0.37	5.53	2.43	3.10	7.02	7.44	−0.42
AT	5.88	5.38	0.49	7.04	7.71	−0.67	6.51	7.79	−1.29	6.78	6.64	0.14
CA	7.75	9.57	−1.82	7.31	9.06	−1.75	6.21	5.98	0.23	6.82	7.18	−0.36
CC	6.77	7.61	−0.84	5.86	6.50	−0.64	5.60	4.14	1.46	5.82	6.05	−0.23
CG	3.76	0.03	3.73	3.98	0.00	3.98	7.44	10.18	−2.74	1.80	0.02	1.78
CT	6.55	7.52	−0.97	5.75	5.66	0.09	5.36	3.93	1.43	7.26	7.72	−0.46
GA	7.61	9.10	−1.49	7.28	8.54	−1.27	6.84	7.43	−0.59	5.93	5.81	0.12
GC	6.32	6.35	−0.04	5.48	4.86	0.62	8.03	12.42	−4.39	4.86	4.58	0.28
GG	6.68	7.11	−0.43	6.31	7.10	−0.80	6.70	6.44	0.26	6.07	6.34	−0.27
GT	4.83	3.12	1.71	4.99	2.80	2.19	5.07	1.91	3.16	5.48	5.02	0.46
TA	3.72	0.05	3.67	5.52	2.80	2.72	4.46	0.07	4.39	5.95	5.40	0.55
TC	5.86	5.80	0.05	5.36	4.39	0.97	5.47	4.34	1.12	6.13	6.03	0.10
TG	7.68	10.17	−2.49	6.90	8.97	−2.07	6.99	9.18	−2.19	7.44	7.95	−0.51
TT	5.57	5.60	0.00	6.45	7.46	−1.00	6.13	7.63	−1.48	9.32	10.32	−0.99

Probability $Q_i(0)$ (%) of the 16 dinucleotides computed in the protein genes of eukaryotes (16371 *genes*, 23620 kb), viruses (5724 genes, 9098 kb), prokaryotes (7271 genes, 8882 kb) and in the eukaryotic introns (3196 genes, 3846 kb). Probability $Q_i(y_{\max})$ (%) of the 16 dinucleotides at the maximal number $y_{\max}$ of substitutions in the protein genes of eukaryotes ($y_{\max}=0.34$), viruses ($y_{\max}=0.44$), and prokaryotes ($y_{\max}=0.50$), and in the eukaryotic introns ($y_{\max}=0.15$). Probability difference $Q_i(0)$–$Q_i(y_{\max})$ identifying the dinucleotides having probabilities after back substitutions lower than the actual ones.

increases. However, the vector $[Q_i(y)]_{1\leqslant i\leqslant 16}$ must remain a probability vector, i.e. the 16 values $Q_i(y)$ must be bounded by 0 and 1 (and of sum 1). Therefore, the condition $0\leqslant Q_i(y)\leqslant 1$ for i in [1, 16] implies a maximum number, $y_{\max}$, of back substitutions. Another difference with $P_i(x)$ lies in the fact that $Q_i(y)$ can only be obtained by analytical solution and not by computer simulation. 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

3.1. PRESENTATION OF THE PROBLEM

The inverse substitution process is studied in four gene populations obtained from release 34 of the EMBL Nucleotide Sequence Data Library as described in previous studies (see e.g. Arquès & Michel, 1987, 1990*a*, 1990*b*, for a description of data acquisitions): the eukaryotic (nuclear) protein (coding) genes (16371 genes, 23620 kb), the viral protein (coding) genes (5724 genes, 9098 kb), the prokaryotic protein (coding) genes (7271 genes, 8882 kb) and the (eukaryotic) introns (3196 genes, 3846 kb).

(Note: Due to the law of large numbers (see Arquès & Michel, 1990*b*: p. 752, section 2.3.3 for the detail), frequencies computed with populations made of several thousands of genes are stable from a statistical point of view.)

The actual dinucleotide probabilities $P_j(\tau)=Q_j(0)$ computed in these four gene populations are given in Table 1. Table 1 extends previous dinucleotide tabulations of organisms (Nussinov, 1981, 1984; McClelland & Ivarie, 1982; Smith *et al*., 1983; Burge *et al*., 1992) to large gene populations. Based on these actual dinucleotide probabilities (see Scheme 2 and Scheme 3 in the Discussion), the formula $Q_i(y)$ allows the dinucleotide probabilities after back substitutions to be determined (i.e. in primitive genes). In order to compute the formula $Q_i(y)$, it is easier to rewrite it as follows:

$$Q_i(y)=\frac{1}{16}\sum_{k=0}^{2}\left(\sum_{j/Id(j,i)=k} P_j(\tau)\right)$$
$$\times(1+3\,e^{4y/3})^k(1-e^{4y/3})^{2-k}.$$

3.2. RESULTS

The analytical solution $Q_i(y)$ applied to the protein gene populations leads to a maximum number, $y_{\max}$, of substitutions equal to 0.34 for the eukaryotes (Fig. 1), 0.44 for the viruses (Fig. 2) and 0.50 for the prokaryotes (Fig. 3). In introns, $y_{\max}$ is lower and equal to 0.15 (Fig. 4). This maximum number of substitutions is related to the dinucleotide CG in the

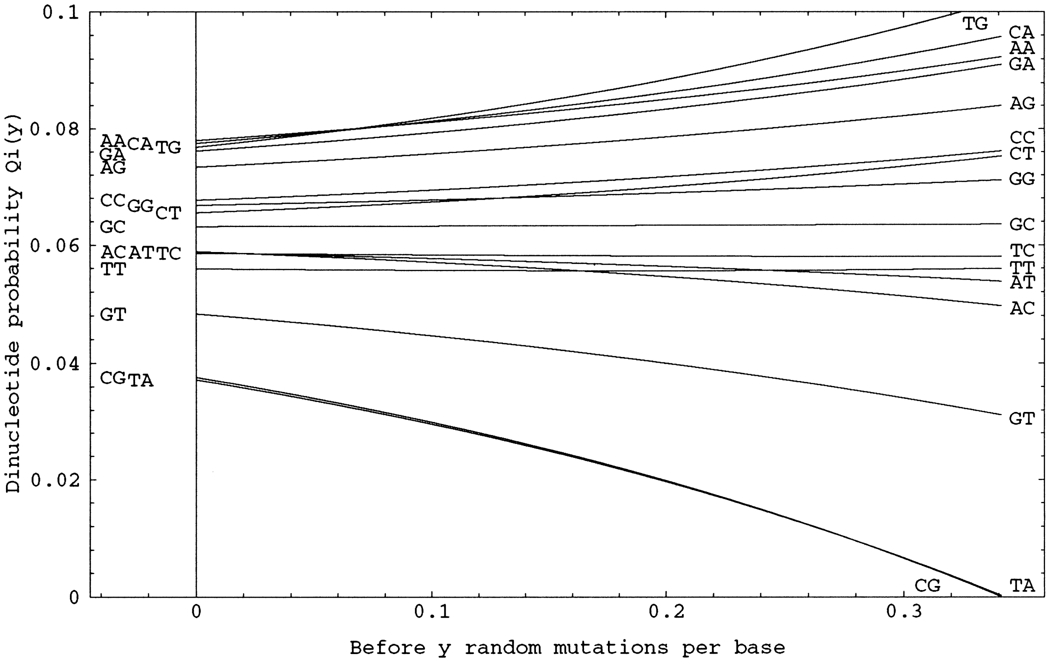


FIG. 1. Occurrence probability of the 16 dinucleotides in the inverse substitution process for the protein genes of eukaryotes. The horizontal axis represents y, the number of back substitutions per base site, $y\in[0, y_{\max}=0.34]$. The vertical axis represents the probabilities of the 16 dinucleotides. The actual dinucleotide probability at $y=0$ is equal to the dinucleotide frequency given in Table 1.

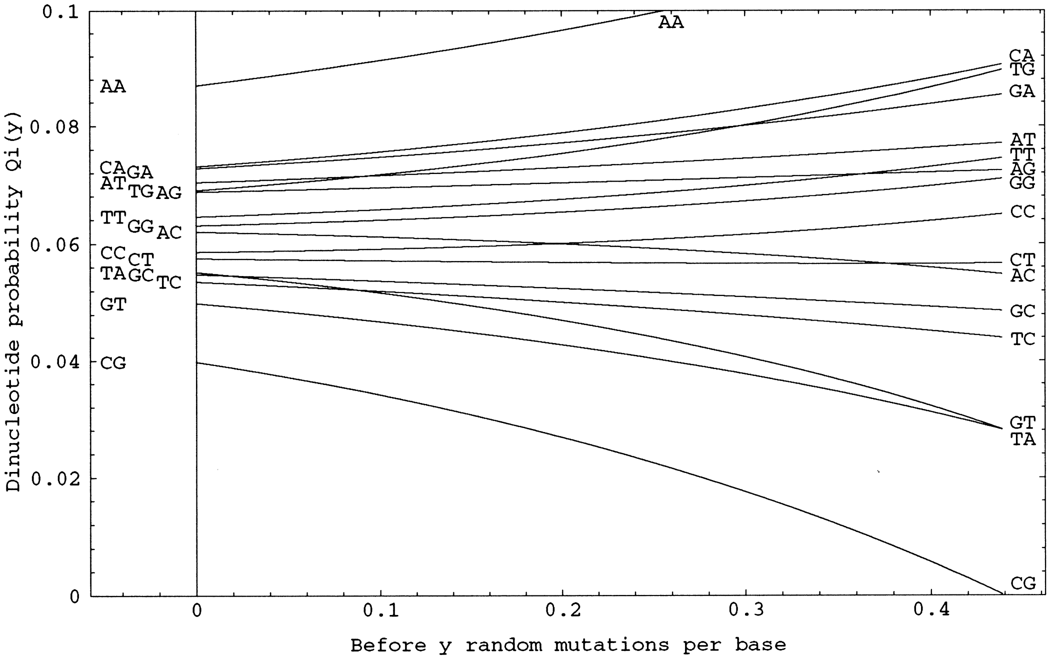


FIG. 2. Occurrence probability of the 16 dinucleotides in the inverse substitution process for the protein genes of viruses. The horizontal axis represents y, the number of back substitutions per base site, $y \in [0, y_{max} = 0.44]$. The vertical axis represents the probabilities of the 16 dinucleotides. The actual dinucleotide probability at $y = 0$ is equal to the dinucleotide frequency given in Table 1.

protein genes of eukaryotes and viruses and in introns, and to the dinucleotide TA in the prokaryotic protein genes. Note that CG and TA have a similar probability curve in the eukaryotic protein genes (Fig. 1).

In the eukaryotic protein genes (Fig. 1), the five dinucleotides with the lowest probabilities at $y_{max} = 0.34$ are CG, TA, GT, AC and AT (ranged in increasing probabilities). These five dinucleotides have

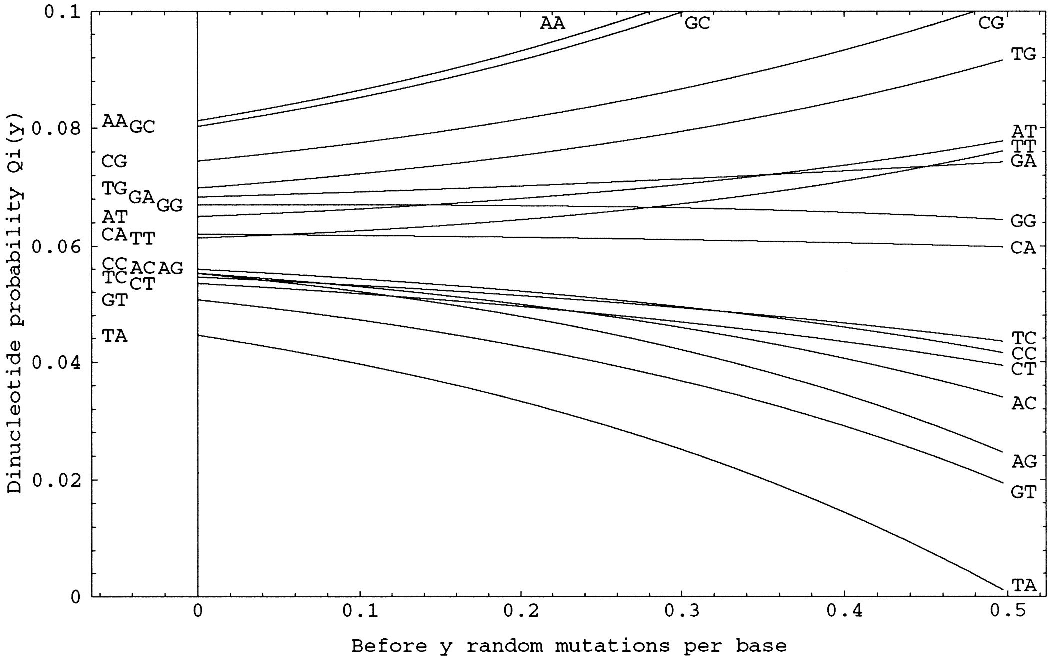


FIG. 3. Occurrence probability of the 16 dinucleotides in the inverse substitution process for the protein genes of prokaryotes. The horizontal axis represents y, the number of back substitutions per base site, $y \in [0, y_{max} = 0.50]$. The vertical axis represents the probabilities of the 16 dinucleotides. The actual dinucleotide probability at $y = 0$ is equal to the dinucleotide frequency given in Table 1.

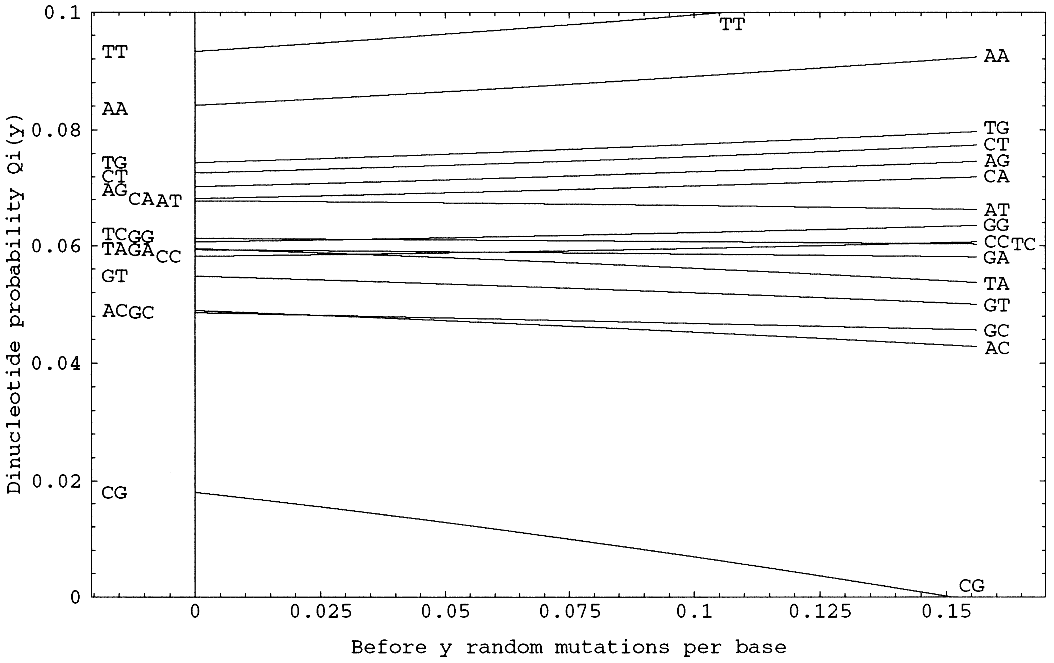


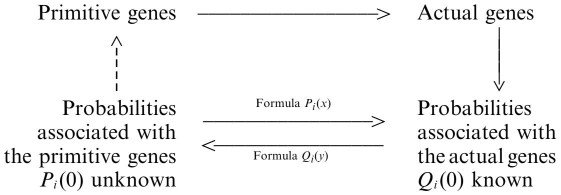
FIG. 4. Occurrence probability of the 16 dinucleotides in the inverse substitution process for the eukaryotic introns. The horizontal axis represents y, the number of back substitutions per base site, $y \in [0, y_{\max} = 0.15]$. The vertical axis represents the probabilities of the 16 dinucleotides. The actual dinucleotide probability at $y = 0$ is equal to the dinucleotide frequency given in Table 1.

probabilities $Q_i(0.34)$ at the maximum number of substitutions lower than their actual probabilities $Q_i(0)$ and are respectively associated with the five highest probability differences $Q_i(0) - Q_i(0.34)$ (Table 1).

In the viral protein genes (Fig. 2 and Table 1), the six dinucleotides with the lowest probabilities at $y_{\max} = 0.44$ and with the six highest probability differences $Q_i(0) - Q_i(0.44)$ are CG, TA, GT, TC, GC and AC. Note that AC decreases faster than GC (Table 1).

In the prokaryotic protein genes (Fig. 3 and Table 1), the four dinucleotides with the lowest probabilities at $y_{\max} = 0.50$ and corresponding respectively to the four highest probability differences $Q_i(0) - Q_i(0.50)$ are TA, GT, AG and AC.

In introns (Fig. 4 and Table 1), the five dinucleotides with the lowest probabilities at $y_{\max} = 0.15$ and with the five highest probability differences $Q_i(0) - Q_i(0.15)$ are CG, AC, GC, GT and TA. Note that TA decreases faster than GT decreasing faster than GC (Table 1). Therefore, the dinucleotides occurring at least three times in these four sets are CG, TA, GT and AC. CG (respectively AT) occurs with a high probability in the protein genes of prokaryotes (respectively, of viruses and prokaryotes) (Figs 2 and 3).


Random substitutions
Primitive genes
Actual genes
Probabilities associated with the primitive genes $P_i(0)$ unknown
Formula $P_i(x)$
Formula $Q_i(y)$
Probabilities associated with the actual genes $Q_i(0)$ known

SCHEME 3. Study of the substitution process in both evolutionary senses with probabilities associated with the genes.

4. Discussion

Two analytical solutions derived here allow us to determine the dinucleotide probabilities after substitutions (in the evolutionary sense: from the past to the present) as well as before substitutions (after back substitutions, in the inverse evolutionary sense: from the present to the past). Different properties and a generalization to motifs of any base length are also derived from these formulas. In particular, the generalization to a motif of base length 3 allows the codon probabilities under substitutions in both evolutionary senses to be analysed. Generalization to d-motifs, i.e. two motifs separated by any d bases such as in the autocorrelation function, is also obvious.

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 (unidirectional arrow between primitive and actual genes in Scheme 3). Unexpectedly, some statistical measures of the substitution process can be inverted, allowing one to go backward in time. If the substitution process is studied, not with the nucleotide ordering, but with probabilities, then it can be inverted. Indeed, the formula $P_i(x)$ giving the dinucleotide probabilities after x substitutions can be inverted and the inverse formula $Q_i(y)$ gives the dinucleotide probabilities before y substitutions (bidirectional arrow in the Scheme 3). The formula $P_i(x)$ can be obtained approximately by computer simulation (simulation of random substitutions in simulated sequences), but not the formula $Q_i(y)$. Therefore, the dinucleotide probabilities in primitive genes can only be determined by using the analytical solution $Q_i(y)$, the dinucleotide probabilities $Q_i(0)$ of actual genes being obtained from gene databases.

The probability curve $Q_i(y)$ cannot be intuitively predicted as the formula $Q_i(y)$ is the sum of several exponential products. For example, the probability curve $Q_i(y)$ of TA in the viral protein genes (Fig. 2), which is the fifth lowest one in actual genes (see also $Q_i(0)$ in Table 1), decreases with a slope crossing the probability curves of GC, TC and GT after back substitutions. Otherwise, the model relies on two basic assumptions: substitutions occur with the same probabilities at all sites and all substitutions are equally probable. These two assumptions can be considered as being verified at the gene population level (an average of several thousands of genes), explaining that the dinucleotide probabilities $Q_i(0)$ need to be computed in gene populations. The maximum number of substitutions with the formula $Q_i(y)$ is related to the two previous assumptions and is the limit of the model developed here to go backward in time.

This model with a unique mutation rate can obviously be extended by considering different mutation rates, for example a rate of transitions and a rate of transversions similarly to the two-parameter model (Kimura, 1980) or different rates according to the motif sites, etc. Finally, the identification of statistical properties in primitive genes, e.g. a few dinucleotides with low probabilities, implies some rules in the nucleotide ordering of these primitive genes (dashed arrow in the Scheme 3).

The application shows that the dinucleotide occurrence is not equiprobable in the four primitive gene populations studied: protein genes of eukaryotes, viruses and prokaryotes and introns after back substitutions. Indeed, four of the 16 dinucleotides—CG, TA, GT and AC—occur with low probabilities in each of these primitive populations, except for CG in the primitive prokaryotic protein genes. AT also has a low probability in the primitive eukaryotic protein genes. The probabilities of these five dinucleotides in the primitive genes are lower than their actual ones so that they may be related to biological signals—for example, one occurrence (ideally) of a signal per primitive gene. These signals are conserved in actual genes by involving additional bases in these dinucleotides (see below).

The low occurrence probability of CG in eukaryotes (Swartz *et al*., 1962; Russel *et al*., 1976) may be explained by the methylation of C (5 mC) in the dinucleotides CG with the MTase enzyme and by the spontaneous deamination of 5 mC to give T (Salser, 1977; Coulondre *et al*., 1978; Bird, 1980). DNA methylation in eukaryotes is involved in gene regulation (methylation leads to gene repression, non-methylation, to gene activation), genomic imprinting and developmental regulation (Bird, 1986; Li *et al*., 1993; reviews in Cedar, 1988; Razin & Cedar, 1991; Bird, 1992; Tate & Bird, 1993). Therefore, the low occurrence probability of CG in primitive eukaryotes (protein genes and introns) may support the idea that the dinucleotide CG would be the gene activity signal of eukaryotes. In agreement, the two methylation mutation dinucleotides, TG and CA (in the other strand), occur with a high probability in primitive eukaryotes (Figs 1 and 4). CG is paired with the same dinucleotide in the complementary strand. The same function for the dinucleotide and its complementary dinucleotide agrees with the methylation of the CG site in both strands of the DNA (Bird, 1978). DNA methylation also exists in prokaryotes. However, the methylation sites in prokaryotes are not dinucleotides. Indeed, cytosines (5 mC) and also adenines (6 mA) can be methylated in three sites: CC(A/T)GG (methylation of the second C by the DcM enzyme), AAA(N)$_6$GTGC (methylation of the second A by the HsdM enzyme) and GATC (methylation of A by the Dam enzyme) (review in Marinus, 1987). In contrast to eukaryotes, DNA methylation in prokaryotes is used by restriction-modification systems to distinghish between host and foreign DNA (review in Wilson & Murray, 1991) and by the mismatch repair system to correct errors in DNA replication (review in Modrich, 1991). The strong difference between eukaryotes and prokaryotes with DNA methylation may explain the absence of low occurrence probability of CG in the primitive prokaryotic protein genes.

The dinucleotide TA is the first dinucleotide of the stop codons TAA and TAG. TAA and TAG are the most important stop codons as the third stop codon TGA can have three other functions which differ from a termination signal. Indeed, TGA can be a tryptophan codon (mitochondria, *Mycoplasma*, *B. subtilis* and *E. coli*), a cysteine codon (*Euplotes octacarinatus*) or a selenocysteine codon (vertebrates and prokaryotes) (reviews in Osawa *et al.*, 1992; Hatfield & Diamond, 1993). TGA is unique in that it can have four functions. Otherwise, TGA can also be a frameshift site (in the release factor-2 gene of *E. coli*: Craigen & Caskey, 1986). Therefore, the low occurrence probability of TA in primitive protein genes (eukaryotes, viruses and prokaryotes) may indicate that the dinucleotide TA would be the stop signal of primitive protein genes. The low occurrence probability of TA in primitive introns is consistent with the proposal that TA is a biological signal of primitive proteins genes. TA is paired with the same dinucleotide in the complementary strand.

The dinucleotide GT is the first dinucleotide in the 5′ splice site of introns (Mount, 1982; Mount *et al.*, 1992). GT pairs with the small nuclear RNA (snRNA) U1 in the intron removal process by the spliceosome (Black *et al.*, 1985). The low occurrence probability of GT in primitive introns may support the idea that the dinucleotide GT would be the start signal of primitive introns. The low occurrence probability of GT in primitive protein genes is consistent with the proposal that GT is a biological signal of primitive introns. Otherwise, the dinucleotide AG which is the last dinucleotide in the 3′ splice site of introns (Mount, 1982; Mount *et al.*, 1992), occurs with a high probability in primitive introns (Fig. 4). In contrast to GT, AG could not have been a biological signal (a stop signal) of primitive introns. In agreement with this hypothesis, AG also occurs with a high probability in primitive protein genes of eukaryotes and viruses (Figs 1 and 2). This result shows an asymmetry of the GT–AG rule and identifies GT as the main intron signal.

Experimental data support this asymmetry. First, no snRNA could have been so far identified to pair with AG. Second, there are two conserved motifs located a few nucleotides upstream AG which may compensate the low specificity of AG: a branch point adenine residue CT$RAY$ (R = A or G, Y = C or T) (Nelson & Green, 1989; Senapathy *et al.*, 1990; Mount *et al.*, 1992) and a polypyrimidine tract (>10) (Mount, 1982) pairing with the snRNA U2 (Black *et al.*, 1985) and the snRNA U2AF (Auxiliary Factor) (Ruskin *et al.*, 1988) respectively. Third, AG is also the last dinucleotide of exons (Mount, 1982; Mount *et al.*, 1992). The asymmetry of the GT–AG rule proposed allows the primitive exons to be distinguished from the primitive introns. This proposition differs from the hypothesis of a "proto-splice site" ancestor leading to a common start signal (GT) and a common stop signal (AG) for the primitive exons and introns (Stephens & Schneider, 1992).

As AC is the complementary dinucleotide of GT and as the probability variation of AC in the inverse evolutionary sense is similar to the one of GT for each of the four gene populations, the dinucleotides GT and AC can be related to the same biological signal, i.e. the start signal of primitive introns.

AT is the first dinucleotide of the initiator codon ATG. The low occurrence probability of AT in primitive eukaryotic protein genes may imply that it would be the start signal of primitive eukaryotic protein genes. AT should have a low probability for similar reasons as those given for GT. AT does not belong to the dinucleotides with the lowest probabilities $Q_i(0.15)$ (Fig. 4); however, its probability $Q_i(0.15)$ is lower than its actual one, $Q_i(0)$ (Table 1). The development of models taking into account different mutation rates in order to increase the maximum number possible of substitutions (the lowest in introns among the four gene populations) may solve this question. In contrast to eukaryotes, AT occurs with a high probability in primitive protein genes of prokaryotes (and viruses) (Figs 2 and 3). The difference between the protein genes of eukaryotes and those of prokaryotes with AT may be explained by the difference in protein synthesis initiation. In eukaryotes, the small ribosomal subunit scans along the mRNA until it reaches the first ATG (Kozak, 1978, 1989; Cigan *et al.*, 1988; review in Merrick, 1992). Therefore, the function of the first dinucleotide AT is important in eukaryotes. In prokaryotes, this dinucleotide has a less important function as the small ribosomal subunit first binds to the Shine–Dalgarno sequence—i.e. a conserved motif of seven nucleotides AGGAGGT located a few nucleotides upstream ATG (Shine & Dalgarno, 1974; Jacob *et al.*, 1987; review in Gualerzi & Pon, 1990).

The five rare dinucleotides, CG, TA, GT, AC and AT, which could be biological signals in primitive genes, become, after substitution, common in actual genes. As the substitution process increases (respectively, decreases) the occurrence probability of the rare (respectively, frequent) dinucleotides toward the random situation (probability equal to 1/16), the biological signals associated with the rare dinucleotides are conserved in actual genes by involving additional bases in these dinucleotides, i.e. by increasing the length (the complexity) of the signals. Nevertheless, these dinucleotides keep the most important functions in the actual conserved motifs associated with the biological signals. For example, in actual genes

(i) The gene activity site is related to several non-methylated dinucleotides, CG, (islands) in the eukaryotic promoter regions (Bird, 1986);

(ii) the start site of eukaryotic coding genes is the conserved codon $\underline{\text{AT}}$G (in 95% of eukaryotic genes: Kozac, 1987; Cigan & Donahue, 1987) in the less well conserved motif RNN|$\underline{\text{AT}}$GG (| is the start site, R = A or G, N = A, C, G or T) (Kozak, 1987, 1989; Cavener & Ray, 1991);

(iii) the stop site of coding genes is related to the conserved codons $\underline{\text{TA}}$A and $\underline{\text{TA}}$G (TGA) in the less well conserved tetranucleotide beginning with a stop codon and ending with T (Brown *et al*., 1990*a*, *b*); and

(iv) the start site of introns is a conserved motif of nine base length MAG|$\underline{\text{GT}}$$R$AGT (| is the splice site, M = A or C) (Senapathy *et al*., 1990; Mount *et al*., 1992) entirely involved in the 5′ cleavage of introns (Aebi *et al*., 1987).

In summary, five dinucleotides could have been biological signals in primitive genes: CG, a gene activity signal of eukaryotes; TA, a stop signal of protein genes; GT or its complementary dinucleotide AC, a start signal of introns; and AT, a start signal of eukaryotic protein genes. In other words, in primitive eukaryotes, there would be three signals in protein genes: AT (a start signal), TA (a stop signal) and CG (a gene activity signal) and one signal in introns: GT or AC (a start signal). In primitive prokaryotes and viruses, it would be one signal in protein genes: TA (a stop signal).

The analytical solutions $P_i(x)$ and $Q_i(y)$ and their generalization provide a new approach for studying gene mutation as they allow the occurrence probability variation of motifs (dinucleotides, trinucleotides, etc) to be analysed after substitutions (in the evolutionary sense) as well as before substitutions (in the inverse evolutionary sense). They are simple enough to be used directly. Nevertheless, we are currently implementing these analytical solutions in the software AGE (Analysis of Gene Evolution) (Arquès *et al*., 1992).

We thank Dr Nouchine Soltanifar and an anonymous referee for their advice. This work was supported by GIP GREG grant (Groupement d'Intérêt Public, Groupement de Recherches et d'Etudes sur les Génomes) and by INSERM grant (Contrat de Recherche Externe No 930101).

REFERENCES

AEBI, M., HORNIG, H. & WEISSMANN, C. (1987). 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU. *Cell* **50,** 237–246.

ARQUÈS, D. G. & MICHEL, C. J. (1987). A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. *J. theor. Biol.* **128,** 457–461.

ARQUÈS, D. G. & MICHEL, C. J. (1990*a*). Periodicities in coding and noncoding regions of the genes. *J. theor. Biol.* **143,** 307–318.

ARQUÈS, D. G. & MICHEL, C. J. (1990*b*). 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. & MICHEL, C. J. (1993). Analytical expression of the purine/pyrimidine codon probability after and before random mutations. *Bull. math. Biol.* **55,** 1025–1038.

ARQUÈS, D. G. & MICHEL, C. J. (1994). Analytical expression of the purine/pyrimidine autocorrelation function after and before random mutations. *Math. Biosci.*, **123,** 103–125.

ARQUÈS, D. G., MICHEL, C. J. & ORIEUX K. (1992). Analysis of Gene Evolution: the software AGE. *Comp. appl. Biosc.* **8,** 5–14.

BIRD, A. P. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J. molec. Biol.* **118,** 49–60.

BIRD, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8,** 1499–1504.

BIRD, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321,** 209–213.

BIRD, A. (1992). The essentials of DNA methylation. *Cell* **70,** 5–8.

BLACK, D. L., CHABOT, B. & STEITZ, J. A. (1985). U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. *Cell* **42,** 737–750.

BROWN, C. M., STOCKWELL, P. A., TROTMAN, C. N. A. & TATE, W. P. (1990*a*). The signal for the termination of protein synthesis in procaryotes. *Nucleic Acids Res.* **18,** 2079–2086.

BROWN, C. M., STOCKWELL, P. A., TROTMAN, C. N. A. & TATE, W. P. (1990*b*). Sequence analysis suggests that tetra-nucleotides signal termination of protein synthesis in eucaryotes. *Nucleic Acids Res* **18,** 6339–6345.

BURGE, C., CAMPBELL, A. M. & KARLIN, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. natn. Acad. Sci. U.S.A.* **89,** 1358–1362.

CAVENER, D. R. & RAY, S. C. (1991). Eukaryotic start and stop translation sites. *Nucleic Acids Res.* **19,** 3185–3192.

CEDAR, H. (1988). DNA methylation and gene activity. *Cell* **53,** 3–4.

CIGAN, A. M. & DONAHUE, T. F. (1987). Sequence and structural features associated with translational initiator regions in yeast—a review. *Gene* **59,** 1–18.

CIGAN, A.M., FENG, L. & DONAHUE, T. F. (1988). $\text{tRNA}_{\text{i}}^{\text{met}}$ functions in directing the scanning ribosome to the start site of translation. *Science* **242,** 93–97.

COULONDRE, C., MILLER, J. H., FARABOUGH, P. J. & GILBERT, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274,** 775–780.

CRAIGEN, W. J. & CASKEY, C. T. (1986). Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **322,** 273–275.

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

GUALERZI, C. & PON, C. L. (1990). Initiation of mRNA translation in prokaryotes. *Biochemistry* **29,** 5881–5889.

HATFIELD, D. & DIAMOND, A. (1993). UGA: a split personality in the universal genetic code. *Trends Genet.* **9,** 69–70.

JACOB, W. F., SANTER, M. & DAHLBERG, A. E. (1987). A single base change in the Shine–Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc. natn. Acad. Sci. U.S.A.* **84,** 4757–4761.

JUKES, T. H. & CANTOR, C. R. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism* (H.N. Munro, ed.) New York: Academic Press. pp. 21–132.

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.

KOZAK, M. (1978). How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* **15,** 1109–1123.

KOZAK, M. (1987). An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15,** 8125–8148.

KOZAK, M. (1989). The scanning model for translation: an update. *J. Cell Biol.* **108,** 229–241.

Li, E., Beard, C. & Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature* **366,** 362–365.

Marinus, M. G. (1987). DNA methylation in *Escherichia coli*. *A. Rev. Genet.* **21,** 113–131.

McClelland, M. & Ivarie, R. (1982). Asymmetrical distribution of CpG in an 'average' mamallian gene. *Nucleic Acids Res.* **10,** 7865–7877.

Merrick, W. C. (1992). Mechanism and regulation of eukaryotic protein synthesis. *Microbiol. Rev.* **56,** 291–315.

Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. *A. Rev. Genet.* **25,** 229–253.

Mount, S. M. (1982). A catalogue of splice junction sequences. *Nucleic Acids Res.* **10,** 459–472.

Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* **20,** 4255–4262.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nelson, K. K. & Green, M. R. (1989). Mammalian U2 snRNP has a sequence-specific RNA-binding activity. *Genes Dev.* **3,** 1562–1571.

Nussinov, R. (1981). Nearest neighbor nucleotide patterns: structural and biological implications. *J. biol. Chem.* **256,** 8458–8462.

Nussinov, R. (1984). Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res.* **12,** 1749–1763.

Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56,** 229–264.

Razin, A. & Cedar, H. (1991). DNA methylation and gene expression. *Microbiol. Rev.* **55,** 451–458.

Ruskin, B., Zamore, P. D. & Green, M. R. (1988). A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell* **52,** 207–219.

Russel, G. J., Walker, P. M. B., Elton, R. A. & Subak-Sharpe, J. H. (1976). Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J. molec. Biol.* **108,** 1–23.

Salser, W. (1977). Globin mRNA sequences: analysis of base pairing and evolutionary implications. *Cold Spring Harbor Symp. Quant. Biol.* **40,** 985–1002.

Senapathy, P., Shapiro, M. B. & Harris, N. L. (1990). Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol.* **183,** 252–278.

Shine, J. & Dalgarno, L. (1974). The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. natn. Acad. Sci. U.S.A.* **71,** 1342–1346.

Smith, T. F., Watermann, M. S. & Sadler, J. R. (1983). Statistical characterization of nucleic acid sequence functional domains. *Nucleic Acids Res.* **11,** 2205–2220.

Stephens, R. M. & Schneider, T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. molec. Biol.* **228,** 1124–1136.

Swartz, M. N., Trautner, T. A. and Kornberg, A. (1962). Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *J. biol. Chem.* **237,** 1961–1967.

Tate, P. H. & Bird, A. P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr. Opinion Genet. Dev.* **3,** 226–231.

Wilson, G. G. & Murray, N. E. (1991). Restriction and modification systems. *A. Rev. Genet.* **25,** 585–627.