J. theor. Biol. (1995) **172,** 279–291

Identification and Simulation of Shifted Periodicities Common to Protein Coding Genes of Eukaryotes, Prokaryotes and Viruses

DIDIER G. ARQUÈS,† JEAN-CHRISTOPHE LAPAYRE† 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* 6 *May* 1994, *Accepted on* 27 *June* 1994)

The distribution of nucleotides in protein coding genes is studied with autocorrelation functions. The autocorrelation function $YRY(N)_iYRY$, analysing the occurrence probability of the i-motif $YRY(N)_iYRY$ (two motifs YRY separated by any i bases N, R = purine = Adenine or Guanine, Y = pyrimidine = Cytosine or Thymine, $N = R$ or Y) in the protein coding genes of eukaryotes, prokaryotes and viruses, reveals the classical periodicity 0 modulo 3 associated with the normal frame 0 (maximal values of the function at $i = 0, 3, 6$, etc). The specification of $YRY(N)_iYRY$ on the alphabet {A, C, G, T} leads to 64 i-motifs: CAC$(N)_i$CAC, CAC$(N)_i$CAT, . . . , TGT$(N)_i$TGT. The 64 autocorrelation functions associated with these 64 i-motifs in protein coding genes have all the periodicity modulo 3, but, surprisingly, not always the expected periodicity 0 modulo 3. Two new types of periodicities are identified: a periodicity 1 modulo 3 associated with the shifted frame +1 (maximal values of the function at $i = 1, 4, 7$, etc) and a periodicity 2 modulo 3 associated with the shifted frame −1 (maximal values of the function at $i = 2, 5, 8$ etc). Furthermore, the classification of i-motifs according to the type of periodicity demonstrates a strong coherence relation between the 64 i-motifs, which is, in addition, common to the three gene populations, as the same i-motifs in the three gene populations have the same periodicities.

The three periodicities 0, 1 and 2 modulo 3 can be simulated by an evolutionary model at two successive processes. The simulated genes are generated by a process of gene construction, with a stochastic automaton followed by a process of gene evolution with random insertions and deletions of trinucleotides simulating RNA editing. For almost all i-motifs, the autocorrelation functions in these simulated genes are strongly correlated with those in protein coding genes, for both the type and the probability level of periodicities.

This paper describes the process of ribosomal frameshifting leading to the shifted periodicities, which may reveal overlapping genes or concatenated genes from different frames. It also presents the evolutionary aspects of the shifted periodicities. The shifted periodicities cannot be associated with the RNY model (Eigen & Schuster, 1978, *Naturwissenschaften* **65,** 341–369) or the RRY model (Crick *et al*., 1976, Origins of Life **7,** 389–397), but are compatible with the oligonucleotide mixing model (Arquès & Michel, 1990, *Bull. math. Biol.* **52,** 741–772). Finally, a variant of the primitive translation model of Crick *et al*. (1976) is proposed to explain the shifted periodicities.

1. Introduction

The distribution of nucleotides in genes is not random. Indeed, several non-random statistical properties were identified with the autocorrelation function $w(N)_iw'$ (defined in Arquès & Michel, 1987 and below in Section 2.1) analysing, in gene populations, the occurrence probability of the i-motif $w(N)_iw'$: for example, two trinucleotides w and w' separated by any i bases N. On the purine/pyrimidine alphabet

§ Author to whom correspondence should be addressed.

0022–5193/95/030279 + 13 $08.00/0

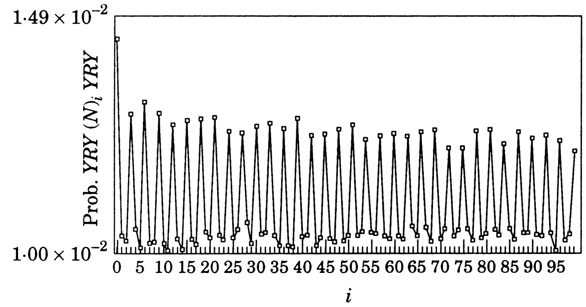


FIG. 1. Periodicity 0 modulo 3 identified with the autocorrelation function $YRY(N)_iYRY$ in the protein coding genes of eukaryotes.

(R = purine = Adenine or Guanine, Y = pyrimidine = Cytosine or Thymine) and with the i-motif $YRY(N)_iYRY$ ($w = w' = YRY$ and $N = R$ or Y), the autocorrelation function $YRY(N)_iYRY$ identifies, in genes, periodicities (modulo 2 and 3), sub-periodicities (modulo 6), maximum, local maxima, etc (Arquès & Michel, 1987, 1990*a*, 1990*b*). For example, in the protein coding genes of eukaryotes, prokaryotes and viruses, the autocorrelation function $YRY(N)_iYRY$ shows the classical periodicity 0 modulo 3 associated with the normal frame 0 (maximal values of the function at $i = 0$, 3, 6, etc; see Fig. 1).

The i-motifs obtained by specifying $YRY(N)_iYRY$ on the alphabet {A, C, G, T} are studied here. With R = A or G and Y = C or T, there are 64 i-motifs: CAC(N)$_i$CAC, CAC(N)$_i$CAT, CAC(N)$_i$TAC, . . . , TGT(N)$_i$TGT. The 64 autocorrelation functions associated with these 64 i-motifs in protein coding genes should also have the periodicity 0 modulo 3 found with $YRY(N)_iYRY$. Surprisingly, this method identifies two new types of periodicities, as well as a strong coherence relation between the 64 i-motifs, which are common to the protein coding genes of eukaryotes, prokaryotes and viruses (Section 2):

(i) Most of the autocorrelation functions have the classical periodicity 0 modulo 3. Unexpectedly, a few autocorrelation functions have a periodicity 1 modulo 3 associated with the shifted frame +1 (maximal values of the function at $i = 1$, 4, 7, etc) and a periodicity 2 modulo 3 associated with the shifted frame −1 (maximal values of the function at $i = 2$, 5, 8, etc).

(ii) The classification of i-motifs according to the type of periodicity demonstrates a strong coherence relation between the 64 i-motifs: If two i-motifs $w(N)_iw'$ and $w'(N)_iw$ have the same periodicity, then they have the periodicity 0 modulo 3. If two i-motifs $w(N)_iw'$ and $w'(N)_iw$ do not have the same periodicity, then they have the periodicities 1 and 2 modulo 3.

The non-random statistical properties identified in genes (periodicities, sub-periodicities, maximum, local maxima, etc) are, in our previous works, simulated by an evolutionary model at two successive processes: (i) a process of gene construction: a duplication of an oligonucleotide (Arquès & Michel, 1992) or an independent mixing of oligonucleotides (Arquès & Michel, 1990*b*) followed by (ii) a process of gene evolution: substitutions (Arquès & Michel, 1993, 1994) or random insertions and deletions of trinucleotides simulating RNA editing (Arquès & Michel, 1992).

In Section 3, we demonstrate that the three periodicities 0, 1 and 2 modulo 3 can be modelled with the same approach. The simulated genes are constructed with a stochastic automaton (a newly developed process of gene construction) followed by random insertions/deletions of trinucleotides (a process of gene evolution already used in Arquès & Michel, 1992). For almost all i-motifs, the autocorrelation functions in these simulated genes are strongly correlated with those in protein coding genes, for both the type and the probability level of periodicities. Otherwise, the method for constructing the automaton is presented step-by-step, allowing the reader to apply it to other situations.

The first part of the Discussion recaps on the properties of the genetic code involved in codon translation and leading to the classical periodicity. Then, the shifted periodicities are related to the process of ribosomal frameshifting, which creates overlapping genes or concatenated genes from different frames by translating frameshift sites. The second part presents the evolutionary aspects of the shifted periodicities. We demonstrate that the shifted periodicities cannot be associated with the RNY model (Eigen & Schuster, 1978) or the RRY model (Crick *et al.*, 1976) deduced from the primitive translation model of Crick *et al.* (1976), but are compatible with the oligonucleotide mixing model (Arquès & Michel, 1990*b*). Then, a variant of the primitive translation model of Crick *et al.* (1976) introducing frameshift site translation, explains the shifted periodicities.

2. Identification of Shifted Periodicities Common to Protein Coding Genes of Eukaryotes, Prokaryotes and Viruses

2.1. METHOD

This method generalizes the previous autocorrelation function definition on the alphabet {R, Y} (Arquès & Michel, 1987) to the alphabet {A, C, G, T}.

Let F be a gene population with $n(F)$ DNA sequences. Let s be a sequence in F with a length $l(s)$.

Let w be a trinucleotide on the alphabet $\{A, C, G, T\}$ (A = Adenine, C = Cytosine, G = Guanine, T = Thymine) obtained by specification of YRY (R = A or G, Y = C or T): $w \in \mathcal{T} = \{CAC, CAT, \ldots, TGT\}$ (eight trinucleotides). Let the i-motif $m_i = w(N)_i w'$ be two trinucleotides w and w' in $\mathcal{T}$ separated by any i bases N, N = A, C, G or T and $i \in [0, 99]$:

$$m_i \in \{CAC(N)_iCAC, CAC(N)_iCAT, \ldots, TGT(N)_iTGT\} \quad (64\ i\text{-motifs}).$$

For each s of F, the counter $c_i(s)$ counts the occurrences of m_i in s. In order to count the m_i occurrences in the same conditions for all $i \in [0, 99]$, only the first $L(s) = l(s) - 104$ $(= l(s) - (99 + 6) + 1)$ bases of s are examined (correction of the side effect induced by the end of the sequence which would have led to a decrease of probabilities). The occurrence probability $o_i(s)$ of m_i for s is then equal to the ratio of the counter to the total number of current bases read, i.e. $o_i(s) = c_i(s)/L(s)$. The occurrence probability $A_{w,w'}(i, F)$ of m_i for F is finally equal to

$$A_{w,w'}(i, F) = \sum_{s \in F} o_i(s)/n(F).$$

For a population F, the function $i \rightarrow A_{w,w'}(i, F)$ giving the mean occurrence probability that w' occurs i bases after w is called autocorrelation function $w(N)_i w'$ (associated with the i-motif $w(N)_i w'$) and is represented as a curve. Note that in order to have a sufficient number of m_{99} occurrences, the autocorrelation function is applied to sequences having a minimal length of 500 bases.

The three gene populations F analysed here with the 64 autocorrelation functions are: the eukaryotic protein coding genes (16366 sequences, 23620 kb); the prokaryotic protein coding genes (8809 sequences, 10770 kb) and the viral protein coding genes (6252 sequences, 9690 kb). They are obtained from release 34 of the EMBL Nucleotide Sequence Data Library in the same way as described in previous studies (see e.g. Arquès & Michel, 1990*a*, for a description of data acquisition). The autocorrelation function $A_{w,w'}(i, F)$ is represented as a curve as follows:

(i) the abscissa shows the number i of bases N between w and w', by varying i between 0 and 99;

(ii) the ordinate gives the mean occurrence probability of $w(N)_i w'$ in a gene population F.

2.2. RESULTS

The 64 autocorrelation functions applied to protein coding genes are all non-random. Each autocorrelation function has a periodicity among the four following types:

(i) Type 1: periodicity 0 modulo 3, for example the autocorrelation function $CAT(N)_iCAT$ in the protein coding genes of eukaryotes [Fig. 2(a)], prokaryotes [Fig. 2(b)] and viruses [Fig. 2(c)];

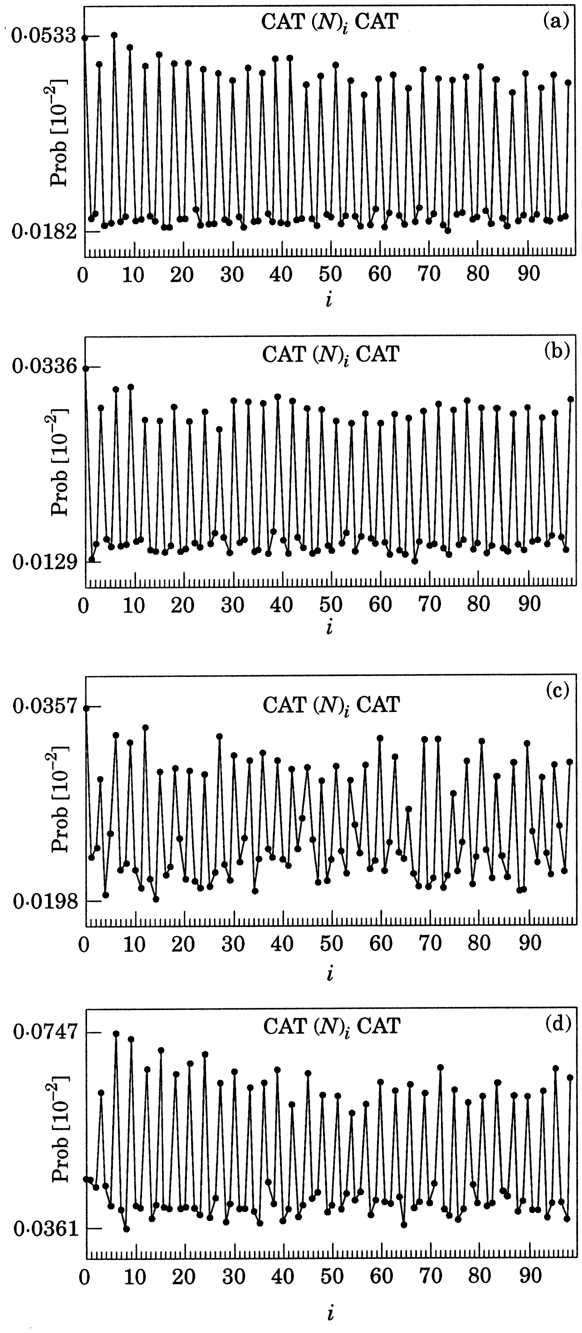


FIG. 2. Periodicity 0 modulo 3 identified with the autocorrelation function $CAT(N)_iCAT$ common to the protein coding genes of eukaryotes, prokaryotes and viruses, and simulated by the stochastic automaton $\mathcal{A}$ (defined in Scheme 2) subjected to random insertions and deletions of trinucleotides. (a) Periodicity 0 modulo 3 in the protein coding genes of eukaryotes; (b) periodicity 0 modulo 3 in the protein coding genes of prokaryotes; (c) periodicity 0 modulo 3 in the protein coding genes of viruses; (d) periodicity 0 modulo 3 simulated.

(ii) Type 2: periodicity 1 modulo 3 (maximal values of the function at $i = 1, 4, 7$, etc), for example the autocorrelation function CAT$(N)_i$TAC in the protein coding genes of eukaryotes [Fig. 3(a)], prokaryotes [Fig. 3(b)] and viruses [Fig. 3(c)];

(iii) Type 3: periodicity 2 modulo 3 (maximal values of the function at $i = 2, 5, 8$, etc), for example

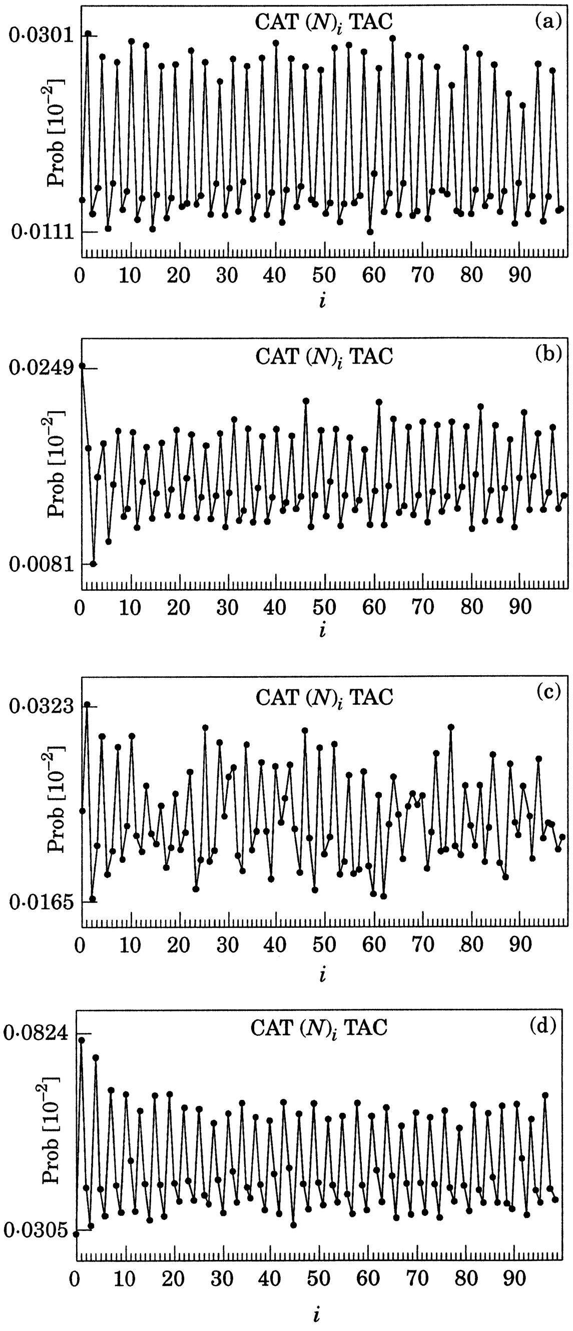


FIG. 3. Periodicity 1 modulo 3 identified with the autocorrelation function CAT$(N)_i$TAC common to the protein coding genes of eukaryotes, prokaryotes and viruses, and simulated by the stochastic automaton $\mathscr{A}$ (defined in Scheme 2) subjected to random insertions and deletions of trinucleotides. (a) Periodicity 1 modulo 3 in the protein coding genes of eukaryotes; (b) periodicity 1 modulo 3 in the protein coding genes of prokaryotes; (c) periodicity 1 modulo 3 in the protein coding genes of viruses; (d) periodicity 1 modulo 3 simulated.

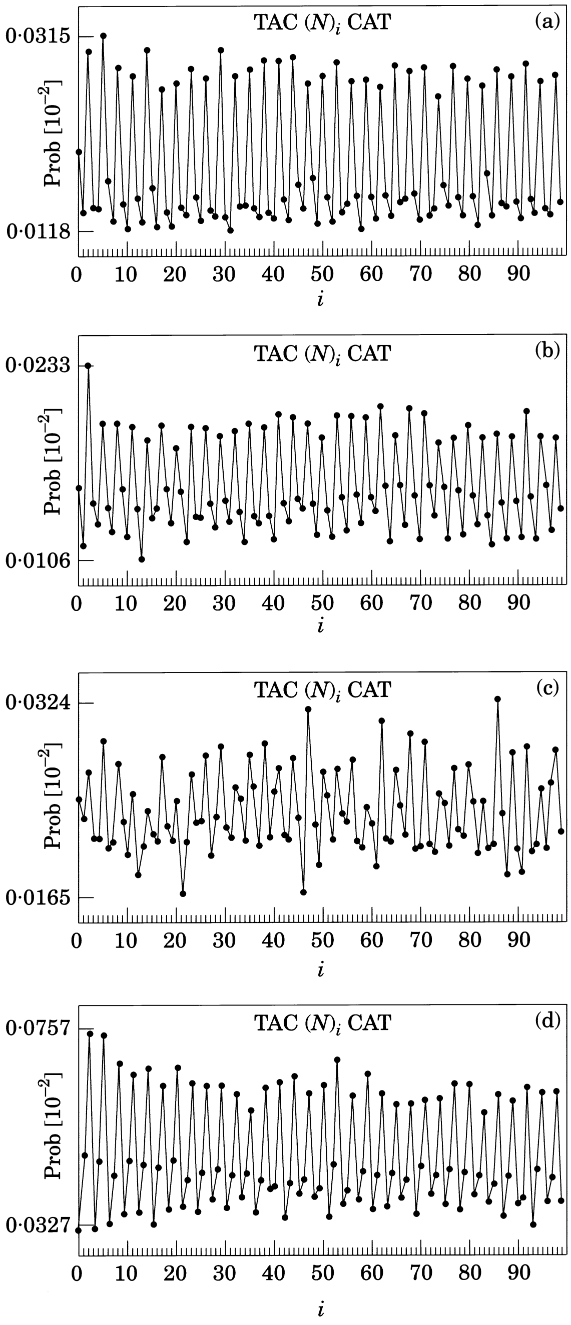


FIG. 4. Periodicity 2 modulo 3 identified with the autocorrelation function TAC$(N)_i$CAT common to the protein coding genes of eukaryotes, prokaryotes and viruses, and simulated by the stochastic automaton $\mathscr{A}$ (defined in Scheme 2) subjected to random insertions and deletions of trinucleotides. (a) Periodicity 2 modulo 3 in the protein coding genes of eukaryotes; (b) periodicity 2 modulo 3 in the protein coding genes of prokaryotes; (c) periodicity 2 modulo 3 in the protein coding genes of viruses; (d) periodicity 2 modulo 3 simulated.

TABLE 1

Periodicities and coherence relation between the i-*motifs common to the protein coding genes of eukaryotes, prokaryotes and viruses*

O modulo 3	1 modulo 3	2 modulo 3
CAC..CAC	CAC..TAC	CAT..TGC
CAC..CAT	CAT..TAC	CGT..TGC
CAC..CGC	CGT..TAC	CGT..TGT
CAC..CGT	TGC..CAT	TAC..CAC
CAT..CAC	TGC..CGT	TAC..CAT
CAT..CAT	TGC..TAT	TAC..CGT
CAT..CGC	TGT..CGT	TAT..TGC
CAT..CGT		
CGC..CAC		
CGC..CAT		
CGC..CGC		
CGC..CGT		
CGC..TGC		
CGT..CAC		
CGT..CAT		
CGT..CGC		
CGT..CGT		
CGT..TAT		
TAC..TAC		
TAC..TAT		
TAT..CGT		
TAT..TAC		
TAT..TAT		
TGC..CGC		
TGC..TGC		
TGC..TGT		
TGT..TGC		
TGT..TGT		

the autocorrelation function TAC$(N)_i$CAT in the protein coding genes of eukaryotes [Fig. 4(a)], prokaryotes [Fig. 4(b)] and viruses [Fig. 4(c)];

(iv) Type 4: a mixing of two periodicities among the three previous types: either 0 and 1 modulo 3 (maximal values of the function at $i = 0, 1, 3, 4, 6, 7$, etc) or 0 and 2 modulo 3 (maximal values of the function at $i = 0, 2, 3, 5, 6, 8$, etc) (data not shown).

Table 1 gives the list of i-motifs (autocorrelation functions) having single periodicities (the first three types): 28 i-motifs have periodicity 0 modulo 3, seven i-motifs have periodicity 1 modulo 3 and seven i-motifs have periodicity 2 modulo 3. Each of these 42 i-motifs has the same type of periodicity in the three populations of protein coding genes (data not shown). The other 22 i-motifs have Type 4 periodicities at least in one population and are only described briefly here.

Table 1 identifies a strong coherence relation between the i-motifs (visualized by correspondence lines):

(i) If two i-motifs $w(N)_i w'$ and $w'(N)_i w$ have the same periodicity then they have periodicity 0 modulo 3. In particular, the i-motifs $w(N)_i w$ have necessarily the periodicity 0 modulo 3.

(ii) If two i-motifs $w(N)_i w'$ and $w'(N)_i w$ do not have the same periodicity then they have periodicities 1 and 2 modulo 3 (mixed with periodicity 0 modulo 3 in the case of the Type 4).

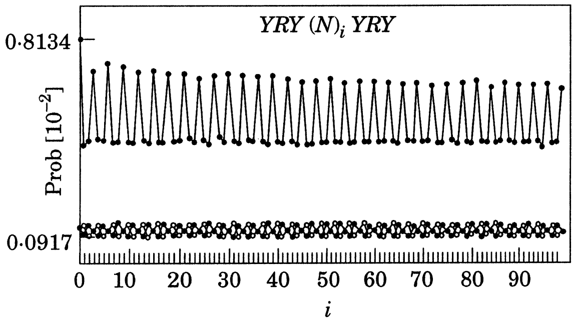


FIG. 5. Periodicity 0 modulo 3 identified with the autocorrelation function $YRY(N)_i YRY$ in the protein coding genes of eukaryotes resulting, according to the Table 1, from the addition of 28 i-motifs with periodicity 0 modulo 3 (top curve) compared to the addition of 14 i-motifs with periodicities 1 and 2 modulo 3 (two mixed bottom curves).

Table 1 also explains why the i-motif $YRY(N)_i YRY$ has periodicity 0 modulo 3 which hides periodicities 1 and 2 modulo 3. Indeed, as the 42 i-motifs obtained by specification of $YRY(N)_i YRY$ on $\{A, C, G, T\}$ have occurrence probabilities of the same level, the addition of 28 i-motifs with periodicity 0 modulo 3 (top curve in Fig. 5), compared to the addition of 14 i-motifs with periodicities 1 and 2 modulo 3 (two bottom curves in Fig. 5), leads to a periodicity 0 modulo 3 with $YRY(N)_i YRY$. The addition of the same number (i.e. seven) of i-motifs with periodicities 1 and 2 modulo 3 explains their same probability level (two mixed bottom curves in Fig. 5). Note that the addition of the 22 i-motifs with a mixing of two periodicities increases the amplitude between the top and bottom curves as periodicity 0 modulo 3 is common to these 22 i-motifs.

In summary, periodicity 0 modulo 3 of $YRY(N)_i YRY$ is the consequence of a greater number of periodicities 0 modulo 3 with the 64 i-motifs compared to the numbers of periodicities 1 and 2 modulo 3.

3. Simulation of the Three Periodicities Identified in Protein Coding Genes

3.1. METHOD

The three periodicities 0, 1 and 2 modulo 3 identified in protein coding genes are simulated by an evolutionary model at two successive processes. The simulated genes are constructed with a stochastic automaton (a newly developed process of gene construction) (explained in Arquès & Michel, 1990*b*, 1992 and below) followed by random insertions/deletions of trinucleotides (a process of gene evolution simulating the RNA editing used in Arquès & Michel, 1992).

Table 2(a)

Correspondence between the computer theory of languages and the gene structure

Gene structure	Computer theory of languages
Four nucleotides (bases):	Letter of the alphabet $B = \{\text{A, C, G, T}\}$ or $P = \{R, Y\}$
$\left.\begin{array}{l}\text{A} = \text{Adenine}\\ \text{G} = \text{Guanine}\end{array}\right\} = R = \text{Purine}$	
$\left.\begin{array}{l}\text{C} = \text{Cytosine}\\ \text{T} = \text{Thymine}\end{array}\right\} = Y = \text{Pyrimidine}$	
Oligonucleotide	Word of a small size (<10 letters) of B^* or P^*
Gene or DNA sequence	Word of a large size (>100 letters) of B^* or P^*
Gene population	Language B^* or P^*

Table 2(b)

Correspondence between the computer theory of languages and the gene construction process

Gene construction process	Computer theory of languages
Duplication of an oligonucleotide	Concatenation of a word of a small size on the alphabet B or P
Mixing of oligonucleotides	Independent or Markov concatenation of words of small sizes on the alphabet B or P

Table 2(c)

Correspondence between the computer theory of languages and the gene evolution process

Gene evolution process	Computer theory of languages
Substitution	Random transformations of letters on the alphabet B
Transversion	Random transformations of letters on the alphabet P: $R \rightarrow Y$ and $Y \rightarrow R$
RNA editing	Random insertions and deletions in words of letters or words of small sizes on the alphabet B or P

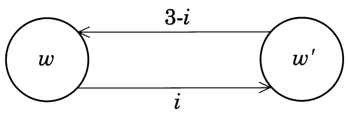
We briefly remind the reader [Tables 2(a)–(c)] the correspondences existing between the computer theory of languages and: (i) the gene structure [Table 2(a)], (ii) the gene construction process [Table 2(b)] and (iii) the gene evolution process [Table 2(c)] which are used in the following.

3.1.1. *Gene construction with a stochastic automaton*

In order to obtain a simulated gene population (a language) having the three periodicities 0, 1 and 2 modulo 3 and the i-motifs correctly associated with the periodicities, an automaton is constructed according to the coherence relation between the i-motifs identified in Table 1. The automaton considered here is stochastic: the choice between the edges issued from a same state is equiprobable. It constructs words having the properties of the automaton and therefore the coherence relation between the i-motifs. The words are constructed by a random path (a series of edges) in the automaton so that the edge crossing concatenates the word associated with the edge to the building word. This gene construction has similarities with translation which concatenates the amino acid of an aminoacyl-tRNA to the growing peptide of a peptidyl-tRNA.

Construction rules of the automaton. (i) Construction rules of the automaton to generate the periodicities 1 and 2 modulo 3 correctly associated with the i-motifs $w(N)_i w'$, i equal to 1 or 2 modulo 3, and the periodicity 0 modulo 3 correctly associated with the i-motifs $w(N)_i w$, i equal to 0 modulo 3, $w, w' \in \mathscr{T} = \{\text{CAC, CAT}, \ldots, \text{TGT}\}$ being a trinucleotide obtained by specification of YRY on $\{\text{A, C, G, T}\}$:

All edges issued from a same state are associated with a word of the type $w(N)_i$ with $w \in \mathscr{T}$ and i equal 1 or 2, i.e. the words wN or wNN ($N =$ A, C, G or T). The automaton is represented as follows. The label of each state is the trinucleotide w common to all edges issued from the state. The label of each edge issued from a state is the number 1 or 2 of any bases N to be added to the trinucleotide w in order to concatenate the word $w(N)_i$ to the building word during the edge crossing. The coherence relation between the i-motifs leading to the periodicities 1 and 2 modulo 3, i.e. the paired i-motifs $w(N)_i w'$ and $w'(N)_{3-i} w$, i equal to 1 or 2 modulo 3, leads to label two opposite edges with the paired words $w(N)_i$ and $w'(N)_{3-i}$, i equal to 1 or 2. Scheme 1 shows such an elementary closed path (of length 2).



SCHEME 1. Elementary closed path deduced from the coherence relation between the i-motifs leading to the periodicities 1 and 2 modulo 3 in Table 1.

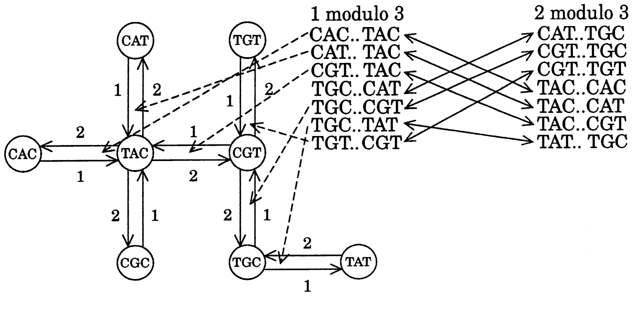
One closed path constructs the word $w(N)_i w'(N)_{3-i} w = w(N)_6 w$ congruent to 0 modulo 3, the word $w(N)_1 w'$ congruent to 1 modulo 3 and the word $w'(N)_2 w$ congruent to 2 modulo 3. Two closed paths construct the word $w(N)_{15} w$ congruent to 0 modulo 3, the word $w(N)_{10} w'$ congruent to 1 modulo 3 and the word $w'(N)_{11} w$ congruent to 2 modulo 3, etc.

(ii) Construction rules of the automaton to generate in addition the periodicity 0 modulo 3 correctly associated with the i-motifs $w(N)_i w'$, i equal to 0 modulo 3:

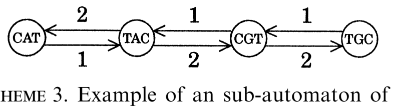
The automaton is constructed by a concatenation of elementary closed paths according to a tree so that the concatenation of two elementary closed paths leads to the word $w(N)_i w''(N)_{3-i} w' = w(N)_6 w'$ (i equal to 1 or 2) congruent to 0 modulo 3.

The automaton $\mathscr{A}$. The final automaton $\mathscr{A}$ (Scheme 2) is obtained by a concatenation of seven elementary closed paths deduced from the coherence relation between the i-motifs leading to the periodicities 1 and 2 modulo 3 in Table 1.

The automaton $\mathscr{A}$ allows the retrieval of several periodicities which have not been directly introduced. For example, the sub-automaton of $\mathscr{A}$ in Scheme 3 (a concatenation of some elementary closed paths of $\mathscr{A}$) generates periodicities 1 (respectively 2) modulo 3 with the i-motif TGC$(N)_i$CAT (respectively CAT$(N)_i$TGC) identified in Table 1 by constructing the word TGCNCGTNTACNNCAT = TGC$(N)_{10}$CAT (respectively, CATNTACNNCGTNNTGC = CAT$(N)_{11}$TGC) which is congruent to 1 (respectively, 2) modulo 3.



SCHEME 2. The automaton $\mathscr{A}$.



SCHEME 3. Example of an sub-automaton of $\mathscr{A}$.

Other simple sub-automata of $\mathscr{A}$ generate the periodicity 0 modulo 3 associated with the i-motifs $w(N)_i w'$ = CGC$(N)_i$CAC, CGC$(N)_i$CAT, CGC$(N)_i$CGT and their associated i-motifs $w'(N)_i w$ identified in Table 1.

3.1.2. *Gene evolution by random insertions and deletions of trinucleotides*

A population S of 400 simulated genes of 3000 base length is generated with the automaton $\mathscr{A}$. Note that the computations obtained with such a sample of 1.2 million bases are precise and stable (i.e. there is no random fluctuations in the probability calculus of i-motifs: a sample having for example 200 simulated genes of 1000 base length leads to similar results).

The 64 autocorrelation functions are computed in this simulated population S (before evolution) as defined in Section 2.1. As expected, they have periodicities 0, 1 and 2 modulo 3 correctly associated with the i-motifs. The gene construction process given in Section 3.1.1 is the most important (and difficult) step in the simulation. However, the amplitudes of periodicities (mean probability difference between the top and bottom curves) are higher than the ones observed in the genes (data not shown). As explained in detail in Arquès & Michel (1990*b*, 1992), a simple process of gene construction, i.e. based on an operation (a duplication, an independent mixing, here a stochastic automaton) involving only a few types of oligonucleotides (here eight trinucleotides), leads to simulated genes with strong statistical properties (e.g. higher amplitudes of periodicities) because some trinucleotides are not used in the simulation. In order to have a higher correlation between real and simulated curves (i.e. not only with the curve shape but also with the curve level), a random process of gene evolution (substitutions or insertions/deletions of trinucleotides) must follow the gene construction process. The trinucleotide insertion/deletion process used here is similar to RNA editing (Benne *et al.*, 1986). We refer the reader to the reviews (Benne, 1989; Feagin, 1990; Simpson, 1990; Cech, 1991; Stuart, 1991; Covello & Gray, 1993) for the biological description of RNA editing, and also the Introduction and Section 2.2 in Arquès & Michel (1992) for the similarities between the insertion/deletion process and RNA editing. In addition to the similarities already mentioned concerning the functions of RNA editing in actual genes such as a correcting mechanism, recent experimental work shows that RNA editing and trinucleotides are also factors of gene evolution (Landweber & Gilbert, 1993, 1994; Morell, 1993; Kuhl & Caskey, 1993).

Therefore, the simulated population S is subjected to a trinucleotide insertion/deletion process with steps such that at each step there are one random (in position and type) insertion of one trinucleotide and one random (in position) deletion of one trinucleotide per simulated gene. Note that the choice of a same number for the inserted bases and the deleted ones allows one to keep genes with the same length during the insertion/deletion process. The 64 autocorrelation functions are computed every ten steps as defined in Section 2.1.

3.2. RESULTS

Figure 2(d) [respectively, Figs 3(d) and 4(d)] shows the autocorrelation functions CAT$(N)_i$CAT (respectively, CAT$(N)_i$TAC and TAC$(N)_i$CAT) after 600 insertions/deletions of trinucleotides per simulated gene in the simulation population S generated with the automaton $\mathscr{A}$. For each of these three i-motifs, the autocorrelation function in these simulated genes is strongly correlated with that in protein coding genes (eukaryotes, prokaryotes and viruses), for the type as well as for the probability level of periodicities. Furthermore, this correlation is verified in almost all i-motifs, as only six i-motifs among 64 have a simulated periodicity different from the periodicities observed in the three gene populations (data not shown). Note that 600 is an average number and represents the middle of a step range in which the statistical properties are similar. Indeed, during the insertion/deletion process, there is a continuous modification of probabilities of simulated periodicities towards the random situation represented as a horizontal line (not observed in protein coding genes).

4. Discussion

The method based on autocorrelation functions with i-motifs on $\{A, C, G, T\}$ has mainly allowed: (i) the retrieval of the classical periodicity 0 modulo 3 which is associated with the normal frame 0 established by the ATG initiation codon; (ii) the identification of two shifted periodicities 1 and 2 modulo 3 which are associated with the shifted frames $+1$ and -1, respectively, and which have escaped from previous statistical analyses as they are hidden by the classical periodicity; and (iii) the demonstration of a strong coherence relation between the i-motifs stated in Section 2.2. Furthermore, these three periodicities, as well as the coherence relation between i-motifs, are common to the protein coding genes of eukaryotes, prokaryotes and viruses. Therefore, they are independent of the evolution of protein coding genes: they are related neither to a particular taxonomy (common to three populations) nor to a particular protein coding function (a population is constituted of thousands of various genes; see Section 2.1).

The three periodicities are generated by a non-random distribution of nucleotides in protein coding genes. The classical periodicity is related to codon translation while the shifted periodicities evolve from frameshift site translation. A reminder of the properties of the genetic code involved in codon translation are given as follows.

The three codon bases do not have the same functions in codon translation: there is a strong difference between the first two bases and the third one. Indeed, in almost all cases (58 codons among 64), the third base has either no function whatever its type in eight groups of four codons (e.g. ACN encodes Thr) or a function if its type is R in six groups of two codons (e.g. AAR encodes Lys, TAR is a stop codon) or Y in seven groups of two codons (e.g. AAY encodes Asn). In some species, this property can be verified in all cases (eight groups of four codons and two times eight groups of two codons): for example, in yeast and mammalian mitochondria, ATA may encode Met instead of Ile and TGA, Trp instead of the stop codon. This genetic code degeneracy is related to the wobble concept (Crick, 1966) which states that the codon–anticodon pairing in the first two codon sites (the last two anticodon sites) follows the classical pairing rules $(A \cdot T, C \cdot G)$ whereas mismatches $(G \cdot T)$ are allowed in the third codon site (first anticodon site or wobble site). The wobble concept is supported by experimental studies with synthetic trinucleotides which demonstrate that single tRNAs can pair with two codons according to the wobble rules. Further analyses, showing that single tRNAs can pair with four codons which have the same first two bases, lead to the "two out of three" rule (Lagerkvist, 1978, 1981) stating that only the first two codon bases are significant in the codon–anticodon pairing. However, the anticodon loop and the proximal stem may also be involved in the codon pairing (Yarus, 1982). Furthermore, unexpected mismatches in the first codon site $(G \cdot T)$ and the third codon site $(C \cdot A)$ have been recently observed (Schultz & Yarus, 1994). Finally, the complexity of this pairing increases with the presence of modified bases in the wobble site, for example inosine $(I \cdot A, I \cdot C, I \cdot T)$.

Frameshift site translation generates the shifted periodicities. There are mainly two ribosomal frameshifting processes that can create overlapping genes or concatenated genes from different frames: hopping and slipping. Hopping can be defined as a ribosomal shift $\geqslant 2$ bases in the $5'$–$3'$ or $3'$–$5'$ direction. It requires a take-off site and a landing site. For example, these two sites can be separated by two bases as in GTGT*R* decoded as Val, by five bases as in AACTCAAT decoded as Asn, by six bases as in CTTTAGCTA decoded as Leu and GTGTAAGTT decoded as Val, up to 50 bases (*E. coli*, Weiss *et al*., 1987, 1990; O'Connor *et al*., 1989). Slipping can be defined as a ribosomal shift of one base either in the $3'$–$5'$ direction (-1 frameshift) or in the $5'$–$3'$ direction ($+1$ frameshift). For example, the sites AAA*N*, CTTG, GGGG and TTTA direct -1 frameshift (retroviruses, Hizi *et al*., 1987; Jacks *et al*., 1988; plants, Prüfer *et al*., 1992). In contrast, the sites AAG*Y*, CCGT, CTTA, CTTT, GGGT and GTTT provoke $+1$ frameshift (*E. coli*, Weiss *et al*., 1987, 1988*a*; *S. typhimurium*, Tuohy *et al*., 1992; yeast transposon, Mellor *et al*., 1985; Belcourt & Farabaugh, 1990). Therefore, only a few specific sites in protein coding genes are frameshift sites (see e.g. an experimental demonstration in Belcourt & Farabaugh, 1990). The site TTTT can induce -1 frameshift (retroviruses, Jacks *et al*., 1988; yeast mitochondria, Fox & Weiss-Brummer, 1980) as well as $+1$ frameshift (yeast mitochondria, Fox & Weiss-Brummer, 1980). Therefore, secondary frameshift sites in protein coding genes associated with the (primary) frameshift sites may influence frameshifting, for example a Shine–Dalgarno-like sequence located three bases upstream the frameshift site (Weiss *et al*., 1988*b*), a stem-loop structure downstream the frameshift site (Jacks *et al*., 1988; Le *et al*., 1989; Prüfer *et al*., 1992). The tRNAs interacting with the non-random frameshift sites to promote slipping and hopping are called "shifty" tRNAs. These shifty tRNAs have often a modified anticodon, for example, a post-transcriptional modification of the wobble base (e.g. Meir *et al*., 1985) and/or an extended anticodon (e.g. Riddle & Carbon, 1973), so that the number and site of bases in the codon–anticodon pairing allow non-standard translation at the frameshift site and frameshifting, such as a two or four base translation in normal frame, a three base translation in -1 or $+1$ shifted frames, etc. To date, few ribosomal frameshifting mechanisms have been reported (e.g. Bruce *et al*., 1986; Jacks *et al*., 1988; O'Connor *et al*., 1989; Belcourt & Farabaugh, 1990). The mechanism found by Jacks *et al*. (1988), which is the most conserved according to the available data, is used to generate the shifted periodicities in the primitive translation model of Crick *et al*. (1976) (see below). Similarly, the frameshift suppressor tRNAs which allow the normal frame to be kept, have also often a modified anticodon interacting with non-random frameshift sites such as AAA*N*, ACC*N*, CCC*N*, CCGT and GGG*N* (e.g. Riddle & Roth, 1972; Riddle & Carbon, 1973; Kohno & Roth, 1978; Donahue *et al*., 1981; Bossi & Roth, 1981; Gaber & Culbertson, 1984).

The non-random distribution of nucleotides in codons and frameshift sites presented above are associated with the codon–anticodon pairing. Other mRNA sites, interacting for example with tRNA bases outside the anticodon, rRNAs or with ribosomal and non-ribosomal proteins, may act on codon and frameshift site translation, tRNA selection and proofreading, which are functions involved in protein synthesis. For example, the Shine–Dalgarno-like sequence, mentioned above, pairs with the $3'$ end of 16S rRNA to influence frameshifting (Weiss *et al*., 1988*b*). Theoretically, 30–40 bases of mRNA may interact at any time in the ribosome with two tRNAs ($\sim 2 \times 75$ bases), four rRNAs in eukaryotes (~ 7000 bases in total) and various proteins, in particular ~ 80 ribosomal proteins in eukaryotes, the initiation, elongation and release factors, etc.

From an evolutionary point of view, the three periodicities can be examined in the primitive translation model proposed by Crick *et al*. (1976). In this model, the primitive tRNAs have an anticodon involving seven conserved (in actual tRNAs) nucleotides (Barrell & Clark, 1974) in two alternative stacking configurations. The aminoacyl-tRNA adopts the $5'$ stacked configuration with the five-base anticodon $YYa_1a_2a_3$, a_1 being the wobble base (Woese, 1970). On the other hand, the peptidyl-tRNA adopts the $3'$ stacked configuration with the five-base anticodon $a_1a_2a_3RN$ (Fuller & Hodgson, 1967). A stacked configuration pairs with a five-base codon. Note that a primitive codon-anticodon pairing involving only three bases, such as an actual one without ribosome, is not stable enough for codon translation (Grosjean *et al*., 1978). Refer to Crick *et al*. (1976) and Eigen & Schuster (1978) for a detailed description of biological concepts and translation steps. The flip mechanism (detailed below for the *RNY* codon) between these two stacked configurations leads to primitive protein coding genes constituted of a series of *RRY* codons (*RRY* model) (Crick *et al*., 1976). These biological concepts are considered by Eigen & Schuster (1978) with an anticodon involving only five conserved nucleotides. Therefore, the aminoacyl-tRNA with the $5'$ stacked configuration has the four-base anticodon $Ya_1a_2a_3$ ($3'a_3a_2a_1Y$) pairing with the four-base codon $c_1c_2c_3R$, and the peptidyl-tRNA with the $3'$ stacked configuration has the four-base anticodon $a_1a_2a_3R$ ($3'Ra_3a_2a_1$) pairing with the four-base codon $Yc_1c_2c_3$ (Scheme 7). The flip mechanism between these two stacked configurations keeps the three anticodon bases $3'a_3a_2a_1$ paired with the three codon bases $c_1c_2c_3$

```
RNY model:                  5′ . R N Y R N Y R N Y R N Y R N Y R N Y R N Y . 3′
                                     | | |       | | |
Autocorrelation function:            Y R Y N N N Y R Y
Mixing model:               5′ . Y R Y N N N Y R Y N N N N N N Y R Y N N N . 3′
                                 | | |       | | |
Autocorrelation function:        Y R Y N N N Y R Y
```

SCHEME 4. Compatibility of the periodicity 0 modulo 3 with the *RNY* model and the oligonucleotide mixing model. The symbol "|" means base pairing between $YRY(N)_3YRY$ and the *RNY* model and between $YRY(N)_3YRY$ and the oligonucleotide mixing model.

```
RNY model:                  5′ . R N Y R N Y R N Y R N Y R N Y R N Y R N Y . 3′
                                     | | |         — | |
Autocorrelation function:            Y R Y N N N N Y R Y
Mixing model:               5′ . Y R Y N N N Y R Y N N N N N N Y R Y N N N . 3′
                                             | | |         | | |
Autocorrelation function:                    Y R Y N N N N Y R Y
```

SCHEME 5. Incompatibility of the periodicity 1 modulo 3 with the *RNY* model and compatibility of the periodicity 1 modulo 3 with the oligonucleotide mixing model. The symbols "|" and "—" mean base pairing and non base pairing, respectively, between $YRY(N)_4YRY$ and the *RNY* model and between $YRY(N)_4YRY$ and the oligonucleotide mixing model.

```
RNY model:                  5′ . R N Y R N Y R N Y R N Y R N Y R N Y R N Y . 3′
                                     | | |           | — —
Autocorrelation function:            Y R Y N N N N N Y R Y
Mixing model:               5′ . Y R Y N N N Y R Y N N N N N N Y R Y N N N . 3′
                                 | | |           | | |
Autocorrelation function:        Y R Y N N N N N Y R Y
```

SCHEME 6. Incompatibility of the periodicity 2 modulo 3 with the *RNY* model and compatibility of the periodicity 2 modulo 3 with the oligonucleotide mixing model. The symbols "|" and "—" mean base pairing and non base pairing, respectively, between $YRY(N)_5YRY$ and the *RNY* model and between $YRY(N)_5YRY$ and the oligonucleotide mixing model.

for codon translation. The sequence $3'Ra_3a_2a_1a_3a_2a_1Y$ of the peptidyl and aminoacyl tRNAs paired with mRNA during the flip mechanism leads to primitive protein coding genes constituted of a series of *RNY* codons (*RNY* model) (Scheme 7). The *RNY* codon includes the *RRY* codon ($N = R$). Note that the *RNY* codon is also more symmetrical than the *RRY* codon (same number of *R* and *Y* on a strand and same codon type on the complementary strand). On the other hand, the oligonucleotide mixing model (Arquès & Michel, 1990*b*) developed from the identification of non-random statistical properties in protein coding and non-coding genes proposes that the primitive genes, and in particular the primitive protein coding genes, are generated by an independent mixing (in contrast to a series) of a few types of oligonucleotides whose two main types are $YRY(N)_3$ and $YRY(N)_6$. The three periodicities observed in the actual protein coding genes can be associated with the oligonucleotide mixing model but not with the *RNY* model (and also not with the *RRY* model which is a particular case of the *RNY* model):

(i) Periodicity 0 modulo 3 with the *i*-motifs on {A, C, G, T} is compatible with the *RNY* model and the oligonucleotide mixing model. Indeed, by choosing a representative of this periodicity on $\{R, Y\}$, for example $YRY(N)_3YRY$, Scheme 4 easily verifies this assertion.

(ii) Periodicity 1 modulo 3 with the *i*-motifs on {A, C, G, T} is incompatible with the *RNY* model but compatible with the oligonucleotide mixing model. Scheme 5 easily verifies this assertion by choosing $YRY(N)_4YRY$ as a representative of this periodicity.

(iii) Periodicity 2 modulo 3 with the *i*-motifs on {A, C, G, T}, as with periodicity 1 modulo 3, is incompatible with the *RNY* model but compatible with the oligonucleotide mixing model. Scheme 6 verifies this assertion by choosing $YRY(N)_5YRY$ as a representative of this periodicity.

The shifted periodicities 1 and 2 modulo 3 cannot be associated with the *RNY* and *RRY* models deduced from the primitive translation model of Crick *et al.* (1976). The shifted periodicities can be easily explained in a variant in which the codon translation is sometimes replaced by frameshift site translation. The mechanism chosen in Scheme 7 involves an anticodon $3'a_3a_2$ pairing with the codon c_3c_1 according to the -1 frameshifting process of simultaneous slippage of the aminoacyl and peptidyl tRNAs found by Jacks *et al.* (1988). This process is

```
            Primitive translation model of Crick et al. (1976)

Codon:     5' . 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 . 3'

Anticodon: 3' R 3 2 1 3 2 1 Y 5'
                    R 3 2 1 3 2 1 Y
                          R 3 2 1 3 2 1 Y
                                R 3 2 1 3 2 1 Y
                                      R 3 2 1 3 2 1 Y
                                            R 3 2 1 3 2 1 Y
                                                  R 3 2 1 3 2 1 Y
                                                        R 3 2 1 3 2 1 Y
                                                              R 3 2 1 3 2 1 Y

Codon pattern:       5' . Y R N Y R N Y R N Y R N Y R N Y R N Y R N Y R . 3'
Autocorrelation function:       Y R Y Y R Y: 0 modulo 3
                                Y R Y N N N Y R Y: 0 modulo 3


     Variant of the primitive translation model of Crick et al. (1976)

Codon:     5' . 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 . 3'

Anticodon: 3' R 3 2 1 3 2 1 Y 5'
                    R 3 2 1 3 2 1 Y
                          R 3 2 1 3 2 1 Y
                                R 3 2 1 3 2 1 Y
                                      R 3 2 1 3 2 1 Y
               Frameshift site translation: 3 2 - 3 2 -
                                                R 3 2 1 3 2 1 Y
                                                      R 3 2 1 3 2 1 Y
                                                            R 3 2 1 3 2 1 Y

Codon pattern:       5' . Y R N Y R N Y R N N R Y N R Y N N Y R N . 3'
Autocorrelation function:       Y R Y Y R Y: 0 modulo 3
                                Y R Y N N N Y R Y: 0 modulo 3
                                            Y R Y N N Y R Y: 2 modulo 3
                                            Y R Y N N N N N Y R Y: 2 modulo 3
                                                      Y R Y Y R Y: 0 modulo 3
```

SCHEME 7. The primitive translation model of Crick *et al.* (1976) with an anticodon involving five conserved nucleotides leading to the RNY model (Eigen & Schuster, 1978) only explains the classical periodicity 0 modulo 3 (see text). A variant of the primitive translation model of Crick *et al.* (1976), introducing frameshift site translation, explains the classical periodicity 0 modulo 3 and the shifted periodicities 1 and 2 modulo 3. For example, the first occurrence of the frameshift site translation according to the −1 frameshifting process of simultaneous slippage of the aminoacyl and peptidyl tRNAs (Jacks *et al.*, 1988), generates the periodicity 2 modulo 3 with the autocorrelation function $YRY(N)_iYRY$. The symbol "—" means non base pairing and the numbers correspond to the sites of bases in the codon and anticodon.

supported by recent experimental results and appears to be the most conserved (Prüfer *et al.*, 1992). It is important to stress that the pattern for the shifted frame is always the same, except at the frameshift site, whatever the mechanism of frameshift site translation (i.e. whatever the number and site of bases involved in the codon-anticodon pairing): theoretically, any mechanism can be chosen. Remember that the shifted periodicities identified here are verified for large values of i (at least until $i = 99$). Precisely, the first occurrence of this frameshift site translation in protein coding genes generates the periodicity 2 modulo 3 with the autocorrelation function $YRY(N)_iYRY$ (Scheme 7), the second occurrence generates the periodicity 1 modulo 3 and the third occurrence restores the periodicity 0 modulo 3 (data not shown). Other types of frameshift site translations in protein coding genes, such as the +1 frameshifting process of the peptidyl tRNA during a translational pause (Belcourt & Farabaugh, 1990), generate the periodicities 1, 2 and 0 modulo 3 successively. Note that most of the amino acids might have been encoded by codon translation; a few amino acids, by frameshift site translation. This feature may also traduce the malleability of the genetic code (Schultz & Yarus, 1994).

The shifted periodicities observed in the actual protein coding genes could have no function and only bear traces of primitive translation. However, they may also reveal overlapping genes or concatenated genes from different frames. Further analyses in protein coding genes and tRNAs are currently in investigation to deepen the relation between the type of periodicity and the codon base according to its alphabet ($\{R, Y\}$ or $\{\mathrm{A, C, G, T}\}$) as well as its site. This approach could be used to locate the frameshift sites in protein coding genes and to characterize the anticodons of shifty tRNAs.

We thank Dr Nouchine Soltanifar for her advice. This work was supported by INSERM grant (contrat de recherche externe No 930101).

REFERENCES

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. (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. & 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.*, in press.

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

Belcourt, M. F. & Farabaugh, P. J. (1990). Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell* **62,** 339–352.

Benne, R., Van Den Burg, J., Brakenhoff, J. P. J., Sloof, P., Van Boom, J. H. & Tromp, M. C. (1986). Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46,** 819–826.

Benne, R. (1989). RNA-editing in trypanosome mitochondria. *Biochem. Biophys. Acta* **1007,** 131–139.

Bossi, L. & Roth, J. R. (1981). Four-base codons ACCA, ACCU, and ACCC are recognized by frameshift suppressor *sufJ*. *Cell* **25,** 489–496.

Bruce, A. G., Atkins, J. F. & Gesteland, R. F. (1986). tRNA anticodon replacement experiments show that ribosomal frameshifting can be caused by doublet decoding. *Proc. natn. Acad. Sci. U.S.A.* **83,** 5062–5066.

Cech, T. R. (1991). RNA editing: world's smallest introns? *Cell* **64,** 667–669.

Covello, P. S. & Gray, M. W. (1993). On the evolution of RNA editing. *Trends Genet.* **9,** 265–268.

Crick, F. H. C. (1966). Codon-anticodon pairing: the wobble hypothesis. *J. molec. Biol.* **19,** 548–555.

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

Donahue, T. F., Farabaugh, P. J. & Fink G. R. (1981). Suppressible four-base glycine and proline codons in yeast. *Science* **212,** 455–457.

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

Feagin, J. E. (1990). RNA editing in kinetoplastid mitochondria. *J. biol. Chem.* **265,** 19373–19376.

Fox, T. D. & Weiss-Brummer, B. (1980). Leaky +1 and −1 frameshift mutations at the same site in a yeast mitochondrial gene. *Nature, Lond.* **288,** 60–63.

Fuller, W. & Hodgson, A. (1967). Conformation of the anticodon loop in tRNA. *Nature, Lond.* **215,** 817–821.

Gaber, R. F. & Culbertson, M. R. (1984). Codon recognition during frameshift suppression in *Saccharomyces cerevisiae*. *Molec. Cell. Biol.* **4,** 2052–2061.

Grosjean, H. J., de Henau, S. & Crothers, D. M. (1978). On the physical basis for ambiguity in genetic coding interactions. *Proc. natn. Acad. Sci. U.S.A.* **75,** 610–614.

Hizi, A., Henderson, L. E., Copeland, T. D., Sowder, R. C., Hixson, C. V. & Oroszlan, S. (1987). Characterization of mouse mammary tumor virus *gag-pol* gene products and the ribosomal frameshifting site by protein sequencing. *Proc. natn. Acad. Sci. U.S.A.* **84,** 7041–7045.

Jacks, T., Madhani, H. D., Marsiarz, F. R. & Varmus, H. E. (1988). Signals for ribosomal frameshifting in the rous sarcoma virus *gag-pol* region. *Cell* **55,** 447–458.

Kohno, T. & Roth, J. R. (1978). A *Salmonella* frameshift suppressor that acts at runs of A residues in the messenger RNA. *J. molec. Biol.* **126,** 37–52.

Kuhl, D. P. A. & Caskey, C. T. (1993). Trinucleotide repeats and genomic variation. *Curr. Opin. Genet. Dev.* **3,** 404–407.

Landweber, L. F. & Gilbert, W. (1993). RNA editing as a source of genetic variation. *Nature, Lond.* **363,** 179–182.

Landweber, L. F. & Gilbert, W. (1994). Phylogenetic analysis of RNA editing: a primitive genetic phenomenon. *Proc. natn. Acad. Sci. U.S.A.* **91,** 918–921.

Lagerkvist, U. (1978). "Two out of three": an alternative method for codon reading. *Proc. natn. Acad. Sci. U.S.A.* **75,** 1759–1762.

Lagerkvist, U. (1981). Unorthodox codon reading and the evolution of the genetic code. *Cell* **23,** 305–306.

Le, S.-Y., Chen, J.-H. & Maizel, J. V. (1989). Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucl. Acids Res.* **17,** 6143–6152.

Meir, F., Suter, B., Grosjean, H., Keith, G. & Kubli, E. (1985). Queuosine modification of the wobble base in tRNA$^{\mathrm{His}}$ influences "in vivo" decoding properties. *EMBO J.* **4,** 823–827.

Mellor, J., Fulton, S. M., Dobson, M. J., Wilson, W., Kingsman, S. M. & Kingsman, A. J. (1985). A retrovirus-like strategy for expression of a fusion protein encoded by yeast transposon Ty1. *Nature, Lond.* **313,** 243–246.

Morell, V. (1993). The puzzle of the triple repeats. *Science* **260,** 1422–1423.

O'Connor, M., Gesteland, R. F. & Atkins, J. F. (1989). tRNA hopping: enhancement by an expanded anticodon. *EMBO J.* **8,** 4315–4323.

Prüfer, D., Tacke, E., Schmitz, J., Kull, B., Kaufman, A. & Rohde, W. (1992). Ribosomal frameshifting in plants: a novel signal directs the −1 frameshift in the synthesis of the putative viral replicase of potato leafroll luteovirus. *EMBO J.* **11,** 1111–1117.

Riddle, D. L. & Carbon, J. (1973). Frameshift suppression: a nucleotide addition in the anticodon of a glycine tRNA. *Nature New Biol.* **242,** 230–234.

Riddle, D. L. & Roth, J. R. (1972). Frameshift suppressors: III. Effects of suppressor mutations on transfer RNA. *J. molec. Biol.* **66,** 495–506.

Schultz, D. W. & Yarus, M. (1994). Transfer RNA mutation and the malleability of the genetic code. *J. molec. Biol.* **235,** 1377–1380.

Simpson, L. (1990). RNA editing—a novel genetic phenomenon? *Science* **250,** 512–513.

Stuart, K. (1991). RNA editing in mitochrondrial mRNA of trypanosomatids. *Trends Biochem. Sci.* **16,** 68–72.

Tuohy, T. M. F., Thompson, S., Gesteland, R. F. & Atkins, J. F. (1992). Seven, eight and nine-membered anticodon loop mutants of tRNA$_2^{\mathrm{Arg}}$ which cause +1 frameshifting. *J. molec. Biol.* **228,** 1042–1054.

Weiss, R. B., Dunn, D. M., Atkins, J. F. & Gesteland, R.n F. (1987). Slippery runs, shifty stops, backward steps and forward hops: −2, −1, +1, +2, +5, and +6 ribosomal frameshifting. *Cold Spring Harbor Symp. Quant. Biol.* **52,** 687–693.

Weiss, R., Lindsley, D., Falahee, B. & Gallant, J. (1988*a*). On the mechanism of ribosomal frameshifting at hungry codons. *J. molec. Biol.* **203,** 403–410.

Weiss, R. B., Dunn, D. M., Dahlenberg, A. E., Atkins, J. F. & Gesteland, R. F. (1988*b*). Reading frame switch caused by base-pair formation between the 3′ end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J.* **7,** 1503–1507.

Weiss, R. B., Dunn, D. M., Atkins, J. F. & Gesteland, R. F. (1990). Ribosomal frameshifting from −2 to +50 nucleotides. *Prog. Nucleic Acids Res. molec. Biol.* **39,** 159–183.

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

Yarus, M. (1982). Translational efficiency of transfer RNAs: uses of an extended anticondon. *Science* **218,** 646–652.