J. theor. Biol. (1992) **156**, 113–127

A Simulation of the Genetic Periodicities Modulo 2 and 3 with Processes of Nucleotide Insertions and Deletions

DIDIER G. ARQUÈS†§ AND CHRISTIAN J. MICHEL‡‖

† *Université de Franche-Comté, Laboratoire d'Informatique de Besançon, Unité Associée CNRS No* 822, 16 *route de Gray, F*-25030 *Besançon, France and* ‡ *Friedrich Miescher Institut, Bioinformatic Group, Mattenstrasse* 22, *P.O. Box* 2543, *CH*-4002 *Basel, Switzerland*

(*Received on* 12 *July* 1991, *Accepted on* 9 *September* 1991)

Recently, a new genetic process termed RNA editing has been identified showing insertions and deletions of nucleotides in particular RNA molecules. On the other hand, there are a few non-random statistical properties in genes: in particular, the periodicity modulo 3 (P3) associated with an open reading frame, the periodicity modulo 2 (P2) associated with alternating purine/pyrimidine stretches, the $YRY(N)_6YRY$ preferential occurrence (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine, $N = R$ or Y) representing a "code" of the DNA helix pitch, etc.

The problem investigated here is whether a process of the type RNA editing can lead to the non-random statistical properties commonly observed in genes. This paper will show in particular that:

—The process of insertions and deletions of mononucleotides in the initial sequence $[YRY(N)_3]^*$ [series of $YRY(N)_3$] can lead to the periodicity modulo 2 (P2).

—The process of insertions and deletions of trinucleotides in the initial sequence $[YRY(N)_6]^*$ [series of $YRY(N)_6$] can lead to the periodicity modulo 3 (P3) and the $YRY(N)_6YRY$ preferential occurrence.

—Furthermore, these two processes lead to a strong correlation with the reality, namely the mononucleotide insertion/deletion process, with the 5′ eukaryotic regions and the trinucleotide insertion/deletion process, with the eukaryotic protein coding genes.

1. Introduction

Recently, a new genetic process termed *RNA editing* has been identified showing *insertions and deletions of nucleotides in particular RNA molecules* (Benne *et al.*, 1986; reviews in Benne, 1989; Feagin, 1990; Simpson, 1990; Stuart, 1991; Cech, 1991). The substitution aspect of RNA editing will not be considered here. Even if the number of actual studies about the insertion/deletion aspect of RNA editing is still too small to state definitive and precise rules, several interesting features have been reported:

§ Author to whom correspondence should be addressed.

‖ Present address: Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie Belfort-Montbéliard, BP 527, 90016 Belfort, France.

0022-5193/92/090113 + 15 $03.00/0

(1) Editing is not universal and is only observed in particular genes such as: in mitochondrial transcripts of the kinetoplastid protozoa (Benne *et al.*, 1986 and others), recently those of *Physarum polycephalum* (Mahendran *et al.*, 1991) and also in a few non-mitochondrial systems, e.g. in paramyxovirus (Thomas *et al.*, 1988). However, editing could have been a general mechanism of gene expression (Benne, 1989: 137) or gene modification (Simpson, 1990) in primitive genetic systems.

(2) Two types of editing have been identified: insertions and deletions (e.g. Feagin *et al.*, 1988; Shaw *et al.*, 1988) or only insertions (e.g. Benne *et al.*, 1986), so far deletions are always accompanied by insertions. At one editing site, one or several nucleotides (up to 8) can be inserted or deleted (e.g. Feagin *et al.*, 1988; Shaw *et al.*, 1988). Guide RNAs may provide the nucleotides for the insertions (shown with the nucleotide U: Blum *et al.*, 1991).

(3) The editing number is variable:
   —Between the protein coding genes: e.g. in *Trypanosoma brucei*, editing of the cytochrome oxidase II entails the addition of only four uridines (Benne *et al.*, 1986) while more than 50% of the cytochrome oxidase III is produced by editing at multiple sites spread across the entire transcript (Feagin *et al.*, 1988).
   —Between the species: e.g. this cytochrome oxidase III extensively edited in *T. brucei* is less edited in *Leishmania tarentolae* and *Crithidia fasciculata* (Shaw *et al.*, 1988).
   —Between protein coding genes and the 5′ and 3′ regions: Editing also occurs in the 5′ and 3′ regions but with a reduced extent compared to the protein coding genes (e.g. Feagin *et al.*, 1988; Shaw *et al.*, 1988).
   —For a given gene, several partially edited forms are observed (Benne, 1989; Decker & Sollner-Webb, 1990).

(4) No specific rule has been identified for the editing site (Benne, 1989) and recently an editing model with random sites has been proposed (Decker & Sollner-Webb, 1990).

(5) Insertions and deletions classically involve the nucleotide U, but recently the nucleotide C was identified in an editing process (Mahendran *et al.*, 1991). An insertion process of the nucleotide G exists in the paramyxovirus P transcript (Thomas *et al.*, 1988). Vaccinia virus RNA leader sequences contain oligo A sequences not encoded in the viral genome (Schwer & Stunnenberg, 1988).

In fact, the biological elements used for RNA editing are common enzymes (endonucleases, exonucleases, transferases of nucleotides and ligases) leading therefore to a great number of possible combinations of nucleotide insertions and deletions, from which probably only a few have yet been observed.

On the other hand, there are a few non-random statistical properties in genes: in particular, the periodicity modulo 3 (P3) associated with an open reading frame (Fickett, 1982; Arquès & Michel, 1987*a*, *b*, 1990*a* and defined below), the periodicity modulo 2 (P2) associated with alternating purine/pyrimidine stretches (Arquès & Michel, 1987*c*, 1990*a* and defined below), the $YRY(N)_6YRY$ preferential occurrence (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine, $N = R$ or Y)

representing a "code" of the DNA helix pitch (Arquès & Michel, 1987*b*, 1990*b* and defined below), etc.

The issue investigated here is whether a process of the type RNA editing can lead to the non-random statistical properties commonly observed in genes. This paper will show in particular that:

—The process of insertions and deletions of mononucleotides in the initial sequence $[YRY(N)_3]$*[series of $YRY(N)_3$] can lead to the periodicity modulo 2 (P2).

—The process of insertions and deletions of trinucleotides in the initial sequence $[YRY(N)_6]$* [series of $YRY(N)_6$] can lead to the periodicity modulo 3 (P3) and the $YRY(N)_6YRY$ preferential occurrence.

—Furthermore, these two processes lead to a strong correlation with the reality, namely the mononucleotide insertion/deletion process, with the 5′ eukaryotic regions and the trinucleotide insertion/deletion process, with the eukaryotic protein coding genes.

2. Method

2.1. PRELIMINARIES: A STATISTICAL FUNCTION ALLOWING THE IDENTIFICATION OF THE GENETIC PERIODICITIES MODULO 2 (P2) AND MODULO 3 (P3)

Let F be a gene population with $n(F)$ DNA sequences. Let s be a sequence in F with a length $l(s)$. Let the i-motif $m_i = YRY(N)_iYRY$ with $i \in [1, 99]$, be two trinucleotides YRY separated by any i bases N. 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, only the first $l(s) - 104$ $[= l(s) - (99 + 6) + 1]$ bases of s are examined ($99 + 6$ is the maximal length of m_i). The occurrence probability $o_i(s)$ of m_i for s, is then equal to $c_i(s)/[l(s) - 104]$, i.e. the ratio of the counter by the total number of current bases read. The occurrence probability $p_i(F)$ of m_i for F, is finally equal to $[\sum_{s \in F} o_i(s)]/n(F)$. *For each population* F, *the statistical function* $i \rightarrow p_i(F)$ *giving the mean probability that* YRY *occurs* i *bases after itself, is represented as a curve* C(F). In order to have a sufficient number of m_{99} occurrences, the function is applied to sequences having a minimal length of 250 bases. This function $p_i(F)$ allows the identification of the periodicities both modulo 2 (P2) and modulo 3 (P3) in genes (Arquès & Michel, 1990*a*). Furthermore, this function $p_i(F)$ is not only important to observe the two periodicities but also to show the preferential occurrence of the motif $YRY(N)_6YRY$ in genes (Arquès & Michel, 1987*b*, 1990*b*).

Two gene populations F: the eukaryotic protein coding genes $F = CEUK$ and the 5′ eukaryotic regions $F = N5EUK$, obtained from the release 21 of the EMBL Nucleotide Sequence Data Library in the same way as described in the previous works (see e.g. Arquès & Michel, 1990*a*), have non-random statistical properties in their curve $C(F)$ which can be surprisingly simulated with simple processes of nucleotide insertions and deletions:

(1) The function $p_i(F)$ applied in eukaryotic protein coding genes $F=CEUK$ (8202 sequences, 7867 kb) shows in the curve $C(CEUK)$ [Fig. 1(a)]:
—The periodicity modulo 3 (P3):

$$p_i(CEUK) > \max\{p_{i-1}(CEUK), p_{i+1}(CEUK)\}$$

with $i \in [1, 98]$ and $i \equiv 0[3]$ ($i=0+3n$, n integer).
—A maximal value at $i=6$ [$YRY(N)_6YRY$ preferential occurrence] associated with this periodicity P3:

$$p_6(CEUK) > p_i(CEUK) \quad \text{with } i \in [1, 99] \text{ and } i \neq 6.$$

(2) The function $p_i(F)$ applied in the 5′ eukaryotic regions $F=N5EUK$ (2172 sequences, 1615 kb) shows in the curve $C(N5EUK)$ [Fig. 1(b)]:

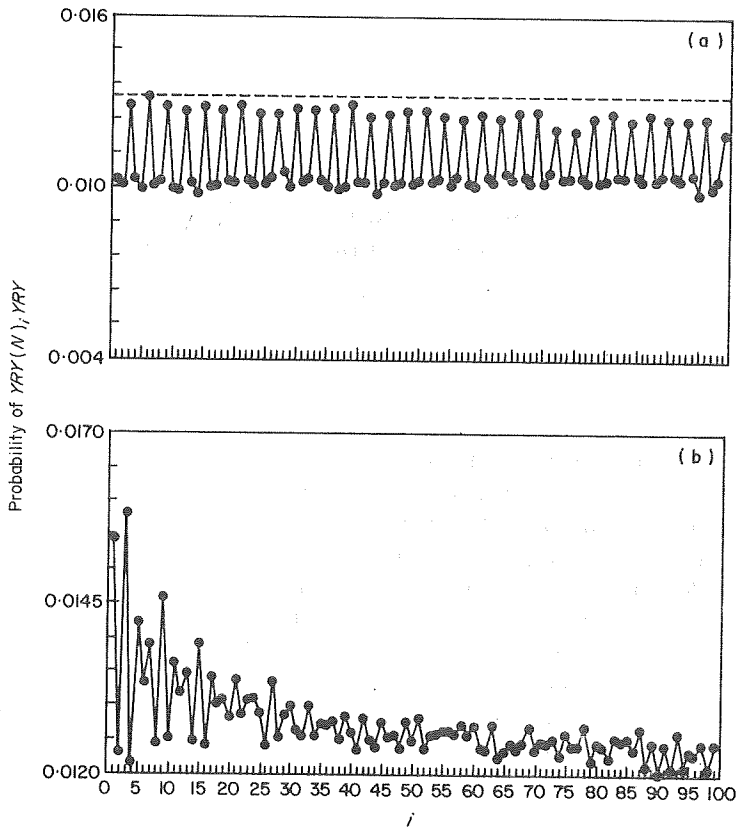


FIG. 1. Identification in genes of the two periodicities modulo 2 (P2) and modulo 3 (P3) with the statistical function $p_i(F)$ (see section 2). The horizontal axis represents the number i of bases N between two trinucleotides YRY [i-motif $YRY(N)_iYRY$] with $i \in [1, 99]$. The vertical axis represents the statistical function $p_i(F)$ in the following populations F: (a) eukaryotic protein coding genes CEUK showing the periodicity P3 (modulo 3) in the range [1, 98] and the maximal value at $i=6$; (b) 5′ eukaryotic regions $N5$EUK showing the periodicity P2 (modulo 2) in the range [1, 23] and the maximal value at $i=3$.

—The periodicity modulo 2 (P2):

$$p_i(N5\text{EUK}) > \max\{p_{i-1}(N5\text{EUK}), p_{i+1}(N5\text{EUK})\}$$

with $i \in [1, 23]$ and $i \equiv 1[2]$ ($i = 1 + 2n$, n integer).

—A maximal value at $i = 3$ associated with this periodicity P2:

$$p_3(N5\text{EUK}) > p_i(N5\text{EUK}) \quad \text{with } i \in [1, 99] \text{ and } i \neq 3.$$

Note: In *N*5EUK, the periodicity P2 with $i \equiv 1[2]$ in the range [1, 23] does not show directly the $YRY(N)_6YRY$ preferential occurrence because $6 \not\equiv 1[2]$, but the motif $YRY(N)_6YRY$ is only hidden as it has the highest frequency in the lowest curve $i \equiv 0[2]$ and the deletion of the large alternating purine/pyrimidine stretches (at the origin of the periodicity P2 with $i \equiv 1[2]$) leads again to the $YRY(N)_6YRY$ preferential occurrence [$p_6(F = N5\text{EUK})$ maximal for all i; Arquès & Michel, 1990*a*].

Remark: Due to the law of large numbers (Arquès & Michel, 1990*b*), the curve $C(F)$ of a gene population F keeps the main and non-random statistical properties (e.g. the periodicities, the maximal and minimal values, etc) even if new genes are available in the population F. For example:

—the periodicity P3 and the $YRY(N)_6YRY$ preferential occurrence were observed in the curve $C(C\text{EUK})$ of the population *C*EUK obtained from the EMBL release 10 [2271 sequences, 1750 kb; Fig. 1(a) of Arquès & Michel, 1987*b*];

—the periodicity P2 and the maximal value at $i = 3$ were observed in the curve $C(N5\text{EUK})$ of the population *N*5EUK obtained from the EMBL release 18 [1808 sequences, 1268 kb: Arquès & Michel, 1990*a*: Fig. 1(a)], etc.

These non-random statistical properties identified are important because they are stable from a statistical point of view. The next research step is naturally their understanding by the development of simulation models.

2.2. PROCESSES OF NUCLEOTIDE INSERTIONS AND DELETIONS STUDIED

Inspired by the existence of RNA editing in genes, a class of models has been investigated analysing the transformations of an initial sequence subjected to operations of nucleotide insertions and deletions. The models developed here are restricted to the following conditions, chosen in order to begin with simple models, to obtain from these simple models properties which can be used later to develop more precise models, to reduce the great number of possible combinations and finally, to take into account some features of RNA editing presented in the introduction:

(1) The initial sequence is the particular sequence $(O)^*$ = series of the oligonucleotide O, an oligonucleotide being a series of a few (about ten) nucleotides, e.g. $[YRY(N)_3]^* = YRYNNNYRYNNN \ldots$.

(2) The initial sequence is subjected to an insertion/deletion process with steps, so that at each step, one insertion of one or several nucleotides and one deletion of one or several nucleotides occur in the sequence. The length of the inserted bases is equal to the length of the deleted ones. This condition allows us to have sequences with the same length during the insertion/deletion

process. An insertion process without deletion has also been studied (see section 4). This condition takes into account the features of RNA editing described in the point 2 of the introduction.

(3) The site of insertions and deletions in the sequence is random (case of a random RNA editing: see the points 3 and 4 of the introduction).

(4) The type (R or Y) of the inserted bases is random (several nucleotides act in RNA editing: see the point 5 of the introduction).

In order to get significant results, a simulated population S having 500 initial sequences (O^*) of 2000 base length, is in fact generated in such a way, that the R percentage is equal to the Y percentage (50%) in any sequence $(O)^*$ of S (step 0). The computations obtained with such a sample of 1 million bases, are precise (i.e. there is no random fluctuations in the calculus of the probabilities: a sample having 100 sequences of 1000 base length leads to similar results). Then, this population S is subjected to an insertion/deletion process (defined above) at k_{max} steps. At each step k, $0 \leq k \leq k_{max}$, this process is studied by applying the statistical function $i \rightarrow p_i(S)$ in S [the same function as defined in section 2.1, i.e. the function analysing the mean occurrence frequency of the motif $YRY(N)_iYRY$ in S] and represented as a curve $C_k(S)$. Any simulated curve $C_k(S)$ can be compared with the real curve $C(F)$ because they result from the same function p_i. Remember that this function p_i is important because it allows us to identify non-random genetic properties which can be associated with a biological meaning, e.g. the periodicity modulo 3 reveals an open reading frame, the periodicity modulo 2, a gene with alternating purine/pyrimidine stretches, the motif $YRY(N)_6YRY$, a "code" of the DNA helix pitch, etc (see Arquès & Michel, 1987*a*, *b*, *c* for the biological details).

These four conditions being chosen, several hundreds of models were analysed by varying the choice of the initial sequence, the type of the process, etc. Two of them, although these insertion/deletion processes have strong random components concerning the site and the type of the bases inserted (the conditions 3 and 4), led to interesting and surprising results which are precisely described in the section 3 below.

3. Results

3.1. SIMULATION OF THE PERIODICITY MODULO 2 WITH THE MONONUCLEOTIDE INSERTION/DELETION PROCESS IN THE INITIAL SEQUENCE $[YRY(N)_3]^*$

A simulated population S, having 500 initial sequences $[YRY(N)_3]^*$ of 2000 base length, is generated by random specification of the $(N)_3$ bases with a R percentage of 66·66% and a Y percentage of 33·33% (step 0) [in order to have the same percentages of R and Y in the sequences $(YRY(N)_3)^*$]. Then, this simulated population S is subjected to a mononucleotide insertion/deletion process, i.e. one insertion of a mononucleotide and one deletion of a mononucleotide per sequence per step.

3.1.1. *Curve* $C_0(S)$ *at step* 0: *Fig.* 2(*a*)

Before the insertion/deletion process, the curve $C_0(S)$ [curve obtained by applying the function $p_i(S)$ in S] is constituted of four horizontal lines Δ_1, Δ_2, Δ_3 and Δ_4 of points in decreasing ordinate:

Δ_1: points $[i, p_i(S)]$ with $i \equiv 3[6]$,
Δ_2: points $[i, p_i(S)]$ with $i \equiv 1, 5[6]$,
Δ_3: points $[i, p_i(S)]$ with $i \equiv 0[6]$,
Δ_4: points $[i, p_i(S)]$ with $i \equiv 2, 4[6]$.

This decomposition is explained in Tables 1 and 2.

There is no maximal value at $i=3$ because the point $[3, p_3(S)]$ on the highest line Δ_1 cannot be differentiated from the other points $[i, p_i(S)]$ with $i \equiv 3[6]$: $[9, p_9(S)]$, $[15, p_{15}(S)]$, etc.

3.1.2. *Curve* $C_1(S)$ *at step* 1: *Fig.* 2(*b*)

After one mononucleotide insertion/deletion, there is a maximal value at $i=3$ in the curve $C_1(S)$ [curve obtained by applying the function $p_i(S)$ in S after one insertion of a mononucleotide and one deletion of a mononucleotide per sequence].

The highest value at $i=3$ and the decreasing slope of the highest line Δ_1 are explained by the fact that 1 mononucleotide insertion (or deletion) in a sequence $[YRY(N)_3]$* destroys only one subsequence $YRY(N)_3YRY$, but two subsequences $YRY(N)_9YRY$, three subsequences $YRY(N)_{15}YRY$, etc (see Table 3). Therefore, the values $p_i(S)$ with $i \equiv 3[6]$ decrease since i increases. A similar reasoning can be used to describe the modifications of the other lines.

3.1.3. *Curve* $C_{15}(S)$ *at step* 15: *Fig.* 2(*c*) (*points joined in one curve*)

By increasing the number of insertions and deletions of mononucleotides, the two lines Δ_1 and Δ_2 become curves with decreasing slopes, while the two lines Δ_3 and Δ_4, curves with increasing slopes. This process keeps the maximal value at $i=3$ and leads to the periodicity P2 (modulo 2) in the range [1, 23].

Furthermore, the simulated curve $C_{15}(S)$ (the points are joined in one curve in order to facilitate the comparison with a real curve) is strongly similar to the real curve $C(N5\text{EUK})$ of the 5′ eukaryotic regions [Fig. 1(b)] as these two curves have the periodicity P2 in the range [1, 23], the maximal value at $i=3$ and also four obvious subcurves modulo 6:

(1) $i=3$, 9, 15, 21, 27 and 33;
(2) $i=1$, 5, 7, 11, 13, 17, 19, 23 and 25;
(3) $i=6$, 12, 18 and 24;
(4) $i=2$, 4, 8, 10, 14 and 16.

All these naturally appearing curves join modulo 6 periodic sets of i values deduced from the four lines Δ_1, Δ_2, Δ_3 and Δ_4.

Note: These four subcurves modulo 6 were not initially observed with the study of the real curve *N*5EUK but were identified with the simulated curve $C_{15}(S)$.

It should be stressed that even if the simulated curve $C_{15}(S)$ has the main statistical properties of the real curve $C(N5\text{EUK})$ (i.e. the most important biological meaning), the curve $C_{15}(S)$ is not necessarily identical to the curve *N*5EUK because:

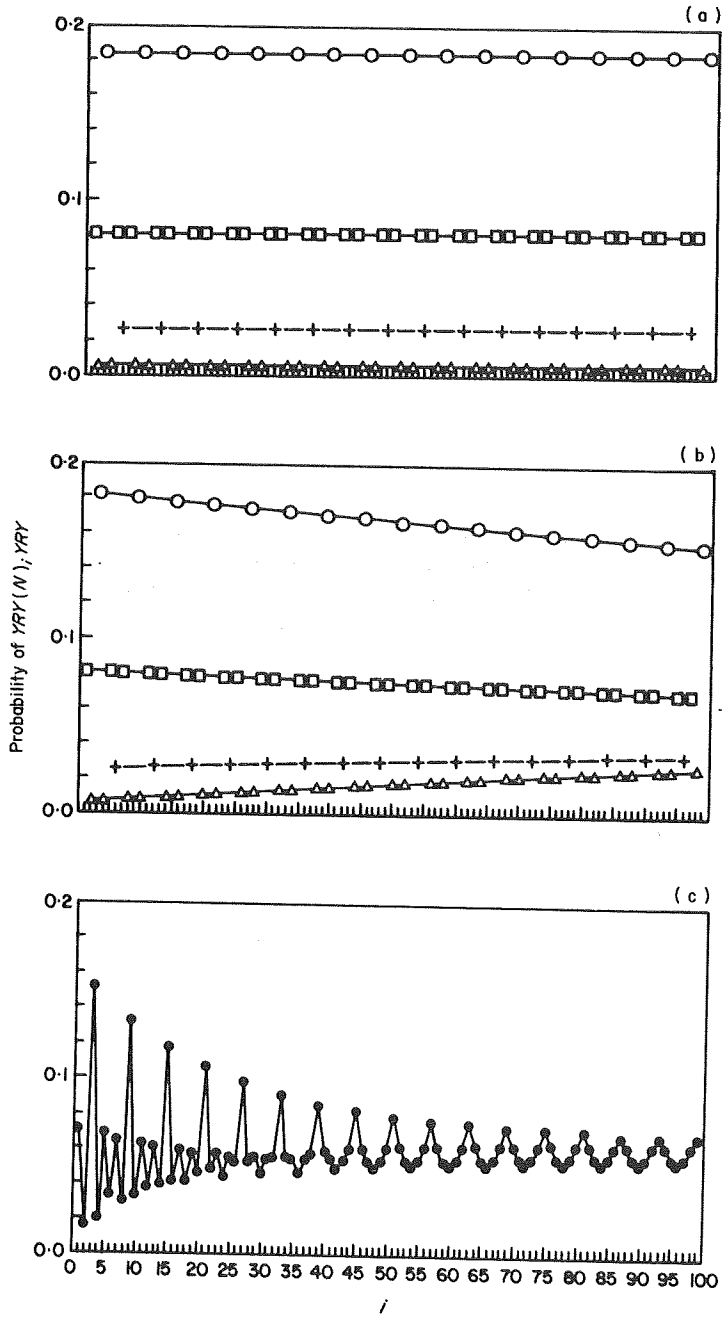


FIG. 2. Process of insertions and deletions of mononucleotides in the initial sequence $[YRY(N)_3]^*$ shown with the statistical function $i \rightarrow p_i(S)$ at the following steps (see section 2): (a) step 0; (b) step 1; (c) step 15 (points joined in one curve): this simulated curve is similar to the real curve *N*5EUK [Fig. 1(b)].

TABLE 1

Probability of $YRY(N)_3YRY$ *in the sequence* $[YRY(N)_3]^*$: *example of a complete calculus (the probability to have* R *in* N *is equal to* 2/3, *to have* Y *in* N, 1/3)

$[YRY(N)_3]^*$:	*YRYNNNYRYNNNYRY*...	Probability
$YRY(N)_3YRY$ location 1	*YRYNNNYRY*	$(1/6)\times 1$
$YRY(N)_3YRY$ location 2	*YRYNNNYRY*	$(1/6)\times 0$
$YRY(N)_3YRY$ location 3	*YRYNNNYRY*	$(1/6)\times(2/3)^2\times(1/3)^2$
$YRY(N)_3YRY$ location 4	*YRYNNNYRY*	$(1/6)\times(2/3)^2\times(1/3)^4$
$YRY(N)_3YRY$ location 5	*YRYNNNYRY*	$(1/6)\times(2/3)^2\times(1/3)^2$
$YRY(N)_3YRY$ location 6	*YRYNNNYRY*	$(1/6)\times 0$
		$\Sigma=(1/6)\times(805/729)$ $\simeq 0{\cdot}184$

TABLE 2

Probability of $YRY(N)_iYRY$ *in the sequence* $[YRY(N)_3]^*$: *final results. The occurrence probability of* $YRY(N)_{i+6k}YRY$ *is equal to the occurrence probability of* $YRY(N)_iYRY$, $i\in[0, 5]$ [*by the modulo* 6 *invariance of the* $(YRY(N)_3)^*$ *sequence*]. *Four different probabilities are obtained, giving the four lines* Δ_1, Δ_2, Δ_3 *and* Δ_4

Probability of *YRYYRY*	in $[YRY(N)_3]^*=(1/6)\times(4/27)$	$\simeq 0{\cdot}025$
Probability of $YRY(N)_1YRY$	in $[YRY(N)_3]^*=(1/6)\times(40/81)$	$\simeq 0{\cdot}082$
Probability of $YRY(N)_2YRY$	in $[YRY(N)_3]^*=(1/6)\times(8/243)$	$\simeq 0{\cdot}005$
Probability of $YRY(N)_3YRY$	in $[YRY(N)_3]^*=(1/6)\times(805/729)$ (see Table 1)	$\simeq 0{\cdot}184$
Probability of $YRY(N)_4YRY$	in $[YRY(N)_3]^*=(1/6)\times(8/243)$	$\simeq 0{\cdot}005$
Probability of $YRY(N)_5YRY$	in $[YRY(N)_3]^*=(1/6)\times(40/81)$	$\simeq 0{\cdot}082$

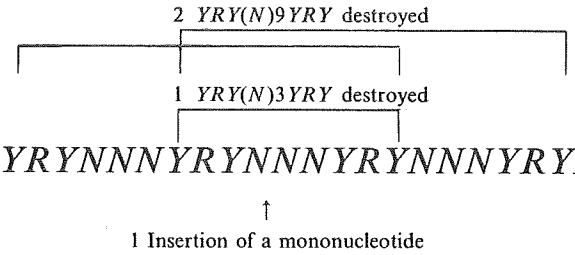
(1) The description of a curve shape using features is a known difficult problem of pattern recognition which cannot be solved by only a few features.

(2) The simulated curve shape is regular with a model using the simple conditions chosen. Such a simple model cannot simulate completely the biological reality depending on a great number of factors, in the same way the first terms of development of a function in series cannot reveal the totality of the function.

In conclusion, a process of insertions and deletions of mononucleotides similar to the RNA editing, can simulate an important feature observed in genes: the periodicity modulo 2. Now the question arises: Is it possible to simulate a periodicity modulo 3 with such a class of processes? On the other hand, RNA editing can create a translatable messenger RNA from the transcript of a gene completely lacking a

TABLE 3

Effects of one inertion of a mononucleotide in the sequence $[YRY(N)_3]^*$



functional open reading frame (Mahendran *et al.*, 1991). Therefore, is it possible to find an insertion/deletion process leading to a sequence with a periodicity modulo 3 (i.e. with an open reading frame) from an initial sequence without periodicity modulo 3 (i.e. without open reading frame)? The section 3.2 below shows the existence of such a process.

3.2. SIMULATION OF THE PERIODICITY MODULO 3 WITH THE TRINUCLEOTIDE INSERTION/DELETION PROCESS IN THE INITIAL SEQUENCE $[YRY(N)_6]^*$

A simulated population S', having 500 initial sequences $[YRY(N)_6]^*$ of 2000 base length, is generated by random specification of the $(N)_6$ bases with a R percentage of 58·33% and with a Y percentage of 41·66% (step 0) [in order to have the same percentages of R and Y in the sequences $(YRY(N)_6)^*$]. Then, this simulated population S' is subjected to a trinucleotide insertion/deletion process, i.e. one insertion of a trinucleotide and one deletion of a trinucleotide per sequence per step.

Note: One trinucleotide insertion (resp. deletion) can be seen as three mononucleotide insertions (resp. deletions) at one site.

3.2.1. *Curve* $C_0(S')$ *at step* 0: *Fig.* 3(*a*)

Before the insertion/deletion process, the curve $C_0(S')$ is made of horizontal points with abscissa invariant modulo 9 [due to the invariance modulo 9 of the $(YRY(N)_6)^*$ sequence]. It is important to stress that a significant modulo 9 periodicity was already identified, but not explained in eukaryotic protein coding genes (Arquès & Michel, 1987*a*). There are five such different horizontal lines D_1, D_2, D_3, D_4 and D_5 in decreasing ordinate:

D_1 : points $[i, p_i(S')]$ with $i \equiv 6[9]$,
D_2 : points $[i, p_i(S')]$ with $i \equiv 4, 8[9]$,
D_3 : points $[i, p_i(S')]$ with $i \equiv 1, 2[9]$,
D_4 : points $[i, p_i(S')]$ with $i \equiv 0, 3[9]$,
D_5 : points $[i, p_i(S')]$ with $i \equiv 5, 7[9]$.

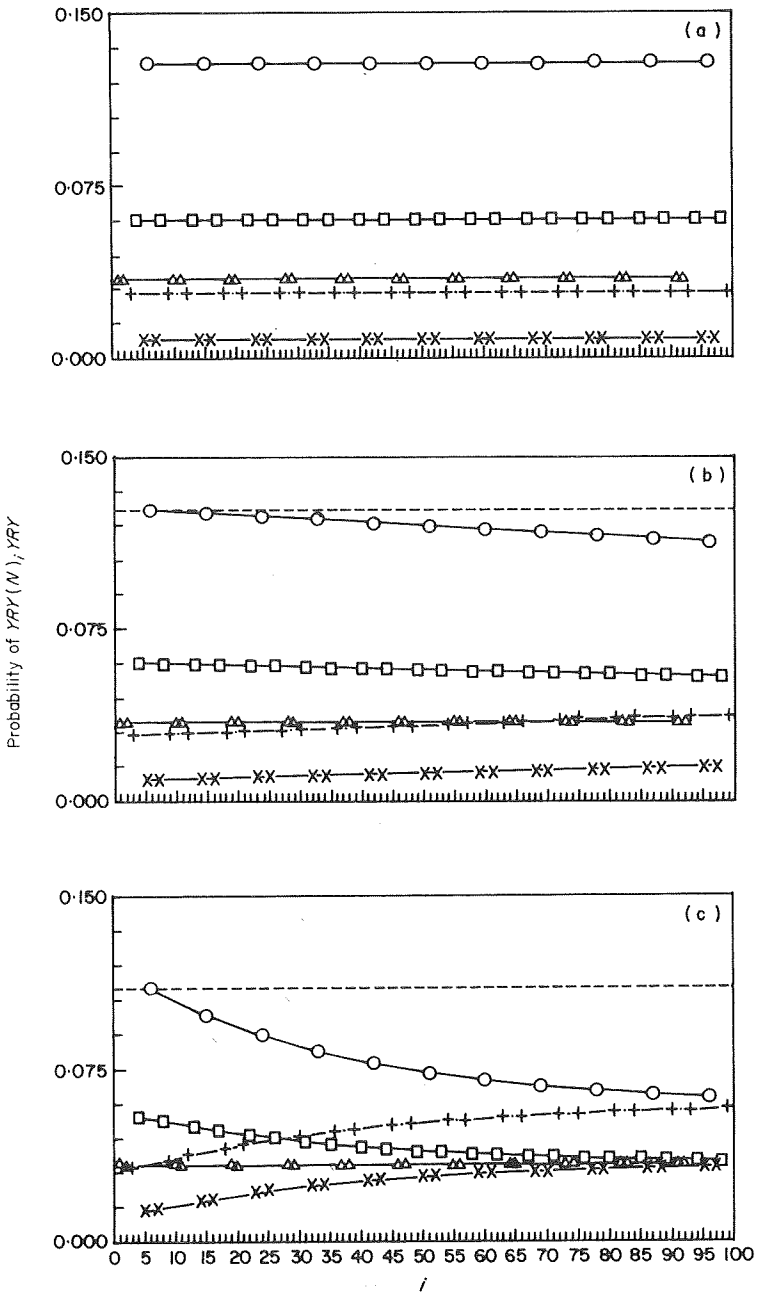


FIG. 3. Process of insertions and deletions of trinucleotides in the initial sequence $[YRY(N)_6]^*$ shown with the statistical function $i \to p_i(S')$ at the following steps (see section 2): (a) step 0; (b) step 1; (c) step 10; (d) step 150 (points joined in one curve): this simulated curve is similar to the real curve CEUK [Fig. 1(a)].

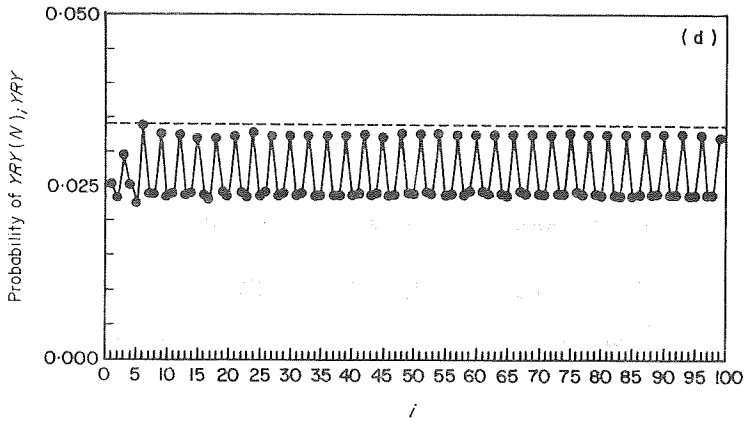


FIG. 3—continued

The proof of this decomposition is identical to the one for the curve $C_0(S)$ given in the section 3.1.1.

There is no maximal value at $i=6$ [$YRY(N)_6YRY$ preferential occurrence] because the point $[6, p_6(S')]$ on the highest line D_1 cannot be differentiated from the other points $[i, p_i(S')]$ with $i \equiv 6[9]$: $[15, p_{15}(S')]$, $[24, p_{24}(S')]$, etc.

There is no periodicity P3 (modulo 3) because the points $[i, p_i(S')]$ with $i \equiv 0, 3[9]$ (D_4) have lower values compared to the points $[i, p_i(S')]$ with $i \equiv 4, 8[9]$ (D_2) and with $i \equiv 1, 2[9]$ (D_3).

3.2.2. *Curve* $C_1(S')$ *at step* 1: *Fig.* 3(*b*)

After one trinucleotide insertion/deletion, a maximal value at $i=6$ is obtained. The highest value at $i=6$ [$YRY(N)_6YRY$ preferential occurrence] and the decreasing slope of the highest line D_1 are explained by the fact that one trinucleotide insertion (or deletion) in a sequence $[YRY(N)_6]^*$ destroys only one subsequence $YRY(N)_6YRY$, but two subsequences $YRY(N)_{15}YRY$, three subsequences $YRY(N)_{24}YRY$, etc (proof identical to the one given in section 3.1.2). Therefore, the values $p_i(S')$ with $i \equiv 6[9]$ decrease since i increases.

At this step, there is still no periodicity P3 because the points $[i, p_i(S')]$ with $i \equiv 0, 3[9]$ (D_4) have lower values compared to the points $[i, p_i(S')]$ with $i \equiv 4, 8[9]$ (D_2).

3.2.3. *Curves* $C_{10}(S')$ *at step* 10: *Fig.* 3(*c*)

By increasing the number of insertions and deletions of trinucleotides, the five lines D_1, D_2, D_3, D_4 and D_5 are gathering into two curves. The top curve is constituted of the two lines D_1 and D_4, i.e. of the points $[i, p_i(S')]$ with $i \equiv 0[3]$. The bottom curve is constituted of the three lines D_2, D_3 and D_5, i.e. of the points $[i, p_i(S')]$ with $i \equiv 1, 2[3]$. This process leads to the maximal value at $i=6$ and to the periodicity P3 in the range [30, 98] (not in the range [1, 30]).

3.2.4. *Curves* $C_{150}(S')$ *at step* 150: *Fig.* 3(*d*) (*points joined in one curve*)

The simulated curve $C_{150}(S')$ is strongly similar to the real curve C(CEUK) of eukaryotic protein coding genes [Fig. 1(a)] as these two curves have the periodicity P3 in the range [1, 98], the $YRY(N)_6YRY$ preferential occurrence and also two different sets of well-separated points [min $\{p_i(F), i \equiv 0[3]\} >$ max $\{p_i(F), i \equiv 1, 2[3]\}$ with $i \in [1, 98]$] which can be joined by nearly horizontal lines.

4. Discussion

The function $p_i(F)$, by varying i between 1 and 99, can lead to 99! (10^{156}) possible curve shapes. On the other hand, the shape of the simulated curve for a given insertion/deletion process is unique. Therefore, the understanding of a curve, or even of only a few points, is a difficult problem.

The maximal value $p_m(F)$ in the curve $C(F)$ can be explained by one (at least one is necessary) insertion (or deletion) of any series of nucleotides (mono, di, tri, etc) in the initial sequence $[YRY(N)_m]^*$ obtained by the concatenation in series of the oligonucleotide $YRY(N)_m$: generalization of the proofs explaining (i) the maximal value $p_3(F)$ with one insertion (or deletion) of a mononucleotide in the initial sequence $[YRY(N)_3]^*$ (see section 3.1.2) and (ii) the maximal value $p_6(F)$ with one insertion (or deletion) of a trinucleotide in the initial sequence $[YRY(N)_6]^*$ (see section 3.2.2.). Consequently, the maximal value $p_m(F)$ is independent on the type of the insertion/deletion process but dependent on the initial sequence which cannot be random.

A periodicity in the curve $C(F)$ depends both on the insertion/deletion process and on the initial sequence, but no clear relation has yet been found (work in progress).

The number of steps in the insertion/deletion process for the simulation of the 5′ eukaryotic regions *N*5EUK (15 steps) is less than the one for the simulation of the eukaryotic protein coding genes *C*EUK (150 steps), in agreement with the experimental results showing a reduced extent of RNA editing in *N*5EUK compared to *C*EUK (see the point 3 of section 1). In fact, the steps shown with the figures represent ranges of steps (in which the statistical properties are similar) because there is a continuous modification of all points in the simulated curves through the insertion/deletion process. The step range agrees with the existence of partially edited forms (see the point 3 of the introduction). Nevertheless, the upper limit of the step range for *N*5EUK is clearly less than the lower limit of the step range for *C*EUK (data not shown). Therefore for *N*5EUK, a substitution aspect of RNA editing, precisely random changes $R \rightarrow Y$ and $Y \rightarrow R$, must be associated with the insertion/deletion process because the absolute values in the curve $C_{15}(S)$ are greater (about ten times more) than the ones of the real curve C(*N*5EUK). In order to reach the real values by keeping the non-random statistical properties (i.e. periodicity P2 in the range [1, 23], maximal value at $i = 3$, etc), substitutions are necessary in the simulated population associated with the curve $C_{15}(S)$ with a maximal rate of order 1/2 substitution per specified base (R or Y) and with any substitution rate per

unspecified base N. The substitution process acts on the absolute values of $p_i(S)$ by decreasing the variations between the points $[i, p_i(S)]$ (by adding "noise") but not on the relative positions of the points $[i, p_i(S)]$ (i.e. not on the curve shape). Consequently, the substitution process can explain neither the periodicities modulo 2 and 3 nor a maximal value, e.g. the $YRY(N)_6YRY$ preferential occurrence cannot be explained by a substitution process.

There is an upper limit for the substitution rate in the specified base (R or Y) (see above) and also for the number of steps in the insertion/deletion process. This limit is reached when all $p_i(S)$ values are equal to 1/64: random situation which is not observed in the reality because the $p_i(F)$ points are not on a horizontal line (periodicities, maximum values at $i=3$ or 6, etc).

For the two insertion/deletion processes presented (which led to results), the insertion process without deletion has been investigated:

—The mononucleotide insertion process in the initial sequence $[YRY(N)_3]^*$ leads to similar results (periodicity modulo 2 and maximal value at $i=3$) compared to the mononucleotide insertion/deletion process in the initial sequence $[YRY(N)_3]^*$ (data not shown).

—The trinucleotide insertion process in the initial sequence $[YRY(N)_6]^*$ leads to similar results (periodicity modulo 3 and maximal value at $i=6$) compared to the trinucleotide insertion/deletion process in the initial sequence $[YRY(N)_6]^*$ (data not shown).

In conclusion, it is important to stress that the simple models developed here do not represent a perfect simulation of RNA editing, at least for two reasons: from a theoretical point of view, these models have simplifications (see the conditions chosen) and from an experimental point of view, the understanding of RNA editing will improve in future. The main purpose of this paper is to show that processes of nucleotide insertions and deletions with strong random components concerning the site and the type of the bases inserted (conditions 3 and 4), can unexpectedly lead to non-random properties observed in genes: in particular, the periodicities modulo 2 and 3 and the $YRY(N)_6YRY$ preferential occurrence. Furthermore, the two insertion/deletion processes presented led to a strong similarity with the eukaryotic protein coding genes and the 5′ eukaryotic regions. The elements of reasoning found here and the combination of different processes (e.g. association of a mononucleotide insertion/deletion process with a dinucleotide or trinucleotide one, etc) should lead to the development of more precise models having a stronger correlation with the genetic reality (better curve shape, additional gene populations simulated, etc).

We thank Professors Max Burger and Jacques Streith, Dr Christoph Nager, Thomas Nyffenegger and Nouchine Soltanifar for their advice. This work was supported by grants from the Unité Associée CNRS No 822 to D.G.A. and from the Friedrich Miescher Institute to C.J.M.

REFERENCES

ARQUÈS, D. G. & MICHEL, C. J. (1987*a*). *Math. Biosc.* **86**, 1–14.
ARQUÈS, D. G. & MICHEL, C. J. (1987*b*). *J. theor. Biol.* **128**, 457–461.

ARQUÈS, D. G. & MICHEL, C. J. (1987*c*). *Nucl. Acids Res.* **15**, 7581–7592.
ARQUÈS, D. G. & MICHEL, C. J. (1990*a*). *J. theor. Biol.* **143**, 307–318.
ARQUÈS, D. G. & MICHEL, C. J. (1990*b*). *Bull. Math. Biol.* **52**, 741–772.
BENNE, R. (1989). *Biochem. Biophys. Acta* **1007**, 131–139.
BENNE, R., VAN DEN BURG, J., BRAKENHOFF, J. P. J., SLOOF, P., VAN BOOM, J. H. & TROMP, M. C. (1986). *Cell* **46**, 819–826.
BLUM, B., STURM, N. R., SIMPSON, A. M. & SIMPSON, L. (1991). *Cell* **65**, 543–550.
CECH, T. R. (1991). *Cell* **64**, 667–669.
DECKER, C. J. & SOLLNER-WEBB, B. (1990). *Cell* **61**, 1001–1011.
FEAGIN, J. E. (1990). *J. biol. Chem.* **265**, 19373–19376.
FEAGIN, J. E., ABRAHAM, J. M. & STUART, K. (1988). *Cell* **53**, 413–422.
FICKETT, J. W. (1982). *Nucl. Acids Res.* **10**, 5303–5318.
MAHENDRAN, R., SPOTTSWOOD, M. R. & MILLER, D. L. (1991). *Nature, Lond.* **349**, 434–438.
SCHWER, B. & STUNNENBERG, H. G. (1988). *EMBO J.* **7**, 1183–1190.
SHAW, J. M., FEAGIN, J. E., STUART, K. & SIMPSON, L. (1988). *Cell* **53**, 401–411.
SIMPSON, L. (1990). *Science* **250**, 512–513.
STUART, K. (1991). *Trends Biochem. Sci.* **16**, 68–72.
THOMAS, S. M., LAMB, R. A. & PARTERSON, R. G. (1988). *Cell* **54**, 891–902.