J. theor. Biol. (1993) **161**, 329–342

Identification and Simulation of New Non-random Statistical Properties Common to Different Populations of Eukaryotic Non-coding Genes

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

† *Equipe de Biologie Théorique, Université de Franche-Comté, Laboratoire d'Informatique de Besançon,* 16 *route de Gray,* 25030 *Besançon,* ‡ *Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie de Belfort-Montbéliard, BP* 527, 90016 *Belfort and* § *Equipe de Biologie Théorique, Université de Franche-Comté, Laboratoire de Mathématique et Informatique,* 4 *rue des Frères Lumière,* 68093 *Mulhouse, France*

(*Received on* 11 *February* 1992, *Accepted in revised form on* 11 *July* 1992)

The autocorrelation function analysing the occurrence probability of the i-motif $YRY(N)_iYRY$ in genes allows the identification of mainly two periodicities modulo 2, 3 and the preferential occurrence of the motif $YRY(N)_6YRY$ (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine, $N = R$ or Y). These non-random genetic statistical properties can be simulated by an independent mixing of the three oligonucleotides $YRYRYR$, $YRYYRY$ and $YRY(N)_6$ (Arquès & Michel, 1990*b*). The problem investigated in this study is whether new properties can be identified in genes with other autocorrelation functions and also simulated with an oligonucleotide mixing model.

The two autocorrelation functions analysing the occurrence probability of the i-motifs $RRR(N)_iRRR$ and $YYY(N)_iYYY$ simultaneously identify three new non-random genetic statistical properties: a short linear decrease, local maxima for $i \equiv 3[6]$ ($i = 3, 9$, etc) and a large exponential decrease. Furthermore, these properties are common to three different populations of eukaryotic non-coding genes: 5′ regions, introns and 3′ regions (see section 2).

These three non-random properties can also be simulated by an independent mixing of the four oligonucleotides R^8, Y^8, $RRRYRYRRR$, $YYYRYRYYY$ and large alternating R/Y series. The short linear decrease is a result of R^8 and Y^8, the local maxima for $i \equiv 3[6]$, of $RRRYRYRRR$ and $YYYRYRYYY$, and the large exponential decrease, of large alternating R/Y series (section 3).

The biological meaning of these results and their relation to the previous oligonucleotide mixing model are presented in the Discussion.

1. Introduction

Non-random genetic statistical properties were able to be identified because a particular statistical function, termed the autocorrelation function, was defined (Arquès & Michel, 1987*b*; generalized below) in order to analyse in gene populations the occurrence probability of two identical trinucleotides T separated by any i bases N, i.e.

‖ Author to whom correspondence should be addressed.

0022-5193/93/070329+14 $08.00/0

the occurrence probability of i-motifs $T(N)_iT$, T being a trinucleotide on the alphabet $\{R, Y\}$ (R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine and $N = R$ or Y). This autocorrelation function is noted T-function. Indeed, the YRY-function, i.e. the autocorrelation function analysing the occurrence probability of the i-motif $YRY(N)_iYRY$ in genes, allows us to identify the two periodicities modulo 2, 3 and the preferential occurrence of the motif $YRY(N)_6YRY$ (Arquès & Michel, 1987*a*, *b*; 1990*a*, *b*).

The biological meanings of these properties were given in detail in our previous study (Arqués & Michel, 1987*b*; 1990*b*). Briefly, the periodicity modulo 2 reveals a gene with alternating R/Y series, the periodicity modulo 3, a gene with an open reading frame, and the preferential occurrence of the motif $YRY(N)_6YRY$, a "code" of the DNA helix pitch.

Although unexpected, we have recently proved that these non-random properties can be simulated by a mixing of the three oligonucleotides $YRYRYR$, $YRYYRY$ and $YRY(N)_6$, i.e. with a concatenation of words of a few ($\leq$ten) letters (Arquès & Michel, 1990*b*). Furthermore, this oligonucleotide mixing has mathematically been demonstrated to be independent (by first using a Markov mixing which is a more general mixing). Therefore, the independent mixing only depends on the probabilities associated with the three oligonucleotides. Finally, it was shown that random nucleotide mutations, i.e. random transformations of a nucleotide into another (in our case $R \rightarrow Y$ and $Y \rightarrow R$), cannot lead to a non-random property (mutations act on the absolute values of the T-function but not on its relative values). By giving a biological meaning to these mathematical/statistical results, a model of DNA sequence evolution has been proposed. Briefly, according to this model, genes first derive from a mixing of primitive oligonucleotides (root of the phylogenetic tree since evolution is commonly accepted to be divergent overall) as an independent mixing is the simplest of the possible mixings. Then, evolution led to the actual gene diversity (leaves of the phylogenetic tree) mainly by random nucleotide processes, such as mutations (see reviews in Kimura, 1987; Nei, 1987), insertions and deletions (a genetic process termed RNA editing which was recently identified; Benne *et al.*, 1986; Benne, 1989; Simpson, 1990; Cech, 1991), specifications (transformations of a nucleotide $\{R, Y\}$ into a nucleotide $\{A, C, G, T\}$ so that $R \rightarrow A$ or $R \rightarrow G$ and $Y \rightarrow C$ or $Y \rightarrow T$; Arquès & Michel, 1990*b*: 742) etc. In other words, the oldest evolutionary process is the independent mixing of only a few types of primitive oligonucleotides, followed later on mainly by random nucleotide processes. These two successive steps are the unavoidable consequence in that random mutations cannot explain the non-random properties existing in genes. We refer the reader to Arquès & Michel (1990*b*) for the mathematical proofs and the biological concepts.

These results further explain investigations to identify new properties with T-functions other than the YRY-function as YRY represents only one among eight of the R/Y trinucleotides (see also Arquès & Michel, 1990*b*: 766, lines 2–3). The RRR- and YYY-functions, i.e. the autocorrelation functions analysing the occurrence probability of the i-motifs $RRR(N)_iRRR$ and $YYY(N)_iYYY$ in eukaryotic non-coding genes (5′ and 3′ regions, introns), simultaneously identify three new non-random properties: a short linear decrease, local maxima for $i \equiv 3[6]$ and a large

exponential decrease (section 2). Furthermore, these properties can also be simulated by an independent mixing of the four oligonucleotides R^8, Y^8, *RRRYRYRRR*, *YYYRYRYYY* and large alternating R/Y series [for example, $(RY)_{50}$]. The short linear decrease is a result of R^8 and Y^8, the local maxima for $i \equiv 3[6]$, of *RRRYR-YRRR* and *YYYRYRYYY*, and the large exponential decrease, of large alternating R/Y series (section 3).

2. Identification of New Non-random Statistical Properties Common to Different Populations of Eukaryotic Non-coding Genes

2.1. INTRODUCTION

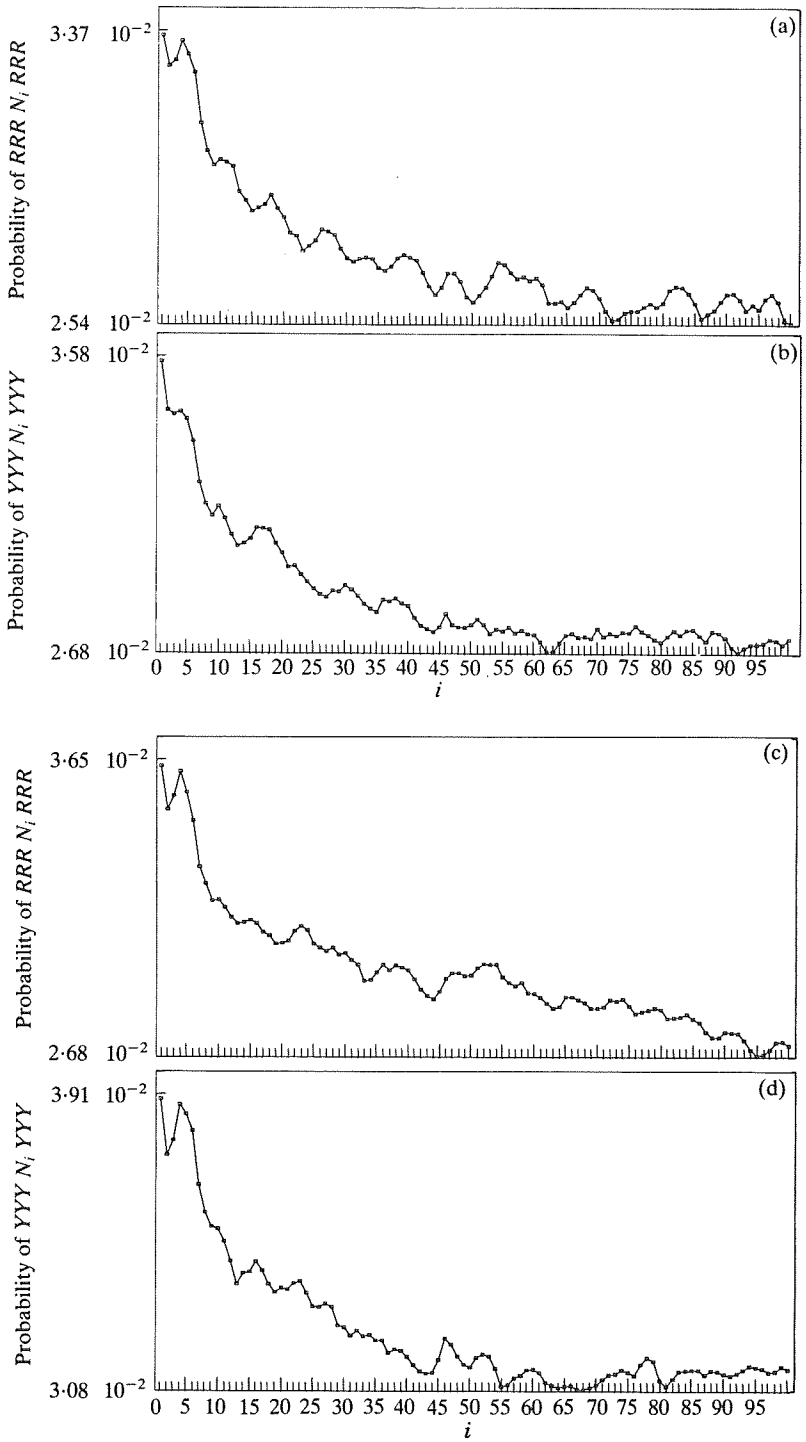
The problem investigated in this section is whether new properties can be identified with T-functions different from the *YRY*-function. The *RRR*- and *YYY*-functions will simultaneously identify the same three non-random properties in three different populations of eukaryotic non-coding genes. The trinucleotides *RRR* and *YYY* are "independent", *RRR* is also "independent" of *YRY* while *YYY* can only overlap *YRY* with one base. Therefore, the results obtained with the *RRR*-, *YYY*- and *YRY* functions cannot be deduced from each other.

2.2. METHOD

This method generalizes the previous one (Arquès & Michel, 1987*b*) to any trinucleotide on the purine/pyrimidine alphabet.

Let F be a gene population with $n(F)$ DNA sequences. Let s be a sequence in F with a length $l(s)$. Let T be a trinucleotide on the alphabet $\{R, Y\}$, R = purine = adenine or guanine, Y = pyrimidine = cytosine or thymine, i.e. $T = \{RRR, \ldots, YYY\}$. Let the i-motif $m_i(T) = T(N)_iT$, $N = R$ or Y and $i \in [0, 99]$, be two identical trinucleotides T separated by any i bases N. For each s of F, the counter $c_i(T; s)$ counts the occurrences of $m_i(T)$ in s. In order to count the $m_i(T)$ 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 $m_i(T)$]. The occurrence probability $o_i(T; s)$ of $m_i(T)$ for s, is then equal to $c_i(T; s)/[l(s) - 104]$, i.e. the ratio of the counter by the total number of current bases read. The occurrence probability $p_i(T; F)$ of $m_i(T)$ for F, is finally equal to $[\sum_{s \in F} o_i(T; s)]/n(F)$. *For a trinucleotide T and a population F, the autocorrelation function* $i \rightarrow p_i(T; F)$ *giving the mean occurrence probability that T occurs i bases after itself, is noted T-function and is represented as a curve* $C(T; F)$. In order to have a sufficient number of $m_{99}(T)$ occurrences, the T-function is applied to sequences having a minimal length of 250 bases.

The populations F of eukaryotic non-coding genes analysed here are: the 5′ eukaryotic regions F = N5EUK (405 sequences, 697 kb), the eukaryotic introns F = NIEUK (1016 sequences, 1181 kb) and the 3′ eukaryotic regions F = N3EUK (615 sequences, 1023 kb). They are obtained from the release 23 of the EMBL Nucleotide Sequence Data Library in the same way as previous studies (see for example, Arquès & Michel, 1990*a* for a description of data acquisitions). The T-function uses the trinucleotides $T = RRR$ and $T = YYY$. The curve $C(T; F)$ is represented as follows: (i) the abscissa shows the number i of bases N between 2 *RRR*, or 2 *YYY*, by varying i between 0


(a)
3·37 10^{-2}
2·54 10^{-2}
Probability of *RRR N*$_i$ *RRR*
(b)
3·58 10^{-2}
2·68 10^{-2}
Probability of *YYY N*$_i$ *YYY*
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
i
(c)
3·65 10^{-2}
2·68 10^{-2}
Probability of *RRR N*$_i$ *RRR*
(d)
3·91 10^{-2}
3·08 10^{-2}
Probability of *YYY N*$_i$ *YYY*
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
i

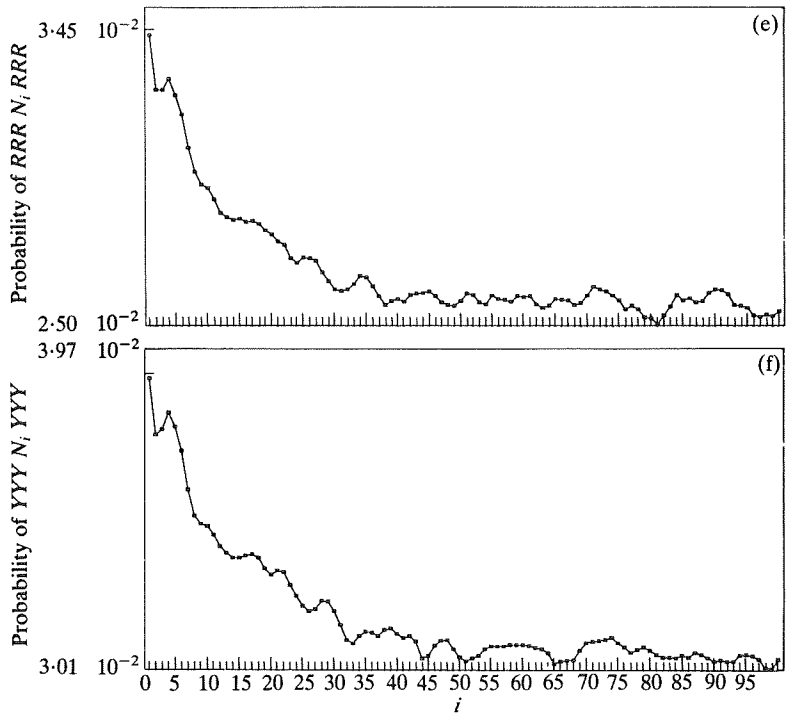


FIG. 1. Identification of new non-random statistical properties common to different populations of eukaryotic non-coding genes: large exponential decrease with a short linear decrease and local maxima for $i \equiv 3[6]$ ($i = 3, 9$, etc). The horizontal axis represents the number i, $i \in [0, 99]$, of any bases N between two identical trinucleotides T, $T = RRR$ or $T = YYY$, i.e. the i-motif $RRR(N)_iRRR$ or the i-motif $YYY(N)_iYYY$. The vertical axis represents the T-function, i.e. the curve $C(T; F)$ shows the occurrence probability of the i-motif $RRR(N)_iRRR$ or $YYY(N)_iYYY$ in the following populations F (see section 2.2). (a) Curve $C(T = RRR; F = N5EUK)$ showing the occurrence probability of the i-motif $RRR(N)_iRRR$ in the 5′ eukaryotic regions F = N5EUK (local maxima for $i = 3, 9$). (b) Curve $C(T = YYY; F =$ N5EUK) showing the occurrence probability of the i-motif $YYY(N)_iYYY$ in the 5′ eukaryotic regions $F =$ N5EUK (local maxima for $i = 3, 9, 15, 21, 27$). (c) Curve $C(T = RRR; F =$ NIEUK) showing the occurrence probability of the i-motif $RRR(N)_iRRR$ in the eukaryotic introns $F =$ NIEUK (local maxima for $i = 3, 9, 15, 21, 27$). (d) Curve $C(T = YYY; F =$ NIEUK) showing the occurrence probability of the i-motif $YYY(N)_iYYY$ in the eukaryotic introns $F =$ NIEUK (local maxima for $i = 3, 9, 15, 21, 27$). (e) Curve $C(T = RRR; F =$ N3EUK) showing the occurrence probability of the i-motif $RRR(N)_iRRR$ in the 3′ eukaryotic regions $F =$ N3EUK (local maxima for $i = 3, 9$). (f) Curve $C(T = YYY; F =$ N3EUK) showing the occurrence probability of the i-motif $YYY(N)_iYYY$ in the 3′ eukaryotic regions $F =$ N3EUK (local maxima for $i = 3, 9, 15, 21, 27$).

and 99; (ii) the ordinate gives the mean occurrence probability of $RRR(N)_iRRR$, or of $YYY(N)_iYYY$, in a gene population F.

2.3. RESULTS

The T-function is applied in:

the 5′ eukaryotic regions $F =$ N5EUK with $T = RRR$ [curve $C(RRR;$ N5EUK): Fig. 1(a)] and $T = YYY$ [curve $C(YYY;$ N5EUK): Fig. 1(b)];

the eukaryotic introns F=NIEUK with $T=RRR$ [curve $C(RRR$; NIEUK): Fig. 1(c)] and $T=YYY$ [curve $C(YYY$; NIEUK): Fig. 1(d)];

the 3′ eukaryotic regions F=N3EUK with $T=RRR$ [curve $C(RRR$; N3EUK): Fig. 1(e)] and $T=YYY$ [curve $C(YYY$; N3EUK): Fig. 1(f)].

A curve with 100 different points can lead to 100! (10^{158}) possible curve shapes. Unexpectedly, these six curves, obtained with two "independent" trinucleotides (*RRR* and *YYY*) and with three "independent" gene populations (N5EUK, NIEUK and N3EUK), have the same main non-random properties: a large decreasing curve (from $i=0$ to 99) of "exponential" type with a short linear decrease (from $i=0$ to 8). This exponential decrease has local maxima for $i\equiv 3[6]$ ($i=3+6n$, n integer): a peak at $i=3$ having the second highest value among the 100 points; small peaks at $i=9$, 15, 21 and 27 [obvious in Fig. 1(b)].

Note: These curves settle at a long distance of i [for $i\geq 50$ except for the Fig. 1(c)] to a variation around a constant value which is closed to the square of the trinucleotide frequency:

the constant value in Fig. 1(a) is $\approx 2{\cdot}54\times10^{-2}$ and the *RRR* frequency in N5EUK is equal to 0·1579, i.e. its square $=2{\cdot}49\times10^{-2}$;

the constant value in Fig. 1(b) is $\approx 2{\cdot}68\times10^{-2}$ and the *YYY* frequency in N5EUK is equal to 0·1609, i.e. its square $=2{\cdot}59\times10^{-2}$;

the constant value in Fig. 1(c) is $\approx 2{\cdot}68\times10^{-2}$ and the *RRR* frequency in NIEUK is equal to 0·1657, i.e. its square $=2{\cdot}75\times10^{-2}$;

the constant value in Fig. 1(d) is $\approx 3{\cdot}08\times10^{-2}$ and the *YYY* frequency in NIEUK is equal to 0·1710, i.e. its square $=2{\cdot}92\times10^{-2}$;

the constant value in Fig. 1(e) is $\approx 2{\cdot}50\times10^{-2}$ and the *RRR* frequency in N3EUK is equal to 0·1453, i.e. its square $=2{\cdot}11\times10^{-2}$;

the constant value in Fig. 1(f) is $\approx 3{\cdot}01\times10^{-2}$ and the *YYY* frequency in N3EUK is equal to 0·1785, i.e. its square $=3{\cdot}19\times10^{-2}$.

Although this relation is less significant in Fig. 1(e), these six curves tend to be a constant function of i at a long distance of i (uniform curve) only depending on the frequency (its square) of the trinucleotide in the gene population. In other words, the curve shape is random at a long distance of i.

The shape of these six curves produced by these three properties is new and has never been observed with the *YRY*-function in any gene populations analysed so far, in particular the *YRY*-function in N5EUK, NIEUK and N3EUK leads to a periodicity modulo 2 (Arquès & Michel, 1990*a*). However, the local maxima for $i\equiv 3[6]$ identified with the *RRR*- and *YYY*-functions although weak, as they are hidden by the large exponential decrease, are in fact an important property because the *YRY*-function in N5EUK shows a periodicity modulo 2 associated with four subcurves modulo 6 whose one is for $i\equiv 3[6]$. The biological meaning of the local maxima for $i\equiv 3[6]$ (pattern observed with the three *RRR*-, *YYY*- and *YRY* functions) is presented in the Discussion.

Finally, owing to the law of large numbers (Arquès & Michel, 1990*b*: 752, Section 2.3.3), the curve $C(T; F)$ obtained with populations F made of several hundreds of genes and for any trinucleotide T, keeps the main and non-random statistical properties (for example, periodicities, a decrease, maximal and minimal values, etc), even

if new genes are further available in the populations F. For example, since the tenth release of the EMBL database, the periodicities and the preferential occurrence of the motif $YRY(N)_6YRY$ were observed in each new release. Therefore, all the non-random properties identified at the gene population level are important as they are stable from a statistical point of view. Thus, the next research step is naturally their understanding with a simulation model (the mathematical and biological reasons of these two steps are explained in Arquès & Michel, 1990*b*).

3. An Oligonucleotide Mixing Model

3.1. INTRODUCTION

The problem which arises in this section is whether the three new non-random properties identified with the *RRR*- and *YYY*-functions in eukaryotic non-coding genes, i.e. the large exponential decrease with a short linear decrease and local maxima for $i \equiv 3[6]$, can also be simulated by an independent mixing of oligonucleotides. This problem, as in the case of the *YRY*-function in Arquès & Michel (1990*b*), is divided into two sub-problems: the identification of oligonucleotides and the determination of probabilities associated with the oligonucleotides. This combinatorial problem may not have any solution as there are 10^{158} possible curve shapes for a curve with 100 different points and as a real curve could not have been simulated by mixing oligonucleotides, i.e. the concept of an oligonucleotide mixing is false. However, the subcurve modulo 6 for $i \equiv 3[6]$ with the *YRY*-function was able to be simulated by an independent mixing of oligonucleotides (Arquès & Michel, 1990*b*), then the local maxima for $i \equiv 3[6]$ with the *RRR*- and *YYY*-functions together with the two other properties should also be simulated by an independent mixing of oligonucleotides. With the help of some previous rules and using the calculus power of the computer, several hundreds of simulation models were analysed by varying the type, the number and the probability of oligonucleotides. We present now a summary of the main research steps which lead to a solution of this problem.

3.2. METHOD

3.2.1. *Creation of a simulated population*

Let an oligonucleotide O be a word of a few ($\leq$ten) letters on the alphabet $\{R, Y\}$. Let a set $\mathscr{S}$ of n oligonucleotides O_i, $1 \leq i \leq n$, be associated with a set $\mathscr{P}$ of n probabilities p_i, $\sum_{1 \leq i \leq n} p_i = 1$. Then, a simulated population $S(\mathscr{S}, \mathscr{P})$ of sequences is created by mixing these oligonucleotides O_i according to an independent concatenation depending only on the probabilities p_i attached to these oligonucleotides O_i. In order to obtain significant statistical results, this simulated population is in fact constituted of 500 sequences of a 2000 base length and generated in order to have the same percentage of R and Y in any sequence of the simulated population. The computations obtained with such a sample of 1 million bases are precise, i.e. there are no random fluctuations in the calculus of probabilities: a sample having 200 sequences of a 1000 base length leads to similar results.

3.2.2. *Choice of the set* $(\mathscr{S}, \mathscr{P})$ *of "probabilized oligonucleotides"*

The set $(\mathscr{S}, \mathscr{P})$ of "probabilized oligonucleotides", i.e. the oligonucleotides and their associated probabilities, will be chosen so that the T-function, $T=RRR$ or $T=YYY$, in the simulated population $S(\mathscr{S}, \mathscr{P})$ leads to a curve $C(T; S(\mathscr{S}, \mathscr{P}))$ called simulated, having the statistical properties of the real curves. Such a simulated curve can be compared with a real one as both result from the same function definition.

The first research step, as mentioned in section 3.1, is to find a set $\mathscr{S}$ of oligonucleotides leading to the properties observed in the real curves. As the real curves are identical using both the *RRR*- and *YYY*-functions, the oligonucleotides O_i will be complementary in R and Y. This strong constraint in the choice of oligonucleotides also leads to sequences with the same percentage of R and Y.

The second research step is then to obtain the best curve shape by scanning in the range [0, 1] and with a step of 0·01, all the probabilities p_i attached to the oligonucleotides O_i of the set $\mathscr{S}$ previously identified, the oligonucleotides O_i being independently mixed.

3.3. RESULTS

3.3.1. *Identification of two oligonucleotides leading to the short linear decrease*

The simulation of a decrease, linear or exponential, short or large, is a new problem as the reasoning used to simulate the periodicities (Arquès & Michel, 1980*b*) cannot be applied to it. The study to simulate a linear decrease led to the identification of the following rule:

Let the two oligonucleotides O_1 (resp. O_2) be a series of R (resp. Y) of the same length l, i.e. $O_1=R^l$ and $O_2=Y^l$, then the independent mixing of O_1 and O_2 with equiprobabilities (0·5) leads to a simulated curve $C(T; S(\mathscr{S}, \mathscr{P}))$, $T=RRR$ or $T=YYY$ and $(\mathscr{S}, \mathscr{P})=\{(R^l, 0{\cdot}5), (Y^l, 0{\cdot}5)\}$, with the following properties:

a quasi-linear decrease for $0\leq i\leq l-1$;

a periodicity modulo l for $i\geq l-1$.

These results obtained by simulation can also be proved by an exact calculus of the occurrence probability p_i of the T-function in the simulated population $S(\mathscr{S}, \mathscr{P})$ (see Appendix).

The two curves $C(RRR; S(\mathscr{S}, \mathscr{P}))$ and $C(YYY; S(\mathscr{S}, \mathscr{P}))$ are identical as O_1 and O_2 are complementary in R and Y and associated with the same probability.

For the simulation of the short linear decrease from $i=0$ to 8, this rule is applied with $l=8$. Indeed, the independent mixing of $O_1=R^8$ and $O_2=Y^8$ with equiprobabilities leads to a simulated curve $C(T; S(\mathscr{S}, \mathscr{P}))$ [Fig. 2(a)], $T=RRR$ or $T=YYY$ and $(\mathscr{S}, \mathscr{P})=\{(R^8, 0{\cdot}5), (Y^8, 0{\cdot}5)\}$, with the following properties [Fig. 2(a) represents the two identical curves $C(RRR; S(\mathscr{S}, \mathscr{P}))$ and $C(YYY; S(\mathscr{S}, \mathscr{P}))$]:

a short linear decrease for $0\leq i\leq 7$;

a periodicity modulo 8 for $i\geq 7$.

The simulated curve [Fig. 2(a)] is not yet similar to the six real curves as:

there is a periodicity modulo 8;

the local maxima for $i\equiv 3[6]$ and the large exponential decrease are missing;

the highest probability in the simulated curve (0·344) is ten times greater than the highest probability in a real curve.
These three facts explain that this simulation is still incomplete.

3.3.2. *Identification of two additional oligonucleotides leading to the short linear decrease and to the local maxima for* $i \equiv 3[6]$

The question here is to identify two complementary oligonucleotides leading to the local maxima for $i \equiv 3[6]$ without destroying the short linear decrease. The local maxima for $i \equiv 3[6]$ is a result of the preferential occurrence of the trinucleotide *RRR* (or *YYY*) 3, 9, etc, bases after itself (see Arquès & Michel, 1980*b* for a reasoning with periodicities). With the two additional oligonucleotides $O_3 = RRRYRYRRR$ and $O_4 = YYYRYRYYY$, there is a preferential occurrence of *RRR* three bases after itself in O_3, of *YYY* three bases after itself in O_4. The independent mixing of the four oligonucleotides O_1, O_2, O_3 and O_4 with equiprobabilities (0·25) leads to a simulated curve $C(T; S(\mathscr{S}, \mathscr{P}))$ [Fig. 2(b)], $T = RRR$ or $T = YYY$ and $(\mathscr{S}, \mathscr{P}) = \{(R^8, 25\%), (Y^8, 25\%), (RRRYRYRRR, 25\%), (YYYRYRYYY, 25\%)\}$, with a short linear decrease and local maxima for $i \equiv 3[6]$, the periodicity modulo 8 being destroyed. The two curves $C(RRR; S(\mathscr{S}, \mathscr{P}))$ and $C(YYY; S(\mathscr{S}, \mathscr{P}))$ are identical as O_3 and O_4 are still complementary in *R* and *Y* and still associated with the same probability. Surprisingly, O_3 and O_4 belong to an interesting and important class of oligonucleotides (see Discussion).

A scanning of probabilities of these four oligonucleotides can give a better curve shape (data not shown), for example, with a peak at $i = 3$ having the second highest value (i.e. greater than the value at $i = 1$), as in the six real curves. Nevertheless, this scanning cannot lead to an exponential decrease. Furthermore, the highest probability in the simulated curve (0·175) is still five times greater than the highest probability in a real curve.
These two facts explain that this simulation is slightly incomplete.

3.3.3. *A simulated curve similar to the six real curves*

We first tried to understand the large exponential decrease in terms of oligonucleotides but we could not find a solution after testing several hundreds of simulations. In fact, the large exponential decrease is related to large alternating *R*/*Y* series, for example, $(RY)_{50}$. Indeed, with the independent mixing of the previous four oligonucleotides R^8, Y^8, *RRRYRYRRR*, *YYYRYRYYY* and $(RY)_{50}$, the best curve shape [Fig. 2(c)] obtained with a scanning of probabilities has the three non-random properties observed in the six real curves:

the large exponential decrease;

the short linear decrease;

the local maxima for $i \equiv 3[6]$ with a peak at $i = 3$ having the second highest value.

Note: The scanning is associated with an algorithm of curve-form recognition (not explained here) in order to compare a simulated curve with a real one and to select automatically the best simulated curve among the 10 000 possible solutions [step = 0·01, prob(O_1) = prob(O_2), prob(O_3) = prob(O_4), prob($(RY)_{50}$) = complement to 1].

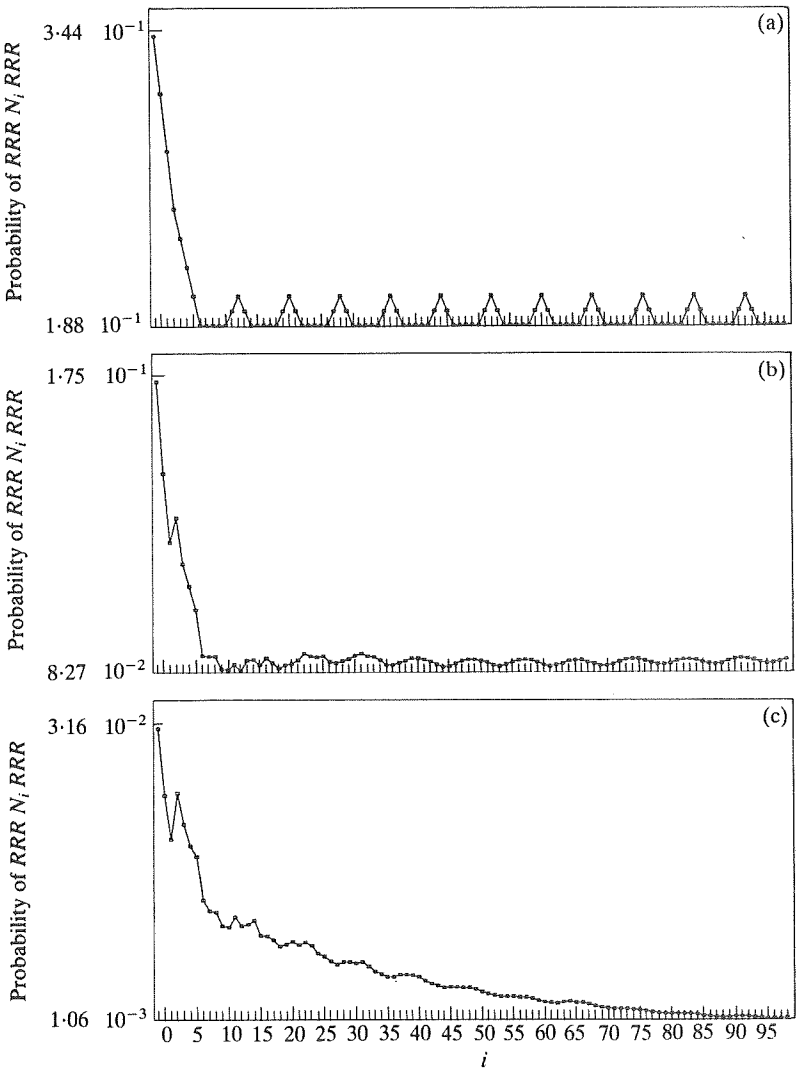


FIG. 2. Simulation of new non-random statistical properties common to different populations of eukaryotic non-coding genes [identified in Fig. 1(a)–(f)]. The horizontal axis represents the number i, $i \in [0, 99]$, of any bases N between two identical trinucleotides T, $T = RRR$ or $T = YYY$, i.e. the i-motif $RRR(N)_iRRR$ or the i-motif $YYY(N)_iYYY$. The vertical axis represents the T-function, i.e. the simulated curve $C(T; S(\mathscr{S}, \mathscr{P}))$ shows the occurrence probability of the i-motif $RRR(N)_iRRR$ or $YYY(N)_iYYY$ in the simulated population $S(\mathscr{S}, \mathscr{P})$ (see section 3.2). (a) Simulation of the short linear decrease. Simulated curve $C(T = RRR; S(\mathscr{S}, \mathscr{P}))$ showing the occurrence probability of the i-motif $RRR(N)_iRRR$ in the simulated population $S(\mathscr{S}, \mathscr{P})$ created by an independent mixing of R^8 and Y^8 with the following probabilities: $(\mathscr{S}, \mathscr{P}) = \{(R^8, 50\%), (Y^8, 50\%)\}$. The simulated curve $C(T = YYY; S(\mathscr{S}, \mathscr{P}))$ is identical. (b) Simulation of the short linear decrease and the local maxima for $i \equiv 3[6]$ ($i = 3, 9$, etc). Simulated curve $C(T = RRR; S(\mathscr{S}, \mathscr{P}))$ showing the occurrence probability of the i-motif $RRR(N)_iRRR$ in the simulated population $S(\mathscr{S}, \mathscr{P})$ created by an independent mixing of R^8, Y^8, $RRRYRYRRR$ and $YYYRYRYYY$ with the

The large alternating R/Y series does not correspond to an oligonucleotide (word less than ten letters) but represents the genetic information not encoded by *RRR* and *YYY* and which must be considered in the simulation. Therefore, any series not containing *RRR* and *YYY* could have been used and the two curves $C(RRR; S(\mathcal{S}, \mathcal{P}))$ and $C(YYY; S(\mathcal{S}, \mathcal{P}))$ are identical.

The simulated curve [Fig. 2(c)] is not completely identical with the six real curves as:

the six real curves differ with non-significant patterns;

the simulated curve [Fig. 2(c)] at a long distance of i has probabilities (for example, $1{\cdot}06 \times 10^{-3}$ at $i=99$) lower than the probabilities observed in the real curves [Fig. 1(a)–(f)]. In Arquès & Michel (1990*b*), we have proved that the problem of frequency level of a simulated curve (either non-random or constant function of i) is not related to the oligonucleotide mixing process, but to another genetic process, precisely to the random base mutation process (see Discussion for details).

The understanding of a curve shape with properties is a known difficult problem of pattern recognition which cannot be solved by only a few properties.

The simple model developed here cannot completely simulate the genetic reality depending on a great number of factors, in the same way that the first terms of the development of a function in series cannot reveal the totality of the function.

Nevertheless, this simulated curve has the significant patterns of the six real curves. Therefore, this simulation model is strongly correlated with the reality observed in eukaryotic non-coding genes.

4. Discussion

The *RRR*- and *YYY*-functions in the 5′ regions, introns and in the 3′ regions of eukaryotes lead to six "independent" curves having the same three non-random properties: a short linear decrease, local maxima for $i \equiv 3[6]$ and a large exponential decrease [Fig. 1(a)–(f)]. As with the periodicities modulo 2, 3 and the preferential occurrence of the motif $YRY(N)_6YRY$, these new properties again demonstrate that the nucleotide distribution in genes is not random and can be simulated: a curve which would have been random could, obviously, not have been simulated. Surprisingly, the local maxima for $i \equiv 3[6]$ found with the *RRR*- and *YYY*-functions have already been identified with the *YRY*-function showing in the 5′ eukaryotic regions four subcurves modulo 6 for $i \equiv 3[6]$, $i \equiv 1, 5[6]$, $i \equiv 0[6]$ and $i \equiv 2, 4[6]$ (Arquès & Michel, 1990*b*). Among these four subcurves, the one for $i \equiv 3[6]$ has the strongest

following probabilities: $(\mathcal{S}, \mathcal{P}) = \{(R^8, 25\%), (Y^8, 25\%), (RRRYRYRRR, 25\%), (YYYRYRYYY, 25\%)\}$. The simulated curve $C(T = YYY; S(\mathcal{S}, \mathcal{P}))$ is identical. (c) Simulation of the short linear decrease, the local maxima for $i \equiv 3[6]$ and the large exponential decrease [identified in Fig. 1(a)–(f)]. Simulated curve $C(T = RRR; S(\mathcal{S}, \mathcal{P}))$ showing the occurrence probability of the i-motif $RRR(N)_iRRR$ in the simulated population $S(\mathcal{S}, \mathcal{P})$ created by an independent mixing of R^8, Y^8, *RRRYRYRRR*, *YYYRYRYYY* and $(RY)_{50}$ with the following probabilities: $(\mathcal{S}, \mathcal{P}) = \{(R^8, 15\%), (Y^8, 15\%), (RRRYRYRRR, 25\%), (YYYRYRYYY, 25\%), ((RY)_{50}, 20\%)\}$. The simulated curve $C(T = YYY; S(\mathcal{S}, \mathcal{P}))$ is identical.

occurrence probability. This result may explain why the local maxima for $i \equiv 3[6]$ are conserved in eukaryotic non-coding genes. Finally, this observation is in agreement with the property 4 and the final remark in Arquès & Michel (1990*b*: 766, lines 10–13 and 771, lines 9–15) stating that there are a few "universal" non-random patterns in genes (i.e. whatever the T-function) from an independent mixing of a few types of oligonucleotides.

The oligonucleotide mixing model developed in section 3 simulates these three non-random properties by an independent mixing of the four oligonucleotides R^8, Y^8, *RRRYRYRRR*, *YYYRYRYYY* and large alternating R/Y series [Fig. 2(c)]. The short linear decrease is a result of R^8 and Y^8, the local maxima for $i \equiv 3[6]$, of *RRRYRYRRR* and *YYYRYRYYY*, and the large exponential decrease, of large alternating R/Y series. The large exponential decrease represents all the genetic information not encoded by *RRR* and *YYY* but containing, in particular, the three oligonucleotides previously identified *YRYRYR*, *YRYYRY* and $YRY(N)_6$ (Arquès & Michel, 1990*b*). Unexpectedly, the two oligonucleotides *RRRYRYRRR* and *YYYRYRYYY* belong to the oligonucleotide class of $YRY(N)_6$: *RRRYRYRRR* is a particular specification of *NNNYRYNNN* and *YYYRYRYYY*, a particular specification of *NNYRYNNNN* (or of *NNNNYRYNN*). This result agrees with property 2 in Arquès & Michel (1990*b*: 765, Section 3.3.6) demonstrating that: "The six bases *N* of the oligonucleotide $YRY(N)_6$ have not to be specified by *YRY* and by *RYR*. Obviously, specification by motifs different from *YRY* and *RYR* could be necessary for a model more general which also considers trinucleotides (e.g. *RRR*) different from *YRY*". Therefore, the two oligonucleotides *RRRYRYRRR* and *YYYRYRYYY* could explain a genetic information encoding simultaneously by *RRR*, *YYY* and *YRY*, i.e. they could explain non-random properties identified with the *RRR*-, *YYY*-, and *YRY*-functions. We are currently testing this hypothesis with a model which independently mixes the six oligonucleotides R^8, Y^8, *RRRYRYRRR*, *YYYRYRYYY*, *YRYRYR* and *YRYYRY* in order to simulate, with a unique model, the non-random properties observed with the *RRR*-, *YYY*- and *YRY*-functions. The choice of these six oligonucleotides is probably not the best one as they are not completely complementary. Nevertheless, it is essential to begin with simple independent models in order to obtain rules which can be used afterwards to develop more general models containing the simple models.

The simulated curve [Fig. 2(c)] has the non-random statistical properties of the six real curves, however, with a frequency level difference. Therefore, to have a complete model, a random base mutation process must be added after the oligonucleotide mixing process as with the previous model mixing the three oligonucleotides *YRYRYR*, *YRYYRY* and $YRY(N)_6$ (Arquès & Michel, 1990*b*). Indeed, while the mixing process acts on the relative values in the simulated curves (i.e. acts on the curve shape and leads to non-random properties such as periodicities), the mutation process acts on the absolute values in the simulated curves as follows: it decreases the values greater than 1/64 (1/64 is the constant value of a population with the same frequency of *R* and *Y*) and increases the values less than 1/64. Random mutations, which are noise in terms of signal processing, have no action on the curve shape. As this random process cannot lead to non-random properties, mutations can

only occur after the mixing of oligonucleotides (see the proposed model of gene evolution below). This paper mainly deals with the mixing process which is a difficult problem owing to its great complexity: a solution is obtained not only if the good oligonucleotides are identified but also if the right probabilities are determined: modifying the length, a base or the probability of an oligonucleotide destroys the non-random properties. In fact, before having the mathematical reasons (described in sections 3.3.1, 3.3.2 and 3.3.3) for the existence of a solution with an independent mixing, several hundreds of Markov mixing models of oligonucleotides were first tested.

Finally, a simple model of gene evolution can be deduced from these results (Fig. 3). First, an independent mixing of a few types of primitive oligonucleotides led to the formation of primitive genes (step 1 in Fig. 3). So far, six oligonucleotides have been identified for this step: R^8, Y^8, *RRRYRYRRR*, *YYYRYRYYY*, *YRYRYR* and *YRYYRY*. Then, random nucleotide processes (mainly mutations) in these primitive genes led to the actual genes and their diversity (step 2 in Fig. 3).

We thank Dr Nouchine Soltanifar and the referee for their advice.

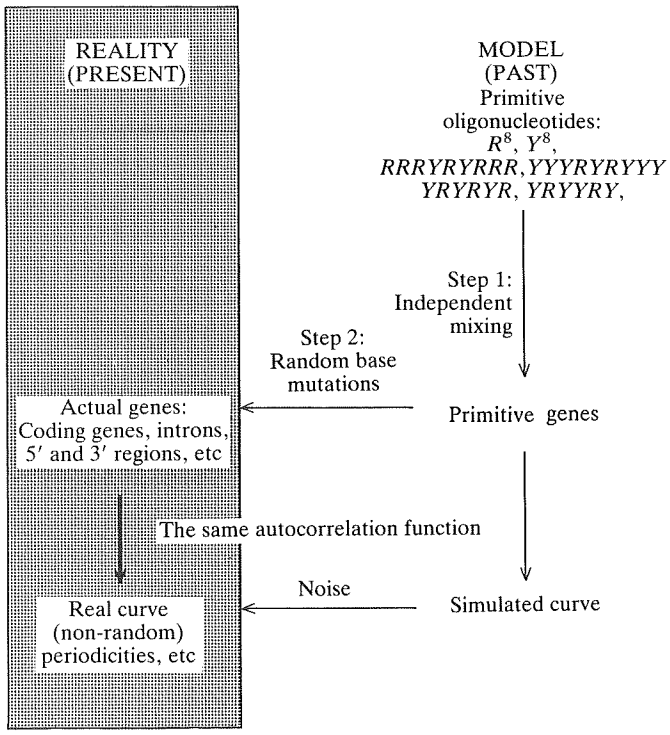


FIG. 3. A model of gene evolution.

REFERENCES

ARQUÈS, D. G. & MICHEL, C. J. (1987*a*). *Math. Biosci.* **86**, 1–14.
ARQUÈS, D. G. & MICHEL, C. J. (1987*b*). *J. theor. Biol.* **128**, 457–461.
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.
CECH, T. R. (1991). *Cell* **64**, 667–669.
KIMURA, M. (1987). *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.
NEI, M. (1987). *Molecular Evolutionary Genetics.* New York: Columbia University Press.
SIMPSON, L. (1990). *Science* **250**, 512–513.

APPENDIX

The occurrence probability p_i of the T-function in the simulated population $S(\mathscr{S}, \mathscr{P})$, $T=RRR$ or $T=YYY$ and $(\mathscr{S}, \mathscr{P})=\{(R^l, 0{\cdot}5), (Y^l, 0{\cdot}5)\}$, can be exactly determined by formulae.

$$p_i=\frac{1}{l}\left[\sum_{\substack{h \text{ integer}/h\in[1,l-2] \text{ and} \\ h+2+i\in[0,l-3] \text{ modulo } l}} 2^{(r-2)^+-r}+\sum_{\substack{h \text{ integer}/h\in[1,l-2] \text{ and} \\ h+2+i\in[l-2,l-1] \text{ modulo } l}} 2^{(r-2)^+-r-1}\right.$$

$$\left.+\sum_{\substack{h \text{ integer}/h\in[l-1,l] \text{ and} \\ h+2+i\in[0,l-3] \text{ modulo } l}} 2^{(r-3)^+-r}+\sum_{\substack{h \text{ integer}/h\in[l-1,l] \text{ and} \\ h+2+i\in[l-2,l-1] \text{ modulo } l}} 2^{(r-3)^+-r-1}\right]$$

with the following notations:

$r=1+\text{Int}\,[(h+2+i)/l]$, (Integer part)

$x^+=\max(x, 0)$

$[a, b]$ modulo l is the union of the ranges $\bigcup_{k\geq 0}[a+kl, b+kl]$.

For $i\geq l-1$, the formulae p_i is invariant for i values congruent modulo l.