Biochimie (1993) 75, 399–407

Identification and simulation of new non-random statistical properties common to different eukaryotic gene subpopulations

DG Arquès[a], CJ Michel[b]

[a]*Equipe de Biologie Théorique, Université de Franche-Comté, Laboratoire d'Informatique de Besançon, 16, route de Gray, 25030 Besançon;*
[b]*Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie de Belfort-Montbéliard, BP 527, 90016 Belfort, France*

(Received 16 December 1992; accepted 29 December 1992)

Summary — The nucleotide distribution in protein coding genes, introns and transfer RNA genes of eukaryotic subpopulations (primates, rodents and mammals) is studied by autocorrelation functions. The autocorrelation function analysing the occurrence probability of the i-motif YRY$(N)_i$YRY (YRY-function) in protein coding genes and transfer RNA genes of these three eukaryotic subpopulations retrieves the preferential occurrence of YRY$(N)_6$YRY (R=purine=adenine or guanine, Y=pyrimidine=cytosine or thymine, N=R or Y). The autocorrelation functions analysing the occurrence probability of the *i*-motifs RRR$(N)_i$RRR (RRR-function) and YYY$(N)_i$YYY (YYY-function) identify new non-random genetic statistical properties in these three eukaryotic subpopulations, mainly: i) in their protein coding genes: local maxima for $i \equiv 6[12]$ (peaks for i=6, 18, 30, 42) with the RRR-function and local maxima for $i \equiv 8[10]$ (peaks for i=8, 18, 28) with the YYY-function; and ii) in their introns: local maxima for $i \equiv 3[6]$ (peaks for i=3, 9, 15) and a short linear decrease followed by a large exponential decrease both with the RRR- and YYY-functions. The non-random properties identified in eukaryotic intron subpopulations are modelised with a process of random insertions and deletions of nucleotides simulating the RNA editing.

eukaryotic genes / non-random statistical properties / nucleotide distribution

Introduction

In [1, 2], we proposed a new definition for the autocorrelation function allowing to analyse in DNA sequences (words of several hundreds of letters on the alphabet {A,C,G,T}: A=adenine, C=cytosine, G=guanine and T=thymine are called nucleotides or bases) the occurrence probability of *i*-motifs without bias, an *i*-motif being two motifs (words of a few letters) separated by any *i* letters. Indeed, this new definition, contrary to the classical one, avoids the decrease of probabilities when the number *i* of bases between the two motifs increases (when *i* increases, a smaller number of *i*-motifs can be analysed in the sequence): the side effect induced by the end of the sequence is corrected (see below and [1, 2]). Therefore, the newly defined autocorrelation function is not the inverse Fourier transform of the power spectra classically used to analyse DNA sequences. This autocorrelation function is simple as it is based on the frequency concept and therefore appropriate for the biological meaning, interesting as it studies not only the frequency of two motifs but also the distance between them, 'general', as a motif is a particular case of an *i*-motif with two motifs separated by 0 base. But most importantly, it is powerful as the particular autocorrelation function, called T-function, analysing in gene populations the occurrence probability of two identical trinucleotides T (motifs of three nucleotides) separated by any *i* nucleotides N, *ie* *i*-motifs T$(N)_i$T, on the alphabet {R,Y} (R=purine=A or G, Y=pyrimidine=C or T, N=R or Y) allows the identification of non-random statistical properties. Indeed, the YRY-function, *ie* the autocorrelation function analysing the occurrence probability of the *i*-motif YRY$(N)_i$YRY in gene populations, allows to identify periodicities (modulo 2 and 3), sub-periodicities, the preferential occurrence of the motif YRY$(N)_6$YRY, local maxima, etc [1–6]. Finally, the autocorrelation function allows the simulation of the non-random properties identified by models of molecular evolution: Markov and independent mixing of oligonucleotides [4, 6], process of random insertions and deletions of nucleotides [5] simulating the RNA editing [7].

In the second part of the article T-functions applied in protein coding genes, introns and transfer RNA genes of eukaryotic subpopulations (primates, rodents and mammals) retrieve the preferential occurrence of YRY$(N)_6$YRY with the YRY-function and also identify new non-random statistical properties: i) in pro-

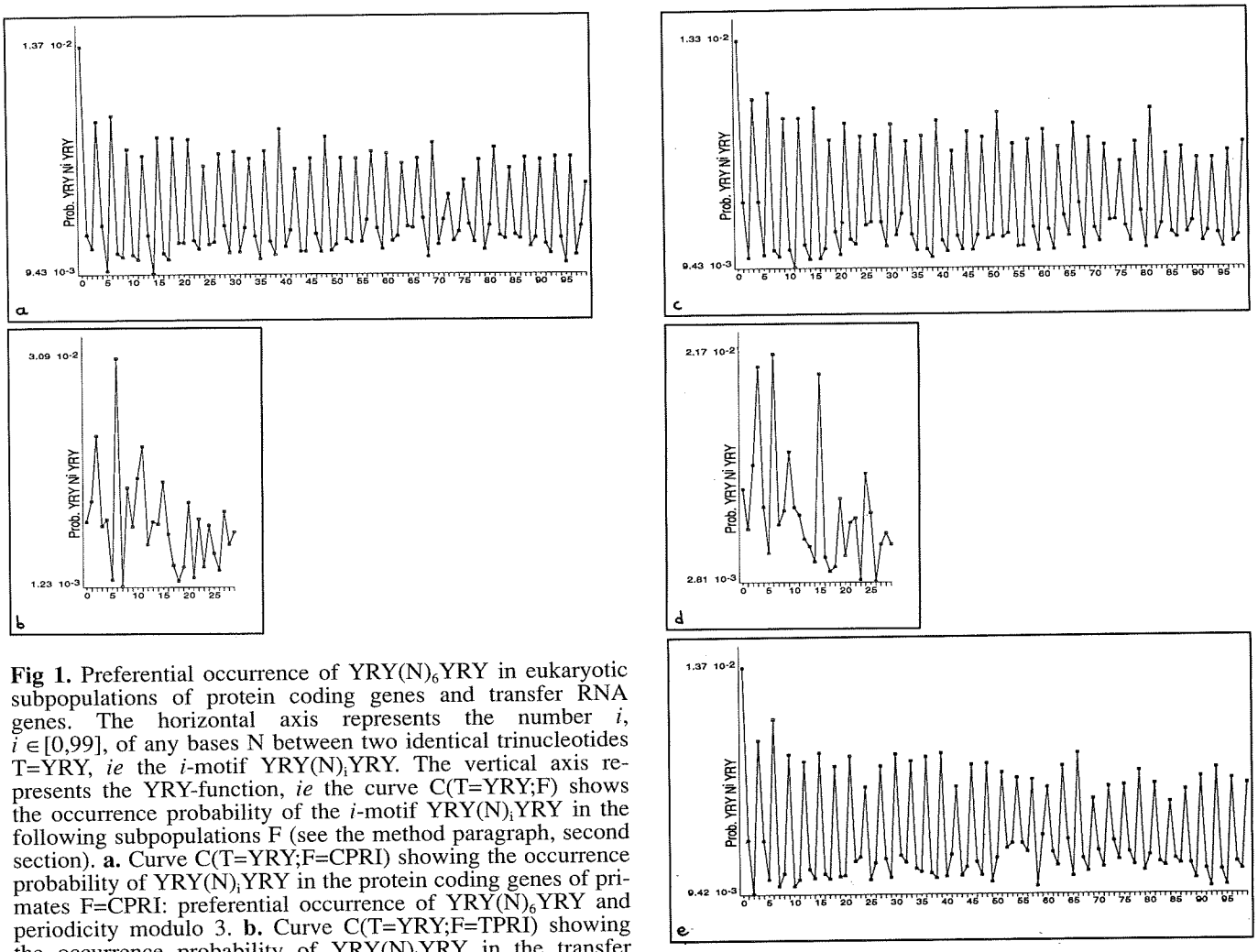


Fig 1. Preferential occurrence of $YRY(N)_6YRY$ in eukaryotic subpopulations of protein coding genes and transfer RNA genes. The horizontal axis represents the number i, $i \in [0,99]$, of any bases N between two identical trinucleotides T=YRY, *ie* the i-motif $YRY(N)_iYRY$. The vertical axis represents the YRY-function, *ie* the curve C(T=YRY;F) shows the occurrence probability of the i-motif $YRY(N)_iYRY$ in the following subpopulations F (see the method paragraph, second section). **a.** Curve C(T=YRY;F=CPRI) showing the occurrence probability of $YRY(N)_iYRY$ in the protein coding genes of primates F=CPRI: preferential occurrence of $YRY(N)_6YRY$ and periodicity modulo 3. **b.** Curve C(T=YRY;F=TPRI) showing the occurrence probability of $YRY(N)_iYRY$ in the transfer RNA genes of primates F=TPRI: preferential occurrence of $YRY(N)_6YRY$. **c.** Curve C(T=YRY;F=CROD) showing the occurrence probability of $YRY(N)_iYRY$ in the protein coding genes of rodents F=CROD: preferential occurrence of $YRY(N)_6YRY$ and periodicity modulo 3. **d.** Curve C(T=YRY;F=TROD) showing the occurrence probability of $YRY(N)_iYRY$ in the transfer RNA genes of rodents F=TROD: preferential occurrence of $YRY(N)_6YRY$. **e.** Curve C(T=YRY;F=CMAM) showing the occurrence probability of $YRY(N)_iYRY$ in the protein coding genes of mammals F=CMAM: preferential occurrence of $YRY(N)_6YRY$ and periodicity modulo 3.

tein coding genes with the RRR-function: local maxima for $i \equiv 6[12]$ followed (for $i \geq 42$) by local maxima for $i \equiv 6[9]$ for each of these three eukaryotic subpopulations; ii) in protein coding genes with the YYY-function: local maxima for $i \equiv 8[10]$ followed (for $i \geq 33$) by local maxima for $i \equiv 3[6]$ for each of these three eukaryotic subpopulations; and iii) in introns with the RRR- and YYY-functions: local maxima for $i \equiv 3[6]$ and a short linear decrease followed by a large exponential decrease for each of these three eukaryotic subpopulations.

In the last part, the non-random properties identified in eukaryotic intron subpopulations will be modelised with a process of random insertions and deletions of nucleotides.

Identification of new non-random statistical properties common to different eukaryotic gene subpopulations

Introduction

In order to retrieve the preferential occurrence of $YRY(N)_6YRY$, a motif that we proposed to be (or have been) related to the pitch of the DNA double

helix, and to increase its statistical significance, the YRY-function is applied in protein coding genes, and transfer RNA genes of eukaryotic subpopulations (primates, rodents and mammals). These subpopulations are obtained from the release 32 of the EMBL Nucleotide Sequence Data Library.

Another problem investigated in this section is whether new properties common to these subpopulations can be identified with T-functions different from the classically used YRY-function. 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 YRY-, RRR- and YYY-functions cannot be deduced from each other.

Method

This method generalizes the previous one [2] 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, *ie* T={RRR,...,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 $c_i(T;s)/[l(s)-104]$, *ie* 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).

Remarks: i) 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; and ii) for transfer RNA genes, *i* is chosen in the range [0,29] as their length is about 75 nucleotides.

The eukaryotic gene subpopulations F analysed here are:
– protein coding genes of primates (CPRI): 5820 sequences (6053 kb).
– Introns of primates (IPRI): 1869 sequences (1665 kb).
– Transfer RNA genes of primates (TPRI): 60 sequences (5 kb).
– Protein coding genes of rodents (CROD): 4401 sequences (5070 kb).
– Introns of rodents (IROD): 1040 sequences (938 kb).
– Transfer RNA genes of rodents (TROD): 64 sequences (5 kb).
– Protein coding genes of mammals (CMAM) (other than primates and rodents): 1390 sequences (1629 kb).
– Introns of mammals (IMAM) (other than primates and rodents): 146 sequences (141 kb).

They are obtained from the release 32 of the EMBL Nucleotide Sequence Data Library in the same way as described in previous studies (see *eg* [4] for a description of data acquisitions). The T-function is based on the classically used trinucleotide T=YRY and on the newly used 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 two YRY, two RRR or two YYY by varying *i* between 0 and 99; ii) the ordinate gives the mean occurrence probability of $YRY(N)_iYRY$, $RRR(N)_iRRR$ or $YYY(N)_iYYY$ in a gene population F.

Results

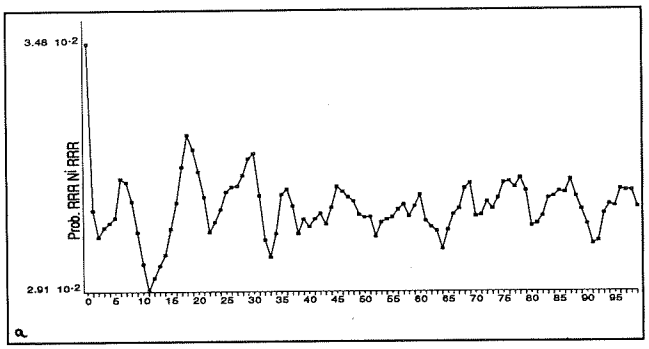
Results with the YRY-function analysing the occurrence probability of the i-motif $YRY(N)_iYRY$ in genes
The YRY-function is applied in:
– protein coding genes of primates F=CPRI: curve C(YRY;CPRI) (fig 1a);
– transfer RNA genes of primates F=TPRI: curve C(YRY;TPRI) (fig 1b);
– protein coding genes of rodents F=CROD: curve C(YRY;CROD) (fig 1c);
– transfer RNA genes of rodents F=TROD: curve C(YRY;TROD) (fig 1d);
– protein coding genes of mammals F=CMAM: curve C(YRY;CMAM) (fig 1e).

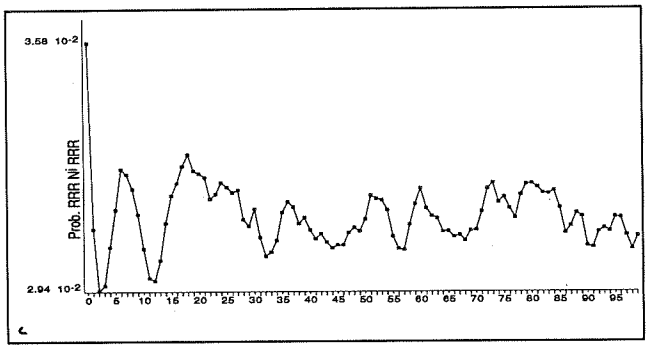
The YRY-function retrieves the preferential occurrence of $YRY(N)_6YRY$ in various eukaryotic subpopulations of protein coding genes (associated with the periodicity modulo 3 and a particular higher value for $i=0$ which is not explained here) and of transfer RNA genes. These non-random statistical properties were modelised [5] with a process simulating the RNA editing.

Results with the RRR-function analysing the occurrence probability of the i-motif $RRR(N)_iRRR$ in genes
The RRR-function is applied in the protein coding genes of:
– primates F=CPRI: curve C(RRR;CPRI) (fig 2a);
– rodents F=CROD: curve C(RRR;CROD) (fig 2b);
– mammals F=CMAM: curve C(RRR;CMAM) (fig 2c).

The RRR-function shows local maxima for $i\equiv 6[12]$ (peaks for i=6, 18, 30, 42) followed (for $i\geq 42$) by local maxima for $i\equiv 6[9]$ (peaks for i=42, 51, 60, 69, 78, 87, 96).

The RRR-function is applied in the introns of:
– primates F=IPRI: curve C(RRR;IPRI) (fig 3a);
– rodents F=IROD: curve C(RRR;IROD) (fig 3b);
– mammals F=IMAM: curve C(RRR;IMAM) (fig 3c).





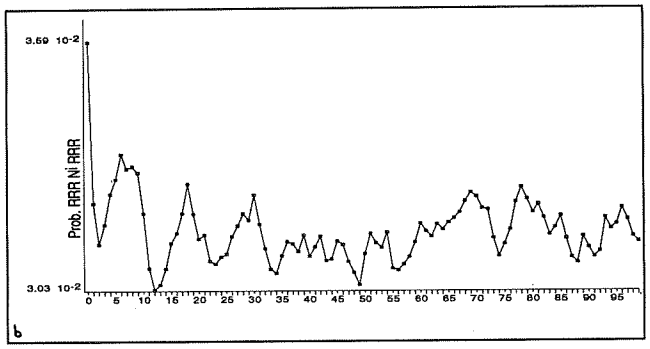


Fig 2. Local maxima for $i \equiv 6[12]$ followed (for $i \geq 42$) by local maxima for $i \equiv 6[9]$ in eukaryotic protein coding gene subpopulations. The horizontal axis represents the number i, $i \in [0,99]$, of any bases N between two identical trinucleotides T=RRR, *ie* the i-motif RRR(N)$_i$RRR. The vertical axis represents the RRR-function, *ie* the curve C(T=RRR;F) shows the occurrence probability of the i-motif RRR(N)$_i$RRR in the following subpopulations F (see the method paragraph, second section). **a.** Curve C(T=RRR;F=CPRI) showing the occurrence probability of RRR(N)$_i$RRR in the protein coding genes of primates F=CPRI: local maxima for $i \equiv 6[12]$ (6, 18, 30, 42) followed (for $i \geq 42$) by local maxima for $i \equiv 6[9]$ (42, 51, 60, 69, 78, 87, 96). **b.** Curve C(T=RRR;F=CROD) showing the occurrence probability of RRR(N)$_i$RRR in the protein coding genes of rodents F=CROD: local maxima for $i \equiv 6[12]$ (6, 18, 30, 42) followed (for $i \geq 42$) by local maxima for $i \equiv 6[9]$ (42, 51, 60, 69, 78). **c.** Curve C(T=RRR;F=CMAM) showing the occurrence probability of RRR(N)$_i$RRR in the protein coding genes of mammals F=CMAM: local maxima for $i \equiv 6[12]$ (6, 18, 30, 42) followed (for $i \geq 42$) by local maxima for $i \equiv 6[9]$ (42, 51, 60, 69).

The RRR-function shows local maxima for $i \equiv 3[6]$ (peaks for i=3, 9, 15) and a short linear decrease followed by a large exponential decrease. These non-random statistical properties are modelised with a process simulating the RNA editing described in the last part of the article.

Results with the YYY-function analysing the occurrence probability of the i-motif YYY(N)$_i$YYY in genes

The YYY-function is applied in the protein coding genes of:
- primates F=CPRI: curve C(YYY;CPRI) (fig 4a);
- rodents F=CROD: curve C(YYY;CROD) (fig 4b);
- mammals F=CMAM: curve C(YYY;CMAM) (fig 4c).

The YYY-function shows local maxima for $i \equiv 8[10]$ (peaks for i=8, 18, 28) followed (for $i \geq 33$) by local maxima for $i \equiv 3[6]$ (peaks for i=33, 39, 45, 51, 57, 63, 69).

The YYY-function is applied in the introns of:
- primates F=IPRI: curve C(YYY;IPRI) (fig 5a);
- rodents F=IROD: curve C(YYY;IROD) (fig 5b);
- mammals F=IMAM: curve C(YYY;IMAM) (fig 5c).

The YYY-function, similar to the RRR-function, shows local maxima for $i \equiv 3[6]$ and a short linear decrease followed by a large exponential decrease (simulated in the last section).

Discussion

Preferential occurrence of YRY(N)$_6$YRY in genes

The YRY-function retrieves in different eukaryotic subpopulations (protein coding and transfer RNA genes of primates, rodents and mammals) obtained from the last release (32) of the EMBL gene database, the preferential occurrence of the motif YRY(N)$_6$YRY which was identified in 1987 with large but various gene populations (eukaryotes, prokaryotes, viruses, chloroplasts, mitochondria). The fact that the YRY-function confirms the previous results, is explained by the law of large numbers ([4] p752, section 2.3.3). Indeed, 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 (*eg* periodicities, maximal and minimal values, etc), even if new genes are further available in the populations F. Therefore, all the non-random properties identified at the gene population level are important as they are stable from a statistical point of view.

New statistical properties common to eukaryotic genes

The existence of statistical properties common to genes was revealed with the motif YRY(N)$_6$YRY. However,

to a lesser extent compared to YRY(N)$_6$YRY (YRY-(N)$_6$YRY is also observed in non-eukaryotic genes), other properties are common to eukaryotic genes: i) local maxima for $i\equiv6[12]$ and $i\equiv8[10]$ in different eukaryotic protein coding gene subpopulations; and ii) local maxima for $i\equiv3[6]$ in different eukaryotic intron subpopulations.

A curve with 100 different points can lead to 100! (10^{158}) possible curve shapes. Therefore, the existence of similar curve shapes obtained either with different gene subpopulations and/or with 'independent' T-functions, is totally unexpected. This result again demonstrates that the nucleotide distribution in genes is non-random and has a common origin. For example, the local maxima for $i\equiv3[6]$ found with the RRR- and YYY-functions in eukaryotic intron subpopulations also exist in the 5' eukaryotic regions with the YRY-function [5]. Furthermore, the simulation models developed have shown that non-random statistical properties common to genes, such as the preferential occurrence of YRY(N)$_6$YRY, can be simulated with operations of insertions and deletions of nucleotides, a process simulating the RNA editing [5]. The following section will show that the non-random properties in the eukaryotic intron subpopulations, identified with the RRR- and YYY-functions, can also be modelised with a process simulating the RNA editing.

Modeling the non-random statistical properties in the eukaryotic intron subpopulations with a process simulating the RNA editing

Introduction

A new genetic process termed RNA editing has been identified showing insertions and deletions of nucleotides in particular genes [7]. Several interesting experimental features have been reported in reviews [7–12]: 1) today, editing is only observed in particular genes such as mitochondrial transcripts of the kinetoplastid protozoa, those of *Physarum polycephalum* and also a few non-mitochondrial systems, *eg* paramyxovirus. However, editing could have been a general mechanism of gene expression or gene modification in primitive genetic systems; 2) at one editing site, most often one nucleotide (but sometimes several nucleotides, up to eight) is inserted or deleted; 3) the editing number is variable: between the protein coding genes, for example, in *Trypanosoma brucei*, editing of the cytochrome oxidase II entails the addition of only four uridines while more than 50% of the cytochrome oxidase III is produced by editing at multiple sites spread across the entire transcript; between the species: for example, cytochrome oxidase III extensively edited in *Trypanosoma brucei* is less edited in *Leish-*

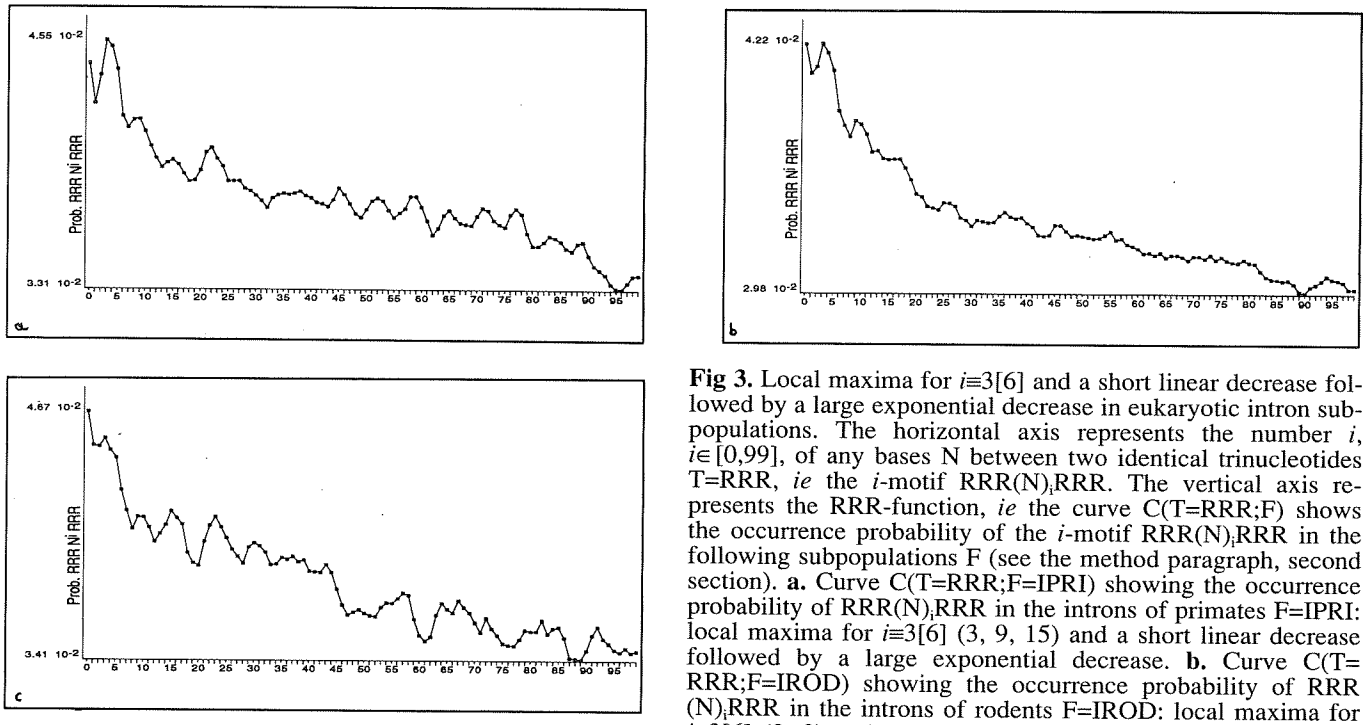


Fig 3. Local maxima for $i\equiv3[6]$ and a short linear decrease followed by a large exponential decrease in eukaryotic intron subpopulations. The horizontal axis represents the number i, $i\in[0,99]$, of any bases N between two identical trinucleotides T=RRR, *ie* the i-motif RRR(N)$_i$RRR. The vertical axis represents the RRR-function, *ie* the curve C(T=RRR;F) shows the occurrence probability of the i-motif RRR(N)$_i$RRR in the following subpopulations F (see the method paragraph, second section). **a.** Curve C(T=RRR;F=IPRI) showing the occurrence probability of RRR(N)$_i$RRR in the introns of primates F=IPRI: local maxima for $i\equiv3[6]$ (3, 9, 15) and a short linear decrease followed by a large exponential decrease. **b.** Curve C(T=RRR;F=IROD) showing the occurrence probability of RRR(N)$_i$RRR in the introns of rodents F=IROD: local maxima for $i\equiv3[6]$ (3, 9) and a short linear decrease followed by a large exponential decrease. **c.** Curve C(T=RRR;F=IMAM) showing the occurrence probability of RRR(N)$_i$RRR in the introns of mammals F=IMAM: local maxima for $i\equiv3[6]$ (3, 9, 15) and a short linear decrease followed by a large exponential decrease.

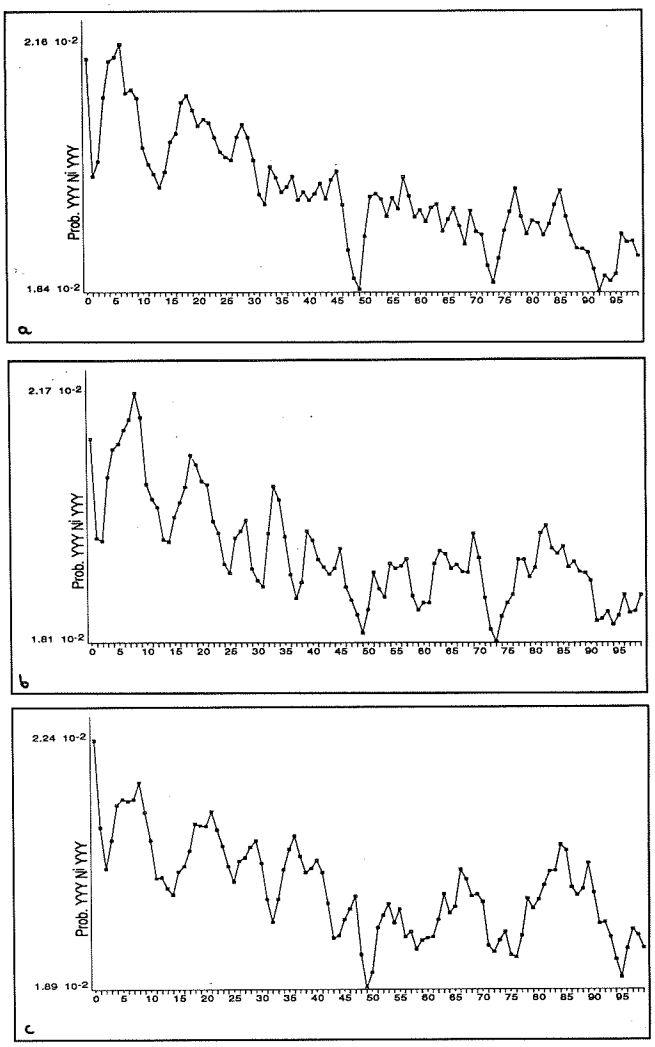


Fig 4. Local maxima for $i \equiv 8[10]$ followed (for $i \geq 33$) by local maxima for $i \equiv 3[6]$ in eukaryotic protein coding gene subpopulations. The horizontal axis represents the number i, $i \in [0,99]$, of any bases N between two identical trinucleotides T=YYY, *ie* the *i*-motif $YYY(N)_iYYY$. The vertical axis represents the YYY-function, *ie*. the curve C(T=YYY;F) shows the occurrence probability of the *i*-motif $YYY(N)_i$ YYY in the following subpopulations F (see the method paragraph, second section). **a.** Curve C(T=YYY;F=CPRI) showing the occurrence probability of $YYY(N)_iYYY$ in the protein coding genes of primates F=CPRI: local maxima for $i \equiv 8[10]$ (18, 28) followed (for $i \geq 33$) by local maxima for $i \equiv 3[6]$ (33, 39, 45, 51, 57, 63, 69). **b.** Curve C(T=YYY; F=CROD) showing the occurrence probability of $YYY(N)_i$ YYY in the protein coding genes of rodents F=CROD: local maxima for $i \equiv 8[10]$ (8, 18, 28) followed (for $i \geq 33$) by local maxima for $i \equiv 3[6]$ (33, 39, 45, 51, 57, 63, 69). **c.** Curve C(T=YYY;F=CMAM) showing the occurrence probability of $YYY(N)_iYYY$ in the protein coding genes of mammals F=CMAM: local maxima for $i \equiv 8[10]$ (8, 18).

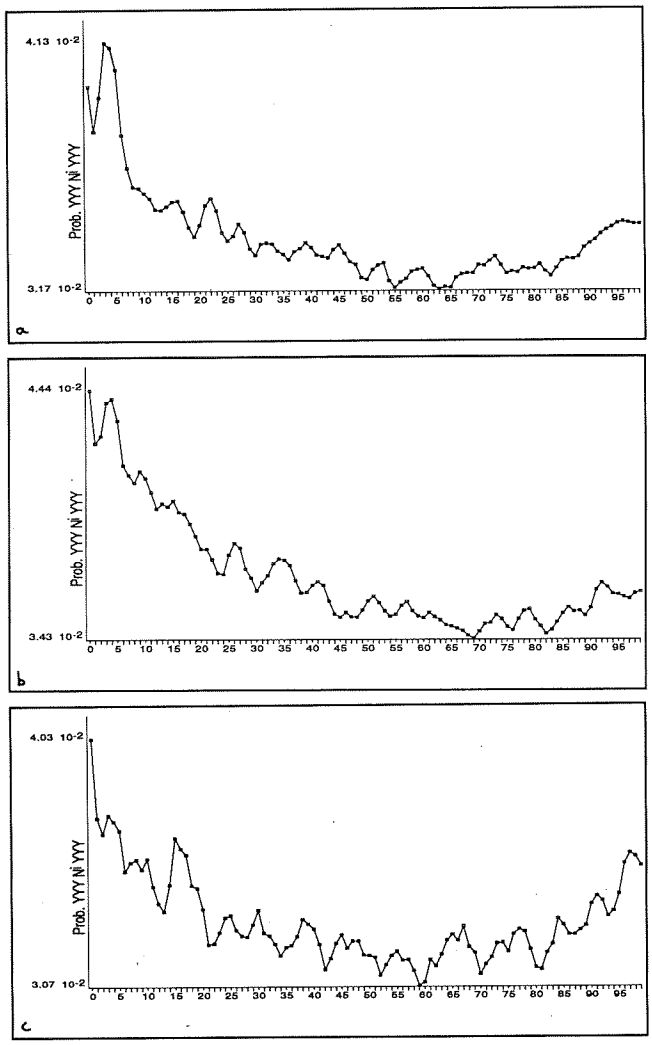


Fig 5. Local maxima for $i \equiv 3[6]$ and a short linear decrease followed by a large exponential decrease in eukaryotic intron subpopulations. The horizontal axis represents the number i, $i \in [0,99]$, of any bases N between two identical trinucleotides T=YYY, *ie* the *i*-motif $YYY(N)_iYYY$. The vertical axis represents the YYY-function, *ie* the curve C(T=YYY;F) shows the occurrence probability of the *i*-motif $YYY(N)_iYYY$ in the following subpopulations F (see the method paragraph, second section). **a.** Curve C(T=YYY; F=IPRI) showing the occurrence probability of $YYY(N)_i$ YYY in the introns of primates F=IPRI: local maxima for $i \equiv 3[6]$ (3, 9, 15, 27, 33, 39, 45) and a short linear decrease followed by a large exponential decrease. **b.** Curve C(T=YYY;F=IROD) showing the occurrence probability of $YYY(N)_iYYY$ in the introns of rodents F=IROD: local maxima for $i \equiv 3[6]$ (3, 9, 15, 21) and a short linear decrease followed by a large exponential decrease. **c.** Curve C(T=YYY;F=IMAM) showing the occurrence probability of $YYY(N)_i$ YYY in the introns of mammals F=IMAM: local maxima for $i \equiv 3[6]$ (3, 15) and a short linear decrease followed by a large exponential decrease.

mania tarentolae and *Crithidia fasciculata*; between protein coding genes and the 5' and 3' regions; for a given gene, several partially edited forms are observed; 4) no specific rule has been identified for the editing site and recently an editing model with random sites has been proposed; 5) insertions and deletions involve the four nucleotides (however, the nucleotide U seems to be more frequently involved according to the data available).

Inspired by the existence of RNA editing in genes, a class of models has been investigated analysing the transformations of an initial population of sequences subjected to operations of nucleotide insertions and deletions.

Method

The models developed here are restricted to the following conditions chosen in order to start 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 consider some features of RNA editing presented in the previous paragraph:

– The initial population of simulated sequences is a language S of infinite sequences (words) obtained by an independent concatenation of given words on the alphabet {R,Y} according to chosen probabilities. Precisely, the chosen simulated population S is constituted by the independent concatenation of the words $RRR(N)_3$, $YYY(N)_3$, R^{30}, Y^{30} according to the probabilities 49%, 49%, 1%, 1% respectively (where N is the base R or Y with probabilities so that R and Y have the same occurrence probability in the whole population).

– The initial sequences are subjected to a nucleotide insertion/deletion process with steps, so that at each step, one insertion of one nucleotide (letter) and one deletion of one nucleotide (letter) occur in the sequence. This condition considers the features of RNA editing described in the point 2 of the previous paragraph.

– The site of insertions and deletions in the sequence is random (see point 4 of the previous paragraph).

– The type (R or Y) of the inserted bases is also random (see point 5 of the previous paragraph).

In order to get significant statistical results, computations have not been made in the theoretical language S but in a simulated version of this language, also called S and made of 500 sequences of 1000 base length generated as previously described. These sequences are generated in such a way that the R percentage is equal to the Y percentage (50%) in any sequence of S (step 0). The computations obtained with such a sample of 0.5 million bases are precise (*ie* there are no random fluctuations in the calculus of probabilities: a sample having 100 sequences of 1000 base length leads to similar results). Then, this population S is subjected to a nucleotide 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 autocorrelation function $i \rightarrow p_i(T;S)$ in S (the same function as defined in the *Method* paragraph, second section) and represented as a curve $C_k(T;S)$. Any simulated curve $C_k(T;S)$ can be compared with the real curve C(T;F) because they result from the same function definition.

Results

As this simulated population S has complementary words in R and Y independently concatenated with identical probabilities, the curves $C_k(RRR;S)$ are identical to the curves $C_k(YYY;S)$.

Curves $C_0(RRR;S)$ or $C_0(YYY;S)$ at step 0 (fig 6a)
Before the insertion/deletion process, the curves $C_0(T;S)$ with T=RRR or T=YYY (curves obtained by applying the T-function in S) are constituted of four subcurves c_1, c_2, c_3 and c_4 of points in decreasing ordinate: c_1, points $(i,p_i(T;S))$ with $i \equiv 3[6]$; c_2, points $(i,p_i(T;S))$ with $i \equiv 2,4[6]$; c_3, points $(i,p_i(T;S))$ with $i \equiv 1,5[6]$; c_4, points $(i,p_i(T;S))$ with $i \equiv 0[6]$.

This decomposition is related to the two words $RRR(N)_3$ and $YYY(N)_3$ (similar to the one described in [5]) and is mainly explained by the fact that the curve c_2 can be obtained when the first or the last base in $(N)_3$ are specified by nucleotides identical to the nucleotides of T, the curve c_3 can be obtained when the two first or the two last bases in $(N)_3$ are specified by nucleotides identical to the nucleotides of T and the curve c_4 can be obtained when the three bases in $(N)_3$ are specified by nucleotides identical to the nucleotides of T.

The decreasing slope of the four subcurves c_1, c_2, c_3 and c_4 until i=30 is related to the two words R^{30} and Y^{30}. Their probabilities decrease because, in a word R^{30} or Y^{30}, the number of i-motifs $T(N)_iT$ decreases since i increases until 30. When i>30, this number becomes constant leading to four horizontal subcurves.

These four decreasing subcurves explain that the maximal value is obtained at i=3, the second highest one, at i=9, etc.

Curves $C_{50}(RRR;S)$ or $C_{50}(YYY;S)$ at step 50 (fig 6b)
After 50 nucleotide insertions/deletions, the four subcurves are gathering into one curve. The maximal value is still at i=3 but the point at i=0 increases (compared to the whole curve which globally decreases).

Curves $C_{180}(RRR;S)$ or $C_{180}(YYY;S)$ at step 180 (fig 6c)
By increasing the number of insertions and deletions of nucleotides, the point at i=0 unexpectedly becomes

higher than the point at i=3, the two values p_3(T;S) and p_9(T;S) being local maxima, such as in reality. Furthermore, these two simulated curves C_{180}(RRR;S) and C_{180}(YYY;S) are strongly similar to the real curves C(RRR;F) (fig 3a–c) and C(YYY;F) (fig 5a–c) of the eukaryotic intron subpopulations (F=IPRI, F=IROD and F=IMAM) because they have also a short linear decrease followed by a large exponential decrease.

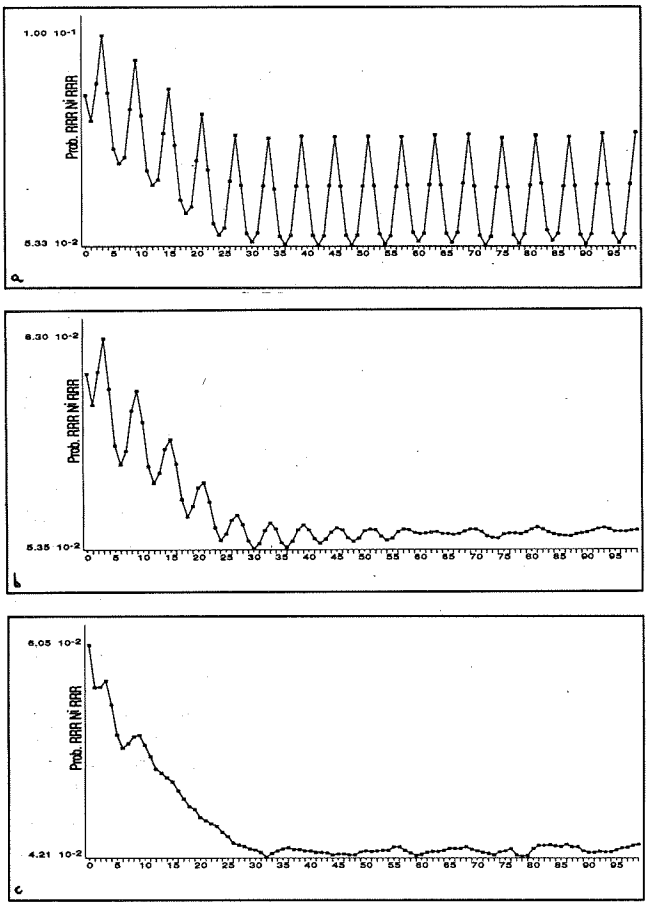


Fig 6. Simulation of the non-random statistical properties in eukaryotic intron subpopulations (local maxima for $i \equiv 3[6]$ and a short linear decrease followed by a large exponential decrease identified in fig 3a–c and 5a–c) with a process of random insertions and deletions of nucleotides in the simulated sequences S={RRR$(N)_3$, YYY$(N)_3$, R^{30}, Y^{30}}* (one random nucleotide insertion and one random nucleotide deletion per sequence per step). The horizontal axis represents the number i, $i \in [0,99]$, of any bases N between two identical trinucleotides T=RRR (or T=YYY), *ie* the i-motif RRR$(N)_i$RRR (or YYY$(N)_i$YYY). The vertical axis represents the RRR- (or YYY-) function, *ie* the curve C(T=RRR;S) (C(T=YYY;S)) shows the occurrence probability of the i-motif RRR$(N)_i$RRR (or YYY$(N)_i$YYY) in the simulated population S at the following process steps (see the method paragraph, last section). **a.** Simulated curve C_0(RRR;S) (or C_0(YYY;S)) at step 0. **b.** Simulated curve C_{50}(RRR;S) (or C_{50}(YYY;S)) at step 50. **c.** Simulated curve C_{180}(RRR;S) (or C_{180}(YYY;S)) at step 180.

Discussion

The T-function can lead to 10^{158} possible curve shapes if all points are different. Furthermore, the shape of the simulated curve for a given insertion/deletion process is unique, *ie* the choice of other initial words, the type of the concatenation (independent, Markov, etc) and finally the use of different numbers of nucleotides which are inserted or deleted, lead to completely different non-random curves [4–6]. These two reasons explain why: i) a real curve which would have been random, could not have been simulated; and ii) the understanding of non-random real curves by the development of simulation models is a difficult problem of pattern recognition. Therefore, initially only the main significant statistical properties can be expected to be simulated with simple models because the genetic reality depends on a great number of parameters.

The steps shown in the figures 6a–c 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. There is an upper limit for the number of steps in the insertion/deletion process. This limit is reached when all p_i(T;S) values are equal to 1/64, *ie* the random situation.

We have recently identified another interesting random insertion/deletion process. Indeed, the process of random insertions and deletions of trinucleotides (words of three letters) leads to the preferential occurrence of YRY$(N)_6$YRY and to the periodicity modulo 3 characterizing the protein coding genes [5].

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 the 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, can unexpectedly lead to non-random properties observed in genes.

Acknowledgment

We thank Dr Nouchine Soltanifar for her advice.

References

1 Arquès DG, Michel CJ (1987) Study of a perturbation in the coding periodicity. *Math Biosc* 86, 1–14

2 Arquès DG, Michel CJ (1987) A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. *J Theor Biol* 128, 457–461

3 Arquès DG, Michel CJ (1990) Periodicities in coding and non-coding regions of the genes. *J Theor Biol* 143, 307–318

4 Arquès DG, Michel CJ (1990) A model of DNA sequence evolution. Part 1: Statistical features and classification of gene populations, 743–753. Part 2: Simulation model, 753–766. Part 3: Return of the model to the reality, 766–770. *Bull Math Biol* 52, 741–772

5 Arquès DG, Michel CJ (1992) A simulation of the genetic periodicities modulo 2 and 3 with processes of nucleotide insertions and deletions. *J Theor Biol* 156, 113–127

6 Arquès DG, Michel CJ, Orieux K (1993) Identification and simulation of new non-random statistical properties common to different populations of eukaryotic non-coding genes. *J Theor Biol,* in press

7 Benne R, Van Den Burg J, Brakenhoff JPJ, Sloof P, Van Boom JH, Tromp MC (1986) Major transcript of the frame-shifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46, 819–826

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

9 Cech TR (1991) RNA editing: world's smallest introns? *Cell* 64, 667–669

10 Feagin JE (1990) RNA editing in kinetoplastid mitochondria. *J Biol Chem* 265, 19373–19376

11 Simpson L (1990) RNA editing – A novel genetic phenomenon? *Science* 250, 512–513

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