Volume 15 Number 18 1987 Nucleic Acids Research

# Periodicities in introns

Didier G.Arquès[1] and Christian J.Michel*

Friedrich Miescher Institut, Bioinformatic group, Mattenstrasse 22, PO Box 2543, CH-4002 Basel, Switzerland and [1]IUT de Belfort, Université de Franche-Comté, rue Engel-Gros, F-90016 Belfort, France

Received June 8, 1987; Revised and Accepted August 18, 1987

## ABSTRACT

The sequence information for the splicing process of introns is found in the consensus sequences at the two splice sites. For long introns, of 300 or more nucleotides, the middle regions may provide additional specificity for splicing which can be investigated by defining an adequate quantitative parameter. This methodology permits to retrieve the coding periodicity in the viral and mitochondrial introns and to identify with a statistical significance, a surprising alternating purine-pyrimidine base sequence -i.e. a modulo 2 periodicity- in the eukaryotic introns, and particularly in the vertebrate introns. This alternating structure suggests that the vertebrate introns do not have the genetic information to code for proteins, they carry structural and regulatory functions.

## INTRODUCTION

The RNA splicing process is related to the genetic information stored in the messenger RNA precursor, and more particularly in the introns which are noncoding sequences. The information in the introns is organized according to the chemical properties of the nucleic acids in order to permit intramolecular interactions, i.e. the self-splicing process [1,2], or cellular factor interactions, e.g. the splicing process with small nuclear ribonucleoprotein particles [3,4]. This information is found in the splice junction sequences, namely in the consensus regions at the 5′ and 3′ splice sites [2,5] and in the branch site [2]. Otherwise, for long introns which range from approximately 300 to over 10000 bases [6], the middle regions should also contain genetic information related to the RNA splicing process [3].

The search of new properties in the introns may also serve as an addition to the existing rules for distinguishing protein coding sequences (PCS) from noncoding sequences (NCS) in eukaryotic organisms. In actual fact, the discrimination between the PCS and the NCS, by making use of probabilistic and statistical methods [7,8], is based on the definition of quantitative parameters which analyse the two following fundamental properties:

i) the existence of a modulo 3 periodicity in the PCS. More precisely, there is a preferential use of the codon RNY [9,10] in the open reading frame [11,12], R being a purine base, Y a pyrimidine one and N any base. We shall name this property "coding periodicity RNY".

ii) the absence of periodicity in the NCS, since the classical parameters, such as the thymine couples separated by i bases, do not reveal any type of periodicity in the NCS [7,8].

Our methodology is based on the definition of a parameter which will be adequate to the study both of the coding periodicity RNY with respect to the undetermined base N and of a modulo 2 periodicity (see below). This approach is applied to the analysis of the middle regions of long introns and allows:

i) to retrieve the coding periodicity RNY in the viral and mitochondrial introns. These results also confirm the reliability of this methodology.

ii) to identify with a statistical significance, a modulo 2 periodicity in the eukaryotic introns.

## METHOD

According to Eigen's theory [9,10], the primordial codon should have the form RNY. If a primitive biochemical process [13,14] led to the selection of the codon RNY, then it can be expected that this codon occurs with the highest probability in all the present-day protein coding genes. Fig. 1 shows the percentages of the codons RNR, RNY, YNR, and YNY calculated over the open reading frame of the eukaryotes (4835 sequences, 1855 kb), of the prokaryotes (953 sequences, 763 kb), of the chloroplasts (182 sequences, 121 kb), of the mitochondria (150 sequences, 112 kb) and of the DNA/RNA viruses (1383 sequences, 1262 kb). These gene populations have been obtained from the EMBL Nucleotide Sequence Data Library (release 10). As a matter of fact, the codon RNY occurs with the highest probability in the open reading frame, which is at least 5% greater than the random occurrence probability. The codon RNR is slightly preferentially used in the open reading frame of the viruses. Similar results have been already shown by chosing a given DNA sequence from the above mentioned populations [11,12].

### Choice of the parameter characterizing the coding periodicity RNY.

Let a motif m be a word on the alphabet {R,Y}. The parameter $P_i(m_1,m_2;s)$ is defined as the occurrence probability for a given sequence s, of the i-motif $m_1(N)_im_2$ constitued of a motif $m_1$ followed by a motif $m_2$ separated

| | RNR | RNY | YNR | YNY |
|---|---|---|---|---|
| Eukaryotic open reading frames | 27.7 | 31.7 | 18.0 | 22.6 |
| Prokaryotic open reading frames | 27.7 | 34.4 | 18.3 | 19.6 |
| Chloroplast open reading frames | 28.0 | 30.7 | 18.5 | 22.8 |
| Mitochondrial open reading frames | 24.6 | 30.7 | 21.7 | 23.0 |
| Viral open reading frames | 30.2 | 30.0 | 18.0 | 21.8 |

Figure 1: Percentages of the codons RNR, RNY, YNR and YNY calculated over the open reading frame of the eukaryotes (4835 sequences, 1855 kb), of the prokaryotes (953 sequences, 763 kb), of the chloroplasts (182 sequences, 121 kb), of the mitochondria (150 sequences, 112 kb) and of the viruses (1383 sequences, 1262 kb).

from $m_1$ by any i bases. For example, the parameter $P_3(YR,YR;s)$ is the occurrence probability of the 3-motif $YR(N)_3YR$ in the sequence s.

An important consequence of a higher occurrence probability of the codon RNY, is the existence of preferential series of consecutive codons RNY, i.e. a series of n codons RNY occurs with the highest probability compared to all the other series of n codons. We lay down herein the problem to find a parameter (see the definition above) which characterizes the coding periodicity RNY with respect to the undetermined base N. In a sequence of consecutive codons RNY, the parameters $P_i(m_1,m_2;s)$ have been compared between themselves in the following cases:

i) $m_1$ and $m_2$ are mononucleotides,<br>
ii) $m_1$ is a mononucleotide and $m_2$ a dinucleotide,<br>
iii) $m_1$ is a dinucleotide and $m_2$ a mononucleotide,<br>
iv) $m_1$ and $m_2$ are dinucleotides.

For any length of a sequence s of codons RNY and for any i index, only the parameter $P_i(YR,YR;s)$ has an occurrence probability which is independent of the base found in the second position of the codon RNY. Particularly, for an infinite sequence s of codons RNY, the parameter $P_i(YR,YR;s)$, with i congruent to 1 modulo 3, is equal to 1/3 and the parameters $P_i(YR,YR;s)$, with i congruent to 0 or 2 modulo 3, are equal to 0.

In addition to the specificity to characterize the coding periodicity RNY, the parameter $P_i(YR,YR;s)$ is also adequate to reveal a modulo 2 periodicity, like an alternating purine-pyrimidine base sequence.

Statistical function and groups of studied introns.

Given an intron group F (see below), let $Q_i(YR,YR;F)$, with i varying between 0 and 39, be the mean of the parameters $P_i(YR,YR;s)$ which are associated to each s of F. F is one of the following groups:

- viral introns (adenovirus), noted F=VI and constituted by 40 sequences (85 kb),
- mitochondrial introns (in plant organisms), noted F=MI and constituted by 22 sequences (26 kb),
- eukaryotic introns, noted F=EI and constituted by 335 sequences (214 kb).

In the EI group, there are the following sub-groups:

- primate introns, noted F=PI and constituted by 112 sequences (92 kb),
- rodent introns, noted F=RI and constituted by 80 sequences (55 kb),
- artiodactyla introns, noted F=AI and constituted by 21 sequences (8 kb),
- insect introns, noted F=II and constituted by 17 sequences (13 kb).

These groups are composed of all the introns whose lengths are greater than 300 bases. In order to avoid the information related to the splice junction sequences (see introduction), the parameter $P_i(YR,YR;s)$ for each intron s is calculated in the range between the 101th base downstream the 5′ splice site and the 100th base upstream the 3′ splice site. For each group F, we represent the autocorrelation function $i \longrightarrow Q_i(YR,YR;F)$ by varying the index i.

## RESULTS

### Periodicity in the introns.

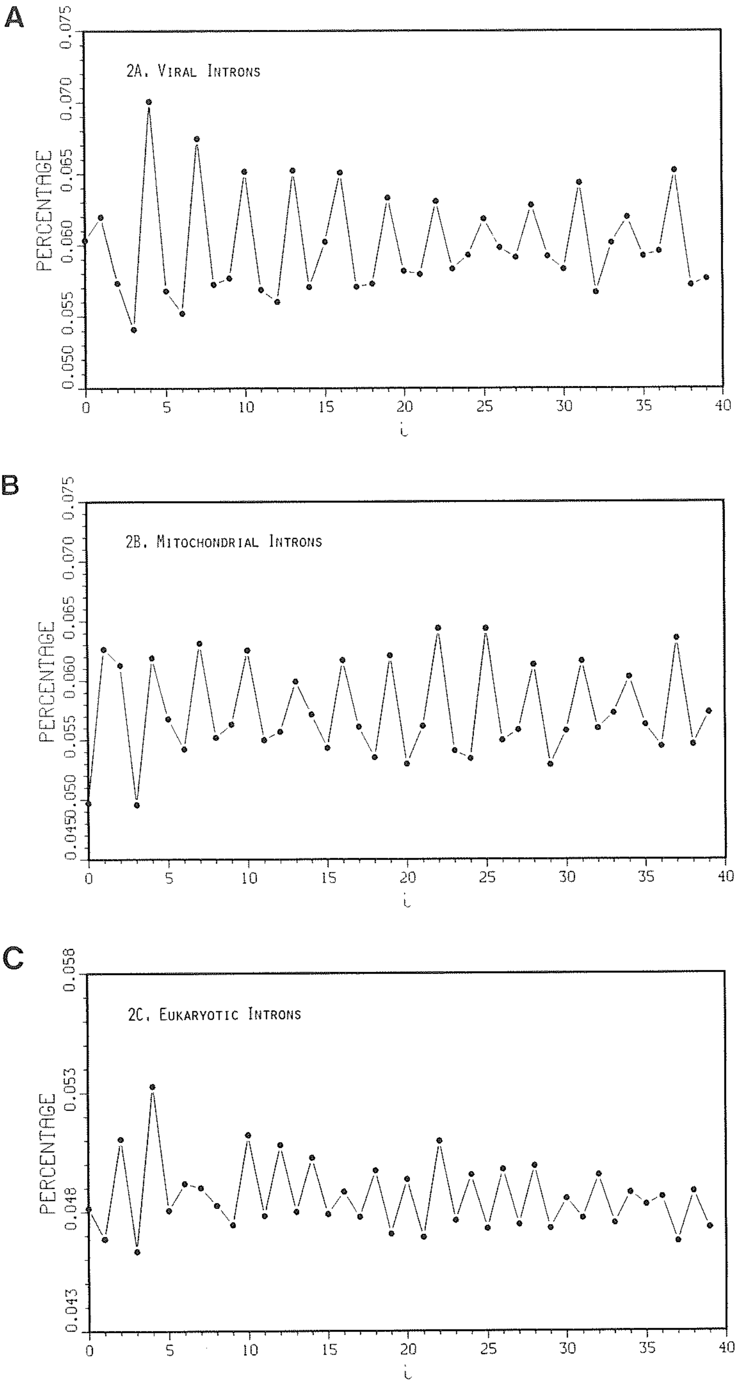
The viral and mitochondrial introns (see fig. 2 (A) and (B)) present a modulo 3 periodicity. The mean $Q_i(YR,YR;F)$, with F=VI and with F=MI, has a higher value (peak) with i congruent to 1 modulo 3 suggesting a coding periodicity RNY (see method).

Surprisingly, the eukaryotic introns show a clear modulo 2 periodicity (see fig. 2 (C)). The mean $Q_i(YR,YR;EI)$ with i congruent to 0 modulo 2 is greater than the one with i congruent to 1 modulo 2, leading to an alternating purine-pyrimidine base sequence, $(YR)_i$.

In order to evaluate the effect due to the large (several tens of nucleic acid bases) alternating purine-pyrimidine runs on this modulo 2 periodicity, one of the referees suggested to delete these large stretches, which are known to be very typical for the introns, in all the sequences in the EI group. In order to estimate the number of stretches $(RY)_n$ and $(YR)_n$ of length n which can be attributed to the random situation, the numbers of $(RY)_n$ and $(YR)_n$ (with $n \geq 2$) in the EI group are compared with the ones obtained with a random group. This random group is generated by randomizing each sequence of the EI group, taking into account its percentage in purine and pyrimidine bases. We shall name F=R this group of randomized sequences. For $n<13$ (resp. $n \geq 13$) the numbers of $(RY)_n$ and $(YR)_n$ in the EI group are smaller (resp. greater) than the ones in the R group. Particularly, for $n>17$, there is no stretch $(RY)_n$ or $(YR)_n$ in the R group, while in the EI group, there are on average two stretches $(RY)_n$ and $(YR)_n$, by varying n between 20 and 50, and one stretch for n=113 (data not shown). Fig. 2 (D) shows the autocorrelation $Q_i(YR,YR;EIM)$ with the EI group modified (EIM) by suppressing all the stretches $(RY)_n$ and $(YR)_n$, with $n \geq 13$. The modulo 2 periodicity is always present even if the variations are slightly reduced. There are four perturbations of this modulo 2 periodicity at i=0,7,16,35. The most important one, which concerns in actual fact three successive points at i=6,7,8, is explained below.

The perturbation in the eukaryotic intron group (see fig. 2(C) and particularly fig. 2(D)), which occurs at i=6 ($Q_6(YR,YR;EI)$ has an unexpected low value), at i=7 ($Q_7(YR,YR;EI)$ has an unexpected high value) and at i=8 ($Q_8(YR,YR;EI)$ has an unexpected low value), is related to a perturbation which exists in all the gene taxonomic groups G [15,16]: in the protein coding genes -of the eukaryotes, of the prokaryotes, of the chloroplasts, of the mitochondria, of the viruses-, in the viral introns, in the ribosomal RNA genes and in the transfer RNA genes. This general perturbation is generated by the mean $Q_6(YRY,YRY;G)$ (associated to the 6-motif $YRY(N)_6YRY$), which has the highest value compared to all the $Q_i(YRY,YRY;G)$ by varying i between 1 and 99. We have put forward that the oligonucleotide $YRY(N)_6$ can be considered as being primitive since it is present with a statistical significance in all the present-day genes and therefore anterior to any molecular evolution. On the other hand, the repetition of this oligonucleotide in series $YRY(N)_6YRY(N)_6...$, the length of this oligonucleotide as well as the properties of the motif YRY could be involved in the spatial structure of the DNA sequences (natural "code" of the helix pitch) [15,16]. This general perturbation also exists in the eukaryotic introns since (see fig. 3):

i) the 6-motif $YR(N)_6YR$ used for the evaluation of $Q_6(YR,YR;EI)$, is in inadequacy with the 6-motif $YRY(N)_6YRY$ (which has the highest occurrence probability, see above) leading to an unexpected low value of $Q_6(YR,YR;EI)$,


A
2A. Viral Introns
PERCENTAGE
0.050
0.055
0.060
0.065
0.070
0.075
0
5
10
15
20
25
30
35
40
i
B
2B. Mitochondrial Introns
PERCENTAGE
0.045
0.050
0.055
0.060
0.065
0.070
0.075
0
5
10
15
20
25
30
35
40
i
C
2C. Eukaryotic Introns
PERCENTAGE
0.043
0.048
0.053
0.058
0
5
10
15
20
25
30
35
40
i

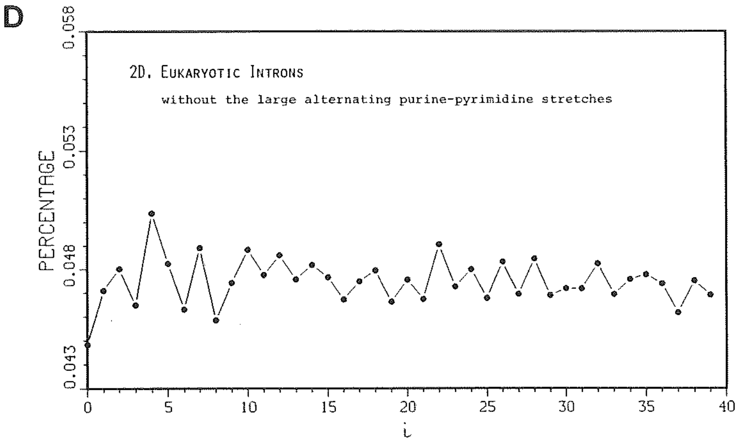


Figure 2: Autocorrelation function i ⟶ $Q_i$(YR,YR;F) (see method) with the viral introns (fig. 2 (A), F=VI, 40 sequences, 85 kb), with the mitochondrial introns (fig. 2 (B), F=MI, 22 sequences, 26 kb), with the eukaryotic introns (fig. 2 (C), F=EI, 335 sequences, 214 kb) and with the eukaryotic introns without the large alternating purine-pyrimidine stretches (fig. 2(D), F=EIM).

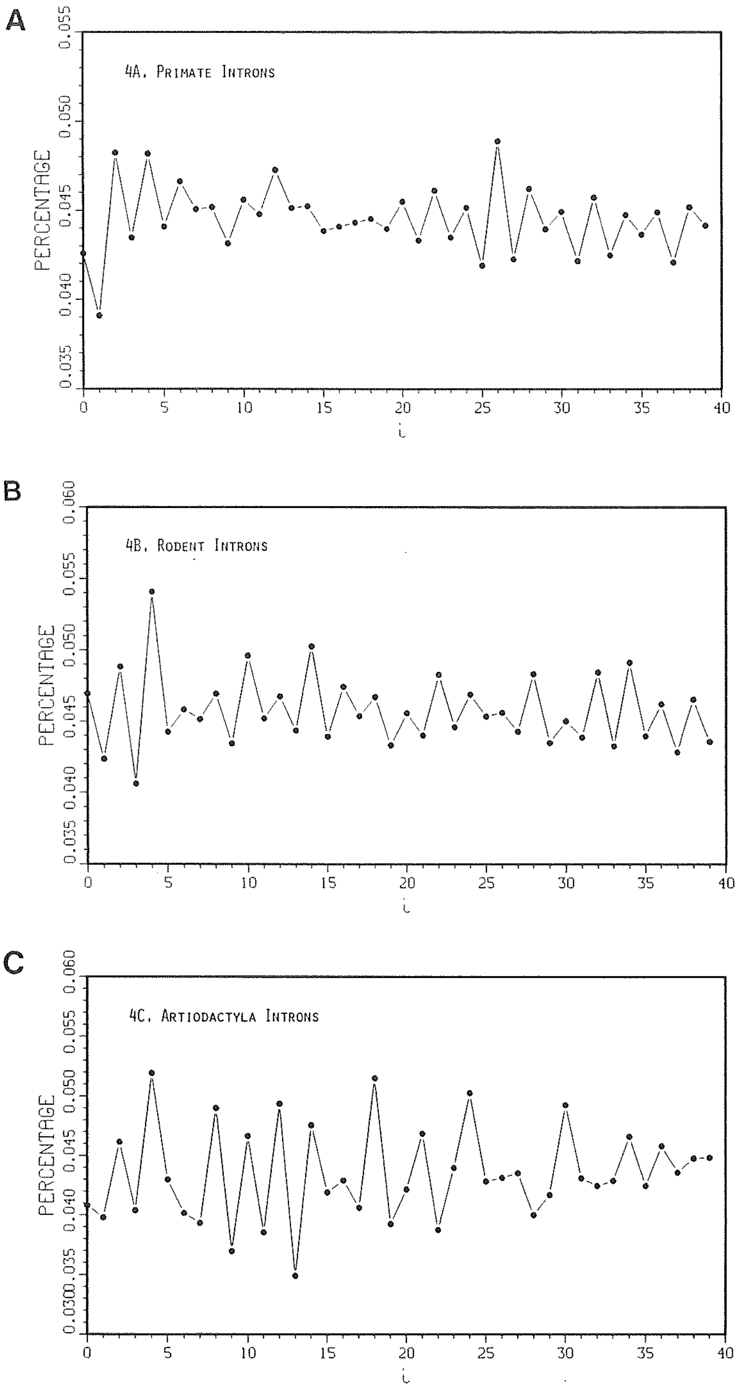
i.e. $m_1$=YR in YR$(N)_6$YR is complementary to the last two bases of $m_1$=YRY in YRY$(N)_6$YRY, $m_2$=YR in YR$(N)_6$YR corresponding to the first two bases of $m_2$=YRY in YRY$(N)_6$YRY,

ii) in the same way, the 7-motif YR$(N)_7$YR of $Q_7$(YR,YR;EI) is in adequacy with the 6-motif YRY$(N)_6$YRY leading to an unexpected high value of $Q_7$(YR,YR;EI) and the 8-motif YR$(N)_8$YR of $Q_8$(YR,YR;EI) is in inadequacy with the 6-motif YRY$(N)_6$YRY leading to an unexpected low value of $Q_8$(YR,YR;EI).

This modulo 2 periodicity is well characterized in the primate and rodent introns (see fig. 4 (A) (B)). For the artiodactyla and insect introns, most of the peaks are always found with an index i congruent to 0 modulo 2

| 6-motif YRY$(N)_6$YRY: | Y R Y . . . . . . Y R Y | General motif |
|---|---|---|
| 6-motif YR$(N)_6$YR: | Y R . . . . . . Y R | Inadequacy |
| 7-motif YR(N)7YR: | Y R . . . . . . . Y R | Adequacy |
| 8-motif YR$(N)_8$YR: | Y R . . . . . . . . Y R | Inadequacy |

Figure 3: Adequacy of the 7-motif YR$(N)_7$YR and inadequacies of the 6-motif YR$(N)_6$YR and of the 8-motif YR$(N)_8$YR with the general 6-motif YRY$(N)_6$YRY (see method and results).


A
4A. Primate Introns
PERCENTAGE
0.035
0.040
0.045
0.050
0.055
0
5
10
15
20
25
30
35
40
i
B
4B. Rodent Introns
PERCENTAGE
0.035
0.040
0.045
0.050
0.055
0.060
0
5
10
15
20
25
30
35
40
i
C
4C. Artiodactyla Introns
PERCENTAGE
0.030
0.035
0.040
0.045
0.050
0.055
0.060
0
5
10
15
20
25
30
35
40
i

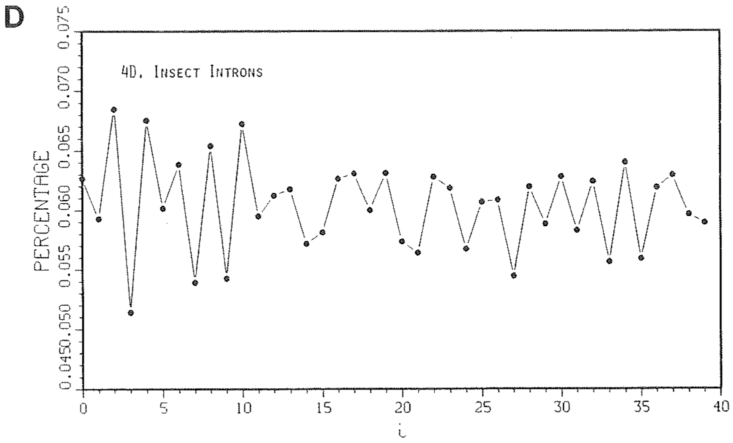


Figure 4: Autocorrelation function i⟶ $Q_i$(YR,YR;F) (see method) with the primate introns (fig. 4 (A), F=PI, 112 sequences, 92 kb), with the rodent introns (fig. 4 (B), F=RI, 80 sequences, 55 kb), with the artiodactyla introns (fig. 4 (C), F=AI, 21 sequences, 8 kb) and with the insect introns (fig. 4 (D), F=II, 17 sequences, 13 kb).

(see fig. 4 (C) (D)). The lesser statistical significance of the results with the AI and II groups is probably due to the small size of these samples. The modulo 2 periodicity is not statistically significant in the chloroplast, plant and fungi introns by making use of this methodology (data not shown).

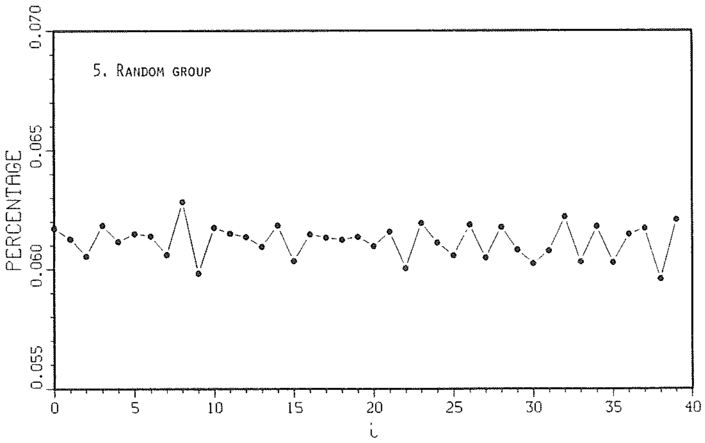


Figure 5: Autocorrelation function i⟶ $Q_i$(YR,YR;R) (see method) with the random group R, which is obtained by randomizing each sequence in the eukaryotic intron group (335 sequences, 214 kb) taking into account its percentage in purine and pyrimidine bases.

Statistical significance of the eukaryotic intron modulo 2 periodicity.

In order to verify the statistical significance of this modulo 2 periodicity, the autocorrelation function i⟶ $Q_i$(YR,YR;R) is computed to the random group R (above defined). The fig. 5 shows the absence of modulo 2 pattern since the distribution of the peaks is congruent to 0 modulo 2 as well as congruent to 1 modulo 2. Futhermore, the variations between two points are smaller than the ones in the fig. 2 (C) and 2 (D). These two facts are also found by making use of an alternating approach which consists of generating a unique, large (e.g. 100000 bases) random sequence, taking into account the percentage in purine and pyrimidine bases calculated over all the eukaryotic introns (data not shown).

In addition, this modulo 2 periodicity is always present by varying the studied range of the eukaryotic introns, which can start between the 40th and the 300th base downstream the 5′ splice site and which can end between the 40th and the 300th base upstream the 3′ splice site (data not shown).

This modulo 2 periodicity is specific to the eukaryotic introns since it is found neither with the protein coding genes nor with the ribosomal and transfer RNA genes (data not shown).

## DISCUSSION

The existence of a coding periodicity RNY in the viral and mitochondrial introns asserts that these groups of introns have the genetic information to code for proteins. Adenoviruses use overlapping genes, both DNA strands and alternative patterns of RNA splicing in order to control the synthesis of several different proteins from the same RNA transcript [17]. These three properties lead to maximizing the functions of a viral genome whose size is small. Many mitochondrial genomes in plants contain introns which are present both in protein coding sequences and in ribosomal RNA genes. These introns encode splicing proteins (maturases) by the intron sequence itself or by an open reading frame which continues through the splice junction into the intron sequence [18-20]. This property seems limited to situations where transcription and translation occur in the same cellular compartment, such as the mitochondria.

The absence of coding periodicity RNY in the eukaryotic introns suggests that this group of introns does not contain the genetic information to code for proteins. The existence of an alternating purine-pyrimidine base sequence, well verified with the vertebrate introns, argues in favour of structural and regulatory functions. We briefly review a few experimental results which suggest such functions for the middle regions of the eukaryotic introns.

An alternating structure may facilitate recombinations in introns which could provide a mechanism for the shuffling of exonic sequences in order to produce new genes from parts of old ones, without sacrificing the original genes [21] (see the examples with the collagen gene [22], with the chicken ovomucoid gene [23] etc.).

An alternating structure may also lead to a Z-DNA form which is involved in the regulatory functions [24]. Therefore, the vertebrate introns could control the alternative patterns of RNA splicing which allow the functional diversity of single genes (see the examples with the rat troponin T gene [25], with the immunoglobulin genes [26] etc.). On the other hand, the vertebrate introns could provide additional specificity for splicing which can be involved in:

i) the hairpin formation in order to shorten the long introns and therefore to a faster exision [27],

ii) the limitations both of the interchangeability of the splice junction signals, which exists with the chimeric introns [28], and of the skipping of an exon,

iii) the discrimination between true splice junctions and other regions that fit the consensus sequences, since new splice sites (called cryptic) can become abnormally activated in some genetic diseases, such as the thalassemias [29],

iv) the interactions with heterogeneous nuclear ribonucleoprotein particles [4,30], ...

This modulo 2 periodicity is not simply stored as stretches $(YR)_i$. As a matter of fact, the autocorrelations $Q_i(R,R;EI)$, $Q_i(R,Y;EI)$, $Q_i(Y,R;EI)$ and $Q_i(Y,Y;EI)$ do not reveal this modulo 2 periodicity (data not shown, see section results and [7,8]). By analogy with Eigen's theory [9,10] which proposes a periodicity $(RNY)_i$ for the primordial genes, we put forward the hypothesis that the primordial primary structure of the eukaryotic introns was built up of stretches $(YR)_i$. These two types of initial information were transformed during evolution since the compositional constraints affect both coding and noncoding sequences [31], but the traces of the perfect coding periodicity $(RNY)_i$ and the traces of the perfect intron periodicity $(YR)_i$ are still significant by making use of statistical methodologies.

Finally, this modulo 2 periodicity can be used as discriminating parameters [8], which can serve as an addition to the existing rules for distinguishing coding sequences from noncoding sequences in eukaryotic organisms.

## ACKNOWLEDGEMENTS

We would like to thank Professors Max Burger, Jacques Streith, Dr Christoph Nager and the referees for their advice, and the Friedrich Miescher Institut bioinformatic group for its assistance. This work was supported by grants from the Friedrich Miescher Institut to C.J.M..

*To whom correspondence should be addressed

## REFERENCES

1. Cech, T.R. and Bass, B.L. (1986) Ann. Rev. Biochem. 55, 599-629.
2. Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S. and Sharp, P.A. (1986) Ann. Rev. Biochem. 55, 1119-1150.
3. Sharp, P.A. (1987) Science 235, 766-771.
4. Maniatis, T. and Reed, R. (1987) Nature 325, 673-678.
5. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
6. Naora, H. and Deacon, N.J. (1982) Proc. Natl. Acad. Sci. USA 79, 6196-6200.
7. Fickett, J.W. (1982) Nucl. Acids Res. 10, 5303-5318.
8. Michel, C.J. (1986) J. Theor. Biol. 120, 223-236.
9. Eigen, M. (1971) Naturwissenschaften 58, 465-523.
10. Eigen, M. (1978) Naturwissenschaften 65, 341-369.
11. Shepherd, J.C.W. (1981) Proc. Natl. Acad. Sci. USA 78, 1596-1600.
12. Shepherd, J.C.W. (1981) J. Mol. Evol. 17, 94-102.
13. Orgel, L.E. (1968) J. Mol. Biol. 38, 381-393.
14. Orgel, L.E. (1986) J. Theor. Biol. 123, 127-149.
15. Arques, D.G. and Michel, C.J. Math. Biosci. (in press).
16. Arques, D.G. and Michel, C.J. J. Theor. Biol. (in press).
17. Ziff, E.B. (1980) Nature 287, 491-499.
18. Lazowska, J., Jacq, C. and Slonimski, P.P. (1980) Cell 22, 333-348.
19. Borst, P. and Grivell, L.A. (1981) Nature 289, 439-440.
20. Davies, R.W., Waring, R.B., Ray, J.A., Brown, T.A. and Scazzochio, C. (1982) Nature 300, 719-724.
21. Gilbert, W. (1978) Nature 271, 501.
22. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, E.V., Sullivan, M., Pastan, I. and De Crombrugghe, B. (1980) Proc. Natl. Acad. Sci. USA 77, 7059-7063.
23. Stein, J.P., Catterall, J.F., Kristo, P., Means, A.R. and O'Malley, B.W. (1980) Cell 21, 681-687.
24. Dickerson, R.E. (1983) Scient. Amer. 249, 86-102.
25. Breitbart, R.E., Nguyen, H.T., Medford, R.M., Destree, A.T., Mahdavi, V. and Nadal-Ginard, B. (1985) Cell 41, 67-82.
26. Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L. and Wall, R. (1980) Cell 20, 303-312.
27. Solnick, D. (1985) Cell 43, 667-676.
28. Chu, G. and Sharp, P.A. (1981) Nature 289, 378-382.
29. Treisman, R., Orkin, S.H. and Maniatis, T. (1983) Nature 302, 591-596.
30. Choi, Y.D., Grabowski, P.J., Sharp, P.A. and Dreyfuss, G. (1986) Science 231, 1534-1539.
31. Bernardi, G. and Bernardi, G. (1986) J. Mol. Evol. 24, 1-11.