Study of a Perturbation in the Coding Periodicity

DIDIER G. ARQUÈS
I. U. T. de Belfort, Université de Franche-Comté, rue Engel-Gros, F-90016 Belfort-Cedex, France

AND

CHRISTIAN J. MICHEL*
Biozentrum, Basel University, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

ABSTRACT

The statistical study of different populations of genes, with principal component analyses (PCA) and with mean curves, shows: (1) The occurrence probability of the dinucleotide $D = RY$ ($D' = YR$), R being a purine base, Y a pyrimidine base, and N any base, after the occurrence of the nucleotide Y (R) in the zero modulo three curve—in the eukaryotic protein coding genes—presents a modulo 9 periodicity with a maximum value nine bases after Y (R). This modulo 9 periodicity is added to the existing coding modulo 3 periodicity RNY, i.e. to the preferential use of the codon RNY in the open reading frame. (2) The occurrence probability of the trinucleotide $T = YRY$ *after* the occurrence of the nucleotide Y—in the protein coding genes of eukaryotes, chloroplasts, and mitochondria, and in the transfer RNA genes—presents a maximum value eight bases after Y. (3) Similar results are obtained (with less statistical significance) with other gene taxonomic groups (viral protein coding genes, viral introns) and with the complementary motif $R(N)_8RYR$. These results may suggest that the motifs $Y(N)_8YRY$, $R(N)_8RYR$, $Y(N)_9RY$, and $R(N)_9YR$ could have a function related to the spatial structure of the DNA sequences.

1. INTRODUCTION

The DNA double helix contains in its nucleotide sequence several types of periodicities:

(1) some well-defined periodicities which are related to the transcription-translation function of the codons,

*To whom correspondence should be addressed. Present address: Friedrich Miescher Institut, Mattenstrasse 22, P.O. Box 2543, CH-4002 Basel, Switzerland; telephone (061) 37 46 48.

 0025-5564/87/$03.50

(2) some lesser-known periodicities which are related to the space constraints assigned by the secondary and tertiary structure of the DNA molecule.

Several studies, which are generally based on the statistical and probabilistic analyses, have permitted characterizing these periodicities [4, 8]. Shepherd [9] determined, by correlation testing, which frame differs the least from a supposed original protein coding sequence, where the codons should have the form *RNY*. This coding periodicity is constant and equal to three nucleotides. Trifonov and Sussman [10] interpreted the mean periodicity of 10.5 for some dinucleotides as the preference of some nucleotides to be in the third position of the codons, in order to facilitate the deformational anisotropy of the DNA molecule. On the other hand, a correlation between the frequencies of some *R-Y* doublets, triplets, and quartets and some DNA structural parameters (angular deviation of torsion, angular deviation of roll, etc.) has been described, but not studied in terms of periodicities [6].

Our statistical study—carried out on DNA sequence samples which have been obtained from the EMBL Nucleotide Sequence Data Library (release 10)—is presented in terms of mean curves and principal component analyses (PCA), whose intuitive interpretation of the graphs is explained in Section 2.

2. PRINCIPAL COMPONENT ANALYSIS (PCA)

The PCA is a graphical statistical method which permits computer study of a large matrix of numbers without any statistical hypothesis. In our study, each line of the matrix represents a DNA sequence, each column a quantitative parameter. The current element of the matrix is the value of the parameter of its column for the sequence associated to its line. This matrix with n rows and p columns is then considered:

(1) as the set of its columns, points in R^n (variable space), or
(2) as the set of its rows, points in R^p (individual space).

The visualization of the two abovementioned clouds of points is not possible under these conditions. Therefore, the goal of the PCA is to seek the plane (or the successive planes), called in this paper the *first factor plane*, which best adjusts these two clouds according to a least squares approximation. Then, the studied cloud is represented on this plane according to an orthogonal projection.

Prior to this geometrical study, the matrix of the initial numbers is modified in order to achieve an invariable analysis for the measure unit. This modification places the cloud of the variable space in the sphere of center 0 and of radius equal to one. This is the reason why, in the graphs below, the cloud of variable points is projected inside a circle (called correlation-circle)

of center 0 and of radius equal to one (trace of the sphere in the plane). The proximities between variable points or groups of variable points are interpreted in terms of correlation. Two nearest points are strongly correlated, i.e., their associated parameters behave in the same way with respect to the set of individuals. On the other hand, two points separated by an angle of 90° with the origin 0 are not correlated, i.e., their associated parameters behave independently with respect to the set of individuals.

We refer the reader to Lebart et al. [5] and Michel [8] for more details about this technique.

3. RESULTS

A. STUDY OF A PERTURBATION IN THE EUKARYOTIC CODING MODULO 3 PERIODICITY RNY

The statistical study presented below has been determined with a population of 1870 eukaryotic protein coding genes (1638 kb) composed of all the sequences whose lengths are greater than 300 bases. The same results have been obtained with the analyses of several samples extracted from this population.

a. Analysis of the Eukaryotic Coding Modulo 3 Periodicity RNY. The parameter T_i, i varying between 0 and 29, is defined as the occurrence probability in the studied sequence of a purine nucleotide R followed by a pyrimidine-purine dinucleotide $D = YR$ separated from R by any i bases, i.e. the motif $R(N)_iYR$.

The first factor plane (which contains 68% of the initial information of the variable point cloud; see Figure 1) shows three groups of well-separated parameters formed of the parameters T_i, with i respectively congruent to 0, 1, and 2 modulo 3. We shall call these three groups G_0, G_1, and G_2 respectively. This result agrees with the existence of a preferential use of the codon RNY [3] in the open reading frame [9].

In a first approach, let we assume that the coding periodicity RNY is perfect (see Figure 2). Then the dinucleotide $D = YR$ is necessarily found in the third position of a codon for Y and in the first position of the following codon for R. Let us fix this dinucleotide $D = YR$ in such admissible position. Then the first nucleotide R of the parameter T_i, necessarily situated upstream of the dinucleotide $D = YR$, can only be found in two types of positions:

(1) If R is in the first position of a codon, then the parameters T_i with i congruent to 1 modulo 3 have an occurrence probability p equal to 1.

(2) If R is in the second position of a codon, then the parameters T_i with i congruent to 0 modulo 3 have an occurrence probability p equal to 0.5.

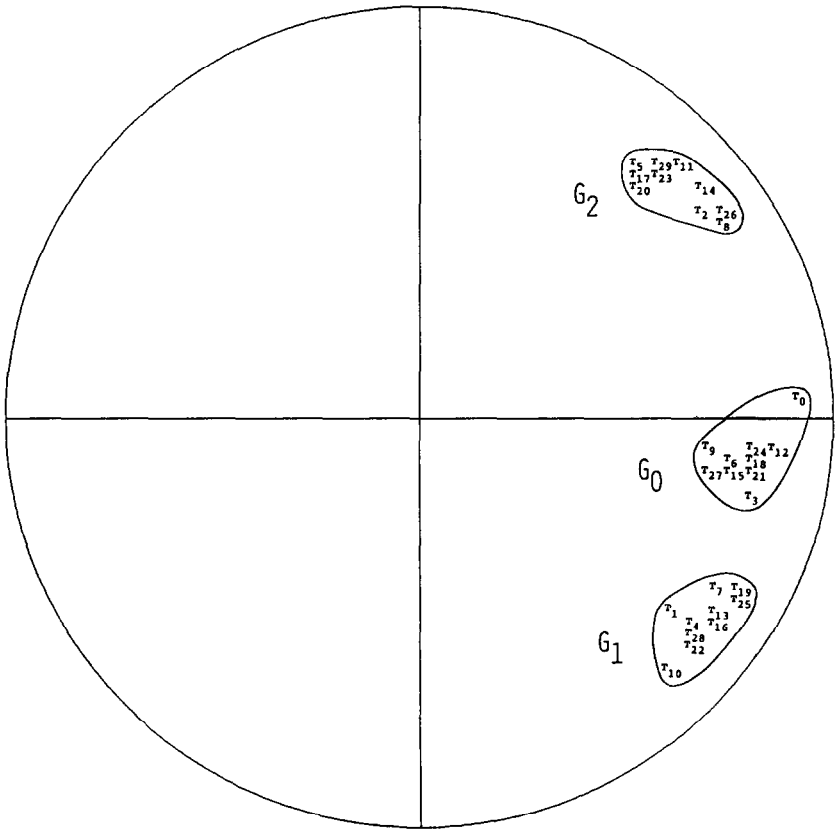

FIG. 1. T_i parameter analysis (see Section 3.A.a). Principal component analysis on the T_i parameter space projected on the first factor plane.

Base position in the codons:	1 2 3 1 2 3 1 2 3	p
Open reading frame:	*RNYRNYRNY*	
T_i, i congruent to 0 modulo 3:	$R \cdot \cdot \cdot YR$	.5
T_i, i congruent to 1 modulo 3:	$R \cdot \cdot \cdot \cdot YR$	1
T_i, i congruent to 2 modulo 3:	impossible	0

FIG. 2. Occurrence probability p of the parameters T_i (see Section 3.A.a), with i congruent to 0, 1, and 2 modulo 3, in the case of a hypothetical perfect coding periodicity *RNY*.

Obviously, the occurrence probability p of the parameters T_i with i congruent to 2 modulo 3 is equal to 0.

In reality, the coding periodicity *RNY* is not perfect. It is only preferential, since the occurrence probability of the codon *RNY* (32.2%) is greater than the occurrence probabilities of the codons *RNR* (27.8%), *YNR* (17.7%), and *YNY* (22.3%). Therefore, the previous probabilistic reasoning must be balanced with the fact that the perfect periodicities of the codons *RNR*, *YNR*, and *YNY* can give different occurrence probabilities of the parameters T_i. Furthermore, the exact values of these probabilities also depend on series of alternating types of codons. In actual fact, the existence of the three groups of parameters, G_0, G_1, and G_2, is related to weak variations of percentages which explain why these three groups are relatively well correlated, i.e. near in the first factor plane (see Figure 1).

b. Analysis of the Group G_0. The existence of the three groups of parameters, G_0, G_1, and G_2, is the consequence of a known rule. The question may be asked whether there is no other rule which can discriminate the parameters belonging to the same group. In this section, we shall demonstrate that the distribution of the parameters T_i in the group G_0 follows a modulo 9 periodicity.

We apply the PCA to the subgroup of parameters T_i in G_0. The first factor plane (which contains 76% of the initial information; see Figure 3) clearly shows that the group G_0 is split up into two subgroups. We shall denote by S the one having the parameters T_i congruent to 0 modulo 9, i.e. T_0, T_9, T_{18}, and T_{27}. In this group S, the parameter T_9 gives a maximum deformation of the cloud of these four points.

c. Interpretation of the Existence of the Group S with the Mean Curves. The existence of the four groups G_0, G_1, G_2, and S, obtained with the PCA, is explained in particular by the study of the curves associating with each parameter T_i its mean t_i calculated on the set of all the studied sequences, i varying between 0 and 44.

Figure 4 shows three separated curves which agree, from the top down, with the groups G_1, G_0, and G_2 respectively. This separation into three curves has been analysed in Section 3.A.a in terms of the preferential use of the codon *RNY* in the open reading frame. The existence of the group S accounts for the four peaks at the means t_9, t_{18}, t_{27}, and t_{36} of the parameters T_9, T_{18}, T_{27}, and T_{36} in the curve associated with G_0. This modulo 9 periodicity in the middle curve G_0 decreases from $i = 36$. The existence of these four peaks is statistically significant at the 1% level. Indeed, if three numbers are randomly generated, then the probability of having a peak (i.e., the second number having the highest value) is obviously equal to $1/3$. If we suppose now that the parameters T_i, with i equal to 0 modulo 3, are independent variables, then the probability of having four

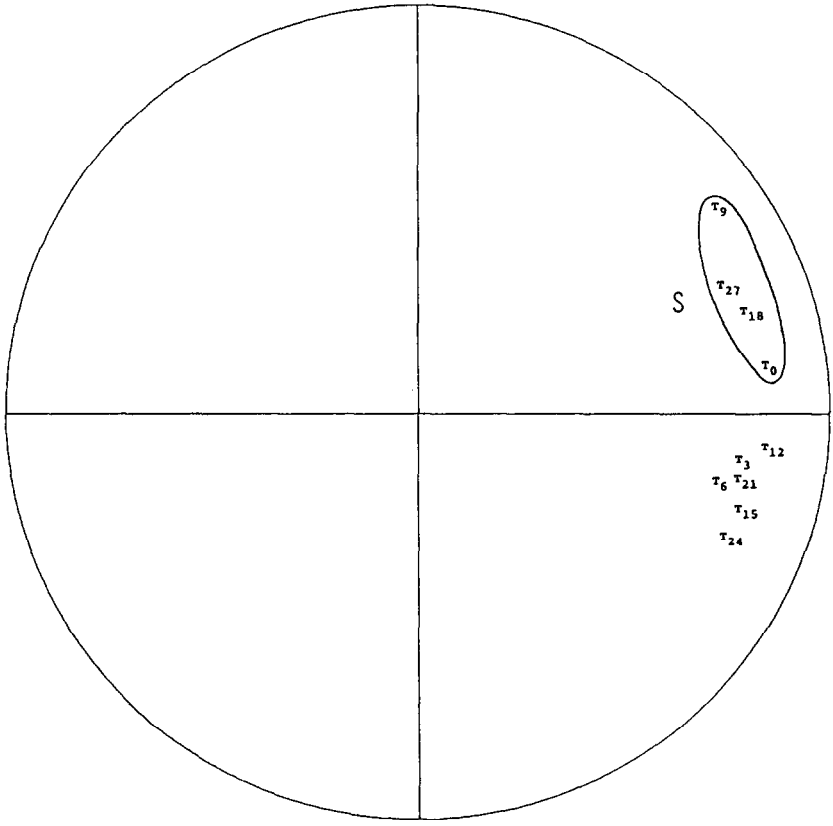

FIG. 3. T_i parameter analysis with i congruent to 0 modulo 3 (see Section 3.A.b). Principal component analysis on the T_i parameter space projected on the first factor plane.

successive peaks (i.e. four independent triplets of numbers, the second one having the highest value) is $3^{-4} = 1.23\%$. In reality, the result is less than 3^{-4} because the statistical reasoning should also take into account the fact that the means t_9, t_{18}, and t_{27} have the greatest values on the curve associated with G_0. As a matter of fact, the statistical test which compares two means of two parameters (see appendix) shows a statistical significance at the 0.1% level for the following six differences: $t_9 - t_6$, $t_9 - t_{12}$, $t_{18} - t_{15}$, $t_{18} - t_{21}$, $t_{27} - t_{24}$, and $t_{27} - t_{30}$.

d. Complementary Results. Similar results are obtained from the study of the parameters T_i' which are deduced from the parameters T_i by permutation in their definition (see Section 3.A.a) of R and Y (data not shown).

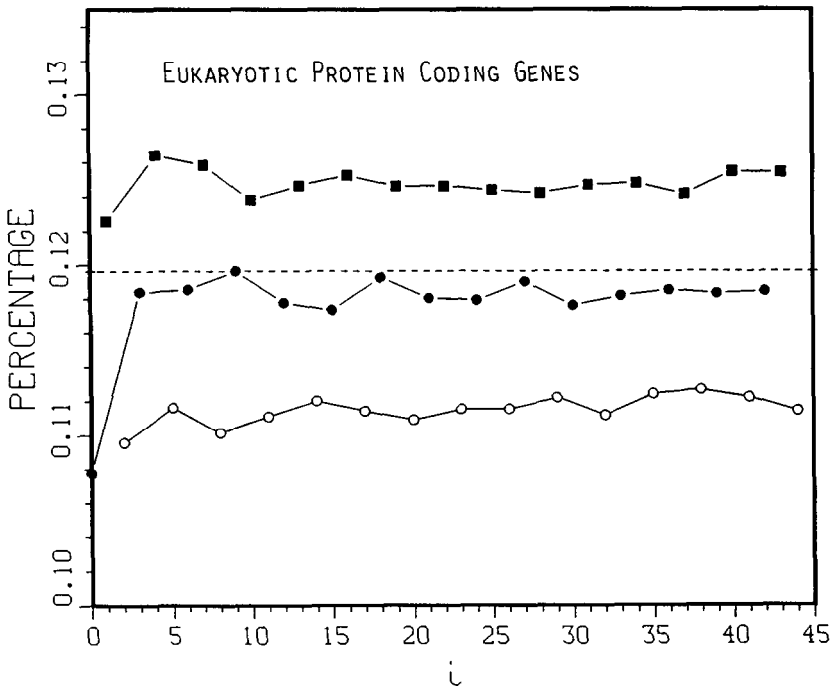

FIG. 4. Occurrence probability of the motif $R(N)_iYR$: curves of the means t_i of the parameters T_i (see Section 3.A.c). A horizontal dashed line goes through the point $(9, t_9)$.

e. Results 1: Perturbation of the Eukaryotic Coding Modulo 3 Periodicity RNY. The occurrence probability of the dinucleotide $D = RY$ ($D' = YR$) after the occurrence of the nucleotide Y (R) in the zero modulo three curve —in the eukaryotic protein coding genes—presents a modulo 9 periodicity with a maximum value nine bases after Y (R), i.e. the motifs $Y(N)_9RY$ and $R(N)_9YR$. Furthermore, the troughs in position i equal 8 (curve associated to G_2) and in position i equal 10 (curve associated to G_1) reflect the inadequacy of the motifs $R(N)_8YR$ and $R(N)_{10}YR$ with the motif $R(N)_9YR$. We shall see in the next section that the variations of the three curves G_0, G_1, and G_2 for i equal to 8, 9, and 10 are generalized to other gene taxonomic groups if we particularize the studied dinucleotide.

B. A MOTIF RELATED TO THE PREVIOUS PERTURBATION

In order to extend the results 1, we have analysed all the trinucleotides T, i.e. $RRR, \ldots, YYY$, which follow either a nucleotide R or a nucleotide Y. Only the motif $Y(N)_8YRY$ allows the generalization of the previous results

1 with the following populations of genes [the complementary motif $R(N)_8RYR$ has less statistically significant results; see below]:

(1) Eukaryotic protein-coding genes, denoted EC and constituted by 2271 sequences (1750 kb).

(2) Chloroplast protein-coding genes, denoted CC and constituted by 121 sequences (116 kb).

(3) Mitochondrial protein-coding genes, denoted MC and constituted by 130 sequences (117 kb).

(4) Transfer *RNA* genes, denoted TR and constituted by 920 sequences (70 kb), which are eukaryotic, prokaryotic, chloroplast, and mitochondrial.

The EC, CC, and MC groups are taken from all the sequences whose lengths are greater than 250 bases, while the TR group is taken from all the sequences whose lengths are greater than 65 bases, because the length of almost all the transfer RNA genes is shorter than 100 bases.

a. Mean Curves of the Motif $Y(N)_iYRY$. We define a new set of parameters as follows: The parameter Q_i is the occurrence probability in the studied sequence of a nucleotide Y followed by a trinucleotide $T = YRY$ separated from Y by any i bases, i.e., the motif $Y(N)_iYRY$. The index i varies between 0 and 99 for the EC, CC, and MC groups, and between 0 and 28 for the TR group. Figure 5(a), (b), (c), (d) shows, for each of the above groups, the mean q_i of the parameter Q_i, calculated on the set of all the sequences in the studied group.

For the graphs concerning the protein coding genes [Figure 5(a), (b), (c)], the points are joined to make three curves, congruent to 0, 1, and 2 modulo 3. The preferential use of the codon RNY in the open reading frame mainly separates the parameters Q_i into two groups: the group H_1, which consists of the parameters Q_i with i congruent to 1 modulo 3 (the lower curve), and the group H, which consists of the parameters Q_i with i congruent to 0 and 2 modulo 3 (the two higher curves). In order to follow the previous reasoning with subsets of points with identical behavior modulo three, we have always grouped the parameters with residue gaps that are equal modulo 3, even if the two higher curves have nearly the same means.

The graphs in Figure 5(a), (b), (d) show that the occurrence probability of the motif $Y(N)_iYRY$ has a maximum value for i equal to 8:

(1) for the EC and CC groups, with i congruent to 2 modulo 3 in the range [0, 99];

(2) for the TR group, with i varying between 0 and 28.

For the MC group [see Figure 5(c)], the q_8 value is the second one after the q_{32} value with i congruent to 2 modulo 3, in the range [0, 99].

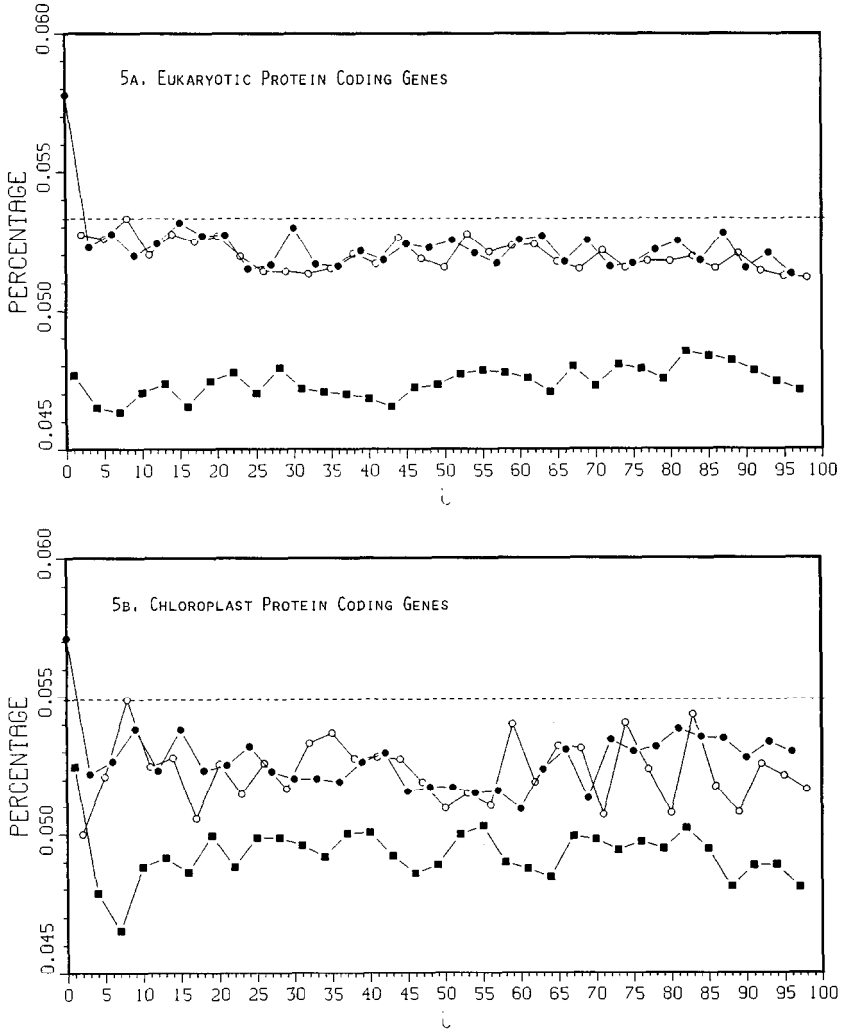

FIG. 5. Occurrence probability of the motif $Y(N)_iYRY$: curves of the means q_i of the parameters Q_i (see Section 3.B.a) for the protein coding genes of the eukaryotes (a), of the chloroplasts (b), and of the mitochondria (c) and for the transfer RNA genes (d). A horizontal dashed line goes through the point $(8, q_8)$.

b. Statistical Significance.

(1) If no rule except the random one had led to the construction of these curves (the 2 modulo 3 curve in the case of the groups of protein coding genes), then the probability that q_8 is one of the two highest values in all

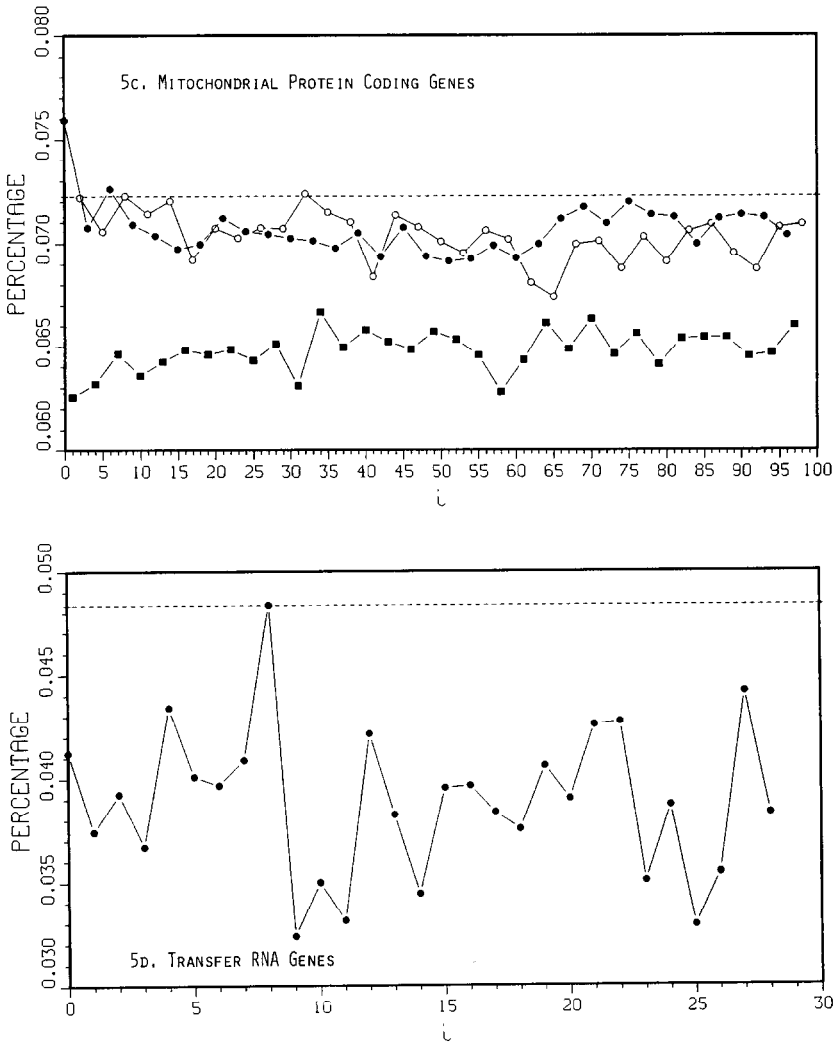

FIG. 5. Continued

these curves (among 33 points for the EC, CC, and MC groups; among 29 points for the TR group) would be

$$\tfrac{2}{33} \times \tfrac{2}{33} \times \tfrac{2}{33} \times \tfrac{2}{29} = 7.6 \times 10^{-6}.$$

(2) In order to prove that q_8 has a maximum value with statistical significance, we again apply the statistical test which compares two means of

two parameters (see appendix). There is statistical significance at the 0.1% level:

(i) for the two differences $q_8 - q_5$ and $q_8 - q_{11}$ with the EC and CC groups,
(ii) for the two differences $q_8 - q_7$ and $q_8 - q_9$ with the TR group.

With the MC group, only the difference $q_8 - q_5$ has statistical significance at the 5% level.

c. Results 2. The occurrence probability of the trinucleotide $T = YRY$ after the occurrence of the nucleotide Y—in the protein coding genes of the eukaryotes, of the chloroplasts, and of the mitochondria, and in the transfer *RNA* genes—presents a maximum value eight bases after Y.

d. Complementary Results. The results remain identical in the study of the parameters Q_i' deduced from the parameters Q_i by permutation in their definition (see Section 3.B.a) of R and Y for the eukaryotic protein coding genes, i.e., q_8' is the first among 33 values (i being congruent to 2 modulo 3, in the range [0, 99]).

The rank of the means q_i' of the parameters Q_i' with i congruent to 2 modulo 3 shows that q_8' is the seventh among 33 values for the chloroplast protein coding genes, the tenth among 33 values for the mitochondria protein coding genes, and the fifth among 29 values for the transfer RNA genes.

e. Additional Results. For the viral protein coding genes of lengths greater than 250 bases (1182 sequences, 1306 kb):

q_8 is the eighth among 33 values,
q_8' is the sixth among 33 values.

For the viral introns of lengths greater than 250 bases (50 sequences, 99 kb):

q_8 is the sixth among 100 values,
q_8' is the twentieth among 100 values.

We may notice that q_8 and q_8' are always among the highest 30% of values for the six previous samples. This statistical result is significant, since the random probability of such a situation is equal to:

$$(0.30)^6 = 7.3 \times 10^{-4}.$$

These results are not found again for the prokaryotic protein coding genes, for the eukaryotic introns, or for the ribosomal RNA genes.

4. DISCUSSION

The effect of the coding periodicity, i.e. the preferential use of the codon *RNY* in the open reading frame, biases any statistical analysis which is not congruent modulo 3. That is the reason why the studies of the dinucleotides and of the trinucleotides after a given nucleotide have been made in relation to the coding periodicity. Therefore, the parameters have been grouped in subsets of parameters with residue gaps that are invariant modulo 3 in order to permit their comparison. This approach allows us to state that the results demonstrated in this paper are independent of any coding periodicity.

These results have been found by reasoning based on the conditional probabilities. This approach is different from the common statistical one, which concerns the highest occurrence probabilities. For example, Fickett [4] showed that the probabilities between the thymine pairs separated by $2+3n$ bases are greater than those between pairs separated by $0+3n$ or $1+3n$ bases, n varying between 0 and 66. In the same way, Trifonov and Sussman [10] explain the mean periodicity 10.5 of the DNA chromatin pitch as due to the modulation of the 3-base pattern (maxima at 9, 21, 30,... bases), which occurs with the highest frequencies. It is important to specify that our results do not assert that the trinucleotide *YRY* has the highest occurrence probability eight bases after *Y*. Indeed, the trinucleotide *RRR* is the one which appears the most frequently (see [5]). What we have demonstrated, with a high degree of significance, are the main following results:

(1) The occurrence probability of the dinucleotide $D = RY$ ($D' = YR$) after the occurrence of the nucleotide Y (R) in the zero modulo three curve —in the eukaryotic protein coding genes—presents a modulo 9 periodicity with a maximum value nine bases after Y (R).

(2) The occurrence probability of the trinucleotide $T = YRY$ after the occurrence of the nucleotide Y—in the protein coding genes of the eukaryotes, of the chloroplasts, and of the mitochondria, and in the transfer *RNA* genes—presents a maximum value eight bases after *Y*.

These results have to be added to the *RNY* rule [3] and clearly differ from Trifonov and Sussman's results [10].

Some connections are even possible between the above results and some experimental conclusions. To begin with, the length of the motifs $Y(N)_8YRY$, $R(N)_8RYR$, $Y(N)_9RY$ and $R(N)_9YR$, i.e. twelve nucleotides, can be related to the DNA double helix pitch. Indeed, the pitch varies from 9.33 to 12 base pairs per turn in the *A*, *B*, *C*, and *Z* DNA forms [1, 7, 12], while its experimental value is estimated to be 10.4 base pairs per turn under physiological conditions [11]. Secondly, the complementary character between these motifs follows the base-pairing rule. Finally, the trinucleotides $T = YRY$ and $T' = RYR$ are those, among all the trinucleotides, which have the maximum values both for the torsion angle and for the propeller twist

[2, 6]. These three properties suggest that these motifs could have a function related to the spatial structure of the DNA sequences. Furthermore, the significant presence of the motif $Y(N)_8YRY$ in several present-day genes suggests that it could be considered as a primitive oligonucleotide or as a nucleotide ring (anterior to any molecular evolution) used in a stacking and linking process and leading to a natural "code" for the helix pitch.

The motifs $Y(N)_8YRY$, $R(N)_8RYR$, $Y(N)_9RY$, and $R(N)_9YR$ may be used as discriminating parameters [8] in order to characterize different taxonomic groups. As noted by one of the referees, these results may serve as an addition to the existing rules for distinguishing eukaryotic coding sequences from noncoding sequences.

APPENDIX

In order to compare two means $\bar{A}$ and $\bar{B}$ of two parameters A_i and B_i computed on the same sample of size n, namely $\bar{A} = (\Sigma_{1 \leqslant i \leqslant n} A_i)/n$ and $\bar{B} = (\Sigma_{1 \leqslant i \leqslant n} B_i)/n$, we consider, for each element i of the sample, the random variable whose value X_i is equal to $A_i - B_i$. For n large ($n \geqslant 30$), the reduced deviation $\varepsilon = \sqrt{n}(\bar{X}/s)$, where $\bar{X}$ and s represent respectively the mean and the standard deviation of the X_i, is known to follow a normal law $\mathcal{N}(0,1)$ if we make the hypothesis that $\bar{A} = \bar{B}$. Therefore, $\bar{A}$ is different from $\bar{B}$ at the 5% statistical level (for example) if ε is greater than 1.96.

We would like to thank Professors Thomas Bickle and Jacques Streith, Dr. Christoph Nager, and the referees for their advice. This work was supported by a Fellowship from the Centre National de la Recherche Scientifique to C.J.M.

REFERENCES

1 S. Arnott, R. Chandrasekaran, D. L. Birdsall, A. G. W. Leslie, and R. L. Ratliff, Left-handed DNA helices, *Nature* 283:743–745 (1980).

2 R. E. Dickerson, Base sequence and helix structure variation in *B* and *A* DNA, *J. Mol. Biol.* 166:419–441 (1983).

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

4 J. W. Fickett, Recognition of protein coding regions in DNA sequences, *Nucleic Acids Res.* 10:5303–5318 (1982).

5 L. Lebart, A. Morineau, and J. P. Fénelon, *Traitement des Données Statistiques*, Dunod, Paris, 1979.

6 G. G. Lennon and R. Nussinov, Eukaryotic oligomer frequencies are correlated with certain DNA helical parameters, *J. Theoret. Biol.* 116:427–433 (1985).

7 A. G. W. Leslie, S. Arnott, R. Chandrasekaran, and R. L. Ratliff, Polymorphism of DNA double helices, *J. Mol. Biol.* 143:49–72 (1980).

8 C. J. Michel, New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation, *J. Theoret. Biol.* 120:223–236 (1986).

9 J. C. W. Shepherd, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification, *Proc. Nat. Acad. Sci. U.S.A.* 78:1596–1600 (1981).

10 E. N. Trifonov and J. L. Sussman, The pitch of chromatin DNA is reflected in its nucleotide sequence, *Proc. Nat. Acad. Sci. U.S.A.* 77:3816–3820 (1980).

11 J. C. Wang, Helical repeat of DNA in solution, *Proc. Nat. Acad. Sci. U.S.A.* 76:200–203 (1979).

12 A. H.-J. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel, and A. Rich, Molecular structure of a left-handed double helical DNA fragment at atomic resolution, *Nature* 282:680–686 (1979).