J. theor. Biol. (1987) **128**, 457-461

A Purine–Pyrimidine Motif Verifying an Identical Presence in Almost All Gene Taxonomic Groups

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

† *I.U.T. de Belfort, Université de Franche-Comté, rue Engel-Gros, F-90016 Belfort-cedex, France and ‡ Friedrich Miescher Institut, Mattenstrasse 22, P.O. Box 2543, CH-4002 Basel, Switzerland*

(*Received 23 March 1987*)

A statistical parameter identifies, with a high degree of significance, a motif which is present in protein-coding sequences of eukaryotes, prokaryotes, chloroplasts, mitochondria, viral introns, ribosomal RNA genes, and transfer RNA genes. The random probability of occurrence of such a situation is 10^{-12}. This motif has the following properties: (i) its significant presence in almost all present-day genes explains why it can be considered as a primitive oligonucleotide, (ii) its nucleotide order is: $YRY(N)_6YRY$, R being a purine base, Y a pyrimidine one and N any base, (iii) its length and its terminal trinucleotides YRY suggest a primordial function related to the spatial structure of the DNA sequences. This motif is found in some viral protein-coding genes, but not in eukaryotic introns.

1. Introduction

A fundamental question with regard to the origin of life (Küppers, 1983; Jantsch, 1980; Bendall, 1983) concerns the production of primitive oligonucleotides during the biochemical evolution phase (Orgel, 1968; Miller & Orgel, 1974), which led, at a later stage, to the formation of primordial genes (Eigen, 1971). At present, laboratory experiments do not suggest the existence of classes of oligonucleotides which would be likely to be more primitive (Orgel, 1986), whereas the various computer methods, e.g. for constructing phylogenetic trees (Zuckerkandl & Pauling, 1965; Fitch & Margoliash, 1967), are limited to the analyses of several different reconstructed primordial genes (Schwartz & Dayhoff, 1978).

In this paper, we study the mean occurrence probability of the i-motif $YRY(N)_iYRY$, i varying between 1 and 99, over all DNA sequences of a gene taxonomic group F among the following, which have been obtained from the EMBL Nucleotide Sequence Data Library (release 10):

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

—Prokaryotic protein-coding genes, noted PC and constituted by 788 sequences, (768 kb).

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

§ To whom correspondence should be addressed.

0022-5193/87/200457+05 $03.00/0

—Mitochondrial protein-coding genes, noted MC and constituted by 130 sequences, (117 kb).
—Viral introns, noted VI and constituted by 50 sequences, (99 kb).
—Ribosomal RNA genes, noted RR and constituted by 92 sequences, (167 kb), which are eukaryotic, prokaryotic, chloroplast as well as mitochondrial.
—Transfer RNA genes, noted TR and constituted by 920 sequences, (70 kb), which are eukaryotic, prokaryotic, chloroplast as well as mitochondrial.
—Viral protein-coding genes, noted VC and constituted by 1182 sequences, (1306 kb).

The choice of these groups permits a universal DNA sequence study, i.e. a study independent of species and of gene function. These groups are composed of all sequences whose lengths are greater than 250 bases, except the sequences of the group TR whose lengths are greater than 65 bases.

2. Method

We note N a purine nucleotide R, or a pyrimidine nucleotide Y, and $l(s)$ the length of a given sequence s. Let m_i be the i-motif $YRY(N)_iYRY$, i.e. two trinucleotides YRY separated by any i bases N, i varying between 1 and 99. For each sequence s of the group F, the counter $c_i(s)$ counts the occurrences of m_i in s. For the i-motif m_i, the number of situations which can be successively tested in s, giving 0 or 1 cumulated in $c_i(s)$, is: $l(s)-(i+6)+1$ (where $i+6$ is the length of m_i). In order to analyse all the parameters $p_i(F)$, as defined below, in the same conditions, we have decided to limit for all i, the computation of $c_i(s)$ to the worst case $i=99$, i.e. for all i, we have only examined the first $l(s)-(99+6)+1$ motifs of the sequence s.

We then consider the occurrence probability $o_i(s)$ for the sequence s, as being the ratio of the counter $c_i(s)$ by the total number of current bases read, i.e. $l(s)-104$. The occurrence probability $p_i(F)$ for the group F is then defined as the mean value of $o_i(s)$ over all the sequences in the considered group F.

The next step, is to represent for each group F the statistical function $i \to p_i(F)$, by varying i. For the groups F of the protein-coding genes, the family of points (i, p_i) is separated into two families of points according to whether i is congruent or not to 0 modulo 3. This result is well known. Indeed, these two families of points are associated with two curves of different mean levels, because there is a preferential use of the codon RNY (Eigen, 1971) in the open reading frame (Shepherd, 1981) which leads to separate into modulo three families (Shepherd, 1981; Trifonov & Sussman, 1980; Fickett, 1982) any quantitative parameters whose definition is of the previous described form (proof not shown).

A minimal length of 250 bases for the sequences has been chosen in order to have a sufficient number of the i-motifs m_i to give meaning to their occurrence probabilities, because the occurrence probability of a i-motif m_i is equal to 1/64 if, in a first approximation, the numbers of R and Y are equal to 0·5 and if the bases are randomly distributed. In the particular case of the TR group, the minimal length for the sequences has been fixed at 65 bases and the maximal value of the index i

has been reduced to 29 (the total number of current bases read for defining $c_i(s)$, is then equal to $l(s)-34$), because the length of almost all the transfer RNA genes is shorter than 100 bases.

3. Results

The first seven graphs (see Fig. 1) show that the occurrence probability of the i-motif m_i has a maximum value for i equal to six. For the viral protein-coding genes (Fig. 1(h)) the p_6(VC) value is the third in the range [1, 99], after the p_{12}(VC) and p_{18}(VC) values. In order to simplify the statistical reasoning presented in section 4, we have omitted this VC group in the probability evaluation. Therefore, we can state the following new rule: The occurrence probability of a trinucleotide YRY *after* the occurrence of the same trinucleotide YRY is not uniform but presents in the range [1, 99], a *maximum after six bases in all the gene taxonomic groups, except in the eukaryotic intron one.*

For the groups of the protein-coding genes, the maximum over 99 points has to be considered as a maximum over 33 points because the occurrence probabilities of the i-motif m_i with i congruent to 0 modulo 3 are greater than the ones with i congruent to 1 or 2 modulo 3 (see above remark). Therefore, in the course of this paper, we only consider, for the groups of the protein-coding genes, the highest curve (0 modulo 3) in which no known rule can discriminate its points between themselves.

4. Why This Result Is Totally Unexpected

If no rule, except the random one, had presided in the construction of these curves (the highest curve in the case of the group of the protein-coding genes), then the probability that the maximum value of the curves (of 33 points for the EC, PC, CC and MC groups of protein-coding genes, of 29 points for the TR group, of 99 points for the VI and RR groups) would be always in position i equal to 6, is:

$$1/(33 \cdot 33 \cdot 33 \cdot 33 \cdot 29 \cdot 99 \cdot 99) = 3 \cdot 10^{-12}.$$

In addition to this statistical proof, we have to make the following remarks:

(1) The data represent almost all the complete information of the EMBL database. Indeed, the statistical study has been applied to 5554 sequences which are split up into eight large taxonomic groups.

(2) The constraint, e.g. a minimal length of 250 bases, has been introduced in order to have a statistical proof. But the same results can be seen by varying the size of the samples, for examples with a choice of a minimal length of 150 bases and/or with a selection of a given maximal length.

(3) No such rule can be identified with this statistical approach based on the studies of the i-motifs which are built with multinucleotides different from *YRY*. The data pertaining to the second and to the third argument are not shown.

5. Discussion

First of all, these statistical results show a pattern almost universal by the identification, with a high degree of signification of the 6-motif $YRY(N)_6YRY$ which

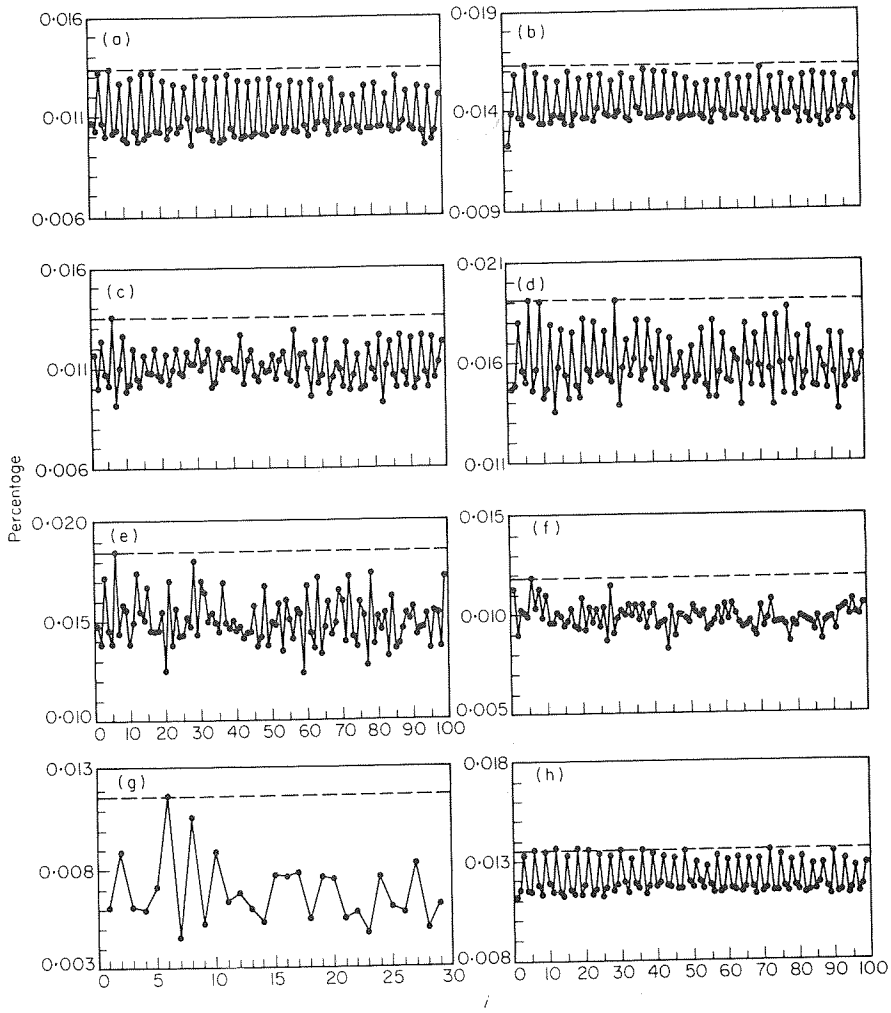

FIG. 1. The statistical results obtained with eight large gene taxonomic groups (see text) are shown in eight graphs. The horizontal axis represents the index i of the i-motif $YRY(N)_iYRY$, i varying between 1 and 99 (29 for the transfer RNA genes). The vertical axis represents the mean occurrence probability $p_i(F)$ of $YRY(N)_iYRY$ over all the sequences of a given gene taxonomic group (see text). A horizontal dashed line goes through the point $(6, p_6(F))$. (a) Eukaryotic, protein-coding genes. (b) Prokaryotic, protein-coding genes. (c) Chloroplast, protein-coding genes. (d) Mitochondrial, protein-coding genes. (e) Viral introns. (f) Ribosomal RNA genes. (g) Transfer RNA genes. (h) Viral, protein-coding genes.

can be considered as a primitive oligonucleotide because it is anterior to any molecular evolution. Indeed, if a primitive biochemical process has led to the selection of a *few* classes of primitive oligonucleotides (Orgel, 1968, page 387), then it can be expected that they are still present, more or less well conserved, in *all* the present-day genes.

We analyse some consequences of these biological results. From a practical point of view, this primitive oligonucleotide can be used as a *reference* for constructing the phylogenetic trees (Fitch & Margoliash, 1967) which may lead to some new insights into the classification of the primordial genes (Schwartz & Dayhoff, 1978) or into the identification of the Universal Ancestor (Bendall, 1983, p. 220, for the concept of progenote). Furthermore, the theoretical aspect, in terms of molecular evolution, is very surprising, because this primitive oligonucleotide has interesting properties. On the one hand, its trinucleotide *YRY*, is the one, among all trinucleotides, which has the maximum values both for the torsion angle and for the propeller twist (Dickerson, 1983). On the other hand, their length of twelve nucleotides can be related to the DNA double helix pitch. Indeed, the pitch varies from 9·33 to 12 base pairs per turn with the A, B, C and Z DNA forms (Leslie *et al.*, 1980; Arnott *et al.*, 1980; Wang *et al.*, 1979) while its experimental value is estimated to be 10·4 base pairs per turn under physiological conditions (Wang, 1979). These two properties suggest that this primitive oligonucleotide could have *a primordial function* related to the spatial structure of the DNA sequences as well *a code of the helix pitch* as *a process to stack and to link nucleotide rings.* We are currently trying to extend this model to the statistical study of the combinations which can link such oligonucleotides. These results, which are both surprising and significant, may lead to new experimental and theoretical works with regard to the origin of the life.

We would like to thank Professors Thomas Bickle, Max Burger, Jacques Streith and Dr Christoph Nager for their advice. Supported by grants from the Friedrich Miescher Institute to C.J.M.

REFERENCES

ARNOTT, S., CHANDRASEKARAN, R., BIRDSALL, D. L., LESLIE, A. G. W. & RATLIFF, R. L. (1980). *Nature* **283,** 743.
BENDALL, D. S. (1983). In: *Evolution from Molecules to Men.* Cambridge: Cambridge University Press.
DICKERSON, R. E. (1983). *J. mol. Biol.* **166,** 419.
EIGEN, M. (1971). *Naturwissenschaften* **58,** 465.
FICKETT, J. W. (1982). *Nucleic Acids Res.* **10,** 5303.
FITCH, W. M. & MARGOLIASH, E. (1967). *Science* **155,** 279.
JANTSCH, E. (1980). In: *The Self-Organizing Universe.* Oxford: Pergamon Press.
KÜPPERS, B-O. (1983). In: *Molecular Theory of Evolution.* Heidelberg: Springer-Verlag.
LESLIE, A. G. W., ARNOTT, S., CHANDRASEKARAN, R. & RATLIFF, R. L. (1980). *J. mol. Biol.* **143,** 49.
MILLER, S. L. & ORGEL, L. E. (1974). In: *The Origins of Life on the Earth.* Englewood Cliffs, NJ: Prentice-Hall, Inc.
ORGEL, L. E. (1968). *J. mol. Biol.* **38,** 381.
ORGEL, L. E. (1986). *J. theor. Biol.* **123,** 127.
SCHWARTZ, R. M. & DAYHOFF, M. O. (1978). *Science* **199,** 395.
SHEPHERD, J. C. W. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 1596.
TRIFONOV, E. N. & SUSSMAN, J. L. (1980). *Proc. natn. Acad. Sci. U.S.A.* **77,** 3816.
WANG, J. C. (1979). *Proc. natn. Acad. Sci. U.S.A.* **76,** 200.
WANG, A. H.-J., QUIGLEY, G. J., KOLPAK, F. J., GRAWFORD, J. L., VAN BOOM, J. H., VAN DER MAREL, G. & RICH, A. (1979). *Nature* **282,** 680.
ZUCKERKANDL, E. & PAULING, L. (1965). *J. theor. Biol.* **8,** 357.