Bulletin of Mathematical Biology Vol. 52, No. 6, pp. 741–772, 1990.
Printed in Great Britain.

0092-8240/90$3.00+0.00
Pergamon Press plc

A MODEL OF DNA SEQUENCE EVOLUTION

■ DIDIER G. ARQUÈS*
Université de Franche-Comté,
Laboratoire d'Informatique de Besançon,
Unité Associée CNRS No 822,
16, route de Gray,
25030 Besançon, France

■ CHRISTIAN J. MICHEL
Friedrich Miescher Institut,
Bioinformatic group,
Mattenstrasse 22,
P.O. Box 2543,
CH-4002 Basel, Switzerland

Statistical studies of gene populations on the purine/pyrimidine alphabet have shown that the mean occurrence probability of the i-motif $YRY(N)_iYRY$ (R = purine, Y = pyrimidine, N = R or Y) is not uniform by varying i in the range [1, 99], but presents a maximum at $i=6$ in the following populations: protein coding genes of eukaryotes, prokaryotes, chloroplasts and mitrochondria, and also viral introns, ribosomal RNA genes and transfer RNA genes (Arquès and Michel, 1987b, *J. theor. Biol.* **128**, 457–461). From the "universality" of this observation, we suggested that the oligonucleotide $YRY(N)_6$ is a primitive one and that it has a central function in DNA sequence evolution (Arquès and Michel, 1987b, *J. theor. Biol.* **128**, 457–461). Following this idea, we introduce a concept of a model of DNA sequence evolution which will be validated according to a shema presented in three parts.

In the first part, using the last version of the gene database, the $YRY(N)_6YRY$ preferential occurrence (maximum at $i=6$) is confirmed for the populations mentioned above and is extended to some newly analysed populations: chloroplast introns, chloroplast 5′ regions, mitochondrial 5′ regions and small nuclear RNA genes. On the other hand, the $YRY(N)_6YRY$ preferential occurrence and periodicities are used in order to classify 18 gene populations.

In the second part, we will demonstrate that several statistical features characterizing different gene populations (in particular the $YRY(N)_6YRY$ preferential occurrence and the periodicities) can be retrieved from a simple Markov model based on the mixing of the two oligonucleotides $YRY(N)_6$ and $YRY(N)_3$ and based on the percentages of RYR and YRY in the unspecified trinucleotides $(N)_3$ of $YRY(N)_6$ and $YRY(N)_3$. Several properties are identified and prove in particular that the oligonucleotide mixing is an independent process and that several different features are functions of a unique parameter.

In the third part, the return of the model to the reality shows a strong correlation between reality and simulation concerning the presence of large alternating purine/pyrimidine stretches and of periodicities. It also contributes to a greater understanding of biological reality, e.g. the presence or the absence of large alternating purine/pyrimidine stretches can be explained as being a simple consequence of the mixing of two particular oligonucleotides.

Finally, we believe that such an approach is the first step toward a unified model of DNA

* To whom correspondence should be addressed.

sequence evolution allowing the molecular understanding of both the origin of life and the actual biological reality.

1. Introduction: Concept of a Model of DNA Sequence Evolution. Our hypothesis is that DNA sequence evolution (Fig. 1) on the *two-letter alphabet* $\{R, Y\}$ (R = purine, Y = pyrimidine) is constituted of *two successive steps* (in first approximation).

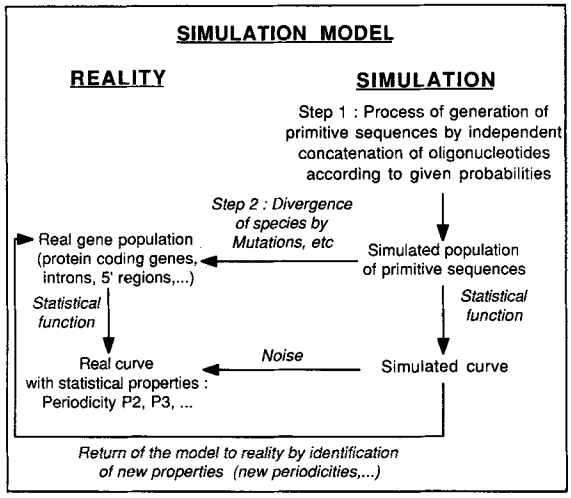


Figure 1. Concept of a model of DNA sequence evolution.

(1) The first step (step 1 in Fig. 1) leads to *primitive sequences* from a *concatenation* process of a few (i.e. preferential type) *primitive oligonucleotides* according to given probabilities. This concatenation process must have been *independent* (due to the absence of any clever process at the primitive stage).

(2) At a later stage, the second step (step 2 in Fig. 1) leads to the actual divergence of species by using *random changes* on the primitive sequences such as mutations (changes R→Y and Y→R), transformation of the alphabet {R, Y} into {A, C, G, T} (A = adenine, C = cytosine, G = guanine, T = thymine), insertion and deletion of bases, etc.

If our hypothesis is true, then the primitive sequences built from a few primitive oligonucleotides must have strong statistical features. One can hope that such features are still statistically significant and still present in *all* (because present before the divergence) actual real gene populations even if random changes (mutations, etc.) have introduced an important "noise" effect. Section 2 will validate this consequence of the hypothesis by identifying

statistical features found in several gene populations (in particular, the $YRY(N)_6YRY$ preferential occurrence is almost universal; N = R or Y), features which can be easily explained by the existence of only a few oligonucleotides.

This DNA sequence evolution process (Fig. 1) can be simulated as follows:

(1) Identification of a few oligonucleotides on the alphabet {R, Y}: $YRY(N)_6$ and $YRY(N)_3$ will be deduced from the results obtained in Section 2.

(2) Concatenation of these oligonucleotides according to a stochastic process (in function of the occurrence probabilities associated with each oligonucleotide) leading to the simulated populations: development of a Markov model in Section 3 which will be proved to be a random one, i.e. the concatenation process is independent.

(3) Determination, for each real population, of an associated simulated population by making use of an appropriate statistical function whose simultaneous application in the real population and in its associated simulated population, gives the same statistical features: results of Section 3.

(4) The simulated curve does not contain the "noise" effect resulting from the random rules (mutations, etc.) and therefore it has stronger statistical features compared to the real curve. Then, the observation of the obvious statistical features in the simulated curve allows the identification of new (but hidden by the "noise" effect) properties in the real populations: return of the model to the reality presented in Section 4.

2. Statistical Features and Classification of Gene Populations.† The transformation of genetic information into statistical information leads to a loss of information which can be minimized by choosing the appropriate statistical function analysing, on the two-letter alphabet {R, Y} (R = purine, Y = pyrimidine), the occurrence probability of the *i*-motif $YRY(N)_iYRY$ (N = R or Y). In gene populations this function can reveal in particular the $YRY(N)_6YRY$ preferential occurrence (Arquès and Michel, 1987b; defined below) and the periodicities P3 and P2 (Shepherd, 1981; Fickett, 1982; Arquès and Michel, 1987a–c, defined below). The gene populations studied are protein coding genes, introns, 5′ regions, ribosomal RNA genes, transfer RNA genes and small nuclear RNA genes (Table 1).

Statistical studies of protein coding genes of eukaryotes, prokaryotes, viruses, chloroplasts, mitochondria and plasmids, have shown a periodicity P3 (called coding periodicity; defined below) (Shepherd, 1981; Fickett, 1982;

† The statistical features of gene populations presented in this section are confirmed by release 21 of the EMBL data base containing about double the amount of sequences described in release 17.

Table 1. Gene populations

1. *Protein coding populations*
   Eukaryotes, noted CEUK (4531 sequences, 4414 kb)
   Prokaryotes, noted CPRO (1753 sequences, 1804 kb)
   Viral, noted CVIR (1895 sequences, 2293 kb)
   Chloroplasts, noted CCHL (182 sequences, 180 kb)
   Mitochondria, noted CMIT (172 sequences, 160 kb)
   Plasmids, noted CPLA (243 sequences, 214 kb)
2. *Intron populations*
   Eukaryotes, noted IEUK (701 sequences, 680 kb)
   Viral, noted IVIR (51 sequences, 102 kb)
   Chloroplasts, noted ICHL (38 sequences, 27 kb)
   Mitochondria, noted IMIT (29 sequences, 37 kb)
3. *5′ Region populations*
   Eukaryotes, noted NEUK (1501 sequences, 1051 kb)
   Prokaryotes, noted NPRO (556 sequences, 285 kb)
   Viral, noted NVIR (266 sequences, 183 kb)
   Chloroplasts, noted NCHL (46 sequences, 20 kb)
   Mitochondria, noted NMIT (39 sequences, 19 kb)
4. *Ribosomal RNA genes*, noted RR (145 sequences, 270 kb)
   (eukaryotes, prokaryotes, chloroplasts, mitochondria)
5. *Transfer RNA genes*, noted TR (1157 sequences, 87 kb)
   (eukaryotes, prokaryotes, viral, chloroplasts, mitochondria)
6. *Small nuclear RNA genes*, noted SNR (97 sequences, 15 kb)

Arquès and Michel, 1987a–c). Then, a statistical study of a perturbation in this periodicity P3 (Arquès and Michel, 1987a) has led to the following result:

The mean occurrence probability of the i-motif $YRY(N)_iYRY$ is not uniform with i in the range [1, 99], but presents a maximum at $i=6$ in the following gene populations: protein coding genes of eukaryotes, prokaryotes, chloroplasts and of mitochondria, viral introns, ribosomal RNA genes and transfer RNA genes (Arquès and Michel, 1987b). The main exception found which is the eukaryotic introns, has been solved by showing that the $YRY(N)_6YRY$ preferential occurrence (maximum at $i=6$) is hidden by a periodicity P2 (called alternating purine/pyrimidine periodicity; defined below) (Arquès and Michel, 1987c). We have already suggested from the "universality" of this observation that the oligonucleotide $YRY(N)_6$ is a primitive one and that it has a central function in DNA sequence evolution (Arquès and Michel, 1987b).

By using the last version of the gene database, the $YRY(N)_6YRY$ preferential occurrence is confirmed for the populations mentioned above and is observed in four newly analysed populations: chloroplast introns, chloroplast 5′ regions, mitochondrial 5′ regions and small nuclear RNA genes. On the other hand, the $YRY(N)_6YRY$ preferential occurrence and the periodicities P3 and P2 are used in order to classify 18 gene populations.

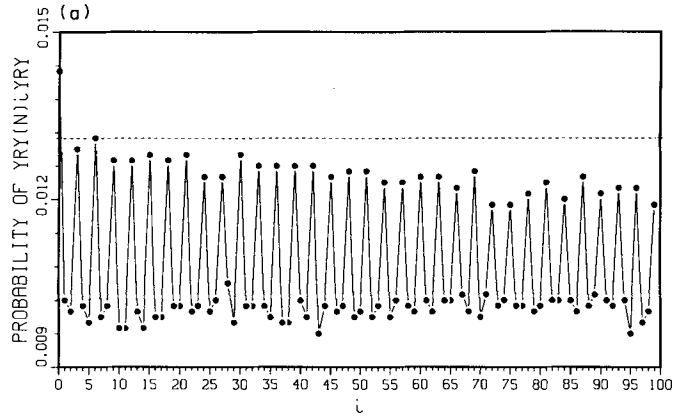
2.1. *Method.* The method was developed previously by Arquès and Michel (1987b). The outlines are briefly stated below.

2.1.1. Data: gene populations. The gene populations obtained from the EMBL Nucleotide Sequence Data Library (release 17) are characterized by their notation, by their number of sequences and by their number of kilobases (kb) (Table 1). The protein coding genes, the introns and the 5′ regions are analysed according to their taxonomic group. The small number of ribosomal RNA genes and the short length of transfer RNA genes actually do not allow an objective statistical analysis per taxonomic group for these two populations. Therefore, the two populations of ribosomal and transfer RNA genes belong to eukaryotes, prokaryotes, chloroplasts and mitochondria (a few transfer RNA genes being also viral). The small nuclear RNA genes are eukaryotic. The 5′ regions are located upstream the open reading frames starting with an initiator ATG codon. The other types of 5′ regions have been excluded from this survey, while the data concerning the 3′ regions are not yet available. A gene population incorporates all the nonduplicated sequences which can be classified, i.e. a sequence with unspecified bases or one with an unmentioned taxonomic group, is excluded, etc.

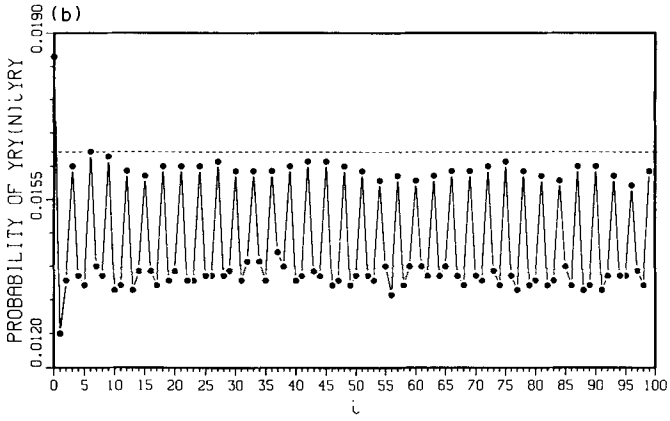
2.1.2. Statistical function. Let F be a gene population with $n(F)$ sequences. Let s be a sequence in F with a length $l(s)$.
Let the i-motif $m_i = \mathrm{YRY(N)}_i\mathrm{YRY}$ (R = purine, Y = pyrimidine, N = R or Y) by varying i in the range [0, 99], be 2 trinucleotides YRY separated by any i bases N (note: compared to the previous methods, this study treats the particular case of the i-motif $\mathrm{YRY(N)}_i\mathrm{YRY}$ at $i=0$). For each s of F, the counter $c_i(s)$ counts the occurrences of m_i in s. In order to count the m_i 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 $m_i)$. Then, the occurrence probability $o_i(s)$ of m_i for s, is equal to $c_i(s)/(l(s)-104)$, i.e. the ratio of the counter by the total number of current bases read. Then, the occurrence probability $p_i(F)$ of m_i for F, is equal to $(\Sigma_{s\in F} o_i(s))/n(F)$. For each population F, the statistical function $i \to p_i(F)$ by varying i, is represented as a curve $C(F)$. In order to have a sufficient number of m_{99} occurrences, the function is applied to sequences having a minimal length of 250 bases. For the transfer and small nuclear RNA genes, the minimal length for the sequences analysed is fixed at 60 bases and the maximal value of i is reduced to 29 because the length of these genes is small (in the range [60, 250]) (Arquès and Michel, 1978b for more details).

2.2. *Results*

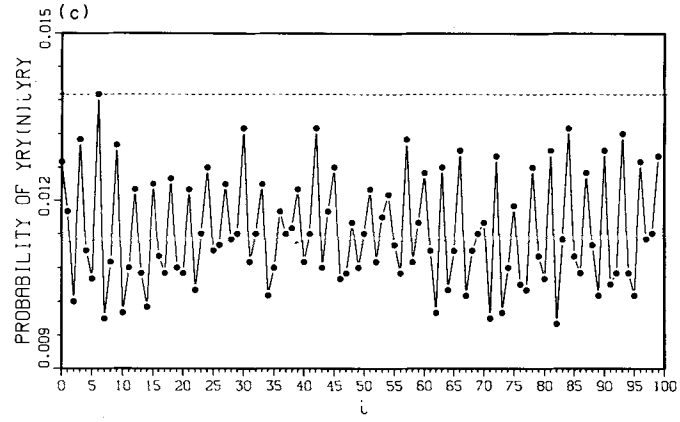
2.2.1. Statistical features of gene populations. The *main* statistically



CEUK

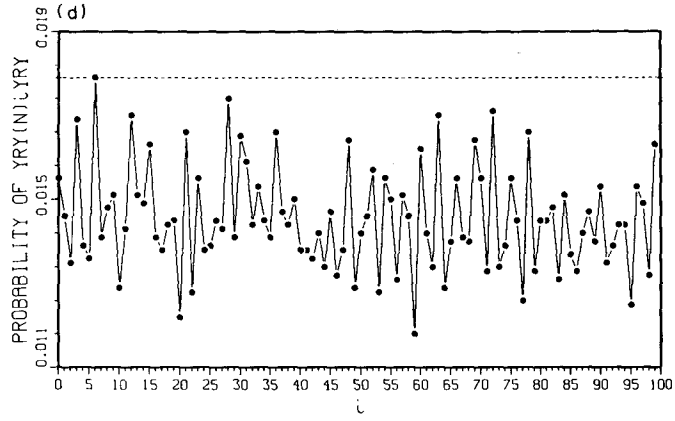


CPRO

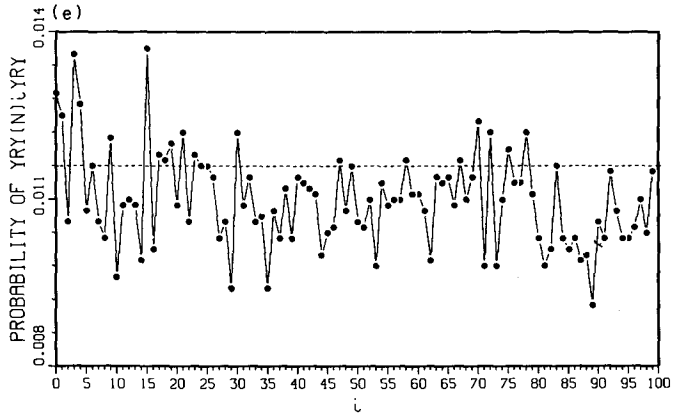


CCHL

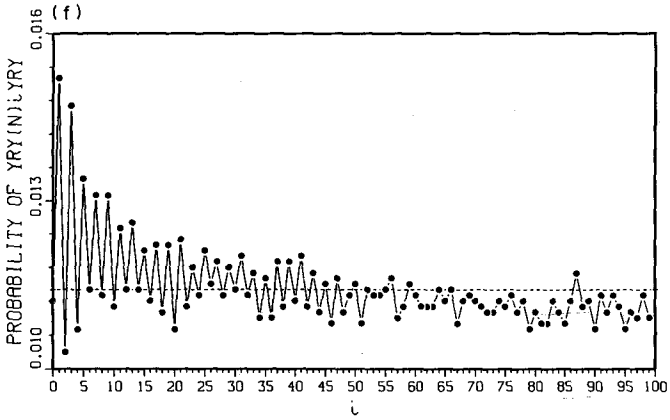
Figure 2



IVIR

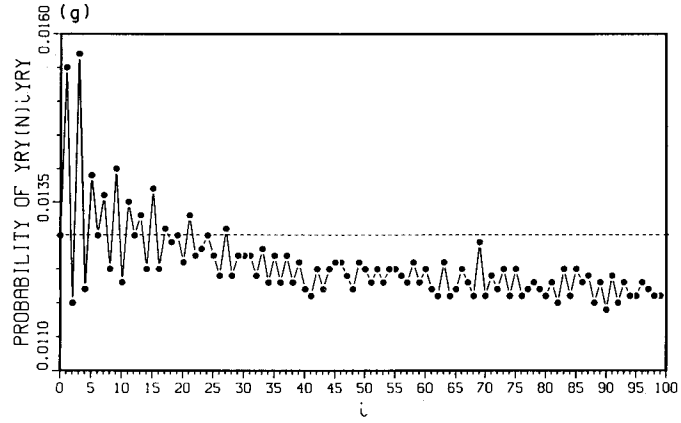


IMIT

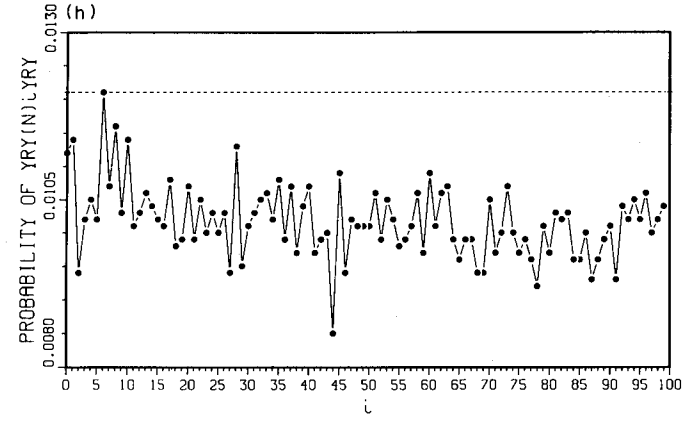


IEUK

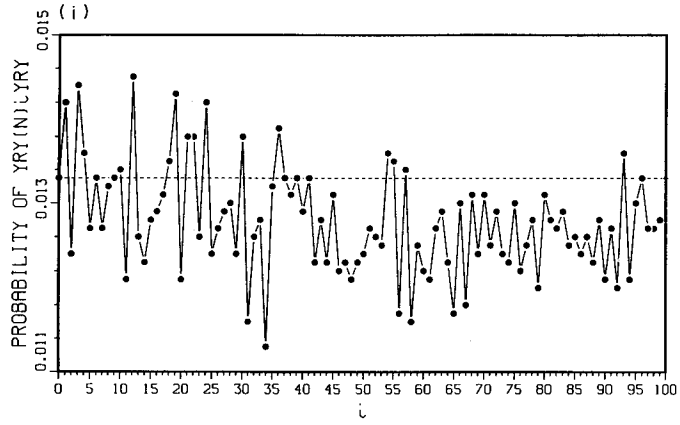
Figure 2—*continued*.



NEUK



RR



NVIR

Figure 2—*continued*.

significant features found in the gene populations and revealed by the curve $C(F)$ are the two periodicities $P3$ and $P2$, and the maximal value of $p_i(F)$ (mainly $p_o(F)$ and $p_6(F)$). The minimal value of $p_i(F)$ is sometimes also considered.

(a) Features concerning the periodicities $P3$ and $P2$.
Feature $P3$: periodicity $P3$ in the range [3, 98]:

$$p_i(F) \geqslant \text{Max } \{p_{i-1}(F), p_{i+1}(F)\} \text{ with } i \equiv 0[3] \text{ and } i \in [3, 98].$$

The periodicity $P3$ is *incomplete* if a few points do not satisfy the above inequality, i.e. a few values i in [3, 98], $i \equiv 0[3]$ and $p_i(F) < \text{Max } \{p_{i-1}(F), p_{i+1}(F)\}$.

Feature $P2$: periodicity $P2$ in the range [0, L]:

$$p_i(F) \geqslant \text{Max}\{p_{i-1}(F), p_{i+1}(F)\} \text{ with } i \equiv 1[2] \text{ and } i \in [0, L].$$

(b) Features concerning the maximal and minimal values of $p_i(F)$ in the range [0, 99].
Feature $M0, 6$:

$$p_0(F) > p_6(F) > p_i(F),\ i = 1, \ldots, 5, 7, \ldots, 99.$$

Feature Mj:

$p_j(F) > p_i(F)$, $i \neq j$ and $i, j \in [0, 99]$. (Mainly M0 and M6.)

Feature mj:

$$p_j(F) < p_i(F),\ i \neq j \text{ and i}, j \in [0, 99].$$

2.2.2. Classification of gene populations according to the statistical features. A figure will be given for each major feature.

(a) Gene populations with the periodicity $P3$ in the range [3, 98].
Gene populations with the features $P3$ and $M0, 6$: CEUK (Fig. 2a), CPRO (Fig. 2b) and CMIT (data not shown). The periodicity $P3$ is

---

Legend for pages 746–748

Figure 2. Mean occurrence probability of the i-motif YRY(N)$_i$YRY in gene populations. The horizontal axis represents the number i of bases N in the i-motif YRY(N)$_i$YRY, with i in the range [0, 99]. The vertical axis represents the mean occurrence probability $p_i(F)$ (see Section 2.1) over all the sequences in the following gene populations F: (a) CEUK, eukaryotic protein coding genes; (b) CPRO, prokaryotic protein coding genes; (c) CCHL, chloroplast protein coding genes; (d) IVIR, viral introns; (e) IMIT, mitochondrial introns; (f) IEUK, eukaryotic introns; (g) NEUK, eukaryotic 5′ regions; (h) RR, ribosomal RNA genes; (i) NVIR, viral 5′ regions. A horizontal dashed line goes through the point (6, $p_6(F)$)).

uniform for CEUK and CPRO: there are two different sets of well separated points so that:

$$\text{Min}\{p_i(F),\ i \equiv 0[3]\} > \text{Max}\{p_i(F),\ i \equiv 1, 2[3]\} \text{ with } i \in [0, 99].$$

Furthermore, for each set, the points can be joined by a nearly horizontal line, except for the top line of CEUK which decreases slightly by increasing i. Finally, for CPRO, p_1(CPRO) is the the lowest value (feature $m1$), whereas for CEUK, p_1(CEUK) is the highest value in the bottom curve $i \equiv 1, 2[3]$.

Gene populations with the features $P3$ and $M0$: CVIR, CPLA, NPRO (data not shown because the features $P3$ and $M0$ are more general than the features $P3$ and $M0$, 6). CVIR has an uniform periodicity $P3$. NPRO has an incomplete periodicity $P3$ and the feature $m1$ (as CPRO). (Note: $p_6(F)$ is in the five first values for these three populations.)

Gene populations with the features $P3$ and $M6$: CCHL (Fig. 2c) and IVIR (Fig. 2d). CCHL and IVIR have an incomplete periodicity $P3$.

Gene population with the features $P3$ and $M15$: IMIT (Fig. 2e). IMIT has an incomplete periodicity $P3$ for $i \leqslant 21$.

(b) Gene populations with the periodicity $P2$ in the range $[0, L]$.

Gene population with the features $P2$ in the range $[0, L=49]$ and $M1$: IEUK (Fig. 2f).

Gene population with the features $P2$ in the range $[0, L=23]$ and $M3$: NEUK (Fig. 2g).

For IEUK, p_1(IEUK) is the highest value and p_3(IEUK) is the second highest, whereas for NEUK, p_3(NEUK) is the highest value and p_1(NEUK) is the second highest. Furthermore, for NEUK, p_0(NEUK), p_6(NEUK), p_{12}(NEUK), p_{18}(NEUK) and p_{24}(NEUK) are nearly equal. Indeed, their associated points can be joined by a horizontal line. Three other obvious sets of points can be joined by regular curves: (1) $i = 3, 9, 15, 21, 27$ and 33; (2) $i = 1, 5, 7, 11, 13, 17, 19, 23$ and 25; (3) $i = 2$, 4, 8, 10, 14, 16, 20 and 22. All these naturally appearing curves join modulo 6 periodic sets of i values (Fig. 2g).

(c) Gene populations with no periodicity $P3$ or $P2$.

Gene populations with the feature $M6$: RR (Fig. 2h), ICHL, NCHL, NMIT, TR ($0 \leqslant i \leqslant 29$), SNR ($0 \leqslant i \leqslant 29$) (data not shown).

Gene population with no feature: NVIR (Fig. 2i).

2.3. *Discussion*

2.3.1. Biological meanings of the statistical features. Statistical studies at the DNA sequence level (Shepherd, 1981) and at the gene population level (Fickett, 1982; Arquès and Michel, 1987a–c) have shown the periodicity $P3$ in

protein coding genes of any taxonomic group: eukaryotes, prokaryotes, viral, chloroplasts, mitochondria and plasmids (Section 2.2.2a). This periodicity *P*3 is found, not only in protein coding genes, but also in introns of viruses and mitochondria (Arquès and Michel, 1987c and Section 2.2.2a). These two populations of introns have the genetic information necessary to code for proteins (Arquès and Michel, 1987c). Indeed, viruses use overlapping genes, both DNA strands and alternative patterns of RNA splicing in order to maximize the functions of a viral genome whose size is small (Ziff, 1980). On the other hand, many mitochondrial introns encode splicing proteins (maturases) (Lazowska *et al.*, 1980). The prokaryotic 5′ regions constitute a newly analysed population having the periodicity *P*3 (Section 2.2.2a). This periodicity *P*3 could be related to one of the reasons mentioned above concerning introns, in particular with a supply protein coding function because the size of the prokaryotic genome is small compared to the eukaryotic one. In summary, *the periodicity* P3 *is related to the protein coding function of a gene and is found in protein coding genes, viral introns, mitochondrial introns and prokaryotic* 5′ *regions.*

A different type of periodicity, i.e. the periodicity *P*2, was identified in eukaryotic introns (Arquès and Michel, 1987c and Section 2.2.2b). *The periodicity* P2 *is related to regulatory functions of a gene* (Arquès and Michel, 1987c) *and is found in eukaryotic introns and in the newly analysed population of the eukaryotic* 5′ *regions* (Section 2.2.2b), *i.e. only in the eukaryotic genome.*

This study shows three families of genes: genes with the periodicity *P*3 (9 populations), genes with the periodicity *P*2 (2 populations) and genes with no particular periodicity (7 populations). *These three families have a common feature: a higher frequency of* $YRY(N)_6YRY$ (Arquès and Michel, 1987b and Section 2.2.2) because one of the two highest frequency of $YRY(N)_iYRY$ is obtained at $i=6$ in 11 populations out of 18 (features *M*6 and *M*0, 6). Only mitochondrial introns and the viral 5′ regions are "real" exceptions. These exceptions may be due to the small size for the IMIT population (see the statistical reason presented in Section 2.3.3) and/or to the presence of some properties which hide the $YRY(N)_6YRY$ preferential occurrence. This latter case is found in eukaryotic introns and in the eukaryotic 5′ regions, where the higher frequency of $YRY(N)_6YRY$ is only true in the bottom curve $i \equiv 0[2]$ (Arquès and Michel, 1987c and Figs 2f and 2g).

2.3.2. Comparison with the current molecular theories of DNA sequence evolution.

(a) *The biological concept.* None of the actual theories of DNA sequence evolution analyses in totality the great genetic variety (in terms of taxonomic group: eukaryotes, prokaryotes, viruses, chloroplasts, mitochondria, plasmids and their subpopulations or in terms of gene

function: protein coding genes, introns, 5′ regions, 3′ regions) according to the existence of a unique process of gene formation. Several observations such as the approximate constancy of the amino acid substitution rate in each protein (Zuckerkandl and Pauling, 1965) and the large amount of genetic polymorphism in many populations (e.g. Lewontin and Hubby, 1966), led to the development of the "mutation" models (reviews in Kimura, 1987; Nei, 1987) without pattern for the primitive genes. The "RNY" model (Eigen and Schuster, 1978), giving a pattern on the purine/pyrimidine alphabet for the primitive protein coding genes, differs from the mutation models.

The molecular evolution must be understood at the entire DNA sequence level and not only at the level of protein coding genes. Indeed, DNA sequences encode more information than protein sequences, as a large proportion of DNA sequences do not code for proteins and as the genetic code is degenerated. Despite of this DNA variety, DNA sequences rely on physical constants: (i) a primary structure with the same four nucleotides: adenine, cytosine, guanine and thymine; (ii) a tertiary structure in double helix, however a few exceptions are found with certain types of viruses. Therefore, a model related to the spatial structure of DNA sequences is more general. Such a model will be developed in Section 3 allowing the simulation of most features which characterize the biological reality. At present, such a model seems to be the most general.

(b) *The great number of biological and statistical hypotheses.* No simple biological rule was identified because the actual theories attempt to find a general model of DNA sequence evolution from the analysis of particular cases. A statistical reason (see Section 2.3.3) explains the limits (even, the impossibility) of such approaches. Furthermore, if a model is developed for each different set of only a few DNA sequences, then some hypotheses are necessary to make the models coherent with each other.

2.3.3. The concept of the population evolution history (or the mean evolution history of a sequence from a population). In order to develop a general model of DNA sequence evolution whose one necessary condition is to be independent of particular cases and of hypotheses, the biological concept of the population evolution history (or the mean evolution history of a sequence) must be introduced. This concept can be statistically studied if a great number of DNA sequences is used to define a population (see Section 3.3). Indeed, if only a few (at the limit, one) sequences are studied, then the population evolution history is confused with the particular history of the chosen sequences. The features of a few sequences (induced by random changes, i.e.

mutations, insertions, deletions, etc.) hide the weaker and more general features characterizing the population evolution history. The statistical reason is obvious and it is deduced from the law of large numbers. An illustration is given with the following example.

Assume, for the sake of simplicity, that all sequences in a population have only two statistical features: a common feature and a specific one. For a given parameter q studied, Fig. 3a shows the statistical function $i \rightarrow q_i$ associated with one sequence of a population F: (i) a common feature related to the F population history, e.g. a small value $q_6 = 0.02$; (ii) a feature specific to the sequence history, e.g. a high value $q_i = 0.5$ for a particular i.

Figure 3b shows the statistical function $i \rightarrow q_i$ which is the mean of the statistical curves associated with 10 sequences: the 10 specific features (assume that all are different) occur with probability of 0.05 (0.5/10), but the common feature q_6 is now only 2.5 (0.05/0.02) times smaller.

Figure 3c shows the statistical function $i \rightarrow q_i(F)$ applied to a great number (10 000) of such sequences (law of large numbers): the common feature $q_6(F)$ is now 4 (0.02/0.005) times larger than the specific features having a probability of approximately 0.005 ($0.5 \times 100/10\,000$). On the other hand, the features specific to the sequence history are no longer identified at the population level because they are on a straight line (Fig. 3c), (i.e. white noise).

In addition, the same reason explains that: (1) the periodicity $P3$ existing at the population level can be absent at the level of a sequence of this population (see for example in Fig. 4, the eukaryotic protein coding gene with the EMBL identification CLCK, starting at position 39 and ending at position 1184); (2) the motifs $YRY(N)_0YRY$ and $YRY(N)_6YRY$ having a high frequency at the population level can occur with a low frequency at the level of a sequence of this population (see the same example in Fig. 4).

In summary, *by hiding the features specific for each sequence, the statistical study of gene populations analyses the population evolution history or the mean evolution history of a sequence from a population.*

3. Simulation Model A simple model is developed in order to simulate a gene population by mixing the two oligonucleotides $YRY(N)_3$ and $YRY(N)_6$ according to a Markov chain (of order 2: 2 matrix parameters), and by varying the percentages of RYR and YRY in the unspecified trinucleotides $(N)_3$ of $YRY(N)_3$ and $YRY(N)_6$ (four percentage parameters). The principle of this model is to determine the values of these six parameters so that the curve $C(S)$ of the simulated population S has similar features to the curve $C(F)$ of a given real population F. The features of the real populations were presented in Section 2. We will demonstrate that a simple model with six parameters can retreive the $YRY(N)_6$ YRY preferential occurrence, the periodicities $P3$ and $P2$ and some particular values of $p_i(S)$. Furthermore, several properties identified

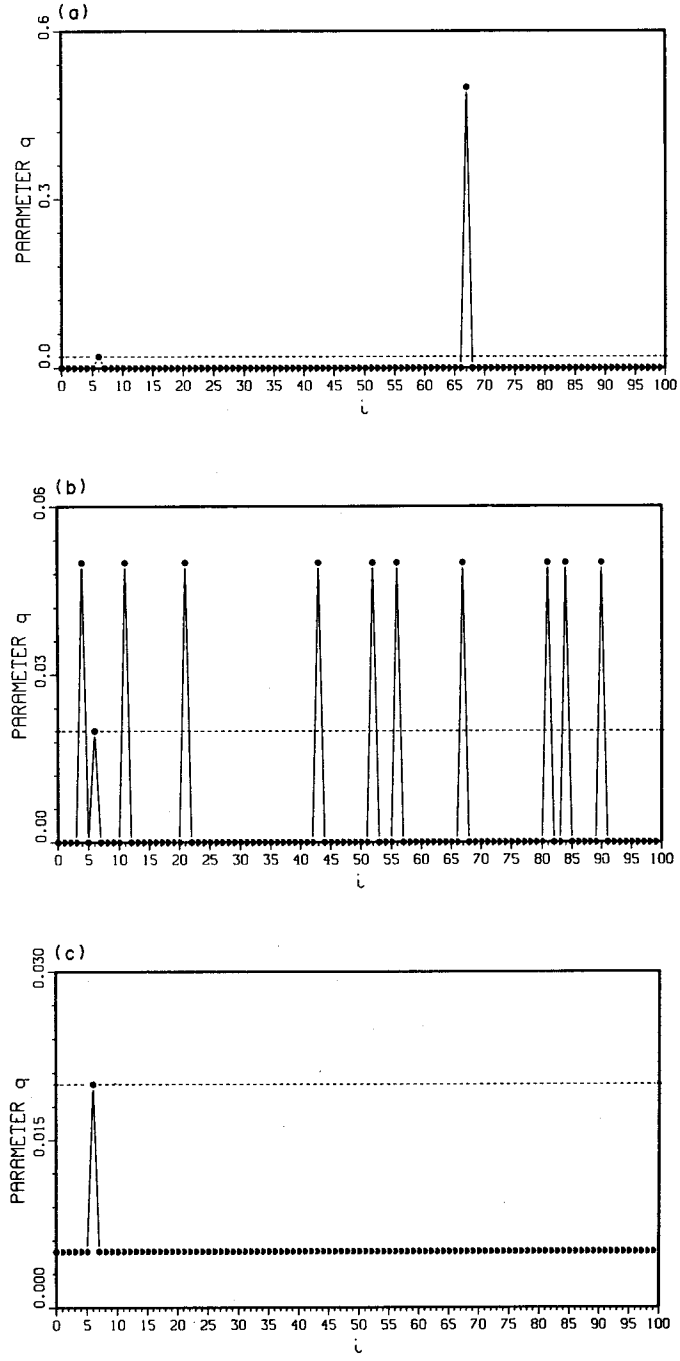


Figure 3. Difference between the statistical study of a few DNA sequences and the statistical study of a large gene population. Mean statistical curve for: (a) 1, (b) 10; (c) 10 000 sequences (see Section 2.3).

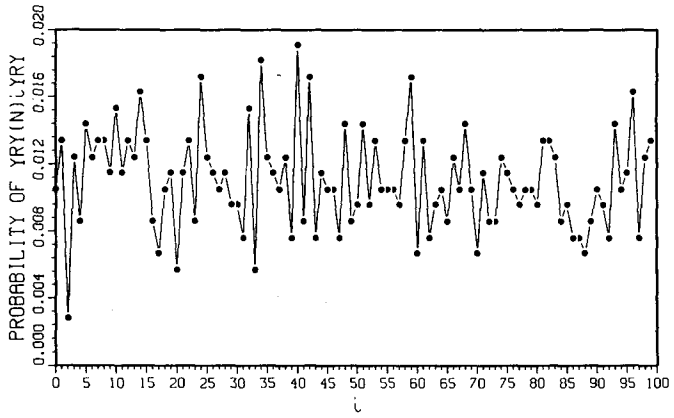


Figure 4. Mean occurrence probability of the i-motif $YRY(N)_iYRY$ in the eukaryotic protein coding gene CLCK (EMBL identification), starting at position 39 and ending at position 1184: absence of the periodicity $P3$ and low frequencies of $YRY(N)_0YRY$ and $YRY(N)_6YRY$. The horizontal axis represents the number i of bases N in the i-motif $YRY(N)_iYRY$ with i in the range [0, 99]. The vertical axis represents the mean occurrence probability p_i(CLCK) (see Section 2.1).

will show that the Markov model is a random independent one and that the $YRY(N)_6YRY$ preferential occurrence and the perioditicities $P3$ and $P2$ are functions of a unique parameter.

3.1. *Method* The model uses 2 oligonucleotides: $O_3 = YRY(N)_3$ and $O_6 = YRY(N)_6$. Then, a simulated sequence is obtained by concatenation of O_3 and O_6 according to the 2 states of a Markov chain having the matrix:

$$\begin{array}{c} \text{State} \quad O_6 \quad O_3 \\ \begin{array}{c} O_6 \\ O_3 \end{array} \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} \end{array}$$

where p (resp. $1-p$, q, $1-q$) is the probability that O_3 follows O_6 (resp. O_6 follows O_6, O_6 follows O_3, O_3 follows O_3) in the concatenation process. In addition to these 2 matrix parameters p and q, four percentage parameters characterize the occurrence probabilities of YRY and RYR in $(N)_3$ of O_3 and O_6: let y_3 (resp. r_3) be the probability of YRY (resp. RYR) in $(N)_3$ of O_3 and let y_6 (resp. r_6) be the probability of YRY (resp. RYR) in the first or in the last three bases in $(N)_6$ of O_6.

When an oligonucleotide O_3 (resp. O_6) is generated, the three bases in $(N)_3$ (resp. the first three bases then the last three bases in $(N)_6$) are specified according to the two following rules.

Rule 1: by YRY and RYR with the probabilities y_3 and r_3 (resp. y_6 and r_6), otherwise by application of the rule 2 with the probabilities $1-y_3-r_3$ (resp. $1-y_6-r_6$).

Rule 2: by three bases R or Y, chosen according to the respective probabilities r and $1-r$ (determined below), so that R and Y have the same percentage 0.5 in the final simulated sequence.

Note: (i) $a_3=y_3+r_3\leqslant 1$ and $a_6=y_6+r_6\leqslant 1$. (ii) $1-a_3$ or $1-a_6$ have to be large enough for balancing the number of R occurrences by application of the rule 2 because O_3 and O_6 have initially a greater number of Y occurrences.

More precisely, from the Markov chain formula (Feller, 1968, p. 375), the proportion of O_3 (resp. O_6) is $p/(p+q)$ (resp. $q/(p+q)$). Let α (resp. β, γ) be the proportion of Y (resp. R, N) in the simulated sequence after having specified $(N)_3$ and $(N)_6$ with the rule 1. Then, the following formulae can be proved:

$$\alpha=\{2(p+q)+p(2y_3+r_3)+q(4y_6+2r_6)\}/(6p+9q)$$
$$\beta=\{p+q+p(y_3+2r_3)+q(2y_6+4r_6)\}/(6p+9q)$$
$$\gamma=\{3p(1-y_3-r_3)+6q(1-y_6-r_6)\}/(6p+9q).$$

(Note: $\alpha+\beta+\gamma=1$.)

Then, the rule 2 specifies the proportion γ of bases N as the sum of a proportion δ of R and of a proportion ε of Y: (i) $\delta+\varepsilon=\gamma$; (ii) $\delta+\beta=\varepsilon+\alpha$ (traduces the equality between the final R proportion and the final Y one).

Then:

$$\delta=\{2p(2-y_3-2r_3)+q(7-4y_6-8r_6)\}/(12p+18q),$$

and:

$$r=\delta/\gamma=\{2p(2-y_3-2r_3)+q(7-4y_6-8r_6)\}/\{6p(1-y_3-r_3)+12q(1-y_6-r_6)\}.$$

The i-motif $YRY(N)_iYRY$ is studied with the statistical function $i\rightarrow p_i(S)$ (curve $C(S)$) by varying i in the range [0, 99] (Section 2.1.2) in a simulated population S constituted by 300 sequences of 1104 (= 1000 + 104) base length which are simulated according to a given sextuplet $(p, q, y_3, y_6, r_3, r_6)$.

In fact, for all i, the probabilities $p_i(S)$ can be exactly determined by formulae. Indeed, $p_i(S)$ is a function of the 6 variables p, q, y_3, y_6, r_3, r_6, for example:

$$p_0(S)=\{2pu+q\{2t+2s(t-y_6)+t^2\}\}/(6p+9q)$$

with:

$$s=r(1-r)(1-y_6-r_6)+r_6,\ t=r(1-r)^2(1-y_6-r_6)+y_6 \text{ and}$$
$$u=r(1-r)^2(1-y_3-r_3)+y_3.$$

These exact values of $p_i(S)$ can also be computed by a tree track algorithm (not

detailed here), but the simulation method described above is sufficient for a small scanning around the exact sextuplets. This simulation method is also simple to be programmed.

The exhaustive search of the sextuplet $(p, q, y_3, y_6, r_3, r_6)$ in $[0, 1]^6$ for identifying a simulated curve $C(S)$ similar to a real curve $C(F)$, needs a high computing time (e.g. more than 20 h with a VAX 8600). Furthermore, a large step for the scanning may not converge towards a solution.

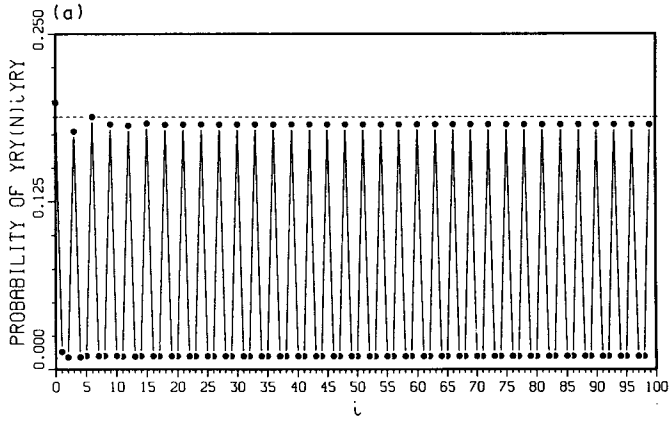
3.2. *Results.* With a complete scanning of 0.01 scale in the range [0, 1], the sextuplets $(p, q, y_3, y_6, r_3, r_6)$ were determined for the real populations $F=\{$CEUK, CPRO, CCHL, IVIR, IMIT, IEUK, NEUK, RR, NVIR$\}$ so that the associated simulated populations noted $S=\{$S-CEUK, S-CPRO, S-CCHL, S-IVIR, S-IMIT, S-IEUK, S-NEUK, S-RR, S-NVIR$\}$ have the same features (Section 2.2 and Table 2):

	p	q	y_3	y_6	r_3	r_6	Fig.	Strongly similar to Fig.
S-CEUK	0.69	0.25	0.92	0.00	0.06	0.00	5a	2a of CEUK
S-CPRO	0.61	0.33	1.00	0.00	0.00	0.00	5b	2b of CPRO
S-CCHL	0.61	0.38	0.74	0.00	0.22	0.00	5c	2c of CCHL
S-IVIR	0.37	0.65	0.00	0.20	0.00	0.00	5d	2d of IVIR
S-IMIT	0.18	0.30	0.05	0.34	0.75	0.24	5e	2e of IMIT
S-IEUK	0.99	0.14	0.13	0.00	0.87	0.00	5f	2f of IEUK
S-NEUK	0.89	0.17	0.32	0.00	0.67	0.00	5g	2g of NEUK
S-RR	0.31	0.76	0.24	0.00	0.70	0.20	5h	2h of RR
S-NVIR	0.67	0.53	0.32	0.08	0.54	0.20	5i	2i of NVIR

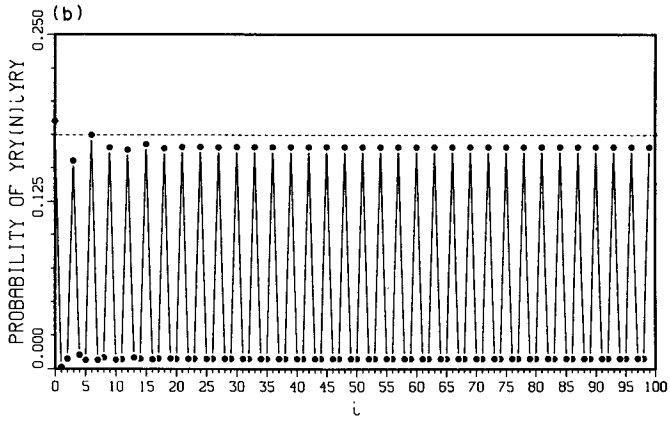
Table 2. Statistical features

Features	Real curves: Fig. 2a–i (Section 2)	Simulated curves: Fig. 5a–i (Section 3)
P3 M0, 6:	CEUK, CPRO, CMIT	S-CEUK, S-CPRO
P3 M0:	CVIR, CPLA, NPRO	see the features *P3 M0, 6*
P3 M6:	CCHL, IVIR	S-CCHL, S-IVIR
P3 M15:	IMIT	S-IMIT
P2 M1:	IEUK	S-IEUK
P2 M3:	NEUK	S-NEUK
M6:	ICHL, NCHL, NMIT, RR, TR, SNR	S-RR
No feature:	NVIR	S-NVIR

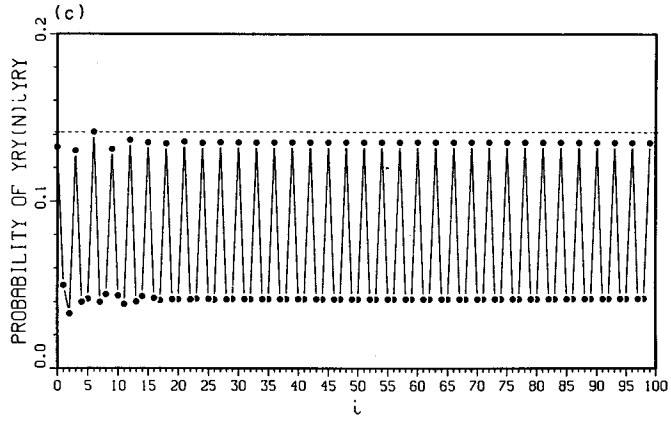
The chosen features being the most statistically significant ones, have the most important biological meanings, e.g. the periodicity *P3* reflects the protein coding function of a gene. However, a simulated curve, even having these features, is not necessarily identical to the real curve because: (1) the description of a curve form by making use of features is a difficult problem of



S-CEUK

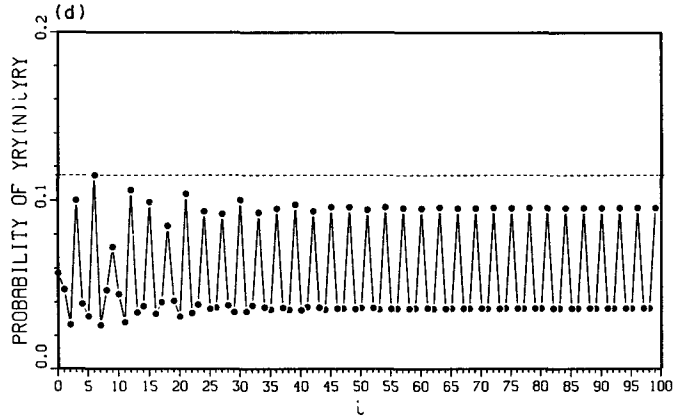


S-CPRO

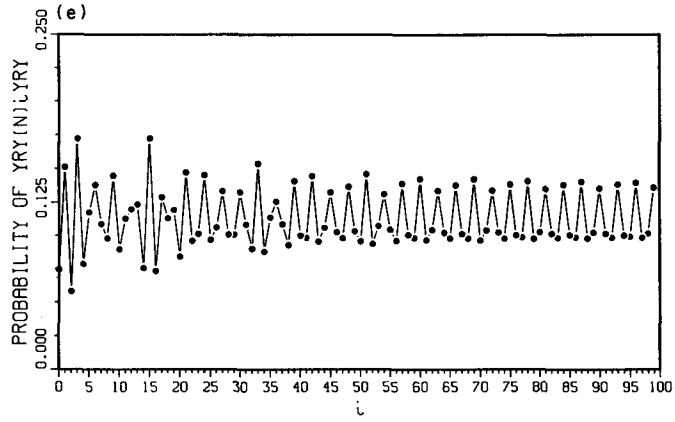


S-CCHL

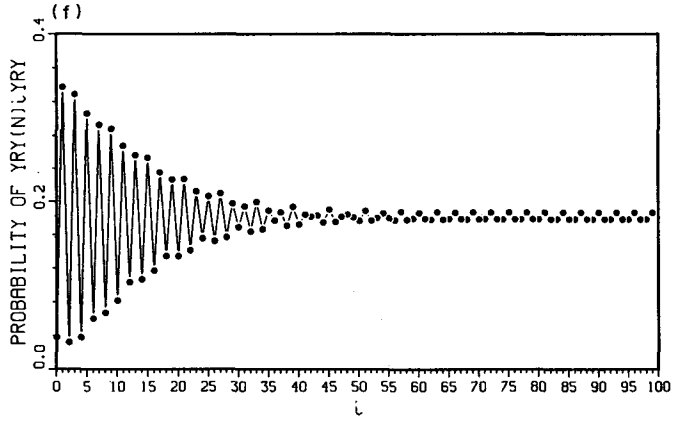
Figure 5



S-IVIR

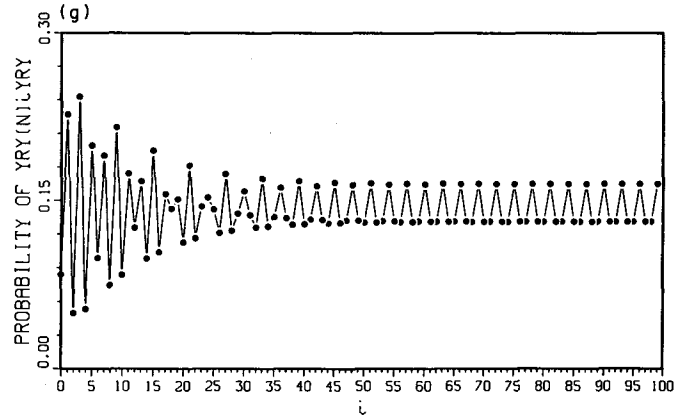


S-IMIT

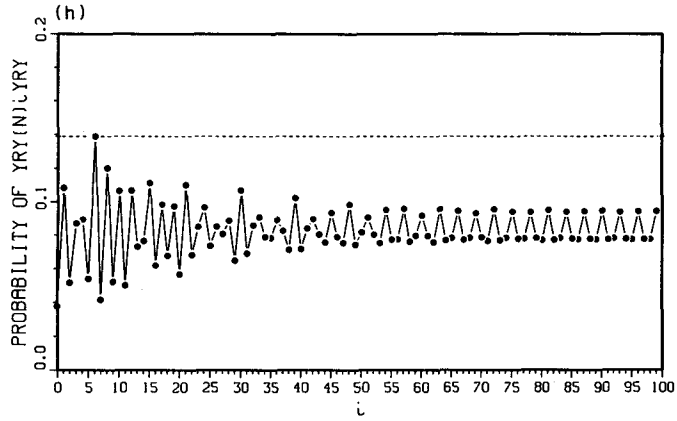


S-IEUK

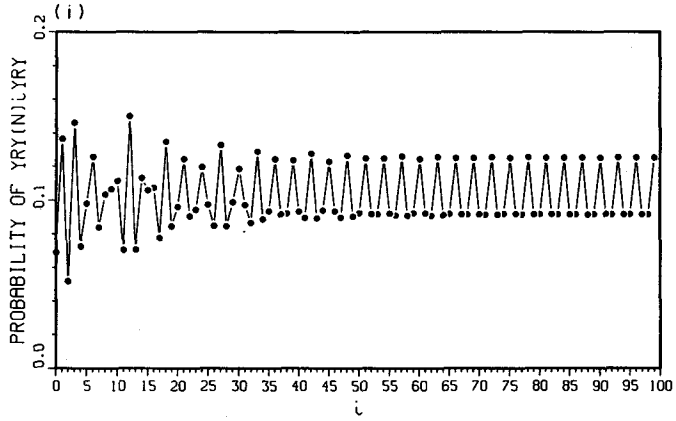
Figure 5—*continued.*



S-NEUK



S-RR



S-NVIR

Figure 5—*continued.*

pattern recognition, which cannot be solved by only a few features, e.g. the periodicity $P3$ can be uniform or incomplete (Section 2.2); (2) the simulated curve form is simple and uniform with a model consisting of six parameters. This model cannot simulate completely the biological reality depending on a great number of factors, in the same way the first terms of development of a function in series cannot reveal the totality of the function. On the other hand, the regularity of the real curves increases with the size of the population (the statistical reason follows from the law of large numbers, Section 2.3.3.). Therefore, all real curve forms obtained from populations of small size, have to be considered with caution, except the features stressed in the Section 2 results.

Another problem concerns the region in $[0, 1]^6$ in which the simulated curves satisfy the given features. This region is limited by a sextuplet of minimal values and by a sextuplet of maximal values. For example for S-CEUK, the rules used in order to obtain such a region, are the periodicity $P3$ and the following inequalities which traduce the features of CEUK observed in Section 2:

$$p_0(\text{S-CEUK}) - p_6(\text{S-CEUK}) > a \times D$$
$$p_6(\text{S-CEUK}) - \text{H} > b \times D$$
$$p_1(\text{S-CEUK}) - \text{B} > c \times D$$
$$\text{H} - p_3(\text{S-CEUK}) < d \times D$$

where D is the difference between the maximum of the top curve (in this case p_0(S-CEUK): feature $M0$) and the minimum of the bottom curve (in this case p_2(S-CEUK): feature $m2$, H (resp. B) is the mean value of the top (resp. bottom) curve and a, b, c, d are real so that, when a, b, c increase and d decreases, the region converges to the limit sextuplet, the convergence being a function of the scanning scale.

Legend for pages 758–760

Figure 5. Mean occurrence probability of the i-motif YRY(N)$_i$YRY in the simulated populations. The horizontal axis represents the number i of bases N in the i-motif YRY(N)$_i$YRY, with i in the range [0, 99]. The vertical axis represents the mean occurrence probability $p_i(S)$ (see Section 2.1) over all the sequences in the following simulated populations S: (a) S-CEUK, simulation of the eukaryotic protein coding genes; (b) S-CPRO, simulation of the prokaryotic protein coding genes; (c) S-CCHL, simulation of the chloroplast protein coding genes; (d) S-IVIR, simulation of the viral introns; (e) S-IMIT, simulation of the mitochondrial introns; (f) S-IEUK, simulation of the eukaryotic introns; (g) S-NEUK, simulation of the eukaryotic 5′ regions; (h) S-RR, simulation of the ribosomal RNA genes; (i) S-NVIR, simulation of the viral 5′ regions. A horizontal dashed lines goes through the point (6, $p_6(F)$).

The region obtained for S-CEUK is:

$$0.68 \leqslant p \leqslant 0.69$$
$$0.25 \leqslant q \leqslant 0.26$$
$$0.90 \leqslant y_3 \leqslant 0.93$$
$$0.00 \leqslant y_6 \leqslant 0.02$$
$$0.06 \leqslant r_3 \leqslant 0.10$$
$$0.00 \leqslant r_6 \leqslant 0.04.$$

Note that the radius of the region obtained is of the same order than the scanning scale 0.01. This same methodology has been used to obtain the regions (data not shown) and the limit sextuplets (given above) for the eight other simulated populations.

For a given feature choice, the simulated curve form is unique and allows some returns to reality (Section 4). As a first example, the curve C(S-NEUK) (Fig. 5g) of the simulated population S-NEUK has four obvious sets of points which can be joined by regular curves: (1) $i=0, 6, 12, 18$ and 24; (2) $i=3, 9, 15, 21, 27$ and 33; (3) $i=1, 5, 7, 11, 13, 17, 19, 23$, and 25; (4) $i=2, 4, 8, 10, 14, 16, 20$ and 22. However no condition was included in the model concerning these obvious curves, these features exist in the curve C(NEUK) (Fig. 2g) of the real population NEUK.

The choice of the two oligonucleotides $YRY(N)_6$ and $YRY(N)_3$ for the model is deduced from our previous studies. Only the three main reasons are given (by convention, a sequence constituted by a concatenation of several identical oligonucleotides O is noted $(O)^*$, e.g. $YRY(N)_6YRY(N)_6 \ldots$ is noted $(YRY(N)_6)^*$): (1) the particular sequence $(YRY(N)_3)^*$ with $(N)_3 = RYR$, alternating purine/pyrimidine stretches, leads obviously to the periodicity *P*2 (Arquès and Michel, 1987c): (2) any mixing of $YRY(N)_6$ and $YRY(N)_3$ leads to the periodicity *P*3; (3) The particular sequence $(YRY(N)_6)^*$ gives a high frequency of the *i*-motif $YRY(N)_iYRY$ at $i=6, 15, 24$, etc. (Arquès and Michel, 1987b) and the insertion of $YRY(N)_3$ in a sequence $(YRY(N)_6)^*$ leads to the highest frequency of the *i*-motif $YRY(N)_iYRY$ at $i=6$ (compared to $i=15, 24$, etc.) because such an insertion destroys one subsequence $YRY(N)_6YRY$ but two subsequences $YRY(N)_{15}YRY$, three subsequences $YRY(N)_{24}YRY$, etc.

3.3. *Discussion*

3.3.1. This "oligonucleotide" model presents the DNA sequence evolution as follows: Two *primitive oligonucleotides* $O_3 = YRY(N)_3$ and $O_6 = YRY(N)_6$, through *concatenations* (step 1 in Fig. 1), led to the *primitive sequences* (here the simulated populations *S*), then through *random changes* (mutations, etc.)

(step 2 in Fig. 1), these primitive sequences led to the *actual sequences* (here the real populations F). Most probably, concatenations and random changes were not separated into two distinct steps, but concatenations were predominant at an early stage and then progressively replaced by random changes. The random changes (i.e. the step 2 in Fig. 1) are necessary in the oligonucleotide model (see Section 3.3.2.).

3.3.2. Relation between the oligonucleotide model and the mutation model (reviews in Kimura, 1987; Nei, 1987). The $p_i(S)$ values, absolute and relative, in the simulated populations S are greater (about 10 times more) than the $p_i(F)$ values in the real populations F. In order to reach the real values by keeping the same features, *mutations are necessary in these simulated populations* S *with a maximal rate of order* 1/2 *mutation per specifiçed base* (*R or Y*) *and with any mutation rate per unspecified base N* (Arquès and Michel, model in preparation). In other words, there is a limit for the mutation rate with the specified bases.

3.3.3. Relation between the oligonucleotide model and the RNY model (Eigen and Schuster, 1978). The oligonucleotide model is more general than the RNY model (the preferential use of the RNY codon leads to preferential series of RNY codons, i.e. (RNY)*) because: (1) fewer bases are specified: 1/3 to 1/2 of the bases are specified in the sequences of the oligonucleotide model while 2/3 of the bases are specified in the (RNY)* sequences of the RNY model; (2) a (RNY)* type sequence can be retrieved from the sequences of the oligonucleotide model by specifying $(N)_3$ by YRN in $YRY(N)_3$ and $YRY(N)_6$; (3) many more features are explained with the ologonucleotide model. In particular, the $YRY(N)_6YRY$ preferential occurrence and the periodicity $P2$ cannot be explained by the RNY model (both models explain the periodicity $P3$).

3.3.4. Concept of the Markov concatenation. From a biological point of view, a Markov concatenation of oligonucleotides is probably too complex for an early stage of DNA sequence evolution. Unexpectedly, some properties (in particular the important property 1 given below) are identified and allow the simplification and the replacement of this Markov model by a random independent model. From a mathematical point of view, the choice of a Markov model allows to prove the existence of a random independent model, i.e. a random independent model is not taken as hypothesis.

3.3.5. For S-CEUK, S-CPRO, S-CCHL, S-IEUK, S-IVIR, S-NEUK, S-NVIR, S-RR:

Property 1: $p+q \approx 1$ (from Section 3.2; Fig. 6a). The projection of these eight

sextuplets in the plane (p, q) shows that these eight points are approximately on a straight line of equation $p+q=1$. With a linear regression, the straight line has the following equation D: $p=-0.94q+1.02$. If S-NVIR is not considered because S-NVIR is the furthest point from D, then the linear regression leads to the straight line of equation D': $p=-1.00q+1.02$.

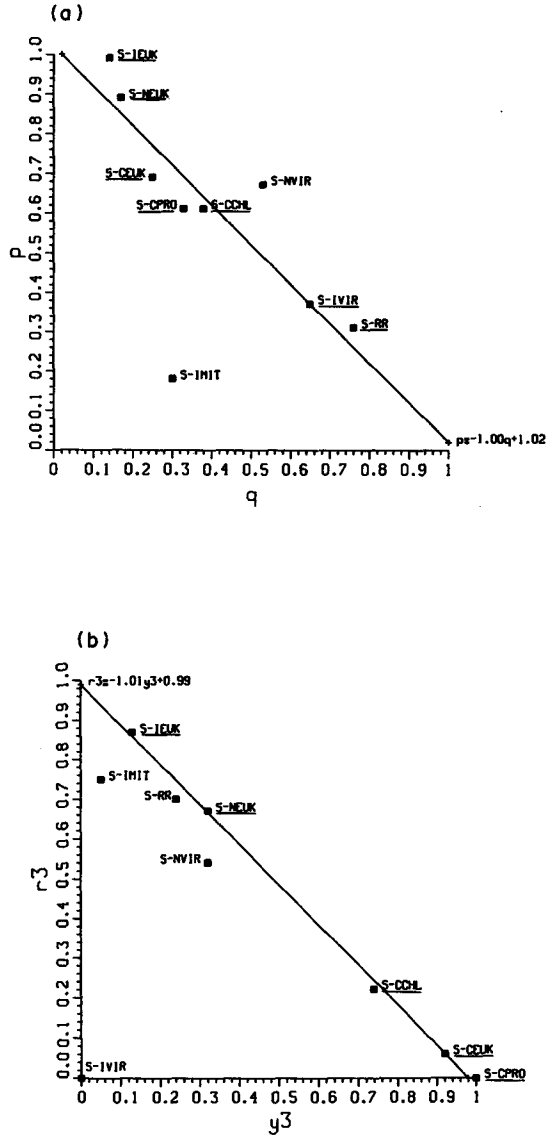


Figure 6.

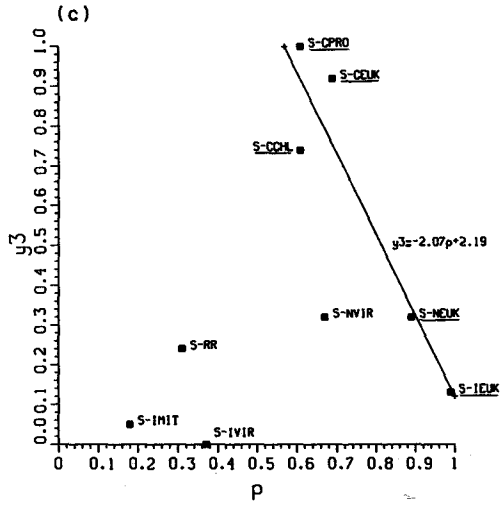


Figure 6. (a) Linear regression between the two parameters p and q proving the independent concatenation of the two oligonucleotides $YRY(N)_3$ and $YRY(N)_6$ for the seven underlined simulated populations S used in the linear regression. (b) Linear regression, for the five underlined simulated populations, between the two parameters r_3 and y_3 proving the complete specification of the oligonucleotide $YRY(N)_3$ as being independently YRYYRY or YRYRYR. (c) Linear regression, for the five underlined simulated populations, leading to a complete specification by the unique parameter p of the probabilities characterizing the independent concatenation of the three oligonucleotides YRYYRY, YRYRYR and $YRY(N)_6$.

Therefore, the Markov model is associated with the simplified matrix:

$$\begin{array}{c} \text{State} \quad O_6 \quad O_3 \\ O_6 \begin{bmatrix} 1-p & p \end{bmatrix} \\ O_3 \begin{bmatrix} 1-p & p \end{bmatrix} \end{array}.$$

Then, the concatenation of O_3 and O_6 is independent because:

$$\begin{aligned} \text{Prob(state } O_3 \cap \text{state } O_6) &= \text{Prob(state } O_3/\text{state } O_6) \times \text{Prob(state } O_6) \\ &= p(1-p) \\ &= \textit{Prob(state } O_3) \times \textit{Prob(state } O_6) \end{aligned}$$

In summary, the simulated populations are generated by an *independent concatenation* of the two oligonucleotides $YRY(N)_3$ and $YRY(N)_6$ with the probabilities p and $1-p$ respectively. This property leads to the important observation that the concatenation is random independent (proved as being a particular case of a Markov model). The absence of a clever concatenation is in agreement with a primitive stage of DNA sequence evolution.

3.3.6. In addition, for S-CEUK, S-CPRO, S-CCHL, S-IEUK, S-NEUK:
Property 2: $y_6 = r_6 = 0$ (from Section 3.2). The six bases N of the oligonucleotide $YRY(N)_6$ *have not to be specified by YRY and by RYR.*

Obviously, specification by motifs different from YRY and RYR could be necessary for a model more general which also considers trinucleotides (e.g. RRR) different from YRY.

Property 3: $y_3 + r_3 \approx 1$ (from Section 3.2; Fig. 6b). With a linear regression, the straight line has the following equation $r_3 = -1.01y_3 + 0.99$. The three bases N of the oligonucleotide YRY(N)$_3$ are *completely specified* either by YRY or by RYR. Then, the simulated populations are generated by an independent concatenation of the *three oligonucleotides* YRYYRY, YRYRYR and YRY(N)$_6$ with the probabilities py_3, $p(1-y_3)$ and $1-p$ respectively.

Property 4: $y_3 = -2.07p + 2.19$ (Fig. 6c). The simulated populations are generated by an independent concatenation of the three oligonucleotides YRYYRY, YRYRYR and YRY(N)$_6$ with the probabilities $-2.07p^2 + 2.19p$, $2.07p^2 - 1.19p$ and $1-p$ respectively, functions of *the unique parameter p*.

3.3.7. Remarks. The real populations CVIR, CPLA and NPRO were not simulated. Indeed, these populations have the features *P*3 and *M*0 which are more general than features *P*3 and *M*0, 6 of the real populations CEUK, CPRO and CMIT (Section 2.2.2a. and Table 2). For the simulated population S-IMIT which does not verify any of the properties mentioned above, the question arises whether this model is sufficient or whether the size of the real population IMIT is too small (see the statistical reason presented in Section 2.3.3). It also reveals the problem of the definition of a gene population. These remarks may also explain that some real populations, e.g. the transfer RNA genes, cannot be simulated satisfactorily, obviously except the feature *M*6.

4. Return of the Model to the Reality. The evaluation of such a model relies on the new information concerning the reality which is returned by the model (new information in the sense that the information was not known before the construction of the model and thereby, not introduced in the model). This model is tested in two biological problems: the distribution of the large alternating purine/pyrimidine stretches and the identification of new hidden periodicities.

4.1. *Study of the large alternating purine/pyrimidine stretches*

4.1.1. Problem. The distribution of the large ($\geqslant 15$ bases) alternating purine/pyrimidine stretches (YR)* (or (RY)*) depends on gene populations. In particular, the (YR)* stretches are large and numerous (i.e. statistically improbable by chance) in eukaryotic introns IEUK and normal in eukaryotic protein coding genes CEUK (Arquès and Michel, 1987c). There is no simple and natural explanation for this observation. In fact, we will demonstrate that the presence or the absence of large stretches (YR)* in a real population *F* is simply a consequence of the characterization (Section 3.2) of this population

by the mixing of the two oligonucleotides $O_3 = YRY(N)_3$ and $O_6 = YRY(N)_6$ and by the specification of YRY and RYR in $(N)_3$ of O_3 and O_6. It is important to stress that no (YR)* stretch was directly introduced in the model.

4.1.2. Method. Let R_F (resp. R_S) be the repartition function of the (YR)* stretches of length $j \geqslant 15$ bases in a real population F (resp. in its associated simulated population S). S is generated according to Section 3 with the sextuplet from Section 3.2. The similarity between R_F and R_S is evaluated with the graphic comparison of the curves, with an identity test and with a correlation test of the length of the largest stretch for which the function reaches its maximum 1.

4.1.3. Results

(a) Curves of R_F and R_S. The functions R_F and R_S are given for the six populations CEUK, NEUK, IVIR, NVIR, RR and IEUK, (Figs 7a–c), whereas for the other populations CPRO, CCHL and IMIT, real as well as simulated, the number of large stretches is not statistically significant for this study. The similarity between R_F and R_S for these six populations is graphically obvious (Figs 7a–c). The function R_S is regular because the size of the simulated population S is large. For the real populations of small size IVIR, NVIR and RR, the function R_F is a repartition function of a discrete random variable (curve with steps) and the function R_S appears to be the "regularized" of the R_F one.

(b) Identity test. The similarity between R_F and R_S can also be evaluated with the Kolmogorov–Smirnov test for two samples (DeGroot, 1986, p. 559). This test accepts (resp. rejects) at the 5% statistical level, the identity hypothesis of R_F and R_S if the value of $D = \mathrm{Max}_j\{R_F(j) - R_S(j)\}$ is less (resp. greater) than $1.36\ \{(m+n)/mn\}^{1/2}$, where m and n are the number of stretches of length $j \geqslant 15$ bases in the populations F and S.

This test accepts the identity hypothesis of R_F and R_S for CEUK, IVIR, NVIR and RR whatever $j \geqslant 15$, and for NEUK with j outside the range [19, 21] (data not shown). For IEUK, the identity of R_F and R_S needs a translation but the real population IEUK and the simulated population S-IEUK have both unexpected large stretch (YR)*.

(c) Correlation test. There is also a strong correlation between the real population F and its associated simulated population S concerning the length of the largest alternating purine/pyrimidine stretch (Table 3). The correlation coefficient r between the length j_F of the largest stretch in F and the length j_S of the largest stretch in S is $r = (\Sigma_{(F,S)} j_F \times j_S)/(\Sigma_F j_F^2 \times \Sigma_S j_S^2)^{1/2}$.

For CEUK, NEUK, IVIR, NVIR, RR and IEUK, r is equal to 94%. If IEUK is again not considered, r is then equal to 99%.

4.2. *Identification of new hidden periodicities*

4.2.1. Introduction. A significant periodicity $P3$ for $i \geqslant 30$ is observed in the simulated populations S, except for S-IEUK. This result was known for the real populations CEUK, CPRO, CCHL and IVIR (Section 2.2.2a) but it was not observed for the real populations NEUK, NVIR, IMIT and RR. In agreement with the simulation results, the presence (resp. the absence) of a periodicity $P3$ will be proved to be statistically significant for NEUK and NVIR (resp. for

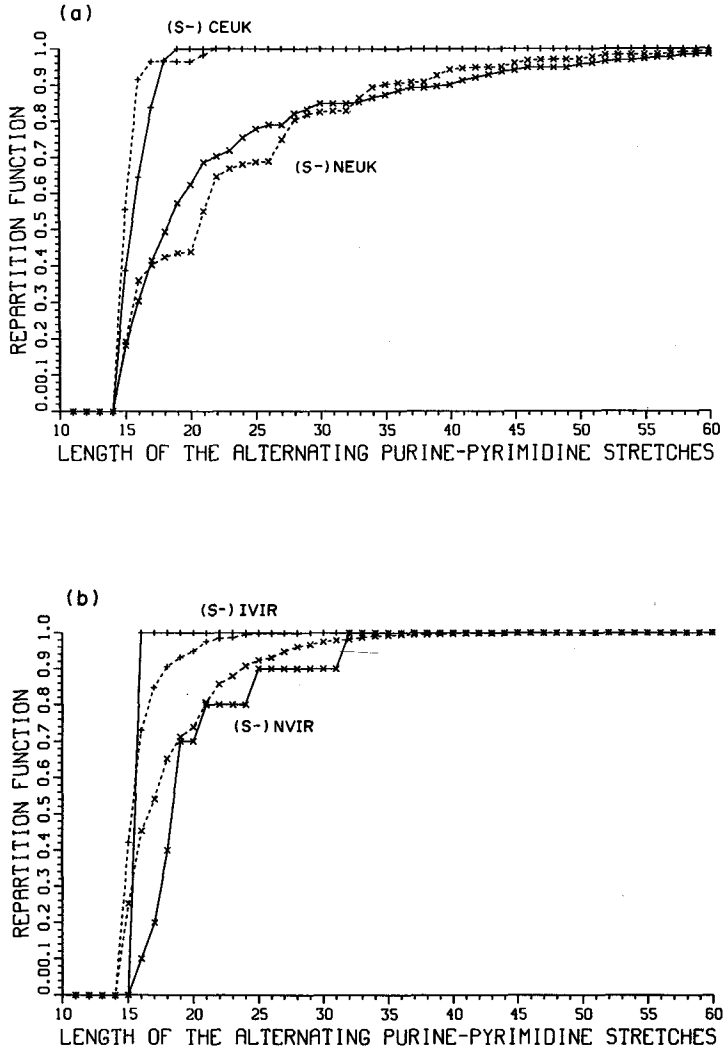


Figure 7.

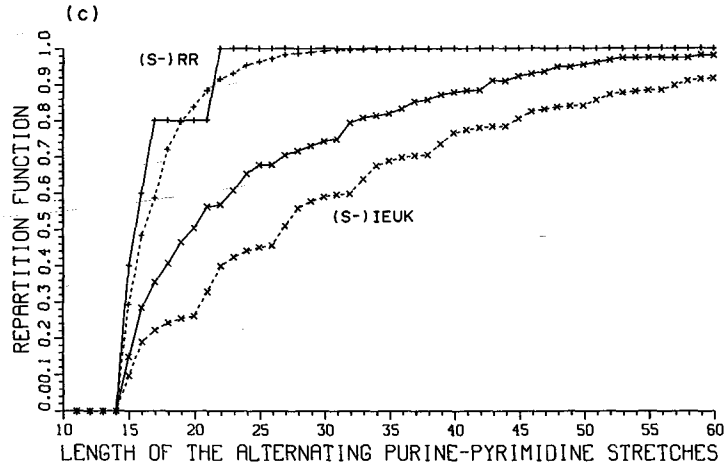


Figure 7. Repartition function of the large alternating purine/pyrimidine stretches (length $\geqslant 15$ bases) (continued curves stand for the real populations and dashed curves, for the simulated populations): (a) for the eukaryotic protein coding genes CEUK and S-CEUK ("+" points), for the eukaryotic 5′ regions NEUK and S-NEUK ("x" points); (b) for the viral introns IVIR and S-IVIR ("+" points), for the viral 5′ regions NVIR and S-NVIR ("x" points); (c) for the ribosomal RNA genes RR and S-RR ("+" points), for the eukaryotic introns IEUK and S-IEUK ("x" points).

Table 3. Length of the largest alternating purine/pyrimidine stretch

	Real population	Simulated population
CEUK	19	22
NEUK	71	118
IVIR	16	29
NVIR	32	54
RR	22	48
IEUK	76	264

IEUK). This result was completely unexpected and it was not observed in the classification in Section 2.2.2.

4.2.2. Method. For a real population F and for $i \equiv 0[3]$ in the range [30, 98], let $X_i(F)$ be the Bernoulli random variable which is equal to 1 if $p_i(F) \geqslant \text{Max}\{p_{i-1}(F), p_{i+1}(F)\}$, and 0 otherwise. The sum of the independent $X_i(F)$ is a Binomial random variable $N(F)$ of unknown parameter p and of order 23, which counts the number of local maxima among the 23 possible

values of i in the range [30, 98]. $N(F)$ is a measure of the periodicity $P3$ (see the definition in Section 2.2.1.): a curve $C(F)$ with a periodicity $P3$ (resp. incomplete periodicity $P3$) is associated to a parameter p equal (resp. close) to 1. To the contrary, a random curve $C(F)$ (having no periodicity) is associated to a parameter $p=1/3$.

For a given real population F, the hypothesis H_0: $p=1/3$ is tested against the hypothesis H_1: $p>1/3$. Since the order is large enough (i.e. $23 \geqslant 5/p(1-p)$), the central limit theorem asserts that under H_0, $Z(F)=(N(F)/23-1/3)/(23^{-1}\times 2\times 9^{-1})^{1/2}$ is close to a reduced centered Gaussian variable. Therefore, the hypothesis H_1 of a (incomplete) periodicity $P3$ is accepted at the 5% statistical level, if $Z(F)>1.645$.

4.2.3. Results. NEUK and NVIR have a periodicity $P3$ because $N(NEUK)=N(NVIR)=16$ and $Z(NEUK)=Z(NVIR)=3.7>1.645$. IEUK has no periodicity $P3$ because $N(IEUK)=9$ and $Z(IEUK)=0.6<1.645$. Note: This statistical test cannot identify a periodicity $P3$ for IMIT and RR because $N(IMIT)=N(RR)=9$ and $Z(IMIT)=Z(RR)=0.6$ (see Section 3.3.7).

4.3. *Conclusion.* The return of the model to the reality allows us to understand the presence or the absence of large alternating purine/pyrimidine stretches as being a simple consequence of the mixing of two particular oligonucleotides. It also identifies a periodicity $P3$ for the two 5′ regions of eukaryotes NEUK and viruses NVIR.

These observations suggest that the 5′ regions have the genetic information for protein coding genes and for introns: (1) in the eukaryotic genome, the 5′ regions (NEUK: periodicities $P2$ and $P3$; Section 2.2.2b. and Section 4.2.3.) have the information for protein coding genes (CEUK: periodicity $P3$; Section 2.2.2a.) and for introns (IEUK: periodicity $P2$; Section 2.2.2b.); (2) in the prokaryotic genome, the 5′ regions (NPRO: periodicity $P3$; Section 2.2.2a) have the information for protein coding genes (CPRO: periodicity $P3$; (Sections 2.2.2a); (3) in the viral genome, 5′ regions (NVIR: periodicity $P3$: Section 4.2.3) have the information for protein coding genes (CVIR: periodicity $P3$; Section 2.2.2a) and for introns (IVIR: periodicity $P3$; Section 2.2.2a). The absence of periodicity $P2$ in virial introns (in opposition to eukaryotic introns) may be related to the absence of periodicity $P2$ in 5′ regions of viruses.

5. Conclusion of Sections 1–4: Consequences of this Model. We have developed a simple model, without making any hypotheses, explaining DNA sequence evolution in terms of primitive oligonucleotides, of primitive sequences, of an oligonucleotide concatenation process and of a mutation process (Fig. 1). Important future applications could be deduced from this model:

More accurate dating, by comparing similar genes to an ancestor issued from this model.
Homologies between various sequences, by considering the existence of specified (R or Y) and unspecified bases (N).
Identification of other oligonucleotides, by obtaining a better shape adequation between simulated and real curves or by reaching the exact real values after the mutation process.
Association of oligonucleotide concatenations with base mutations, etc.

Finally, the development of a mathematical function which analyses motifs also different from YRY (but according to the statistical results, YRY has a central role in DNA sequence evolution) should lead to a unified model of DNA sequence evolution allowing the molecular understanding of both the origin of life and the actual biological reality. This view is strongly supported by the fact that several different features are functions of a unique parameter (see property 4 in Section 3.3.6).

We thank Professors Max Burger and Jacques Streith, Dr Christoph Nager, Thomas Nyffenegger, John Olsen and Nouchine Soltanifar for their advice and the bioinformatic group for its assistance. This work was supported by grants from the Unité Associée CNRS No 822 to D.G.A. and from the Friedrich Miescher Institute to C.J.M.

LITERATURE

Arquès, D. G. and C. J. Michel. 1987a. Study of a pertubation in the coding periodicity. *Math. Biosci.* **86**, 1–14.
Arquès, D. G. and C. J. Michel. 1987b. A purine–pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. *J. theor. Biol.* **128**, 457–461.
Arquès, D. G. and C. J. Michel. 1987c. Periodicities in introns. *Nucl. Acids Res.* **15**, 7581–7592.
DeGroot, M. H. 1986. *Probability and Statistics*. Reading, MA: Addison-Wesley.
Eigen, M. and P. Schuster. 1978. The hypercycle: a principle of natural self-organization. Part C: the realistic hypercycle. *Naturwissenschaften* **65**, 341–369.
Fickett, J. W. 1982. Recognition of protein coding regions in DNA sequences. *Nucl. Acids Res.* **10**, 5303–5318.
Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*. New York: Wiley.
Kimura, M. 1987. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
Lazowska, J., C. Jacq and P. P. Slonimski. Sequence of introns and flanking exons in wild-type and box3 mutants of cytochrome *b* reveals an interlaced splicing protein coded by an intron. 1980. *Cell* **22**, 333–348.
Lewontin, R. C. and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila Pseudoobscura*. *Genetics* **54**, 595–609.
Nei, M. 1987. *Molecular Evolutionary Genetics*. Washington, DC: Columbia University Press.
Shepherd, J. C. W. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. natl Acad. Sci. U.S.A.* **78**, 1596–1600.

Ziff, E. B. 1980. Transcription and RNA processing by the DNA tumour viruses. *Nature* **287**, 491–499.

Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. *J. theor. Biol.* **8**, 357–366.

Received 12 August 1989
Revised 9 February 1990