International Journal of Modelling & Simulation, Vol. 13, No. 3, 1993

# A MODEL OF GENE EVOLUTION BASED ON RECOGNIZABLE LANGUAGES AND ON INSERTION AND DELETION OPERATIONS

Didier G. Arquès* and Christian J. Michel**

* Université de Franche-Comté, Laboratoire d'Informatique, U.R.A. CNRS No 822,
16 route de Gray, 25030 Besançon, France
** Equipe de Biologie Théorique, Université de Franche-Comté
Institut Universitaire de Technologie de Belfort-Montbéliard, BP 527, 90016 Belfort, France

## Abstract

Each gene population F (i.e. a set of DNA sequences) can be considered as a language F on the two letter alphabet {R,Y} (R=purine: adenine or guanine; Y=pyrimidine: cytosine or thymine) issued from the EMBL gene database (release 21). For each language F, an autocorrelation function p is defined associating to each i in the range [1,99] the probability $p_i(F)$ of occurrence of the two same factors YRY separated by i letters over all words in F. The associated curve C(F): $i \to p_i(F)$ reveals in the first part two properties naturally associated with the biological features of the gene population F:
- Property 1: a periodicity P3 (modulo 3 periodicity) associated with the maximal value $p_6(F)$ if F is a protein coding gene population,
- Property 2: a periodicity P2 (modulo 2 periodicity) associated with the maximal value $p_3(F)$ if F is a noncoding gene population.
In the second part, we prove that:
(1) The rational language $(YRY(N)_6)^*$ (where $(N)_6$ is any word of length 6) associated with random insertions and deletions of factors of three letters leads to the property 1.
(2) The rational language $(YRY(N)_3)^*$ (where $(N)_3$ is any word of length 3) associated with random insertions and deletions of single letter leads to the property 2.
Therefore, the two types of words $YRY(N)_6$ and $YRY(N)_3$ constitute small primitive DNA strings which may have a central function in DNA sequence evolution.

Keywords: Modelling, Languages, Statistical methods, Bioinformatic, DNA sequence evolution.

## 1. General Introduction

The biological determination of nucleotides (bases or letters) which increases exponentially (today, more than 80 million nucleotides are known) their storage and description in gene databases and the recent development in computer sciences of algorithms, calculus power, data structures, together with mathematics and statistics, allows us to study precisely the distribution of nucleotides in genes (words of several letters). The identification of non-random statistical properties will imply the development of simulation models issued from the language theory in order to understand the organization of nucleotides in the actual genes and its evolution from primitive nucleotide sequences.

In the first part, the definition of a statistical function allows us to identify two types of periodicities in genes: a periodicity modulo 2 and a periodicity modulo 3. Furthermore, each periodicity is associated with a particular maximal value. These non-random statistical properties allow a simple classification of gene populations. The biological meaning of these properties are given in the discussion section of the part 1.

In the second part, the genetic reality observed in the first part is simulated. Indeed, we show that the periodicity P3 (resp. P2) and its maximal value can be simulated by applying a random insertion/deletion process of trinucleotides, i.e. words of 3 letters, (resp. of mononucleotides, i.e. letters) in sequences (called simulated) generated by the repetition of a particular oligonucleotide (word of length less than 10 letters). The properties and the biological meaning of this simulation model are given in the discussion section of the part 2.

## 2. Periodicities and Classification of Gene Populations

### 2.1. Introduction

The appropriate statistical function analysing, on the two-letter alphabet {R,Y} (R=purine: adenine or guanine, Y=pyrimidine: cytosine or thymine) the occurrence probability of the i-motif $YRY(N)_iYRY$ (2 trinucleotides (factors of length 3) YRY separated by any i bases (letters) N; N=R or Y) in gene populations, reveals the periodicity P3 (modulo 3 periodicity called coding periodicity; defined below; [8,6,1,2,3,4]) and the periodicity P2 (modulo 2 periodicity called alternating purine/pyrimidine periodicity; defined below; [3,4]). Each periodicity is associated with a maximal value: at i=6 ($YRY(N)_6YRY$ preferential occurrence; defined below; [2]) for the periodicity P3 and at i=3 for the periodicity P2.

### 2.2. Method

#### 2.2.1. *Statistical Function*

The gene populations are languages made of words associated with DNA sequences which are obtained from the EMBL Nucleotide Sequence Data Library (release 21). 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 $m_i=YRY(N)_iYRY$ (R=purine, Y=pyrimidine, N=R or Y) by varying i in the range [1,99], be 2 trinucleotides YRY separated by any i bases N. 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.

#### 2.2.2. *Periodicities*

Two types of periodicities are revealed in the curves C(F):
(1) **The periodicity P3 in the range [1,98]**:
$p_i(F)>\mathrm{Max}\{p_{i-1}(F),p_{i+1}(F)\}$ with $i \in [1,98]$ and $\mathbf{i \equiv 0[3]}$.
This periodicity P3 is associated with the maximal value $p_6(F)$ (called **$YRY(N)_6YRY$ preferential occurrence**):
$p_6(F)>p_i(F)$ with $i \in [1,99]$ and $i \neq 6$;
(2) **The periodicity P2 in the range [1,L]**:
$p_i(F)>\mathrm{Max}\{p_{i-1}(F),p_{i+1}(F)\}$ with $i \in [1,L]$ and $\mathbf{i \equiv 1[2]}$.
This periodicity P2 is associated with the maximal value $p_3(F)$:
$p_3(F)>p_i(F)$ with $i \in [1,99]$ and $i \neq 3$;

### 2.3. Results: Classification of Gene Populations

By using the periodicities, the gene populations of large size are classified as below. One figure is given for each periodicity which will be simulated in part 3.

2.3.1. *Gene Populations Having the Periodicity P3 in the Range [1,98] and the Maximal Value $p_6(F)$*

- Eukaryotic protein coding genes (8202 sequences), noted CEUK: Fig. 1(a).
- Prokaryotic protein coding genes (2890 sequences).
- Viral protein coding genes (2613 sequences).
- Chloroplast protein coding genes (305 sequences).
- Mitochondrial protein coding genes (280 sequences).
- 5' prokaryotic regions (883 sequences).
- 3' prokaryotic regions (476 sequences).

2.3.2. *Gene Populations Having the Periodicity P2 in the Range [1,L] and the Maximal Value $p_3(F)$*

- in the range [1,L=23]: 5' eukaryotic regions (2172 sequences), noted N5EUK: Fig. 1(b).
- in the range [1,L=49]: eukaryotic introns (1790 sequences) (note: $p_3$(IEUK) is the second highest value).
- in the range [1,L=99]: 3' eukaryotic regions (3966 sequences).

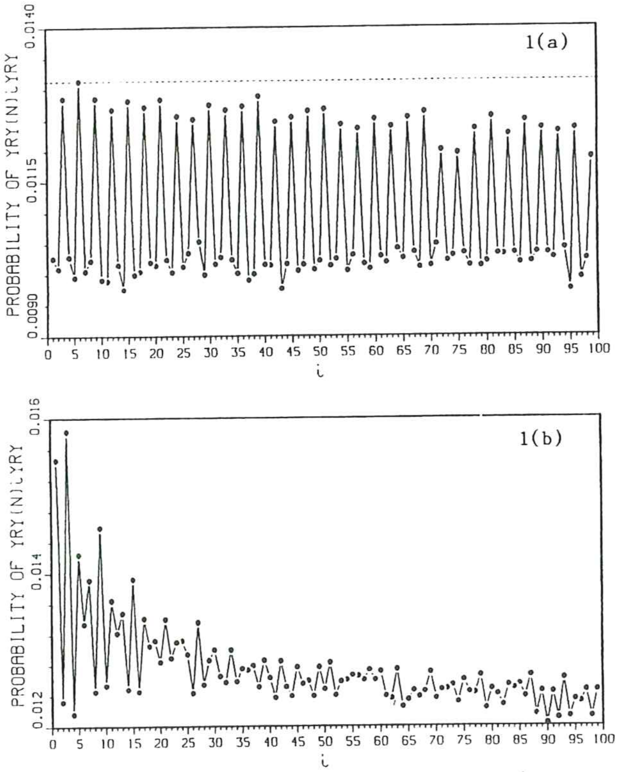


Fig. 1. Mean occurrence probability of the i-motif $YRY(N)_iYRY$ in gene populations. The horizontal axis represents the number i of bases N in the i-motif $YRY(N)_iYRY$, with i in the range [1,99]. The vertical axis represents the mean occurrence probability $p_i(F)$ (see part 1 method) over all the sequences in the following gene populations F: (a) CEUK: eukaryotic protein coding genes; a horizontal dashed line goes through the point $(6,p_6(\text{CEUK}))$, (b) N5EUK: 5' eukaryotic regions.

## 2.4. Discussion

The periodicity P3 is related to the protein coding function of a gene [5] and is found in protein coding genes of any taxonomic group [8,6,1,2,3,4].

The periodicity P2 is related to regulatory functions of a gene [3] and is found in eukaryotic introns and in the 5' and 3' regions of eukaryotes, i.e. only in the eukaryotic genome [3,4].

The $YRY(N)_6YRY$ preferential occurrence is related to the spatial structure (pitch) of the DNA double helix [2]. It is "universal" because: (i) it is found in gene populations having no periodicity: ribosomal RNA genes (196 sequences), transfer RNA genes (1081 sequences) and small nuclear RNA genes (129 sequences) [2]; (ii) it is hidden by the periodicity P2 as the deletion of the large alternating purine/pyrimidine stretches leads to a value $p_6(F)$ maximal for all i [3,4].

# 3. Simulation of the Periodicities

## 3.1. Introduction

Despite the huge combinatory of motifs, the periodicities and their respective maximal values presented in part 1 can be simulated with a simple model using:
(1) the language $(O)^*$ generated by the repetition of the oligonucleotide O (i.e. $(O)^*=\{(O)^n, n\geq 0\}$) associated with
(2) an evolutionary process classified by Kimura [7] in: i) a process of base mutations, ii) a process of insertions and deletions of one and more bases.
We will demonstrate that:
(1) The periodicity P3 and the maximal value $p_6(F)$ can be simulated with the language $(YRY(N)_6)^*$ associated with the random trinucleotide insertion/deletion process (defined below).
(2) The periodicity P2 and the maximal value $p_3(F)$ can be simulated with the language $(YRY(N)_3)^*$ associated with the random mononucleotide insertion/deletion process (defined below).

## 3.2. Method

A language called "simulated population S" made of 500 words called "sequences" $(YRY(N)_6)^*$ or $(YRY(N)_3)^*$, of 2000 base length, is generated in such a way that the R percentage is equal to the Y percentage (50%) in any sequence of S (step 0). Note: The computations obtained with such a sample of 1 million bases, are precise (i.e. not related to random fluctuations: a sample having 100 sequences of 1000 base length leads to similar results). Then, this population S is subjected to an insertion/deletion process in k steps so that, at each step, one random insertion and one random deletion (noted insertion/deletion) of nucleotides are applied to each sequence of S. Two processes are considered:
- The (random) mononucleotide insertion/deletion process is 1 mononucleotide (letter) insertion and 1 mononucleotide deletion per sequence per step.
- The (random) trinucleotide insertion/deletion process is 1 trinucleotide (word of 3 letters) insertion and 1 trinucleotide deletion per sequence per step.
The location of insertions and deletions in the sequence and the type (R or Y) of the inserted bases are random (no hypothesis was taken).

For a step k, **the statistical function $i \rightarrow p_i(S)$, the same as defined in the part 1 method 2.1, is represented as a curve** $C_k(S)$. The curve $C_k(S)$ is given for a step k but is in fact, representative of a step range.

## 3.3. Results

3.3.1. *Simulation of the Periodicity P3 in the Range [1,98] with the Maximal Value $p_6(F)$ ($YRY(N)_6YRY$ Preferential Occurrence) by Using the $(YRY(N)_6)^*$ Sequence Associated with the Trinucleotide Insertion/deletion Process*

A simulated population S, having 500 sequences $(YRY(N)_6)^*$ of 2000 base length, is generated by random specification of the $(N)_6$ bases with an R percentage of 58.33% and with an Y percentage of 41.66% (step 0) (in order to have the same percentages of R and Y in the sequences).

(a) Curve $C_0(S)$ at Step 0: Fig. 2(a)

Before the insertion/deletion process, the curve $C_0(S)$ is made of horizontal points with abscissa invariant modulo 9 (due to the invariance modulo 9 of the $(YRY(N)_6)^*$ sequence). It is important to stress that a significant modulo 9 periodicity was already found in eukaryotic protein coding genes [1].
There are five such different horizontal lines $D_1$ , $D_2$ , $D_3$ , $D_4$ and $D_5$ in decreasing ordinate:

$D_1$: points $(i,p_i(S))$ with $i\equiv 6[9]$
$D_2$: points $(i,p_i(S))$ with $i\equiv 4,8[9]$
$D_3$: points $(i,p_i(S))$ with $i\equiv 1,2[9]$
$D_4$: points $(i,p_i(S))$ with $i\equiv 0,3[9]$
$D_5$: points $(i,p_i(S))$ with $i\equiv 5,7[9]$
This decomposition can be explained by a trivial statistical calculus.

There is no $YRY(N)_6YRY$ preferential occurrence because the point $(6,p_6(S))$ on the highest line $D_1$ cannot be differentiated from the other points $(i,p_i(S))$ with $i\equiv 6[9]$: $(15,p_{15}(S))$, $(24,p_{24}(S))$, etc. There is no periodicity P3 because the points $(i,p_i(S))$ with $i\equiv 0,3[9]$ are on the second lowest line $D_4$ .

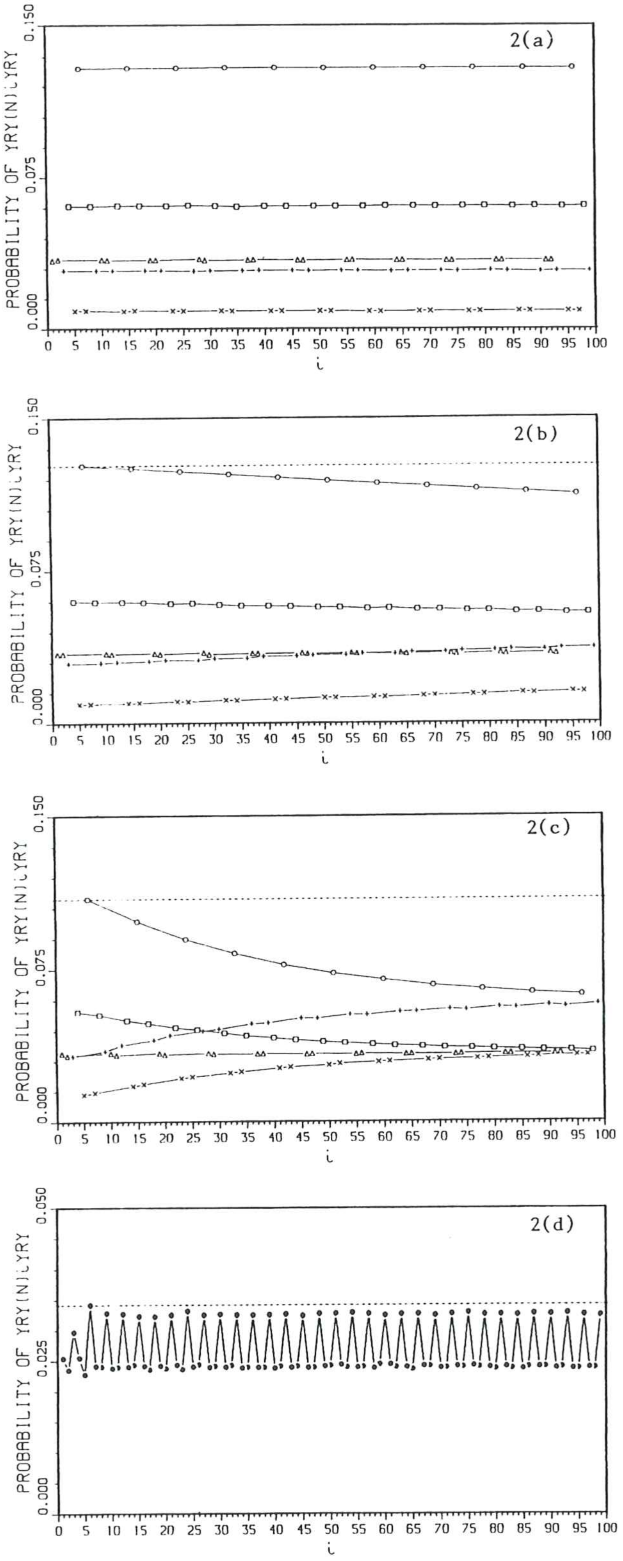


Fig. 2. Trinucleotide insertion/deletion process in the $(YRY(N)_6)^*$ sequence (1 trinucleotide insertion and 1 trinucleotide deletion per sequence per step). The horizontal axis represents the number i of bases N in the i-motif $YRY(N)_iYRY$, with i in the range [1,99]. The vertical axis represents the mean occurrence probability $p_i(S)$ over all the $(YRY(N)_6)^*$ sequences in the simulated population S. A horizontal dashed line goes through the point $(6,p_6(S))$. This statistical function $i \rightarrow p_i(S)$ is constituted of five curves (see part 3 method) shown at the following process steps: (a) Step 0, (b) Step 1, (c) Step 10, (d) Step 150 with one curve.

(b) Curve $C_1(S)$ at Step 1: Fig. 2(b)

There is the $YRY(N)_6YRY$ preferential occurrence. Indeed, the point $(6,p_6(S))$ has the highest ordinate because one trinucleotide insertion and/or one trinucleotide deletion destroy only one subsequence $YRY(N)_6YRY$, but two subsequences $YRY(N)_{15}YRY$, three subsequences $YRY(N)_{24}YRY$, etc. The values $p_i(S)$ with $i \equiv 6[9]$ decrease all the more since i increases. At this step, there is no periodicity P3.

(c) Next Curves at Steps 10 and 150: Fig. 2(c-d)

By increasing the number of steps, the five lines $D_1$ , $D_2$ , $D_3$ , $D_4$ and $D_5$ are gathered into two curves. The top curve is constituted of the two lines $D_1$ and $D_4$ , i.e. of the points $(i,p_i(S))$ with $i \equiv 0[3]$. The bottom curve is constituted of the three lines $D_2$ , $D_3$ and $D_5$ , i.e. of the points $(i,p_i(S))$ with $i \equiv 1,2[3]$. This process leads to the $YRY(N)_6YRY$ preferential occurrence with the periodicity P3.

Furthermore, the simulated curve $C_{150}(S)$ (Fig. 2(d) in which the points are joined in one curve) is strongly similar to the real curve C(CEUK) of eukaryotic protein coding genes (Fig. 1(a)) because in both cases there are:

- the $YRY(N)_6YRY$ preferential occurrence with the periodicity P3,
- two different sets of well separated points,
- for each set, the points can be joined by a nearly horizontal line.

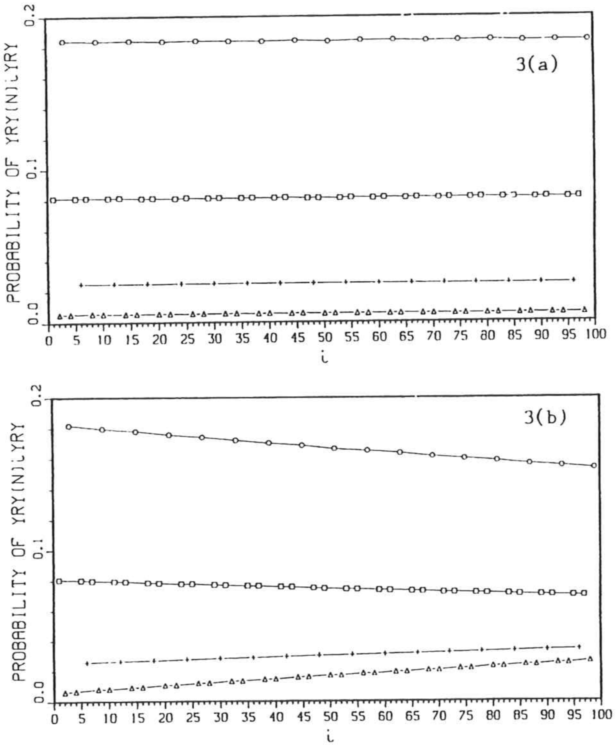
### 3.3.2. *Simulation of the Periodicity P2 in the Range [1,23] with the Maximal Value $p_3(F)$ by Using the $(YRY(N)_3)^*$ Sequence Associated with the Mononucleotide Insertion/deletion Process*

A simulated population S', having 500 sequences $(YRY(N)_3)^*$ of 2000 base length, is generated by random specification of the $(N)_3$ bases with an R percentage of 66.66% and with a Y percentage of 33.33% (step 0) (in order to have the same percentages of R and Y in the sequences).

(a) Curve $C_0(S')$ at Step 0: Fig. 3(a)

Before the insertion/deletion process, the curve $C_0(S')$ is constituted of four horizontal lines $\Delta_1$ , $\Delta_2$ , $\Delta_3$ and $\Delta_4$ of points in decreasing ordinate (proof identical to the one for the curve $C_0(S)$):

$\Delta_1$: points $(i,p_i(S'))$ with $i \equiv 3[6]$
$\Delta_2$: points $(i,p_i(S'))$ with $i \equiv 1,5[6]$
$\Delta_3$: points $(i,p_i(S'))$ with $i \equiv 0[6]$
$\Delta_4$: points $(i,p_i(S'))$ with $i \equiv 2,4[6]$



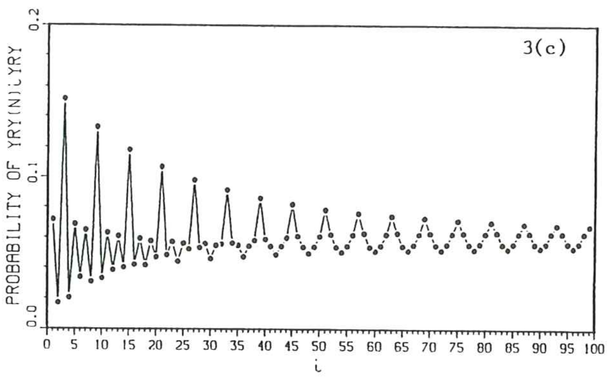


Fig. 3. Mononucleotide insertion/deletion process in the $(YRY(N)_3)^*$ sequence (one mononucleotide insertion and one mononucleotide deletion per sequence per step). The horizontal axis represents the number i of bases N in the i-motif $YRY(N)_iYRY$, with i in the range [1,99]. The vertical axis represents the mean occurrence probability $p_i(S')$ over all the $(YRY(N)_3)^*$ sequences in the simulated population S'. This statistical function $i \rightarrow p_i(S')$ is constituted of four curves (see part 3 method) shown at the following process steps: (a) Step 0, (b) Step 1, (c) Step 15 with one curve.

(b) Next Curves at Steps 1 and 15: Fig. 3(b-c)

By increasing the number of steps, the two lines $\Delta_1$ and $\Delta_2$ become curves with decreasing slope, while the two lines $\Delta_3$ and $\Delta_4$ , curves with increasing slope. This process leads to the periodicity P2.

Furthermore, the simulated curve $C_{15}(S')$ (Fig. 3(c) in which the points are joined in one curve) is strongly similar to the real curve C(N5EUK) of 5' eukaryotic regions (Fig. 1(b)) because in both cases there are:

- the periodicity P2 in the range [1,23],
- the highest value $p_i(S')$ at i=3 (proof identical to the one given in the part 2 results 3.1.2 for the $YRY(N)_6YRY$ preferential occurrence),
- four obvious sets of points which 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=6, 12, 18 and 24 and (4) i=2, 4, 8, 10, 14 and 16. All these naturally appearing curves join modulo 6 periodic sets of i values (Fig. 1(b) and 3(c)) and are deduced from the four lines $\Delta_1$ , $\Delta_2$ , $\Delta_3$ and $\Delta_4$ .

## 3.4. Discussion

### 3.4.1. *Properties of this Model*

This model is **simple** because it uses a unique type of sequence ($(YRY(N)_6)^*$ or $(YRY(N)_3)^*$) associated with an evolutionary process based on trivial operations such as random insertions and deletions of mono(di,tri)nucleotides. Therefore, it is surprising that such a simple model not only simulates the two periodicities P3 and P2 with their respective maximal values found in the DNA word, but also leads to a strong similarity with the real situation concerning the eukaryotic protein coding genes and the 5' eukaryotic regions.

**The maximal value $p_k(F)$ in C(F) is specifically associated with the oligonucleotide $YRY(N)_k$ in series $(YRY(N)_k)^*$** and is independent of the type of the insertion/deletion process (proof identical to the one given in the part 2 results 3.1.2 for the maximal value $p_6(F)$ in the study of $(YRY(N)_6)^*$ or for the maximal value $p_3(F)$ in the study of $(YRY(N)_3)^*$). Consequently, a nonspecific sequence or a random sequence cannot explain the maximal value $p_k(F)$.

Furthermore, the $YRY(N)_6YRY$ preferential occurrence seems only to be explained by the oligonucleotide $YRY(N)_6$ and by the sequence $(YRY(N)_6)^*$ (several hundreds of models tested in this field did not lead to a contradiction). The sequence $(YRY(N)_6)^*$ being chosen, any insertion/deletion process (of mononucleotides, or dinucleotides, or trinucleotides) leads to the $YRY(N)_6YRY$ preferential occurrence (results 3.3.1 and data not shown).

**The periodicities in C(F) are specifically associated with the insertion/deletion process**, but are less dependent of the oligonucleotide: The periodicity P3 is explained by the insertion/deletion of trinucleotides, but not by the insertion/deletion of mono(di)nucleotides (part 2 results 3.1 and data not shown). The insertion/deletion of trinucleotides being chosen, the periodicity P3 can be obtained from the sequence $(YRY(N)_6)^*$ (part 2 results 3.1) or with the sequence (RNY)* [5] for example (data not shown).

The periodicity P2, alternating purine/pyrimidine periodicity, can be obtained from a sequence having initially series of trinucleotides (results 3.3.2).

**A mutation process (random changes R→Y and Y→R) is necessary**. The absolute values of $p_i(S)$ in the simulated populations S are greater (up to 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 specified base (R or Y) and with any mutation rate per unspecified base N (Arquès and Michel, mutation model in preparation).

The mutation process acts on the absolute values of $p_i(S)$ by decreasing the variations between points $(i,p_i(S))$ (by adding "noise") but not on the relative positions of the points $(i,p_i(S))$. Consequently, the mutation process can explain neither the periodicities P3 and P2 nor a maximal value.

### 3.4.2. *Biological Meaning of this Model*

One possible hypothesis is that DNA sequence evolution on the **two-letter alphabet {R,Y}** is constituted of **two successive steps** (in first approximation):

| | Step 1:<br>Insertion/Deletion | | Step 2:<br>Mutation | |
|---|---|---|---|---|
| Primitive Oligonucleotides<br>(concatenated in series) | —————> | Primitive Sequences | ————> | Actual Sequences |

(1) The first step led to **primitive sequences** from a few (i.e. preferential type) **primitive oligonucleotides in series** mainly through an insertion/deletion process. This insertion/deletion process must have been **random** (due to the absence of any clever process at the primitive stage).

(2) At a later stage, the second step led to the **actual sequences** from the primitive sequences mainly through the **mutation process**, the transformation of the alphabet {R,Y} into {A,C,G,T} (A=adenine, C=cytosine, G=guanine, T=thymine), 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 gene populations even if mutations have introduced an important "noise" effect. Part 2 has identified common statistical features (in particular, the $YRY(N)_6YRY$ preferential occurrence is almost universal; N=R or Y). These features were explained in part 3 by two oligonucleotides associated with an insertion/deletion process.

## REFERENCES

[1] D.G. Arquès and C.J. Michel, "Study of a Pertubation in the Coding Periodicity." *Math. Biosc.*, *86*, 1987, pp. 1-14.

[2] D.G. Arquès and C.J. Michel, "A Purine-Pyrimidine Motif Verifying an Identical Presence in almost all Gene Taxonomic Groups." *J. Theor. Biol.*, *128*, 1987, pp. 457-461.

[3] D.G. Arquès and C.J. Michel, "Periodicities in Introns." *Nucl. Acids Res.*, *15*, 1987, pp. 7581-7592.

[4] D.G. Arquès and C.J. Michel, "Periodicities in Coding and Noncoding Regions of the Genes." *J. Theor. Biol.*, *143*, 1990, pp. 307-318.

[5] M. Eigen and P. Schuster, "The Hypercycle. A Principle of Natural Self-Organization. Part C: The Realistic Hypercycle." *Naturwissenschaften*, *65*, 1978, pp. 341-369.

[6] J.W. Fickett, "Recognition of Protein Coding Regions in DNA Sequences." *Nucl. Acids Res.*, *10*, 1982, pp. 5303-5318.

[7] M. Kimura, *The neutral theory of molecular evolution.* (Cambridge: Cambridge University Press, 1987).

[8] 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. Natl. Acad. Sci. USA*, *78*, 1981, pp. 1596-1600.