Identification of protein coding genes in genomes with statistical functions based on the circular code

Didier G. Arquès\textsuperscript{a,1}, Jérôme Lacan\textsuperscript{b,2}, Christian J. Michel\textsuperscript{c,*}

\textsuperscript{a} Equipe de Biologie Théorique, Institut Gaspard Monge, Université de Marne la Vallée, 2 rue de la Butte Verte 93160 Noisy le Grand, France
\textsuperscript{b} Département de Mathématiques Appliquées et d’Informatique, ENSICA, 1 Place Emile Blouin, 31056 Toulouse Cedex 5, France
\textsuperscript{c} Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France

Received 23 May 2001; received in revised form 3 April 2002; accepted 30 May 2002

Abstract

A new statistical approach using functions based on the circular code classifies correctly more than 93\% of bases in protein (coding) genes and non-coding genes of human sequences. Based on this statistical study, a research software called ‘Analysis of Coding Genes’ (ACG) has been developed for identifying protein genes in the genomes and for determining their frame. Furthermore, the software ACG also allows an evaluation of the length of protein genes, their position in the genome, their relative position between themselves, and the prediction of internal frames in protein genes. © 2002 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Protein coding genes; Genomes; Circular code; Statistical functions; Research software

1. Introduction

The concept of code ‘without commas’, introduced by Crick et al. (1957) for the protein (coding) genes, is a code readable in only one out of three frames. Such a theoretical code without commas, called circular code in the theory of codes (e.g. Béal, 1993; Berstel and Perrin, 1985), is a particular set \( X \) of trinucleotides such that a concatenation (a series) of trinucleotides of \( X \) leads to sequences that cannot be decomposed in another frame with a concatenation of trinucleotides of \( X \).

For example, suppose that \( X \) is the following set of trinucleotides: \( X = \{ \text{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC} \} \). Some trinucleotides of \( X \) are concatenated randomly, for example as follows:
If the set $X$ of trinucleotides is a circular code, then there is an unique solution:

This unique solution is obtained by choosing a window (sufficiently large) in any position in the sequence and then verifying the belonging of the trinucleotides of the window to $X$:

- $\ldots$ CAG, GCC, TTC, AAT, ACC, ACC, CAG, GAA, GAG, GTA, ATT, ACC, AAT, GTA, AAC, TAC, TTC, ACC, ATC $\ldots$

The commas between the trinucleotides show the frame of construction (reading frame in biology). Suppose now that the commas are ‘lost’, leading to the sequence:

- $\ldots$ CAGGCCCTTCAATACCACCCAGGAAGAGGTAAATTACCAATGTAAACTACTTACCTACCA

The problem is to retrieve the original frame of construction. There are three obvious possibilities:

- $\ldots$ C, AGG, CCT, TCA, ATA, CCA, CCC, AGG, AAG, AGG, TAA, TTA, CCA, ATG, TAA, ACT, ACT, TCA, CCA, TC $\ldots$
- $\ldots$ CA, GGC, CTT, CAA, TAC, CAC, CCA, GGA, AGA, GGT, AAT, TAC, CAA, TGT, AAA, CTA, CTT, CAC, CAT, C $\ldots$
- $\ldots$ CAG, GCC, TTC, AAT, ACC, ACC, CAG, GAA, GAG, GTA, ATT, ACC, AAT, GTA, AAC, TAC, TTC, ACC, ATC $\ldots$

The first decomposition proposed is rejected immediately as the first trinucleotide AGG in the window does not belong to $X$. The second decomposition proposed is rejected with a window of 13 nucleotides. Indeed, the first nucleotide A in the window may belong to several trinucleotides of $X$, e.g. GTA. The trinucleotides GGT, AAT, and

- $\ldots$ C, AGG, CCT, TCA, ATA, CCA, CCC, AGG, AAG, AGG, TAA, TTA, CCA, ATG, TAA, ACT, ACT, TCA, CCA, TC $\ldots$
- $\ldots$ CA, GGC, CTT, CAA, TAC, CAC, CCA, GGA, AGA, GGT, AAT, TAC, CAA, TGT, AAA, CTA, CTT, CAC, CAT, C $\ldots$
- $\ldots$ CAG, GCC, TTC, AAT, ACC, ACC, CAG, GAA, GAG, GTA, ATT, ACC, AAT, GTA, AAC, TAC, TTC, ACC, ATC $\ldots$

This unique solution is obtained by choosing a window (sufficiently large) in any position in the sequence and then verifying the belonging of the trinucleotides of the window to $X$:

- $\ldots$ CAGGCCCTTCAATACCACCCAGGAAGAGGTAAATTACCAATGTAAACTACTTACCTACCA

The commas between the trinucleotides show the frame of construction (reading frame in biology).
TAC following A belong to $X$. The next trinucleotide CAA does not belong to $X$ as the 13th nucleotide A (from the beginning of the window) differs from the unique possibility G of CAG belonging to $X$. The third decomposition is the original one as all the trinucleotides in the window belong to $X$ and the original decomposition of the sequence is deduced automatically.

Such a code was proposed by Crick et al. (1957) in order to explain how the reading of a series of nucleotides in the protein genes could code for the amino acids constituting the proteins. The two problems stressed were: why are there more trinucleotides than amino acids and how to choose the correct reading frame? Crick et al. (1957) proposed that only 20 among 64 trinucleotides code for the 20 amino acids. However, the determination of a set of 20 trinucleotides forming a circular code $X$ depends on a great number of constraints:

i) A trinucleotide with identical nucleotides (AAA, CCC, GGG or TTT) must be excluded from such a code. Indeed, the concatenation of AAA with itself does not allow the retrieval of the reading (original) frame as there are three possible decompositions:

...AAA,AAA,AAA,...
...A,AAA,AAA,AA...
and...AA,AAA,AAA,A...

ii) Two trinucleotides related to circular permutation, e.g. ATC and TCA, must be excluded from such a code. Indeed, the concatenation of ATC with itself does not allow the retrieval of the reading (original) frame as there are two possible decompositions: ...

...ATC,ATC,ATC,... and ...

Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides so that, in each class, the three trinucleotides are deduced from each other by circular permutations, e.g. ATC, TCA, and CAT, a circular code, has only one trinucleotide per class and, therefore, contains at most 20 trinucleotides (maximal circular code). This trinucleotide number is identical to the amino acid number suggesting a circular code assigning one trinucleotide per amino acid.

No set of 20 trinucleotides leading to a circular code has been found at this time. Furthermore, the two discoveries that the trinucleotide TTT, an ‘excluded’ trinucleotide in the concept of circular code, codes for phenylalanine (Nirenberg and Matthaei, 1961) and that the protein genes are placed in the reading frame with a particular trinucleotide, namely the start trinucleotide ATG, have led to giving up the concept of circular code on the alphabet \{A,C,G,T\}. For several biological reasons, in particular the interaction between mRNA and tRNA, the concept of circular code has been resumed subsequently regarding the alphabet \{R,Y\} (R = purine = A or G, Y = pyrimidine = C or T) with two trinucleotide models for the primitive protein genes: RRY (Crick et al., 1976) and RNY (N = R or Y) (Eigen and Schuster, 1978).

Unexpectedly, a maximal circular code has been identified recently in the protein genes of both eukaryotes and prokaryotes on the alphabet \{A, C, G, T\} (Arqué’s and Michel, 1996). This circular code has been obtained by two methods:

i) by computing the occurrence frequencies of the 64 trinucleotides AAA,...,TTT in the three frames of protein genes and then, by assigning each trinucleotide to the frame associated with its highest frequency (Arqué’s and Michel, 1996);

ii) by computing the 12 288 (3 $\times$ 64$^2$) autocorrelation functions analysing the probability that a trinucleotide in any frame occurs any $i$ bases N after a trinucleotide in a given frame of protein genes and then, by classifying these autocorrelation functions according to their modulo 3 periodicity for deducing a frame for each trinucleotide (Arqué’s and Michel, 1997a).

The maximal circular code identified is the set $X_0 = \{AAC,AAT,ACC,ATC,ATT, CAG,CTC,CTG,GAA,GAC, GAG,GAT,GCC, GCC,GGT,GTA,GTC,GTT,TAC,TTC\}$ of 20 tri-
### Table 1

List per frame and in lexicographical order of the trinucleotides of the complementary circular code identified in protein coding genes of eukaryotes and prokaryotes (Arques and Michel, 1996)

<table>
<thead>
<tr>
<th>Frame 0</th>
<th>AAA AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame 1</td>
<td>AAG ACA ACG ACT AGG ATA ATG CCA CCC CCG GCG GTG TAG TCA TCC TCG TCT TGC TTA TTG</td>
</tr>
<tr>
<td>Frame 2</td>
<td>AGA AGT CAA CAC CAT CCT CGA CGC CGG CGT CTA CTT GCA GCT GGA GGG TAA TAT TGA TGG TGT</td>
</tr>
</tbody>
</table>

**Circularity property with the three circular codes $X_0$, $X_1$, and $X_2$ of 20 trinucleotides identified in protein coding genes of eukaryotes and prokaryotes**

<table>
<thead>
<tr>
<th>Frame 0</th>
<th>AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame 1</td>
<td>ACA ATA CCA TCA TTA AGC TCC TGC AAG ACG AGG ATG CCG GCG GTG TAG TCG TTG ACT TCT</td>
</tr>
<tr>
<td>Frame 2</td>
<td>CAA TAA CAC CAT TAT GCA CCT GCT AGA CGA GGA TGA CGC CGG TGG AGT CGT TGT CTA CTT</td>
</tr>
</tbody>
</table>

**Complementarity property with the three circular codes $X_0$, $X_1$, and $X_2$ of 20 trinucleotides identified in protein coding genes of eukaryotes and prokaryotes. This property is verified with $T_0$ (AAA and TTT) and $T_1$ and $T_2$ (CCC and GGG)**

<table>
<thead>
<tr>
<th>Frame 0</th>
<th>AAA AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame 1</td>
<td>AAG ACA ACG ACT AGG ATA ATG CCA CCC CCG GCG GTG TAG TCA TCC TCG TCT TGC TTA TTG</td>
</tr>
<tr>
<td>Frame 2</td>
<td>CTT TGT CGT AGT GCT CTC TAT CAT TGG GGG CGG CGC CAC CTA TGA GGA CGA AGA GCA TAA CAA</td>
</tr>
</tbody>
</table>

Three subsets of trinucleotides can be identified: $T_0 = X_0 \cup \{AAA, TTT\}$ in frame 0, $T_1 = X_1 \cup \{CCC\}$ in frame 1 and $T_2 = X_2 \cup \{GGG\}$ in frame 2. The three sets $X_0$, $X_1$, and $X_2$ of 20 trinucleotides are maximal circular codes.
nucleotides in frame 0 of protein genes (reading frame). Furthermore, the two sets $X_1$ and $X_2$ of 20 trinucleotides identified in the frames 1 and 2, respectively, (frames 1 and 2 being the frame 0

misplaced trinucleotides in the shifted frames is equal to 24.6%. If the trinucleotides of $X$ are concatenated randomly, for example as follows:

...GAA,GAG,GT A,GT A,ACC,AAT,GT A,CT C,TA C,T T C,ACC,ATC...
then, the trinucleotides in frame 1:
...G,AAG,AGG,TA A,CCA,ATG,TA C,TCT,ACT,TCA,CCA,TC...
and the trinucleotides in frame 2:
...GA,AGA,GG T,AAC,CAA,TGT,ACT,CT A,CT T,CAC,CAT,C...

shifted by one and two nucleotides respectively in the 5′–3′ direction) by these two methods, are also maximal circular codes (Table 1). These three circular codes have several important properties:

i) circularity: $X_0$ generates $X_1$ by one circular permutation and $X_2$ by another circular permutation (one and two circular permutations of each trinucleotide of $X_0$ lead to the trinucleotides of $X_1$ and $X_2$ respectively) (Table 1).

ii) complementarity: $X_0$ is self-complementary (ten trinucleotides of $X_0$ are complementary to the ten other trinucleotides of $X_0$) and, $X_1$ and $X_2$ are complementary to each other (the 20 trinucleotides of $X_1$ are complementary to the 20 trinucleotides of $X_2$) (Table 1). Note that this property is also verified with $T_0 = X_0 \cup \{AAA,TTT\}$, $T_1 = X_1 \cup \{CCC\}$ and $T_2 = X_2 \cup \{GGG\}$ (Table 1).

iii) rarity: the occurrence probability of $X_0$ is equal to $6 \times 10^{-8}$. As there are 20 classes of three trinucleotides (see above), the number of potential circular codes is $3^{20} = 3486784401$. The computed number of complementary circular codes with two shifted circular codes (called $C^3$ codes), such as $X_0$, is 216. Therefore, its probability is $216/3^{20} = 6 \times 10^{-8}$.

iv) flexibility: the lengths of the minimal windows to automatically retrieve the frames 0, 1, and 2 with the three circular codes $X_0$, $X_1$, and $X_2$ respectively, are all equal to 13 nucleotides and represent the largest window length among the 216 $C^3$ codes. The frequency of

belong mainly to $X_1$ and $X_2$, respectively. A few trinucleotides are misplaced in the shifted frames. With this example, in frame 1, nine trinucleotides belong to $X_1$, one trinucleotide (TAC) to $X_0$ and one trinucleotide (TAA) to $X_2$. In frame 2, eight trinucleotides belong to $X_2$, two trinucleotides (GGT, AAC) to $X_0$ and one trinucleotide (ACT) to $X_1$. By computing exactly, the average frequencies of misplaced trinucleotides in frame 1 are 11.9 for $X_0$ and 12.7% for $X_2$. In frame 2, the average frequencies of misplaced trinucleotides are 11.9 for $X_0$ and 12.7% for $X_1$. The complementarity property explains on the one hand that the frequency equality of $X_0$ in frames 1 and 2 and on the other hand, the frequency equality of $X_2$ in frame 1 and $X_1$ in frame 2. The sum of percentages of misplaced trinucleotides in frame 1 ($X_0$ and $X_2$) is equal to the sum of percentages of misplaced trinucleotides in frame 2 ($X_0$ and $X_1$) and is equal to 24.6%. This value is close to the highest frequency (27.9%) of misplaced trinucleotides among the 216 $C^3$ codes. The four types of nucleotides occur in the three trinucleotide sites with the three circular codes $X_0$, $X_1$, and $X_2$ (Table 1).

v) evolutionary: an evolutionary analytical model at three parameters ($p$, $q$, $t$) based on an independent mixing of the 20 trinucleotides of $X_0$ with equiprobability (1/20) followed by $t \approx 4$ substitutions per trinucleotide according to the proportions $p \approx 0.1$, $q \approx 0.1$ and $r = 1 - p - q \approx 0.8$ in the three trinucleotide sites,
respectively, retrieves the frequencies of \(X_0\), \(X_1\), and \(X_2\) observed in the three frames of protein genes.

The proof that \(X_0\), \(X_1\), and \(X_2\) are circular codes, the detailed explanation of the properties (i–iv) and the different biological consequences, in particular on the two-letter genetic alphabets, the genetic code and the amino acid frequencies in proteins, are given in Arqué’s and Michel (1996, 1997a). The property (v) is described in Arqué’s et al. (1998, 1999).

Note: a non-complementary circular code has been identified recently in the mitochondrial protein genes (Arqué’s and Michel, 1997b).

As the circular code is a strong structural property of protein genes, different statistical functions based on the circular code are investigated in this paper in order to discriminate between coding and non-coding genes. Indeed, the sets of 20 trinucleotides based on a circular code, i.e. the 216 \(C^3\) codes and in particular \(X_0\), \(X_1\), and \(X_2\), have a lesser number of misplaced trinucleotides in the shifted frames compared with the vast majority of sets without particular property. This low number implies that the three circular codes \(X_0\), \(X_1\), and \(X_2\) can clearly be associated with the three frames 0, 1, and 2, respectively, (detailed in method).

After having validated this statistical approach with the human sequences from the EMBL database, research software has been developed for identifying protein genes in genomes and for determining their frame. Furthermore, this software also allows an evaluation of the length of protein genes, their position in the genome, their relative position between themselves, and the prediction of internal frames. These possibilities are presented with five examples taken from human chromosomes: a large protein gene, a complementary protein gene, a series of five exons, a protein gene with four internal frames, and a possible coding region in the human DNA sequence. An example with a prokaryotic genome is also given.

2. Method

2.1. Introduction

The method developed is based on a strong structural property of protein genes, i.e. the circular code, and in particular its properties of circularity and complementarity. This method differs from the classical methods, such as the codon usage methods and the HMM methods, at least for the following reasons:

i) The circular code is observed in protein genes of eukaryotes as well as of prokaryotes and is not found in the non-coding genes (Arqué’s et al., 1998). Therefore, a method based on this circular code can be applied independently of the type of eukaryotic/prokaryotic organism under investigation. In contrast, the codon usage methods use codon frequencies that depend on the species and the functional classes of protein genes (see e.g. Karlin et al., 1998).

<table>
<thead>
<tr>
<th>Codon in frame 0</th>
<th>Frequency (%)</th>
<th>Codon in frame 1</th>
<th>Frequency (%)</th>
<th>Codon in frame 2</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATG</td>
<td>2.31</td>
<td>ATG</td>
<td>3.08</td>
<td>ATG</td>
<td>0.57</td>
</tr>
<tr>
<td>CAA</td>
<td>1.65</td>
<td>CAA</td>
<td>1.55</td>
<td>CAA</td>
<td>3.71</td>
</tr>
<tr>
<td>CCA</td>
<td>1.66</td>
<td>CCA</td>
<td>2.91</td>
<td>CCA</td>
<td>2.04</td>
</tr>
<tr>
<td>GGA</td>
<td>1.76</td>
<td>GGA</td>
<td>1.27</td>
<td>GGA</td>
<td>3.49</td>
</tr>
<tr>
<td>GTC</td>
<td>1.60</td>
<td>GTC</td>
<td>0.81</td>
<td>GTC</td>
<td>1.05</td>
</tr>
<tr>
<td>GTT</td>
<td>1.55</td>
<td>GTT</td>
<td>0.75</td>
<td>GTT</td>
<td>1.35</td>
</tr>
<tr>
<td>TCC</td>
<td>1.63</td>
<td>TCC</td>
<td>1.85</td>
<td>TCC</td>
<td>1.40</td>
</tr>
</tbody>
</table>

Table 2
A few examples taken from the Table 1 of Arqué’s and Michel (1996) showing that the codons GTC and GTT belonging to \(X_0\) occur in frame 0 with lower frequencies compared with the codons ATG belonging to \(X_1\) and CAA belonging to \(X_2\), etc.
ii) The circular code $X_0$ (respectively, $X_1$ and $X_2$) contains the 20 codons having a preferential occurrence in the frame 0 (respectively, 1 and 2). It is important to stress that the set $X_0$, for example, does not necessarily represent the common codons in frame 0, i.e., the 20 codons having the highest frequencies in frame 0 (see a few examples in Table 2).

iii) The $216 \cdot C^3$ codes have a low number of misplaced trinucleotides in the shifted frames, 27.9% in the worst case and 24.6% for $X_0$. This number is close to $2/3 \approx 66.6\%$ with the vast majority of trinucleotide sets without particular property. Indeed, by excluding AAA, CCC, GGG, and TTT, there is one chance out of three to observe, for example, a codon of $X_1$ in frame 1, i.e., two chances out of three to observe a codon of $X_0$ or $X_2$ in frame 1. In summary, the method developed according to the circular code allows to associate clearly the three sets of trinucleotides $X_0$, $X_1$, and $X_2$ with the three frames 0, 1 and 2 respectively of protein genes.

iv) The complementarity property of these three sets $X_0$, $X_1$, and $X_2$ is used for identifying protein genes on the direct strand but also on the complementary strand (see the definition of the four functions below).

v) The method developed is based on the global probabilities of $X_0$, $X_1$, and $X_2$ and not on the individual codon probabilities that are used in the codon usage methods.

2.2. Definition of statistical functions

Let $t$ be a trinucleotide in the set \{AAA, ..., TTT\} (64 trinucleotides). Let $F$ be a population with $m(F)$ sequences $S$. Each sequence $S$ has a base length $l(S)$. Let $w_i$ be a window of $n$ trinucleotides starting at the base position $i$, $i = 1, \ldots, l(S) - 3n + 1$, in a sequence $S$ of $F$, i.e., $w_i = t_i \ldots t_n$ where $t_j$ is the $j$th trinucleotide in the window $w_i$. Let $T_g$, $g \in \{0, 1, 2\}$, be the three subsets of trinucleotides constituting the three circular codes in the protein coding genes of eukaryotes and prokaryotes, $T_0$ in the open reading frame (frame 0) and, $T_1$ and $T_2$, in the shifted frames 1 and 2, respectively (Table 1). In a given window $w_i$, the function

$$\delta_g(t_j) = \begin{cases} 1 & \text{if } t_j \in T_g \\ 0 & \text{if } t_j \notin T_g \end{cases}$$

determines whether or not if the trinucleotide $t_j$ at the position $j$ in $w_i$ belongs to $T_g$ with $g \in \{0, 1, 2\}$. Next, the occurrence frequency $P(T_g, w_i)$ of a subset $T_g$ in $w_i$, is $P(T_g, w_i) = \Sigma_{i=1}^{n} \delta_g(t_j)/n$ where $n$ is the total number of trinucleotides in the window $w_i$.

Several statistical functions based on the properties of the circular code, are defined:

$$F_1(i) = P(T_0, w_i)$$

$$F_2(i) = P(T_0, w_i) - P(T_2, w_i)$$

$$F_3(i) = \frac{(21/22)2P(T_0, w_i)}{P(T_1, w_i) + P(T_2, w_i))}$$

$$F_4(i) = \sum_{j=0}^{2} P(T_j, w_{i+j})$$

These four statistical functions use different properties of the circular code, in particular the properties of circularity and complementarity.

The function $F_1$ is the simplest, and is based on the circular code $X_0$ (extended to $T_0$) in each window $w_i$. In a protein gene, $F_1(i)$ associated with the reading frame of the sequence (i.e., $w_i$ in reading frame and, therefore, $w_{i+1}$ and $w_{i+2}$ in the shifted frames 1 and 2, respectively) is in general greater than $F_1(i+1)$ and $F_1(i+2)$ as the occurrence probability of $T_0$ is by definition maximum in the reading frame (see point (ii) of Section 2.1 explaining the misplaced trinucleotides).

The function $F_2$ considers the two circular codes $X_0$ and $X_2$ (extended to $T_0$ and $T_2$). The probability difference $P(T_0, w_i) - P(T_2, w_i)$ is maximum among the 18 possible probability differences in the 3 frames. Indeed, the average probabilities of $T_0$, $T_1$, and $T_2$ in the frame 0 (respectively, 1, 2) of protein genes are 49% (respectively, 26.5%, 32%), 28.5% (respectively, 43%, 23%), and 22.5% (respectively, 30.5%, 45%) (Arqué`s et al., 1998). By consequence, the max-
imum probability difference in frame 0 (respectively, 1, 2) is 26.5% with \( \text{Prob}(T_0)/C_{28}/\text{Prob}(T_2) \) (respectively, 16.5% with \( \text{Prob}(T_1)/C_{28}/\text{Prob}(T_0) \), and 22% with \( \text{Prob}(T_2)/\text{Prob}(T_1) \)).

The functions \( F_3 \) and \( F_4 \) are based on the three circular codes \( X_0, X_1, \) and \( X_2 \) (extended to \( T_0, T_1, \) and \( T_2 \)). The function \( F_3 \) tests a ratio that is maximum in the reading frame. The functions \( F_1, F_2, \) and \( F_3 \) favor the circular code \( X_0 \) characterising the reading frame, while the function \( F_4 \) considers the three circular codes in their three associated frames.

Finally, as \( T_0 \) is self-complementary, and as \( T_1 \) and \( T_2 \) are complementary to each other, the four
functions have a similar behaviour on both strands.

These statistical functions are represented by curves as follows. As the value of $F_k(i)$ is a mean value computed from the bases of the window $w_i$, i.e. the bases $i, i+1, \ldots, i+3n-1$, the point associated with $F_k(i)$ is represented graphically in the abscissa by the base position $i + \lfloor 3n/2 \rfloor$ (greatest integer less or equal) and in the ordinate by $F_k(i)$. As the protein coding genes have three frames, a function $F_k$ is represented by three curves where the points are joined modulo 3: by convention, the curve $C_j$, $j = 0, 1, 2$, joins the points associated with $F_k(i)$ with $i = (j+1) \mod 3$, i.e. the curve $C_0$ (respectively, $C_1$, $C_2$) joins the points in base position 1 (respectively, 2, 0) modulo 3 (Fig. 1a) and, therefore, is related to the base position in frame 0 (respectively, 1, 2) determined from the beginning of the sequence.

The four functions defined above are extended to the trinucleotide concept as follows: the maximum of a given previous function $F_k$ in a series of three successive bases and the difference between the maximum and the second highest of a given previous function $F_k$ in a series of three successive bases.

Let $\tilde{t}_c$ be the $c^{th}$ trinucleotide in the genome sequence, i.e. constituted by the three bases in position $i$ such that $c = \lceil i/3 \rceil$ (smallest integer greater or equal). Then,

$$F^M_k(c) = \max_{j=0,1,2} F_k(3c - j)$$
$$F^D_k(c) = F_k(3c - j_0) - F_k(3c - j_1)$$

so that $j_0, j_1, j_2 \in \{0, 1, 2\}$ are defined by the inequality $F_k(3c - j_0) \geq F_k(3c - j_1) \geq F_k(3c - j_2)$.

Note: The two functions $F^M_k$ and $F^D_k$ always have values greater than or equal to 0.

The statistical significance of the two functions $F^M_k$ and $F^D_k$ is evaluated according to the parameter $s$ based on the discrete sum (called ‘surface’) of the values of a function in a given range $R$ of trinucleotides. The function $F^M_k$ (respectively, $F^D_k$) identifies the three bases of the trinucleotide $\tilde{t}_c$ as coding bases if $\sum_{r \in R} F^M_k(r) > s$ (respectively, $\sum_{r \in R} F^D_k(r) > s$) where $R$ is the greatest range containing $c$ such that $\forall r \in R$, $\max_{j=0,1,2} F_k(3r - j) = F_k(3r - j_0)$ where $j_0 \in \{0, 1, 2\}$ is constant. In order to visualise this concept, these two discrete sums $\sum_{r \in R} F^M_k(r)$ (respectively, $\sum_{r \in R} F^D_k(r)$) are represented in the Fig. 1b (respectively, Fig. 1c).

In summary, eight functions are analysed with the parameter $s$, with $F^M_k$ and $F^D_k$ by varying $k$ between 1 and 4. This statistical method is the main scientific part of the research software that is presented below.

2.3. Development of a research software called ACG

The main functionalities of the research software called ACG are the statistical analyses of...
different functions based on the circular code in sequence populations, the identification of protein genes in genomes, and the determination of their frame. Furthermore, several patterns of protein genes can be evaluated: their length, their position in the genome, their relative position between themselves, and the presence of internal frames. Several examples of these possibilities are given in the Section 3.

The software is written with three units: a sequence analysis unit, a statistical function unit and an interface unit.

The sequence analysis unit reads the sequences and computes the occurrence frequency \( P(T_g, w_j) \) in a window according to the algorithm described below. This unit calls the statistical function unit for computing a chosen function \( F_k^M \) or \( F_k^D \). Precisely, the four functions \( F_k \) and their trinucleotide evaluation \( F_k^M \) or \( F_k^D \) are implemented in this statistical function unit, which allows statistical numerical results on a sequence population \( F \) (eventually on one sequence). The interface unit allows the choice of different statistical parameters: the EMBL sequence file (population \( F \) or sequence \( S \)), the statistical function \( F_k^M \) or \( F_k^D \), the window length \( n \) in trinucleotides, and the statistical surface parameter \( s \). It also has a graphical functionality for displaying the graphical curves: the start base position in the sequence, the curve display window length in bases, the left/right scroll of a curve allowing to display a curve again, and a coloured curve associated with the frame for a direct interpretation. The curve display window can be printed on a broad range of printing devices. The statistical numerical results are stored in text files.

This structure in units easily allows modifications and extensions of the software ACG. ACG has been developed to be interactive and user-friendly. ACG is written in Pascal Delphi and implemented on IBM compatible microcomputers. It can be used without any computer knowledge.

The algorithm for computing the occurrence frequency \( P(T_g, w_j) \) is constructed such that the different bases in each sequence are read only one time.

A window \( w_i \) of \( n \) trinucleotides runs from the first \((i = 1)\) base to the base position \( i = (I(S) - 3n + \)
1 in the sequence. At each step of the algorithm, a base is read in the sequence and treated.

From the base position \( k = 1 \) to \( k = 3n + 2 \), the algorithm computes the values \( P(T_g, w_i) \) for \( g = 0, 1, 2 \) and \( i = 1, 2, 3 \).

From the base position \( k = 3n + 3 \) to \( l(S) \), the algorithm computes for \( i = k - 3n + 1 \), the values \( P(T_g, w_i) \) for \( g = 0, 1, 2 \), i.e. from \( i = 4 \) to \( l(S) - 3n + 1 \). The value \( P(T_g, w_i) \) is deduced from \( P(T_g, w_{i-3}) \). Indeed, the two windows \( w_i \) and \( w_{i-3} \) differ only from one trinucleotide (Fig. 2). Let the ‘destroyed trinucleotide’ \( t_d \) be the trinucleotide belonging to \( w_{i-3} \) but not to \( w_i \), i.e. the trinucleotide from the position \( i-3 \) to \( i-1 \). Similarly, let the ‘constructed trinucleotide’ \( t_c \) be the trinucleotide belonging to \( w_i \) but not to \( w_{i-3} \), i.e. the last trinucleotide in \( w_i \) that starts in position \( i + 3(n-1) \) and ends in position \( i + 3n - 1 \) (Fig. 2). Suppose that \( t_d \in T_g \) and \( t_c \in T_g \). If \( g \neq g' \), then \( P(T_g, w_i) = P(T_g, w_{i-3}) - 1/n \) and \( P(T_g, w_i) = P(T_g', w_{i-3}) + 1/n \). If \( g = g' \), then \( P(T_g, w_i) = P(T_g', w_{i-3}) \) for \( g'' = 0, 1, 2 \).

This algorithm is implemented with three words indexed from 0: two words \( d \) and \( c \) of length three associated with the destroyed and constructed trinucleotides respectively, and \( w \) of length \( 3n \), with the current window. At the step treating the \( k \)th base in the sequence, the base in the position \( k \) modulo \( 3n \) in the word \( w \) is moved in the position \( k \) modulo 3 in the word \( d \). Then, the \( k \)th base in the sequence is read and placed at the position \( k \) modulo 3 in the word \( c \) and at the position \( k \) modulo \( 3n \) in the word \( w \). In this way, the three words contain correctly the series of bases of the sequence which is read only one time.

Example of computation (Fig. 3): The subsequence of \( S \) that is analysed comprises the base positions between 9 and 19. The base length of the window \( w \) is chosen as \( 3n = 6 \). The proposed computation starts at the step \( k = 17 \), treating the 17th base in \( S \). The window \( w_i \) is in position \( i = k - 3n + 1 = 17 - 6 + 1 = 12 \) and then contains the bases \( w = \text{CCCCCC} \). The ‘destroyed trinucleotide’ \( t_d \) is the trinucleotide starting at the position \( i = 12 - 3 = 9 \): the word \( d \) contains \( d = \text{AAA} \). The ‘constructed trinucleotide’ \( t_c \) is the trinucleotide starting at the position \( i + 3(n-1) = 12 + 3 = 15 \): the word \( c \) contains \( c = \text{CCC} \). Note that at this step \( w = w_{12}, d = t_d, \) and \( c = t_c \). At the step \( k = 18 \) treating the 18th base in the sequence, the base in the position \( (k \mod 3n) = 0 \) in the word \( w \) is moved in the position \( (k \mod 3) = 0 \) in the word \( d \). Therefore, \( d = \text{CAA} \). Then, the 18th base in the sequence is read and placed at the position \( (k \mod 3) = 0 \) in the word \( c \) and at the position \( (k \mod 3n) = 0 \) in the word \( w \). Therefore, \( c = \text{GCC} \) and \( w = \text{GCCCCC} \). Note that at this step, \( w \neq w_{13}, d \neq t_d, \) and \( c \neq t_c \). The words \( w, d, \) and \( c \) can be obtained by circular permutations from \( w_i, t_d, \) and \( t_c \), respectively. This method avoids base shifting in the words. This process is reiterated at the next steps (see Fig. 3 for the step 19).

2.4. Data acquisition

The gene population \( F \) used for the statistical analysis is made of all human sequences (84222 sequences, 303 124 560 bases) obtained from release 57 (December 1998) of the EMBL Nucleotide Sequence Data Library. This large population leads to stable frequencies for the different functions analysed (law of large numbers). Therefore, these functions can be compared in order to identify the most interesting. The protein coding genes are extracted according to the keyword CDS without discarding particular sequences. In this population, 9.4% of bases are annotated as coding.

After the validation of the statistical approach, the research software aCG has been developed for identifying protein genes and used with the human chromosomes (Sanger Centre, March 1999).

3. Results

3.1. Statistical results

The different functions are evaluated with the software aCG according to the classical parameter Simple Matching Coefficient (SMC) (Bürset and Guigo, 1996), which considers the proportion of bases (according to the EMBL release) identified correctly by the function. Let \( \Sigma_{x \in P} l(S) = n_F \) be the total number of bases in the gene population \( F \). Let True Positives (TP) (respectively, True Nega-
that the total number of bases identified as coding (respectively, non-coding) bases by a function (defined above) in the coding (respectively, non-coding) genes in the gene population $F$. The coefficient SMC is then defined as $SMC = (TP + TN) / n_F$ (Burset and Guigoé, 1996).

The eight functions $F_k^M$ and $F_k^D$ defined above are analysed with the coefficient SMC. These
functions are evaluated with the parameter surface \( s \) between 6 and 130 with a step of 2. They are calculated with a window length \( n \) varying between 33 and 169 trinucleotides with a step of 17 trinucleotides. For each function, a maximum value of the coefficient SMC is obtained for given values of \( s \) and \( n \). The eight curves associated with the eight maximum values of the eight functions are represented in Fig. 4a and b by varying \( s \) for a given \( n \). Fig. 4a (respectively 4b) gives the four curves \( F_k^M \) (respectively, \( F_k^D \)).

The Fig. 4a and b show that the coefficient SMC is maximum with the function \( F_2^D \) with \( n/C_{30}/50 \) and \( s/C_{30}/38 \), which identifies 93.32% of bases correctly.

For the function giving the maximum value of the coefficient SMC (\( F_2^D \) with \( n/C_{50}/50 \) and \( s/C_{38}/38 \)), four other classical measures are computed, \( S_n \), \( S_p \), \( S_p' \), and CC, as follows (Burset and Guigó, 1996). Let False Positives (FP) (respectively, False Negatives (FN)) be the total number of bases identified as coding (respectively, non-coding) bases by a function (defined above) in the non-coding (respectively, coding) genes in the gene population \( F \). Note: \( TP + TN + FP + FN = n_F \).

The definitions and results of these four measures are:

i) The Sensitivity \( S_n \) is the proportion of coding bases identified correctly by the function:

\[
S_n = \frac{TP}{TP + FN} = 39.75\%
\]

ii) The Specificity (Sp) is the proportion of non-coding bases identified correctly by the function:

\[
Sp = \frac{TN}{TN + FP} = 98.88\%
\]

iii) Another definition of the Specificity \( S_p' \) is the proportion of coding bases among the bases identified as coding by the function:

\[
S_p' = \frac{TP}{TP + FP} = 78.71\%
\]

iv) The correlation coefficient CC is a measure of global accuracy where the value 1.00 corresponds to a perfect prediction and where the value 0.0 is expected for a random prediction:

\[
CC = \frac{(TP \times TN) - (FN \times FP)}{(TP + FN)(TN + FP)(TP + FP)(TN + FN))^{1/2} = 0.53
\]

3.2. Applications of the research software ACG with the human chromosomes

3.2.1. Three examples leading to classical results

The research software ACG identifies protein genes and their frames as follows:

The identification of a protein gene (called CDS according to the EMBL syntax) results from a curve that is significantly greater than the two others that lead to a large surface \( s \) (notion introduced in Section 2). The existence of a top curve is justified by the fact that the associated function \( F_k \) is based on the circular code, which is a strong property of the protein genes (see Section 1). The intersection of the two highest curves allows for predicting a beginning and end regions of protein genes.

The identification of a frame of a protein gene is deduced from the frame of the top curve determined from the beginning of the sequence (see the Section 2).

Three examples of identification of protein genes (CDS) listed in the EMBL human chromosomes with the software ACG, are given.

Fig. 5a identifies a large CDS (2295 bases in the human DNA sequence from clone 512B11 on chromosome 6p24–25). Indeed, the curve becomes greater than the two others. Note that in a random sequence, e.g. a sequence generated with the four bases with equiprobability, leads to three similar horizontal curves. These three curves are gathered together before the beginning of the CDS (5'
regions) and after the end of the CDS (3' regions), close to random sequences. As the top curve is $C_0$, the predicted frame of the CDS from the beginning of the sequence, is 0. This is in agreement with the EMBL frame (frame $(44\,614-1) \mod 3 = 0$).

Fig. 5b identifies a complementary protein coding gene (cCDS in the human DNA sequence from clone E146D10 on chromosome 22) of middle size (595 bases). The cCDS chosen as an example, starts at the complementary base 17\,772 and ends at the complementary base 17\,178. As the top curve is $C_0$, the predicted frame of the cCDS is 0. The circular code $T_0$, being self-complementary, a function of $F_k$ based on $T_0$, leads to the same results on the DNA complementary strand. Therefore, the association of a frame with a curve, and in particular the top curve, is identical with the CDS and the cCDS. The predicted frame 0 agrees with the EMBL frame. The determination of the frame, from the beginning of the sequence, of a cCDS starting at the complementary base position $j$ and ending at the complementary base position $k$.
is obtained with the value \( j \) modulo 3. The EMBL frame with this example is then 0, as 17 772 mod 3 = 0.

**Fig. 5c** identifies a series of five CDS (exons in the human DNA sequence from cosmid B2046 on chromosome 6), three among them have a small size (about 200 bases). Their frames determined by the software ACG are 1, 2, 1, and 1 respectively. All the predicted frames agree with the frames deduced from the EMBL data:

- the frame of the first CDS is equal to 1 as \((29 039 - 1) \mod 3 = 1\);
- the frame of the second CDS is equal to 2 as \(((29 397 - 1) + 1 - (29 245 - 1) \mod 3 = 2\) (this expression is obtained from the beginning of the CDS, i.e. 29 397, the frame of the previous CDS, i.e. 1, and the end of the previous CDS, i.e. 29 245);
- the frame of the third CDS is equal to 2 as \(((30 373 - 1) + 2 - (30 228 - 1) \mod 3 = 2\);
- the frame of the fourth CDS is equal to 1 as \(((30 740 - 1) + 2 - (30 584 - 1) \mod 3 = 1\);
- the frame of the 5th CDS is equal to 1 as \(((31 110 - 1) + 1 - (30 953 - 1) \mod 3 = 1\).

**Fig. 5c** gives an analysis of the relative positions of the different exons between themselves.

### 3.2.2 Two examples leading to unexpected results

Two unexpected results obtained in the EMBL human chromosomes with the software ACG, are given: the prediction of internal frames in the
protein genes and the prediction of a coding region in the genomes.

**Fig. 6a** predicts internal frames in a CDS (1334 bases in the human DNA sequence from clone 1189B24 on chromosome Xq25–26.3). Indeed, there are three intersections of the two first highest curves, which are associated with four assumed internal frames 0, 1, 0 and 2. The frameshift of 1 (respectively, 2) base can be associated with 1 (respectively, 2) base insertion (modulo 3) or 2 (respectively, 1) base deletions (modulo 3). The internal frames can also be explained with the concatenation of coding regions whose lengths are not all multiple of 3. This CDS is mentioned as pseudogene in the EMBL file.

**Fig. 6b** predicts a coding region in the human DNA sequence from clone 1048E9 on chromosome 22q11.2–12.2 and its frame 2. This region is associated with the primary transcript starting at 18411 and ending at 18946. Surprisingly, the predicted frame by the software ACG is equal to the EMBL frame \((18411–1) \text{ mod } 3 = 2\).

### 4. Discussion

A new statistical approach using functions based on the circular code identifies 93.32% (coefficient SMC) of bases in the human sequences correctly, i.e. classifies the bases in coding and
non-coding genes correctly. This approach has been evaluated with the coefficient SMC, which represents a good compromise between the Sensitivity Sn of coding bases identified correctly (39.75%) and the Specificity Sp of non-coding bases identified correctly (98.88%) (in the population studied, 90.6% bases are non-coding). These frequencies are retrieved by varying the size of the sequence population, e.g. by eliminating the short sequences (data not shown). Indeed, a large quantity of data used for the computation leads to stable values. The parameter s (surface) used for evaluating the statistical significance is a concept extending the natural and simplest parameter v based on the value of a function for a given base position. The statistical results obtained with this parameter v lead to a coefficient SMC that is significantly low than 93.32% (data not shown). The choice of the coefficient SMC for evaluating this new statistical approach in order to identify protein genes in genomes is confirmed by the Correlation Coefficient (CC) whose maximum (0.53) is also reached with the function $F_2^D$ with $n=50$ and $s=38$. Note that if all bases

Fig. 6. (a) Prediction of internal frames in a protein coding gene (CDS) and their internal frames by the software ACG. This example is observed in the human DNA sequence from clone 1189B24 on chromosome Xq25–26.3. (b) Prediction of a coding region in the human DNA sequence from clone 1048E9 on chromosome 22q11.2–12.2.
are predicted as non-coding then CC is equal to 0.

The main purpose of this paper is to propose a completely new approach for identifying protein coding genes in genomes by using a gene model based on the circular code. Therefore, the method developed allows the global location of regions that are coding for proteins or not. The start and end of the coding region can be predicted by the intersection of the two highest curves. Obviously, the exact location of the boundaries can be improved in the future by analysing in detail the start regions and the end regions of coding genes by considering, for example, the start codon ATG, the stop codons TAA, TAG, and TGA, the splicing sites, the TATA box, etc. It can also be associated with other methods for identifying protein genes in genomes, such as the codon usage methods, the methods based on the hidden Markov model (HMM), etc. (e.g. Shulman et al., 1981; Shepherd, 1981; Staden and McLachlan, 1982; Fickett, 1982; Smith et al., 1983; Blaisdell, 1983; Staden, 1984; Borodovsky and Mc Ininch, 1993; Krogh et al., 1994; Burge and Karlin, 1997; Lukashin and Borodovsky, 1998; Salzberg et al., 1998; Pavy et al., 1999; Shmatkov et al., 1999, etc.). However, it should be stressed that the method in its actual state gives interesting results. Indeed, the Correlation Coefficient (CC) is equal to 0.53 with a data set containing 303 124 560 bases without discarding particular sequences. Obviously, these values become significantly better.
if there is a data selection before the statistical analysis, e.g. by discarding particular sequences: the sequences for which the exact location of the protein genes is determined ambiguously, the sequences encoding pseudogenes, etc. For example, the removal of pseudogenes (784,274 bases representing about 2.92% of the gene population studied) increases the coefficient SMC from 93.32 to 93.53%.

In summary, the research software ACG using functions based on the circular code, constitutes a new approach for identifying protein genes in genomes and for determining their frame. As it is based on the circular code of protein genes of both eukaryotes and prokaryotes, it can be applied independently of the type of eukaryotic/prokaryotic organism under investigation. An example of the use of the software ACG on a prokaryotic organism (Escherichia coli) is presented in Fig. 7. Furthermore, it also allows an evaluation of the length of protein genes, their position in the genome, their relative position between themselves, and the prediction of internal frames. It can be used without prerequisite knowledge: interactivity, graphical tools, possibilities of varying the parameters (the function, the length of the window, the surface level), etc. As the user-friendly software ACG is based on a new concept (circular code), the genomes can easily be investigated for obtaining new results in this research field.

Fig. 7. Identification of several protein coding genes (CDS) and their frames by the software ACG. This example is observed in the E. coli K-12 MG1655 complete genome.
Acknowledgements

We thank the Editor-in-Chief and the Referee for their advice.

References


