Universidade de Aveiro
Ano 2013

Departamento de Electrónica, Telecomunicações e Informática

João André Rodrigues Antunes

Distâncias Inter-Simbólicas no ADN

Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática
2013

João André Rodrigues Antunes

Distâncias Inter-Simbólicas no ADN

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Electrónica e Telecomunicações, realizada sob a orientação científica do Doutor Carlos Alberto da Costa Bastos, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e da Doutora Vera Mónica Almeida Afreixo, Professora auxiliar do Departamento de Matemática da Universidade de Aveiro

o júri

presidente	Prof. Doutor Armando José Formoso de Pinho Professor Associado c/ Agregação da Universidade de Aveiro
vogais	Prof. Doutor Rui Carlos Camacho de Sousa Ferreira da Silva Professor Associado do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto
	Prof. Doutor Carlos Alberto da Costa Bastos Professor Auxiliar da Universidade de Aveiro (Orientador)

agradecimentos

Agradeço quer ao Professor Carlos quer à Professora Vera pelo acompanhamento e orientação durante a realização desta Dissertação, bem como a constante disponibilidade para ajudar quando necessitei.

À minha família, amigos e namorada agradeço o apoio e o carinho constantes, permitindo assim a estabilidade pessoal e a motivação necessária para a realização deste trabalho.

palavras-chave segmentação, distâncias inter-simbólicas, ilhas CpG, processamento de sinal.

resumo

O presente trabalho visou estudar o contributo das distâncias-inter simbólicas na segmentação do ADN. Para esse efeito, foi estudada a segmentação das sequências genómicas em código e não código e em ilhas e não ilhas CpG. Desenvolveu-se um estudo das distâncias inter-trinucleótidas no contexto da identificação de regiões codificantes e das distâncias inter-dinucleótidas para a identificação de ilhas CpG. Com base nestas distâncias foi analisado o desempenho de um algoritmo para discriminação de regiões de código e não código, tendo os resultados evidenciado haver ainda margem para aperfeiçoamento e foi desenvolvido um algoritmo para identificação de ilhas CpG tendo as taxas de boa classificação atingido valores elevados.

keywords segmentation, inter-symbolic distances, CpG islands, signal processing

abstract

The present work aimed to study the contribution of the inter-symbolic distances in DNA segmentation. To this end, the segmentation of genomic sequences into coding and non coding regions and CpG islands and non CpG islands was studied. A study of the inter-trinculeotide distances in the context of identifying coding regions and of the inter-dinucleotide distances for identifying CpG islands was developed. Based on these distances the performance of an algorithm to discriminate coding and non coding regions was analyzed, with the results showing there is still room for improvement and an algorithm for identification of CpG islands was designed, resulting in high values of good classification rates.

Table of contents

1. Introduction

The human genome contains information about how cells are organized in our body and how the body interacts with the surrounding environment. In the last years there has been a considerable interest in the study of the human genome, which has already been sequenced.

The DNA of each species can be seen as a long sequence of letters of an alphabet made up of four symbols, A, C, G and T, which make up the DNA. The inter-symbolic distances are believed to be an important contribute in providing relevant information stored in the DNA, namely information about the three-dimensional DNA structure [1].

The present study aims at exploring the capability of using the inter-symbolic distances to identify genomic regions, of different species, which display features with biological interest, such as coding and non coding regions and CpG islands.

In order to explore this capability, some tools were developed, using MATLAB, that allow the study of the inter-symbolic distances between nucleotides, dinucleotides (CpG islands) and trinucleotides (coding and non-coding regions) in an easy and efficient way, representing a contribution to the genomic signals processing.

1.1. History of DNA study

Curiosity and academic interest in DNA has been stimulating its study, experiments and research since the 19th century. The pioneering experiments were conducted by Gregor Mendel, a Czech monk, who after some observations and tests with peas, came to the conclusion that their shape and color were acquired according to different packages, which we now identify as their genes [2].

Figure 1 - Gregor Mendel (1822 - 1884) [3].

In 1868, Friedisch Miescher, a Swiss physician, achieved a significant breakthrough. He managed to isolate a compound which he named as “nuclein”, which is now known as nucleic acid (NA) that is part of DNA (deoxyribo-nucleic-acid) and RNA (ribo-nucleic-acid) [4].

However, DNA was still a source of great controversy among distinguished scientists in the 1940s. In spite of their awareness that DNA might well be the molecule of life, some of them found it hard to recognize it because of its simplicity. Furthermore, they still could not figure out how the molecule was likely to look like, although they knew that the four bases (adenine, thymine, guanine and cytosine), were part of DNA.

Further information had to be gathered, so that all the pieces of the puzzle might fit, such as finding out that the phosphate bases were on the inside, while its backbone was on the outside; understanding the reasons why the two strands could run in both directions; checking that the molecule had a unique base pairing and was a double helix [5].

Many people had to be involved in this painstaking research. Stick-and-ball models were used by Watson and Crick in order to confirm their general speculations about DNA structure [6]. On the other hand, another group of scientists like Rosalind Franklin and Maurice Wilkins used X-ray diffraction so that they might get aware of the physical structure of the DNA molecule. They even tried a three-helical model in 1951 but without success.

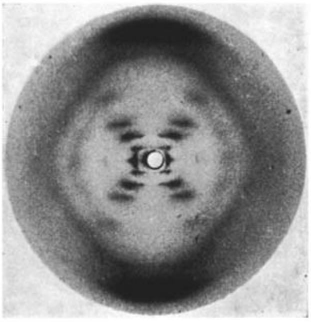


Figure 2 - Rosalind Franklin "Photograph 51"[7].

Linus Pauling, who had already published an article on a triple-helical structure for DNA in 1953, was also trying to discover the real DNA shape but Franklin's famous "photograph 51" was the one that helped Watson and Crick to understand the double-helical structure of DNA.

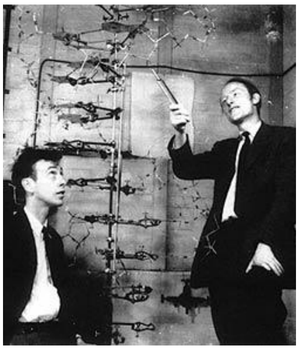


Figure 3 - First photo of James Watson's and Francis Crick's double helix DNA model. May 1953 [8].

In what concerns the base-paring question, Erwin Chargaff, a biochemist was capable, in 1949, of demonstrating that the quantity of adenine and thymine are always the same, even knowing that the length of the DNA sequences varies in different organisms. He was also able to prove that the adenine-thymine link had precisely the same length as the cytosine-guanine and the bases were paired according to this pattern [9].

1.2 Some biological concepts

The DNA (deoxyribonucleic acid) is an informational molecule encoding the genetic instructions used in the development and functioning of all known living organisms and many viruses. It is composed of two polynucleotide chains twisted around each other forming a double helix. These nucleotides are made up of a phosphate linked to a sugar (known as deoxyribose) to which a base is attached. There are four different bases in a DNA molecule, adenine (A), cytosine (C), guanine (G) and thymine (T), which are joined together in pairs, with each base from one chain being hydrogen-bonded to a base from the other chain, lying side by side. The bounding can only occur between a purine and a pyrimidine so only specific pairs of bases can bond together, adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine). The backbone of each strand of the helix is composed of alternating sugar and phosphate residues.

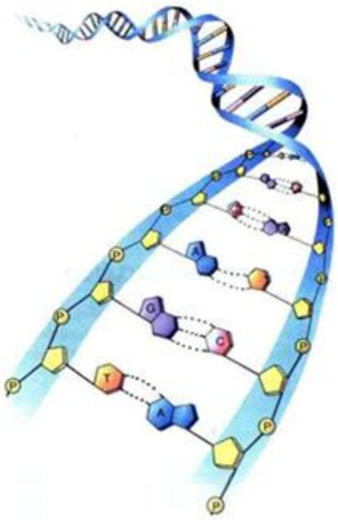


Figure 4- Representation of the DNA molecule [10].

The sequence of nucleotides that make up DNA can be split into two different categories, coding and non-coding regions. The coding regions consist of sets of relevant sequences in terms of protein production. The number of nucleotides in these regions is multiple of three because each triplet (codon) represents the code of an amino acid, the structural unit of a protein. However, on the eukaryotic species, those zones are only a small part of the whole sequence of DNA (approximately 2% for the *Homo sapiens*) and are contained in the so called genes.

On those organisms, most genes have a sequential structure of alternating parts, exons (constituting the code for proteins), and introns, which are non coding sequences.

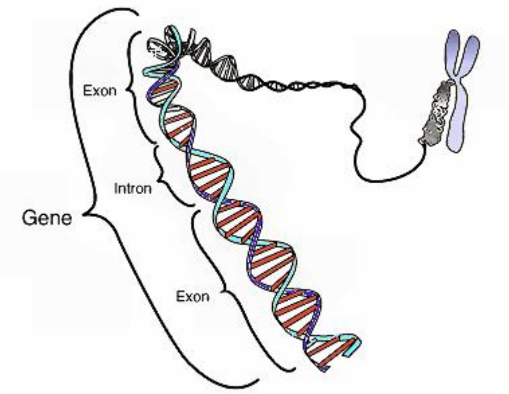


Figure 5 - The diagram of a gene (constituted by exons and introns) [11].

The RNA (ribonucleic acid) is a family of large biological molecules that perform multiple vital roles in the coding, decoding, regulation, and expression of genes. Like DNA, RNA is assembled as a chain of nucleotides, but it is usually single-stranded and has uracil instead of thymine in its structure. Cellular organisms use messenger RNA to convey genetic information that directs synthesis of specific proteins, while many viruses encode their genetic information using one RNA genome.

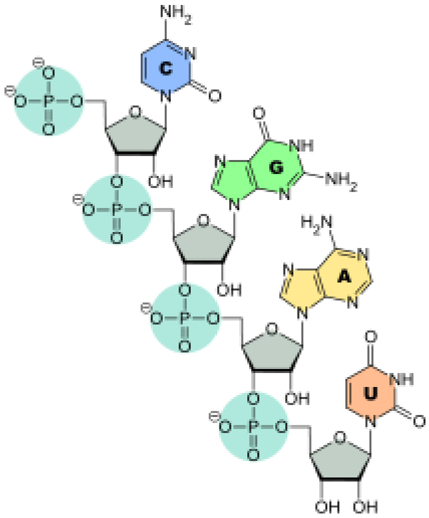


Figure 6 - Representation of a RNA molecule [12].

1.3. Dissertation structure

This work is organized in five chapters, four appendices and a glossary.

The first chapter starts with an introduction setting a global framework as well as the motivations and the objectives to be achieved, followed by the history of the DNA study and some biological concepts such as the structure of both DNA and RNA.

Moreover, in chapter two, it is described not only the importance and the goals of DNA segmentation, but also some state of the art algorithms used to discriminate coding and non coding regions and to find CpG islands.

The third chapter covers the methods used to apply the inter-symbolic distances to DNA segmentation. It starts by introducing some concepts related to the inter-nucleotide distances previously studied by other investigators and it is followed by the description of an algorithm used to discriminate coding and non coding regions. Finally, in this chapter the whole process that led to the development of a new algorithm using inter-symbolic distances to find CpG islands is described.

Besides, in chapter four, the DNA data used and the MATLAB tools developed are described. It continues by showing the experimental results obtained during this work starting at those related to the exploratory studies on the stop codons distribution in coding and non coding regions and the CG symbol distances distribution in CpG and non CpG islands. Furthermore, in this chapter, the results of the discrimination of coding and non coding regions are displayed, ending with the evaluation of the performance of the developed CpG distances algorithm. The performance of this algorithm was also compared to that of the Hidden Markov Model, which is one of the most frequently used models to find CpG islands.

Also, in the fifth chapter, some important conclusions are drawn and some possible future work is proposed.

Finally, the described chapters are followed by four appendices and a glossary with the definition of the most relevant biological terms used in this work.

2. DNA segmentation methods

The segmentation of DNA represents an important issue for scientists, as it allows to extract information of useful genomic regions and to help understand the organization of the genetic process.

It is known that a typical DNA sequence is not homogeneous and that some segments that reveal a certain homogeneity as well as regions with varying statistical properties may have biological meanings (such as regulatory elements, structural features of the DNA, CpG islands, coding and non coding regions) [13]. The computational methods used to identify these homogeneous regions are called segmentation methods [14].

Moreover, the comparison of sequences between species requires methods of determining similarities in evolution or function. Today, large chunks of the genome are sequenced but the role played by many of the sequences remains unknown.

In order to face this challenge the DNA segmentation methods are used to divide this unknown sequences into a number of segments, where each segment has a certain degree of internal homogeneity, so they can be compared with previously well studied small sequences and provide useful characterizations [15].

There are many segmentation techniques such as the Moving Window, the Maximum Likelihood Estimation, the Hidden Markov Models and Recursive Segmentation (see for example: [16], [17], [18], [19]).

This work will focus on methods to discriminate coding and non coding regions and to detect CpG islands.

2.1. Detection of coding and non coding regions

The computational recognition of genes and coding regions is one of the major challenges for the molecular biology in the analysis of newly sequenced genomes. However, there are two basic issues in gene identification: detection of protein-binding sites of the genes and finding out regions that code for proteins [20]. There are numerous segmentation methods with the goal of finding borders between coding and non coding regions, some of which are briefly described below (the enumeration and description of the methods does not pretend to be exhaustive).

The method described by Bernaola et al, employed a 12-symbol alphabet to identify the borders between coding and non coding regions based on nucleotide statistics inside codons [21].

Later, Nicorici and Astola segmented the DNA sequence into coding and non coding regions using recursive entropic segmentation (based on Jensen-Shannon and Jensen- Rényi divergences) and stop-codon statistics [20].

Moreover, signal processing techniques based on the period-3 property (periodicity of DNA in exons, with the period being equal to 3 nucleotides) [22] can play an important role for gene finding. Thus, Tiawari used the discrete Fourier transform (DFT) spectrum to achieve this goal, where the DFT energy at a central frequency is calculated for a fixed length window, and the window is slid across the numerical sequence [23].

Plus, Vaidyanathan identified protein coding regions using an anti-notch filter which magnified regions with period-3 property [24] and Akhtar applied time domain algorithms, average magnitude difference function and time domain periodogram algorithms to identify eukaryotic gene locations [25].

2.2. Detection of CpG islands

Over the last decades there has been an increasing interest in the study of CpG islands or CG islands, genomic regions where a cytosine nucleotide occurs next to a guanine nucleotide (connected by a phosphodiester bond) with high frequency, as they are often located around the promoters of genes that are essential for general cell functions [26]. These islands are useful markers for genes and play important roles during X-chromosome inactivation, imprinting and silencing of intragenomic parasites.

In the last twenty five years, some studies have tried to find a precise definition for these regions. Gardiner-Garden and Frommer in 1987 [27] considered CpG island as a DNA sequence with at least 200bp, with a C+G content greater than 50% and an observed-to-expected CpGratio ($\frac{Number\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Total\ number\ of\ nucleotides\ in\ the\ sequence$) greater than 60%.

Moreover, in 2002, Takai and Jones [28] revised this definition in order to discriminate other genomic sequences with rich GC content, such as Alu repeats, considering sequences with more than 500bp, G+C content greater than 55% and an observed-to-expected CpG ratio greater that 65% as more likely to be real CpG islands.

Although there are many models available to find CpG islands, one of the most used and that has a satisfactory performance is the Hidden Markov Model [26]. Hence, a version of this model will be presented in the next section and later its performance will be compared with the algorithm developed in this study.

2.2.1 Hidden Markov Model

Because pairs of consecutive nucleotides are important in this context, a model in which the probability of one symbol depends on the probability of its predecessor is necessary. To capture this dependency, it can be applied an Hidden Markov Model (HMM).

An HMM is a system $H = (\alpha, Q, \Lambda, e)$ consisting of:

- an alphabet α
- a set of states Q
- a matrix $\Lambda = \{a_{kl}\}$ of transition probabilities a_{kl} for k, $l \in Q$, and
- an emission probability $e^k(b)$ for every $k \epsilon Q$ and $b \in \alpha$

For CpG-islands a possible model is:

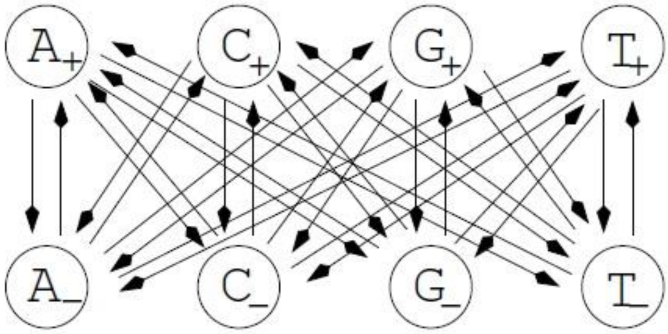


Figure 7 - Possible transitions between states for the HMM model [29].

Using the alphabet $\alpha = \{A, C, G, T\}$, if the symbol comes from a CpG island, the states are A^+, C^+, G^+, T^+, otherwise, the states are A^-, C^-, G^-, T^-. The remaining state (not represented in the picture) is 0, representing the begin/end state. Thus, there are nine possible states [29].

The transition probability matrix Λ used in this model is displayed in Appendix A, in section 7.1.

Besides, the model emits the letters A, C, G, T, but for each letter there are two states from which the letter can come from. The emission probabilities matrix is also presented in Appendix A.

On the other hand, having observed a sequence of symbols generated by an HMM, it is important to decode the sequence states from it. The most common way to do this is using the Viterbi algorithm [30].

The advantage of the Viterbi algorithm is that it does not blindly accept the most likely state at each instant i, but in fact takes a decision based on the whole sequence. This is useful, if there is an unlikely event at some point in the sequence.

3. DNA segmentation with inter-symbolic distances

The inter-symbolic distances were brought up by Nair and Mahalakshmi in 2005 and provide a new approach to explore the correlation structure of DNA [31].

If this method is used, each symbol constituting a given DNA sequence will be converted into a number corresponding to the distance to the next equal symbol. Therefore, if the sequence is considered to be circular, the length of the new numeric sequence keeps the same as the original.

Given the alphabet $\alpha = \{A, C, G, T\}$ and considering a word, μ, defined in α with any length chosen, there can be a numerical sequence, d^{μ}, that represents the distance between the first occurrence of the word μ and the next one in the DNA sequence. For example, given the sequence:

A C A C G A A T T T A T T C G A A T T C A A C T T A A C

considering a word μ ={AA}, the distance sequence d^{μ} of the word μ assuming that this is a circular sequence and there is overlapping in the word, is:

$$d^{\mu} = \{1,9,5,5,8\}$$

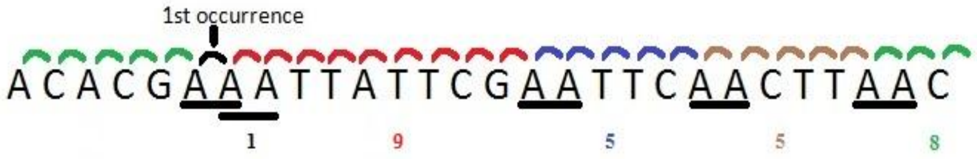


Figure 8 - Illustration of the distances vector for the word μ ={AA} with overlapping.

On the other hand, considering that there is no overlapping, the distance sequence d^{μ} is:

$$d^{\mu} = \{9,4,4,7\}$$

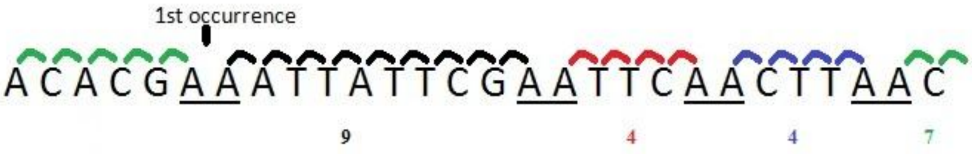


Figure 9 - Illustration of the distances vector for the word μ ={AA} without overlapping.

Other way of studying the inter-symbolic distances that can be important in many cases is using different reading frames. There are as many reading frames as the word length in study and the distance unit is the word.

Considering the same word, μ ={AA}, and a circular sequence, the distances vectors of this word with two different frames are:

$$\boldsymbol{d}_{F_1}^{\mu} = (\mathbf{14}, \dots)$$

$$\boldsymbol{d}_{F_2}^{\mu} = (\mathbf{5}, \mathbf{5}, \mathbf{4}, \dots)$$

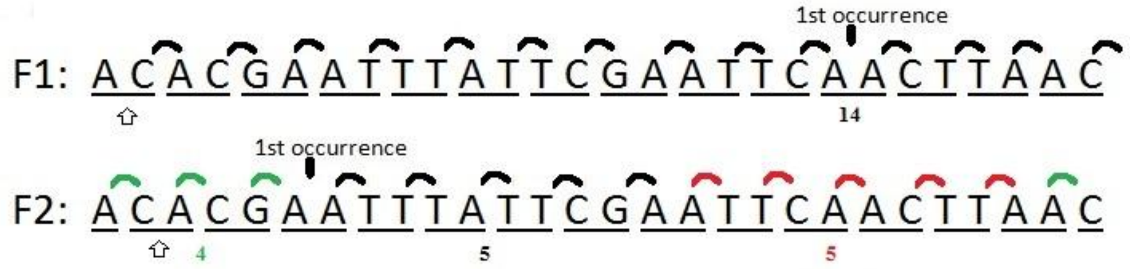

1st occurrence
F1: A C A C G A A T T T A T T C G A A T T C A A C T T A A C
14
1st occurrence
F2: A C A C G A A T T T A T T C G A A T T C A A C T T A A C
4
5
5

Figure 10 - Illustration of the distances vector for the word μ ={AA} considering two different reading frames.

3.1. Inter-nucleotide distances

Later, in 2009, the inter-symbolic distances method was explored [1] and applied to the inter-nucleotide distances, introducing four new sequences, one for each nucleotide (d^A, d^C, d^G, d^T), in order to analyze the behaviour of the distance vector of the four nucleotides and of the global sequence.

In the next figure, the four inter-nucleotide distance sequences are represented, for a short sequence of the *Homo sapiens* chromosome 1:

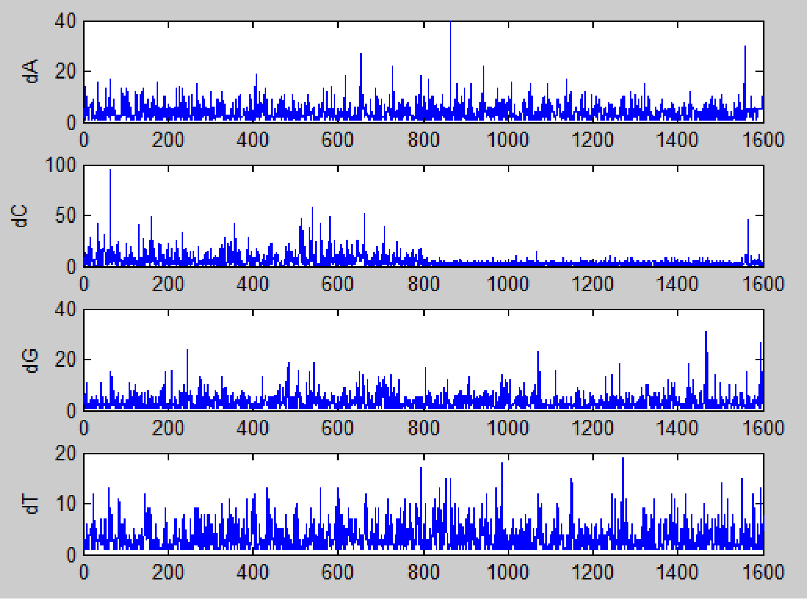


Figure 11 - Inter-nucleotide distances for the first 1600 distances of each nucleotide of the gi|157811749|ref|NW_001838563.2| *Homo sapiens* chromosome 1.

The length (*N*) of the global distance sequence, *d*, can be calculated by the sum of the lengths of the four inter-nucleotide distance (N^A, N^C, N^G, N^T).

Furthermore, the positions of all the nucleotides in the complete sequence may be determined if the position of the first occurrence, k_0^x, of each nucleotide is known:

$$k_j^x = \sum_{i=1}^{j} d_i^x + k_0^x$$

whence, $k_j^x - k_{j-1}^x = d_j^x$ and $N = \sum_{i \in N^x} d_i^x \, , x \in A$

With a view to study some statistical properties of the DNA of different species, the inter-nucleotide distance distribution was used.

In addition, it was considered that if the nucleotide sequences were generated by an independent and identically distributed random process, then each of the inter-nucleotide distance sequences, d^x, would follow a geometric distribution. The probability functions are:

$$f^x = P(d^x = k) = P^x (1 - p^x)^{k-1}$$

$$F^x = P(d^x \leq k) = 1 - (1 - p^x)^k$$

where the expected value is $1/p^x$ and the variance is $(1 - p^x)/(p^x)^2$.

Moreover, to estimate the nucleotide occurrence probability, p^x, the relative frequencies of each nucleotide in the DNA segment, N^x/ N, were calculated.

3.2. Detection of coding and non coding regions

Despite of the well known non homogenous distribution of the DNA sequence [32] and the existence of many published algorithms ([33] [34] [35]) to detect borders between coding and non coding regions, there is still room for improvement, as their performance is not ideal.

In order to analyze the behavior of the distance distribution between stop codons: TAA, TAG and TGA (as any of these symbols signals the end of genes), in coding and non coding regions, an exploratory study was carried out. In this study, the distribution of distances between stop codons in known coding and non coding regions was evaluated using three different reading frames.

The results of this study are displayed in section 4.2.1 and constitute an important indicator of the capability of this method to identify coding regions as it was found that not only the distributions of stop symbols in coding and non coding regions are different, but also that, in the correct reading frame of coding regions, the stop symbol occurs only at the end (as expected).

This way, there was the expectation that the distance between stop symbols could have high potentiality in improving DNA segmentation and setting better limits for coding regions. This concept was used and extended to develop an algorithm to achieve the discrimination of coding and non coding DNA regions using the inter-stop distance sequences [36].

3.2.1 Inter-stop symbols distance sequences

The inter-stop symbols distance sequence is determined considering a word, µ, composed by three nucleotides, that represent the stop codons and calculating the trinucleotides distance between them. From a single genomic sequence, it is possible to generate three trinucleotide sequences, one for each reading frame.

- **Reading frames**

For example, considering a genomic sequence starting by

AAACAAACTGACACAAAACACTAATAGTTTAAAATAATAATGA . . . .

Then, the three trinucleotide reading frames (R_1 , R_2 *and* R_3) produce the following trinucleotide sequences,

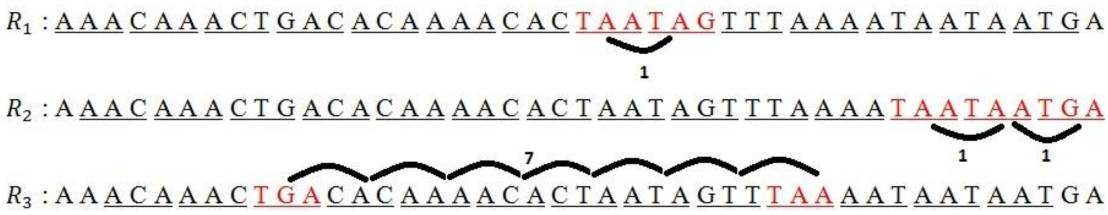

R_1 : AAA CAA ACT GAC ACA AAA CAC TAA TAG TTT AAA ATA ATA ATG A
1
R_2 : A AAC AAA CTG ACA CAA AAC ACT AAT AGT TTA AAA TAA TAA TGA
1
1
7
R_3 : AA ACA AAC TGA CAC AAA ACA CTA ATA GTT TAA AAT AAT AAT GA

Figure 12 - Reading frames for the given genomic sequence.

The distance vector represents the number of trinucleotides between the STOP symbols, and not the number of nucleotides, producing the following inter-STOP distance sequences:

$$\boldsymbol{d}_{R_1}^{\boldsymbol{STOP}} = (\mathbf{1}, \dots)$$

$$\boldsymbol{d}_{R_2}^{\boldsymbol{STOP}} = (\mathbf{1}, \mathbf{1}, \dots)$$

$$\boldsymbol{d}_{R_3}^{\boldsymbol{STOP}} = (\mathbf{7}, \dots)$$

The inter-STOP distance distribution of a sequence of random and independently placed nucleotides is given by

$$f^{STOP}(k) = P(d^{STOP} = k) = p^{STOP}(1 - p^{STOP})^{k-1}, k = 1, 2, \dots$$

with $p^{STOP} = p^{TAA} + p^{TAG} + p^{TGA}$. Considering that the four nucleotides have the same probability, $p^{STOP} = {}^{3}/_{64}$ and the expected distance, assuming independence and equal probability for the nucleotides, value is ${}^{64}/_{3}$. This function is a specific application of the probability function displayed in section 3.1 to the stop symbols.

3.2.2 Chi-square statistic

With the objective of measuring the lack of homogeneity of the inter-STOP distance distribution between the three possible reading frames a chi-square statistic was used. Moreover, with the aim of computing the chi-square statistic along the trinucleotide sequences a sliding window of fixed length (w) was used in each frame, and the distances within each window were separated into two different categories: short distance and long distance. The measure used to separate the short and long distances was called cut-off.

Furthermore, an extra category with the number of nonstop symbols within the window was also included. For each DNA sequence, contingency tables at the position of each trinucleotide were constructed, with a window of w trinucleotides:

	Frame 1	Frame 2	Frame 3	Total
non STOP	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
short distance	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
long distance	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

Table 1 - Contingency table for each window with N nucleotides ($n_{\cdot 1} = n_{\cdot 2} = n_{\cdot 3} = w$) [36].

In order to evaluate the homogeneity between reading frames a chi-square statistic was used, being in this case defined by:

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{\cdot j} n_{i\cdot}}{N}\right)^2}{\frac{n_{\cdot j} n_{i\cdot}}{N}} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot}}{3}\right)^2}{\frac{n_{i\cdot}}{3}}$$

Note that when one of the categories (non stop, short distance or long distance) does not occur in the three frames, the X^2 is set at 0.

3.2.3 Experimental procedure

The chi-square statistic for each symbol of the three reading frames was obtained for a sliding window with fixed length (1000 symbols) and a cut-off between 100 and 400. Besides, it was applied a ROC (receiver operating characteristic) curve and the area under curve, (AUC), was computed so that it was possible to evaluate the discrimination accuracy of the chi-square statistic and to establish the cut-point for prediction purposes (higher AUC values mean better discrimination performance). The point of the ROC curve closest to (0,1) is the “optimal point” in terms of sensibility, specificity and global accuracy of the prediction. The method used relies on the occurrence of two conditions:

- the existence of a long distance (greater than a certain threshold value)
- the value of the chi-square statistic being above a certain reference value.

If a certain DNA position verifies the two previous conditions then it is expected that there will be a start codon near the STOP codon in that reading frame.

Hence, starting at the STOP codon, it is necessary to search the next ATG codon (the most frequent initiation codon) in the same reading frame and consider it as the beginning of the coding regions.

Therefore, the symbols between the STOP codon and the beginning of the coding region are marked as non coding symbols. On the other hand, the symbols between the beginning of the coding region and the next STOP codon, in the same reading frame, are marked as coding symbols.

3.3. Detection of CpG islands

Aiming at developing an algorithm based on the inter-symbolic distances to find CpG islands, an exploratory study was also carried out to ascertain how the distribution of the distances between CG symbols varies in segments considered as CpG islands by the Takai and Jones definition and in segments that are not CpG islands.

The results of this study are shown in section 4.2. and represented an important motivation to develop an algorithm based on the inter C/G distances and on the concept of short and long distances.

3.3.1 Inter-CG symbols distance sequences

In this section the distance between dinucleotides was considered without overlapping and without reading frames.

This distance between CG symbols is calculated by the number of nucleotides between them, considering that the sequence is circular, as illustrated next:

$$d_{CG} = \{4,5,8\}$$

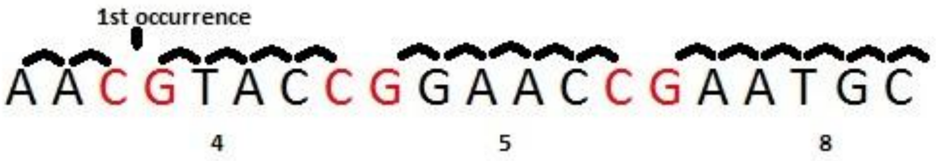


Figure 13 - Illustration of the distance between CG symbols.

Moreover, as it will be presented in the next section, it can be important to consider a set of related symbols as a single symbol and calculate the distance between them. So, considering the symbol S as the set of symbols where only cytosine and guanine are present, S = { CC, CG, GC, GG }, the respective distance vector of this symbol is:

$$d_S = \{3,1,3,5,3\}$$

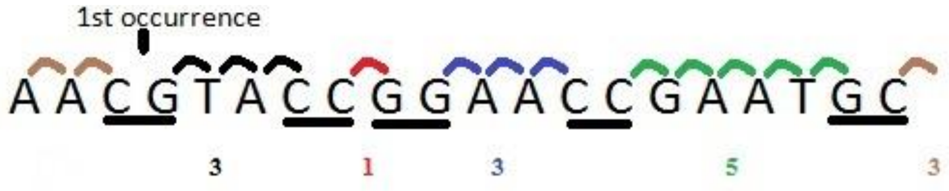


Figure 14 - Illustration of the distance between S= { CC, CG, GC, GG } symbols.

3.3.2 Methods

Based on the results of the exploratory study and the behavior of the inter-CG distances distribution, an algorithm to find CpG islands was developed.

- **First approach**

In the first approach, two parameters were defined in order to try to find significant differences between segments of DNA that constituted CpG islands and those that were not CpG islands.

These parameters were defined using the concept of short and long distances between CG symbols, considering that a certain cutpoint ∂ separates them.

The first parameter, meant to discriminate CpG islands in terms of the observed-to-expected ratio of symbol occurrence is defined by,

$$\mathrm{k}_1 = \frac{nd_c}{\mathrm{E}}$$

where nd_c is the number of short distances between CG symbols, and E the expected value that is calculated by

$$\mathrm{E} \ = \ nCpG * \varepsilon$$

where $nCpG$ represents the number of CG symbols in the sequence and ε the probability. Considering a sequence of nucleotides generated by an independent and identically distributed random process (the dinucleotides distances follows a geometric distribution), the probability ε is determined as follows:

$$\varepsilon = \sum_{\mathrm{d}=1}^{\mathrm{d}=\partial} (1 - \hat{p}\ )^{\mathrm{d}-1} \times \hat{p},$$

where $\hat{p}$ represents an estimative of the probability of the CG symbol,

$$\hat{p} \ = \frac{nCpG}{\theta}$$

with θ representing the length of the segment.

The second parameter, related to the C+G content,

$$\mathrm{k}_2 = \frac{nd_c}{nd_l}$$

represents the ratio between short and long distances of CG symbols detected in the sequence.

A DNA segment may be considered as CpG island, if k_1 and k_2 are greater than r_1 and r_2 which are threshold values to be found experimentally.

After exhaustive tests were made, between the possible ranges for each parameter, $]0, nCpG]$ for k_1 and $[1, nCpG]$ for k_2, the values that led to the best performance of the algorithm were r_1 = 0.80 and r_1 = 1.20.

However, the results obtained with this approach were not those desired in terms of sensibility (see Appendix D) and in spite of an exhaustive attempt to adjust the parameters, the results did not improve.

Thus, a new approach was tried, in order to improve sensibility, using information of other inter-dinucleotide distances.

- **Second approach: CpG distances algorithm**

Considering the symbol S as the combination of the 4 dinucleotides where only C and G are present, S = { CC, CG, GC, GG } and assuming that the nucleotides are generated by an independent and identically distributed random process with the same probability for each nucleotides the expected average distance between S symbols is 4 (4 out of 16 dinucleotides). Therefore, the first parameter developed in this approach, was

$$\bar{d}_S < 4$$

where $\bar{d}_S$ is the average distance between S symbols so that the segment has the necessary G+C content to be considered as a CpG island. However, when the algorithm was tested, it was found that this criterion was too restrictive, and the condition was changed to

$$\bar{d}_S < 4.5$$

Furthermore, the second parameter considered, so that the segment may have the necessary observed-to-expected value, considering $S^* = \{CG\}$, was,

$$\Pi = \frac{nd_c(S^*)}{nd_c{}'(S)} > k, \qquad k \in [\frac{1}{16}, 1]$$

where the number of short distances of the symbol S, $nd_c{}'(S)$, is divided by 4, the number of di-nucleotides constituting that symbol.

This parameter represents the ratio between the number of short distances of the symbol S^* and the number of short distances between S symbols. Different cutpoints were exhaustively tested between the range {4,...,16} and the one that conducted to best results was 8, so this value has been set. Furthermore, the value of k that led to the best performance was $k = 0.25$.

Therefore, in order to be considered as CpG island, a certain DNA segment has to verify $\bar{d}_S < 4.5$ and $\Pi > 0.25$.

3.3.3 Experimental procedure

This algorithm was then applied to different species and the results were compared with the formal definition of CpG island by Takai and Jones [28], so the length of each segment, θ, considered was 500. In order to evaluate its performance, three statistics were considered: Accuracy, Sensibility and Specificity that were calculated using four measures:

- i_i : number of segments considered as CpG islands by the Takai and Jones definition and by the algorithm in study.

- i_n : number of segments considered as CpG islands by the Takai and Jones definition and not by the algorithm.

- n_i : number of segments not considered as CpG islands by the Takai and Jones definition but considered by the proposed algorithm.

- n_n : number of segments not considered as CpG islands by the Takai and Jones definition neither by this algorithm.

CpG distances algorithm		Takai and Jones	
		Yes	No
	Yes	i_i	n_i
	No	i_n	n_n

Table 2 - Table representing the parameters used to determine the performance of the algorithm.

so that,

$$\text{Accuracy} = \frac{i\_i + n\_n}{num_{seq}} \times 100$$

where,

num_{seq} is the number of segments of length 500 in the considered chromosome.

Moreover,

$$\text{Sensibility} = \frac{i\_i}{i\_i + i\_n} \times 100$$

And finally,

$$\text{Specificity}=\frac{n\_n}{n\_n+n\_i}\times 100$$

4. Experimental results

4.1. Materials

4.1.1 DNA data

The DNA sequences used in this work were the *Homo sapiens* (annotation release 105) available in the National Center for Biotechnology Information (NCBI) [37], the eukaryotes *Saccharomyces cerevisae* [38] and *Encephalitozoon cuniculi* [39], the bacterias *Bifidobacteruim asteroides* [40], *Haemophilus influenzae* [41] and *Thermotoga maritima* [42], the phage *Aeromonas phage 65* [43] and the organelle *Calliarthtron tuberculosoum* [44] all available in the European Bioinformatics Institute database.

4.1.2 Developed MATLAB tools

In order to study and test the performance of the application of the inter-symbolic distances algorithm, in an easy and efficient way, a set of MATLAB tools were developed.

The first tool has two input parameters, a DNA sequence and a word (with any length), and provides as output the distances vector of that word in the DNA sequence.

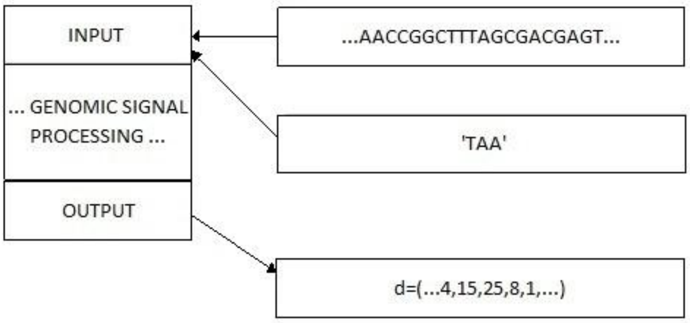


Figure 15 - Diagram of the first tool designed.

Aiming to test the performance of detecting CpG islands, another tool was developed. It receives as input a DNA sequence and has three incorporated functions, one to determine the CpG islands using the Takai and Jones definition (section 2.2), the second one using the distances algorithm and the third one using the HMM model. This program outputs the results in terms of accuracy, sensibility and specificity of the two algorithms when compared with the Takai and Jones definition.

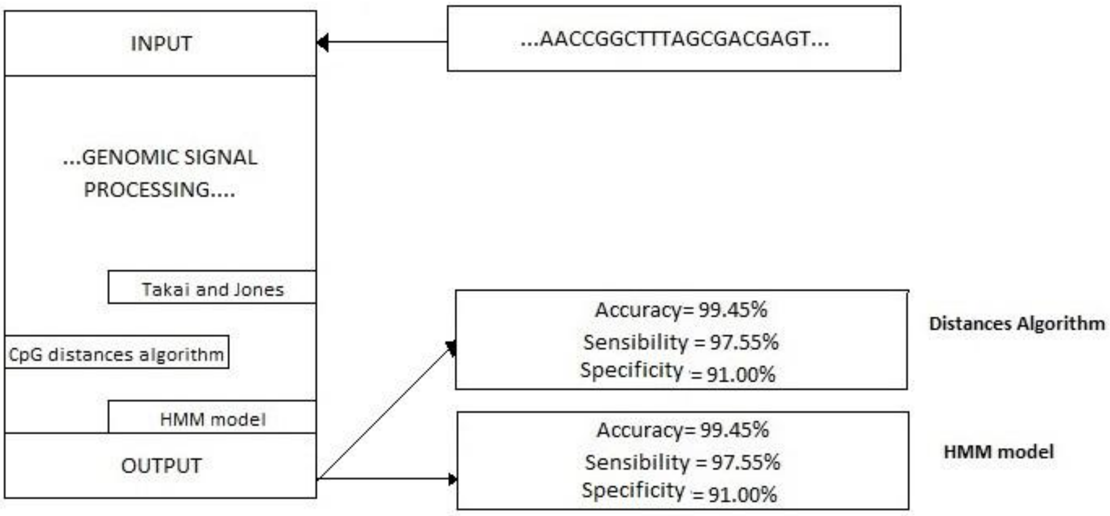


Figure 16 - Diagram describing the second projected tool to test the performance of the related algorithms.

Finally, the diagram of the tool used to test the performance of the Inter-stop symbols distances [36], applied to different species to discriminate coding and non coding regions is displayed in figure 17:

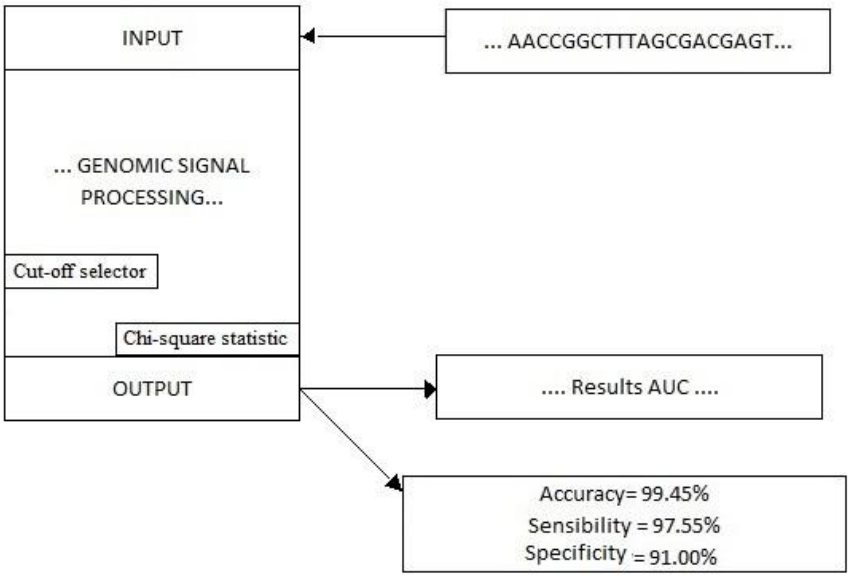


Figure 17 - Diagram describing the tool used to discriminate coding and non coding regions.

This algorithm receives as input a DNA data sequence and using an internal function selects the cut-off that conducts to the best AUC result. Using this cut-off that discriminates between short and long distances and a chi-square statistic it returns as output not only the results of the area under curve (AUC), but also the accuracy, sensibility and specificity of the algorithm.

4.2. Inter-symbolic distances

4.2.1 Inter-stop symbols distances distribution

To conduct the exploratory study of the distribution of distances between stop symbols, the annotated coding regions for each species were used. For each species, the distance distribution in three reading frames were determined by counting the occurrences of each distance, both in coding and non coding regions. In coding regions, frame 1 was always "the correct" reading frame, the one that results in the correct translation of codons into aminoacids.

Figures 18 to 20 show the histograms for the distance distribution, in each reading frame, between stop codons for the *Aeromonas phage 65,* in non coding regions:

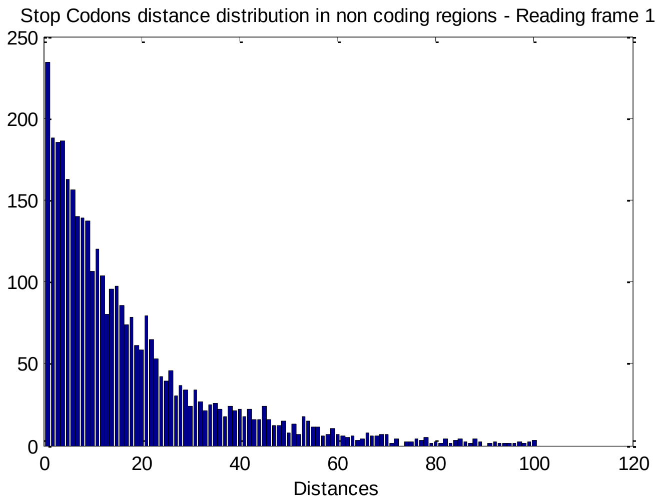


Figure 18 - Distribution of the distances between stop codons in non coding regions for the *Aeromonas phage 65*- Frame 1.

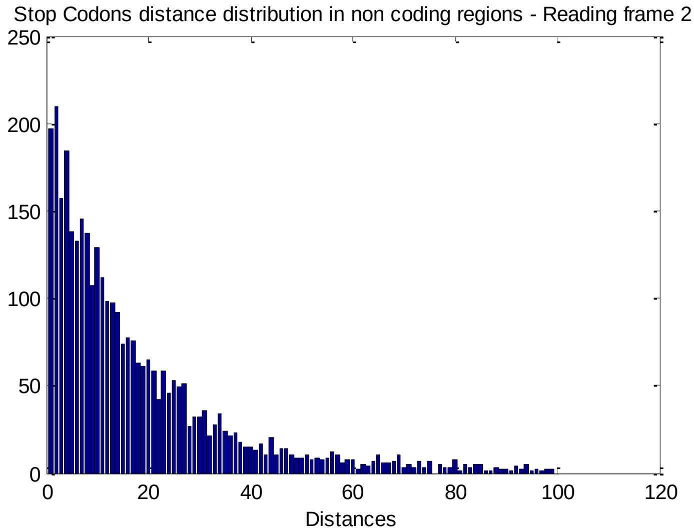


Figure 19 - Distribution of the distances between stop codons in non coding regions for the *Aeromonas phage 65*- Frame 2.

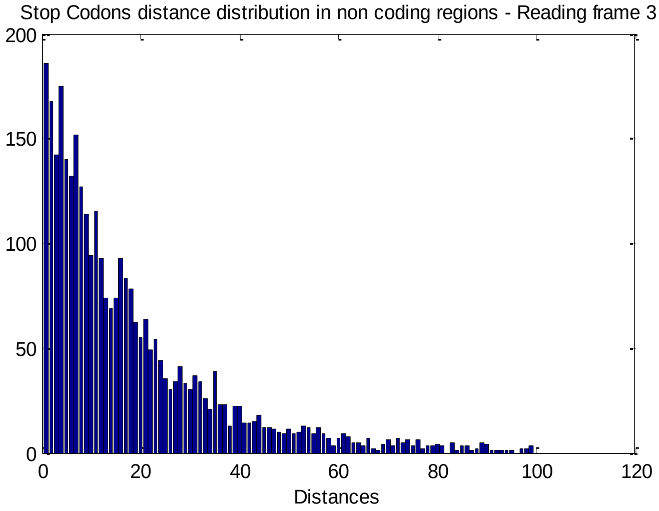


Figure 20 - Distribution of the distances between stop codons in non coding regions for the *Aeromonas phage 65*- Frame 3.

Observing the graphics, there are not any clear differences between each frame. So, the distribution of the distances between stop codons is similar in the three reading frames of the non coding regions.

On the other hand, figures 21 to 23, show the corresponding histograms for coding regions:

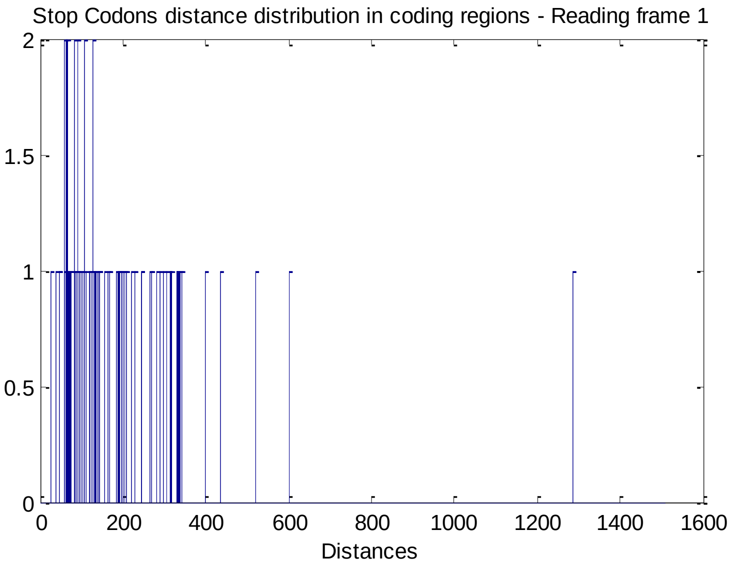


Figure 21 - Distribution of the distances between stop codons in coding regions for the *Aeromonas phage 65*- Frame 1.

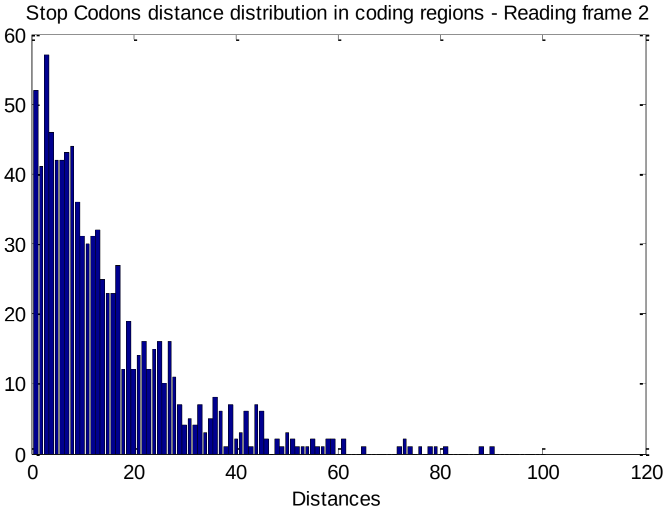


Figure 22 - Distribution of the distances between stop codons in coding regions for the *Aeromonas phage 65*- Frame 2.

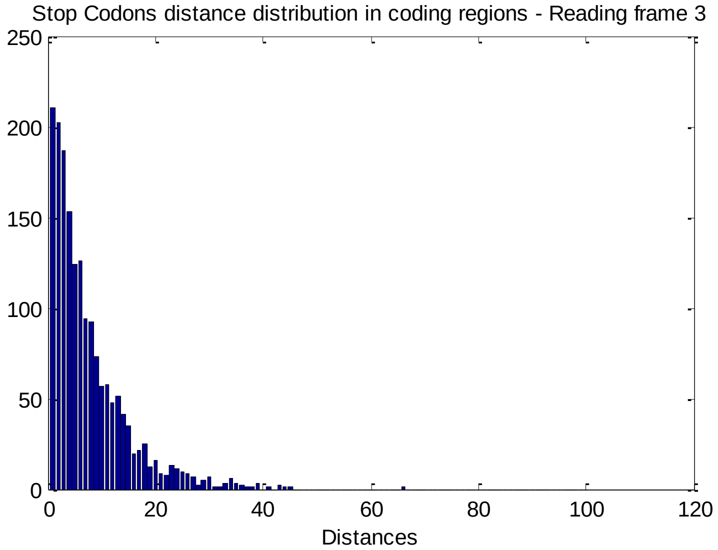


Figure 23 - Distribution of the distances between stop codons in coding regions for the *Aeromonas phage 65*- Frame 3.

Contrary to what was observed in non coding regions, in this case there is a frame totally different from the others, frame 1. In this frame, figure 21, there is only one stop codon in each coding sequence, and therefore one distance, signaling the end of the gene. As each coding sequence is considered circular, this distance represents the length of the gene. The smallest distance was 27 and the maximum distance was 1287. This behavior was similar in all the species which were tested and the results are displayed in Appendix B.

Hence, these observations constituted an important motivation in order to develop an algorithm based on distances between stop codons to discriminate coding and non coding regions.

4.2.2 Inter-CG symbol distances distribution

An exploratory study was carried out to characterize the inter CG symbol distance distribution in sequential segments of 500 nucleotides for the genome of all eukaryotes studied in this work. As the bacterias, the phage and the organelle do not have or have only a residual number of CpG islands, this study was not applied to them.

If a segment was considered, by the Takai and Jones definition, as CpG island the distance counts were accumulated in a vector corresponding to CpG islands. If the segment was not considered as CpG island, the counts were accumulated in a different vector.

The histograms of each distances vector, for *Homo sapiens*, are displayed in Figures 24 and 25:

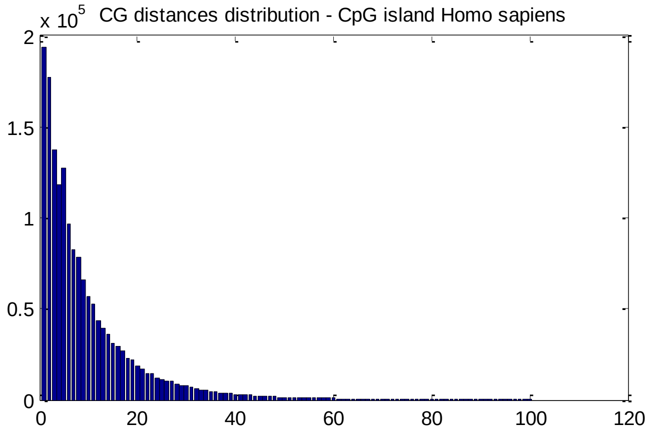


Figure 24 - Distribution of CG distances in CpG islands of *Homo sapiens* genome.

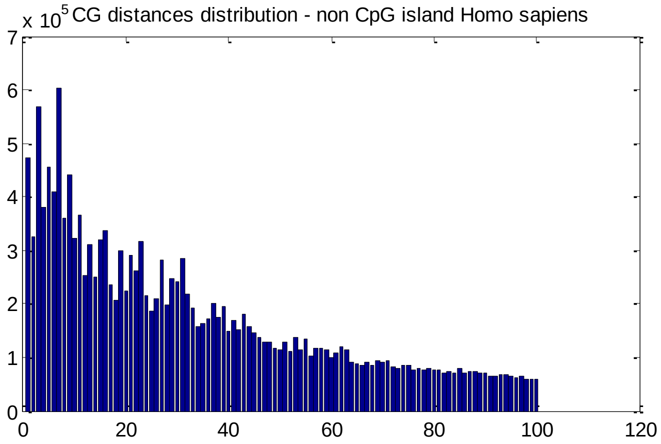


Figure 25 - Distribution of CG distances in non CpG islands of *Homo sapiens* genome.

When the graphics are analyzed, it is possible to verify that the distribution of the CG distances in segments from CpG islands has a steeper slope than in segments from non CpG islands .and thus, revealing a higher percentage of short distances in these segments.

The percentage of short distances between CG symbols was computed considering a cutpoint $\delta = 8$ (as set in 3.3.2). In the human genome the percentage of short distances in CpG islands was 60.64% and in non CpG islands was 20.70%.

In what concerns to the other two eukaryotes, *Sacccharomyces cerevisae* and *Encephalitozoon cuniculi* the behavior of the distribution of distances between CG symbols was similar to the *Homo sapiens* and the corresponding graphics are displayed in Appendix C.

Thereby, this exploratory study demonstrated that these characteristics could be the base of an algorithm which could be able to find CpG islands.

4.3. Detection of coding and non coding regions

An adjustment to the algorithm was initially tried in order to improve its performance, varying the cut-off and the window length between larger ranges.

However, as this has not led to any significant improvement, the algorithm was tested as it was initially developed [36].

It is important to note that this algorithm was applied to the *Homo sapiens* but the results were very poor. This is probably due to the presence of introns.

The algorithm was also applied to three bacterias (*Bifidobacteruim asteroides*, *Haemophilus influenzae* and *Thermotoga maritima*), one phage (*Aeromonas phage 65*), an organelle (*Calliarthtron tuberculosoum*) and to the chromosomes of two eukaryotes (*Sacccharomyces cerevisae* and *Encephalitozoon cuniculi*).

For each case, the cut-off that led to the best AUC result was calculated and the performance values, for each species, are displayed in tables 3, 4 and 5:

Species	Cut-off	AUC	X^2_{cut}	Accuracy (%)	Sensibility (%)	Specificity (%)
Aeromonas phage 65	190	0,89	0	83,49	85,60	82,83
Calliarthtron tuberculosoum	170	0,76	130,43	71,19	96,11	55,58
Thermotoga maritima	340	0,79	8,05	52,70	84,44	67,45
Haemophilus influenzae	210	0,82	0	77,68	95,22	64,40
Bifidobacteruim asteroides	420	0,72	0	70,75	72,23	69,93

Table 3 - Results of the inter-stop symbol distances algorithm for five different species.

From the results presented in table 3 it is possible to conclude that while the sensibility was good (average 86,72%), its accuracy (average 71.16% and 75,80% when the worst result for the *Thermotoga maritima* is excluded) and specificity (average 68,04%) were poorer. In terms of AUC, the results were good, with average 0.8 (the best possible value for the AUC is 1).

For the two eukaryotes, tables 4 and 5 show the results of the performance of the algorithm for segmentation of coding and non coding regions:

Chromosome	Cut-off	AUC	X_{cut}^2	Accuracy (%)	Sensibility (%)	Specificity (%)
1	240	0,81	8,64	80,09	88,78	77,22
2	230	0,82	0	80,01	88,51	76,09
3	270	0,77	0	78,01	79,78	77,14
4	320	0,79	0	65,50	83,96	77,05
5	210	0,81	5,11	75,71	91,23	68,09
6	350	0,79	20,18	79,91	78,78	81,36
7	250	0,81	0	78,85	87,71	73,03
8	320	0,77	0	77,88	82,87	74,77
9	280	0,75	0	79,40	83,55	73,25
10	310	0,78	0	79,20	84,30	71,30
11	230	0,79	3,25	78,55	92,75	70,47
12	220	0,81	12,99	78,67	87,19	74,17
13	320	0,78	0	78,12	91,38	71,08
14	230	0,79	0	79,19	85,30	75,69
15	310	0,78	47,33	79,02	83,80	75,92
16	260	0,79	0	77,37	86,13	72,00

Table 4 - Results of the inter-stop symbol distances algorithm for all chromossomes of the *Sacccharomyces cerevisae.*

Chromosome	Cut-off	AUC	X_{cut}^2	Accuracy (%)	Sensibility (%)	Specificity (%)
1	300	0,78	0	72,62	84,28	66,09
2	410	0,78	0	78,40	76,92	79,75
3	370	0,77	0	75,11	74,09	76,18
4	310	0,70	26,96	71,11	73,87	68,82
5	320	0,75	0	74,53	83,45	64,82
6	340	0,75	0	73,27	76,89	70,88
7	410	0,74	12,08	70,97	76,13	64,99
8	330	0,70	15,61	70,58	75,71	67,05
9	360	0,79	0	77,46	75,18	78,92
10	410	0,78	0	77,08	71,69	81,20
11	270	0,71	15,14	65,23	89,67	51,00

Table 5 - Results of the inter-stop symbol distances algorithm for all chromossomes of the *Encephalitozoon cuniculi.*

From the analyze of tables 4 and 5 tables it is possible to assess that, in terms of accuracy, the results were better in the case of the *Sacccharomyces cerevisae*, with an average of 77.84% in the whole genome sequence (78,72% excluding the worst result in chromosome 4), while the *Encephalitozoon cuniculi* had an average accuracy of 73.30% (74.11% also excluding the worst result).

Relating to the sensibility, the results were also good for the *Sacccharomyces cerevisae* (average 86,00%) and a little worse for the *Encephalitozoon cuniculi* (average 77,99%).

As in Table 3, the specificity was the worst parameter for the two eukaryotes (average 74.3% and 69.97%).

Finally, in what concerns to the AUC values, the results were again good, averaging 0,79 and 0,75, respectively.

These results are generally good but there is still room to improve the performance of the algorithm, mainly in terms of specificity, which had the worst results.

4.4. Detection of CpG islands

The results of the performance of the developed CpG distances algorithm, when compared to the Takai and Jones definition, are shown, for each chromosome of the *Homo sapiens*, in table 6.

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	99,21	95,27	99,24
2	99,38	94,83	99,40
3	99,39	95,34	99,41
4	99,42	94,61	99,44
5	99,36	95,25	99,39
6	99,27	94,78	99,29
7	99,24	93,55	99,28
8	99,43	94,44	99,46
9	99,21	94,18	99,25
10	99,28	94,30	99,31
11	99,25	94,50	99,29
12	99,16	93,87	99,19
13	99,36	93,04	99,39
14	99,26	95,14	99,29
15	99,23	96,51	99,25
16	98,94	94,00	99,00
17	98,69	95,53	98,74
18	99,34	93,68	99,37
19	97,94	93,62	98,05
20	99,09	95,17	99,13
21	99,23	96,23	99,25
22	98,78	93,92	98,85
X	99,46	94,89	99,48
Y	99,44	87,69	99,47

Table 6 - Performance of the new approach of the distances algorithm for each chromosome of the *Homo sapiens* (with $\bar{d}_S < 4.5$ and $k > 0.25$).

Moreover, the algorithm was applied to the *Saccharomyces cerevisiae* and to the *Encephalitozoon cuniculi* and the results are shown in the tables 7 and 8 ('*' means there were no CpG island by the definition of Takai and Jones detected on that chromosome) :

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	97,83	100,00	97,80
2	98,89	*	98,89
3	97,31	*	97,31
4	98,96	*	98,96
5	97,75	50,00	97,83
6	97,96	100,00	97,96
7	99,04	100,00	99,04
8	99,02	*	99,02
9	97,72	100,00	97,72
10	98,12	*	98,12
11	98,27	*	98,27
12	97,91	100,00	97,91
13	98,70	100,00	98,70
14	99,04	100,00	99,04
15	98,72	100,00	98,72
16	98,42	*	98,42

Table 7 - Performance of the distances algorithm for each chromosome of *Saccharomyces cerevisiae* (with $\bar{d}_S < 4.5$ and $k > 0.25$).

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	84,00	48,72	87,63
2	93,15	37,50	94,30
3	96,39	100	96,36
4	92,67	55,56	93,44
5	94,79	28,57	95,90
6	96,36	100	96,36
7	96,69	83,33	96,87
8	94,12	50	95,45
9	95,20	66,67	96,08
10	96,38	82,35	96,85
11	93,83	70	94,29

Table 8 - Performance of the distances algorithm for each chromosome of *Encephalitozoon cuniculi* (with $\bar{d}_S < 4.5$ and $k > 0.25$).

Observing the results in tables 6, and 7 it is possible to verify that the CpG distances algorithm shows good performance in all parameters which were considered, having improved significantly in terms of sensibility, when compared to the first approach (Table 13 in Appendix D).

However, there are some poor results in terms of sensibility in table 8. This happens because, in this species, there are only a small number of segments considered as CpG islands by the Takai and Jones definition and every increment in the i_n parameter (number of segments considered as CpG islands by the Takai and Jones definition and not by the algorithm) has a very significant weight in the sensibility value.

4.4.1. Comparison with HMM model

In order to better evaluate the performance of the developed distances algorithm, it is important and challenging to see how one of the most used state-of-art algorithms in finding CpG islands (the HMM model described in 2.2.1 section) performs.

This model was applied to the same DNA data and the results, once again, were compared to the Takai and Jones definition.

For the chromosomes of the *Homo sapiens*, the results were:

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	95,68	98,93	95,66
2	96,79	98,30	96,78
3	97,56	99,12	97,55
4	97,43	98,59	97,42
5	97,24	99,04	97,23
6	97,21	99,25	97,19
7	95,71	98,36	95,70
8	96,60	97,44	96,60
9	94,83	98,61	94,80
10	96,18	98,79	96,16
11	95,26	98,71	95,23
12	96,16	98,76	96,15
13	97,22	98,09	97,21
14	95,89	99,40	95,86
15	96,25	99,66	96,23
16	91,84	98,63	91,76
17	91,02	99,07	90,89
18	97,03	98,42	97,03
19	85,04	99,22	84,68
20	93,96	98,61	93,91
21	93,67	98,45	93,63
22	87,39	98,19	87,23
X	97,89	99,10	97,88
Y	98,07	96,92	98,07

Table 9 - Performance of the HMM model for each chromosome of the *Homo sapiens.*

Analyzing the results for the *Homo sapiens*, it is possible to verify that while the HMM model has better sensibility than the distances algorithm, but it has poorer accuracy and specificity (Tables 6 and 9).

The results, for the two eukaryotes, are shown in tables 10 and 11:

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	98,91	80,00	99,12
2	99,20	*	99,20
3	99,37	*	99,37
4	99,12	*	99,12
5	98,87	50,00	98,96
6	98,52	100,00	98,52
7	99,36	100,00	99,36
8	99,73	*	99,73
9	99,20	100,00	99,20
10	98,99	*	98,99
11	99,10	*	99,10
12	98,65	100,00	98,65
13	99,03	100,00	99,03
14	99,43	100,00	99,43
15	98,85	100,00	98,85
16	99,05	*	99,05

Table 10 - Performance of the HMM model for each chromosome of *Saccharomyces cerevisiae.*

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	90,69	79,49	91,84
2	97,21	75	97,67
3	97,16	100	97,14
4	97,94	66,67	98,59
5	97,16	71,43	97,59
6	97,95	100	97,95
7	97,13	100	97,09
8	95,59	71,43	96,32
9	97,80	73,33	98,56
10	97,71	94,12	97,83
11	97	90	97,14

Table 11 - Performance of the HMM model for each chromosome of *Encephalitozoon cuniculi.*

In the case of the *Saccharomyces cerevisiae*, the performance of the two algorithms was similar, although there were some worst results of the HMM model in terms of sensibility, visible on chromosomes 1 and 5 (Tables 7 and 10).

Finally, in relation to the *Encephalitozoon cuniculi* both the developed algorithm and the HMM had some problems in terms of sensibility, because of the reason explained before in relation to this species, which is worse in the CpG distances algorithm, having a similar good performance in the other two parameters (Tables 8 and 11).

5. Conclusion

Firstly, the inter-nucleotide distances revealed that the inter-symbolic distances algorithm could be very important in the characterization of DNA sequences, as well as for finding relevant patterns that are distinguishing characteristics of each species [1].

This work aimed at exploring new possible applications for the inter-symbolic distances, and its capability of discriminating coding and non coding regions and detecting CpG islands.

Thus, an exploratory study was carried out in order to analyze the behavior of the distribution of the distances between stop codons in coding and non coding regions, in three different reading frames. As expected, in one frame of the coding regions, there was only one stop symbol per coding region, and its distance represented the length of that gene.

This behavior was the basis of an algorithm to discriminate coding and non coding regions [36], and as it was not possible, in this work, to successfully adjust it to improve its performance, this algorithm was evaluated. The performance results showed that there is still room for improvement. Neverthless, it has revealed that the inter-symbolic distances can also play an important role in finding coding regions.

Moreover, a study was also conducted to analyze some characteristics of the CG dinucleotide distances distribution in CpG islands and in non CpG islands. From the results of this study, it was possible to verify that the percentage of short distances in the segments from CpG islands tend to be higher than in other segments, representing an important indicator that the inter-symbolic distances can be important to find these regions. Therefore, the inter-symbolic distances were used to develop an algorithm to find CpG islands.

From the results of the developed CpG distances algorithm, it is possible to state that the developed algorithm constitutes an important tool to detect CpG islands and performing in some cases better than the Hidden Markov Model.

Finally, the tools developed in MATLAB provide a contribution to the processing of genomic signals, which was another of the objectives of this work.

5.1. Future work

As it has not been possible yet to apply the presented algorithm [36] to discriminate coding and non coding regions to more complex organisms, such as the *Homo sapiens*, possibly because of the presence of introns, this represents an important future challenge in the study of the inter-symbolic distances.

Moreover, other possible application of the inter-symbolic distances may be finding transposons elements, and this subject may constitute another future challenge.

6. References

1. Afreixo, V., et al., *Genome analysis with inter-nucleotide distances.* Bioinformatics, 2009. **25**(23): p. 3064-3070.
2. Mendel, G. *Experiments in Plant Hybridization* 1865; Available from: http://www.biologie.uni-hamburg.de/b-online/e08_mend/mendel.htm.
3. *Gregor Mendel*. Available from: http://www.biography.com/people/gregor-mendel-39282.
4. Dahm, R., *Discovering DNA: Friedrich Miescher and the early years of nucleic acid research.* Hum Genet, 2008. **122**(6): p. 565-81.
5. Klug, A., *The Discovery of the DNA Double Helix.* Journal of Molecular Biology, 2004. **335**(1): p. 3-26.
6. Watson, J.D. and F. Crick, *Molecular structure of nucleic acids.* Nature, 1953. **171**: p. 737–738.
7. Franklin, R. *Photo 51*. 1952 Available from: http://www.pbs.org/wgbh/nova/body/DNA-photograph.html.
8. *Watson and Crick with their DNA model*. 1953; Available from: http://www.nobelweekdialogue.org/?attachment_id=364.
9. Elson, D. and E. Chargaff, *On the desoxyribonucleic acid content of sea urchin gametes.* Experientia, 1952. **8**(4): p. 143-145.
10. *DNA structure*. Available from: http://manipulacaogeneticaemlinha.blogspot.pt/2010/05/estrutura-do-adn.html.
11. *Gene* Available from: http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/Images/gene.gif.
12. Life, B.f.; Available from: http://www.biologyforlife.com/d-evolution.html.
13. Pevzner, P., *Computational Molecular Biology: An Algorithmic Approach*2000: MIT Press.
14. Szpankowski, W., W. Ren, and L. Szpankowski, *An optimal DNA segmentation based on the MDL principle.* Int J Bioinform Res Appl, 2005. **1**(1): p. 3-17.
15. V.Braun , J. and H.-G. Müller *Statistical Methods for DNA Sequence Segmentation.* Statistical Science, 1998. **13**.
16. Elhaik E Fau - Graur, D., K. Graur D Fau - Josic, and K. Josic, *Comparative testing of DNA segmentation algorithms using benchmark simulations.* (1537-1719 (Electronic)).
17. Li, W., et al., *Applications of recursive segmentation to the analysis of DNA sequences.* Computers & Chemistry, 2002. **26**(5): p. 491-510.
18. Marioni, J.C., N.P. Thorne, and S. Tavaré, *BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data.* Bioinformatics, 2006. **22**(9): p. 1144-1146.
19. Schwartzkopf, W.C., A.C. Bovik, and B.L. Evans, *Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images.* Medical Imaging, IEEE Transactions on, 2005. **24**(12): p. 1593-1610.
20. Nicorici, D. and J. Astola, *Segmentation of DNA into Coding and Noncoding Regions Based on Recursive Entropic Segmentation and Stop-Codon Statistics.* EURASIP Journal on Advances in Signal Processing, 2004. **2004**(1): p. 832471.

21. Bernaola-Galván, P., et al., *Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method.* Physical Review Letters, 2000. **85**(6): p. 1342-1345.

22. Eskesen, S., et al., *Periodicity of DNA in exons.* BMC Molecular Biology, 2004. **5**(1): p. 12.

23. Tiwari, S., et al., *Prediction of probable genes by Fourier analysis of genomic sequences.* Computer applications in the biosciences : CABIOS, 1997. **13**(3): p. 263-270.

24. Vaidyanathan, P.P. and B.-J. Yoon, *The role of signal-processing concepts in genomics and proteomics.* Journal of the Franklin Institute, 2004. **341**(1–2): p. 111-135.

25. Akhtar, M., J. Epps, and E. Ambikairajah, *Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction.* Selected Topics in Signal Processing, IEEE Journal of, 2008. **2**(3): p. 310-321.

26. Wu, H., et al., *Redefining CpG islands using hidden Markov models.* Biostatistics, 2010. **11**(3): p. 499-514.

27. Gardiner-Garden, M. and M. Frommer, *CpG Islands in vertebrate genomes.* Journal of Molecular Biology, 1987. **196**(2): p. 261-282.

28. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proceedings of the National Academy of Sciences, 2002. **99**(6): p. 3740-3745.

29. Durbin, R., et al., *Biological sequence analysis (Chapter 3)*1998: Cambridge University Press.

30. Hui-Ling, L., *Implementing the Viterbi algorithm.* Signal Processing Magazine, IEEE, 1995. **12**(5): p. 42-52.

31. Nair, A.S.S. and T. Mahalakshmi, *Visualization of genomic data using inter-nucleotide distance signals.* Proceedings of IEEE Genomic Signal Processing., 2005.

32. Abbasi, O., A. Rostami, and G. Karimian, *Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform.* BMC Bioinformatics, 2011. **12**(1): p. 430.

33. Wei, W. and D.H. Johnson, *Computing linear transforms of symbolic signals.* Signal Processing, IEEE Transactions on, 2002. **50**(3): p. 628-634.

34. Tsonis, A.A., et al., *Wavelet analysis of DNA sequences.* Physical Review E, 1996. **53**(2): p. 1828-1834.

35. Deng, S., et al., *Detecting the borders between coding and non-coding DNA regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics.* BMC Genomics, 2012. **13**(Suppl 8): p. S19.

36. Bastos, C.C., et al., *Segmentation of DNA into Coding and Noncoding Regions Based on Inter-STOP Symbols Distances*, in *7th International Conference on Practical Applications of Computational Biology & Bioinformatics*, M.S. Mohamad, et al., Editors. 2013, Springer International Publishing. p. 23-28.

37. Information, N.C.f.B. *Release name: NCBI Homo sapiens Annotation Release 105*. 2013 Release date: 13 August 2013; Available from: ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/.

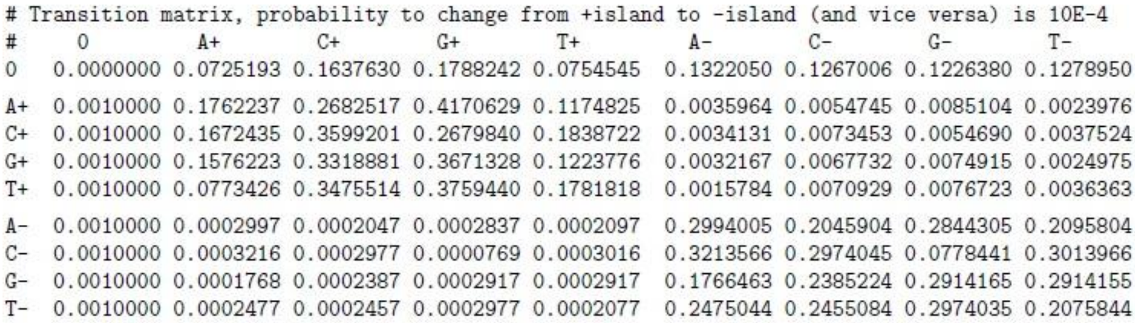
38. Institute, E.B. *Saccharomyces cerevisiae*. 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:4932.

39. Institute, E.B. *Encephalitozoon cuniculi*. 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:6035.

40. Institute, E.B. *Bifidobacterium asteroides*. 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:1684.
41. Institute, E.B. *Haemophilus influenzae*. 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:727.
42. Institute, E.B. *Thermotoga maritima* 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:2336.
43. Institute, E.B. *Aeromonas phage 65*. 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:260149.
44. Institute, E.B. *Calliarthron tuberculosum*. 2013; Available from: http://www.ebi.ac.uk/ena/data/view/Taxon:48942.

7. Appendix

7.1. Appendix A

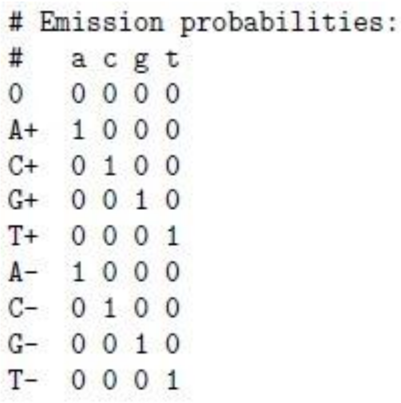


```
# Transition matrix, probability to change from +island to -island (and vice versa) is 10E-4
#     0         A+        C+        G+        T+         A-        C-        G-        T-
0   0.0000000 0.0725193 0.1637630 0.1788242 0.0754545  0.1322050 0.1267006 0.1226380 0.1278950

A+  0.0010000 0.1762237 0.2682517 0.4170629 0.1174825  0.0035964 0.0054745 0.0085104 0.0023976
C+  0.0010000 0.1672435 0.3599201 0.2679840 0.1838722  0.0034131 0.0073453 0.0054690 0.0037524
G+  0.0010000 0.1576223 0.3318881 0.3671328 0.1223776  0.0032167 0.0067732 0.0074915 0.0024975
T+  0.0010000 0.0773426 0.3475514 0.3759440 0.1781818  0.0015784 0.0070929 0.0076723 0.0036363

A-  0.0010000 0.0002997 0.0002047 0.0002837 0.0002097  0.2994005 0.2045904 0.2844305 0.2095804
C-  0.0010000 0.0003216 0.0002977 0.0000769 0.0003016  0.3213566 0.2974045 0.0778441 0.3013966
G-  0.0010000 0.0001768 0.0002387 0.0002917 0.0002917  0.1766463 0.2385224 0.2914165 0.2914155
T-  0.0010000 0.0002477 0.0002457 0.0002977 0.0002077  0.2475044 0.2455084 0.2974035 0.2075844
```

Figure 26 - Transition matrix for HMM applied to CpG islands [29].



```
# Emission probabilities:
#   a c g t
0   0 0 0 0
A+  1 0 0 0
C+  0 1 0 0
G+  0 0 1 0
T+  0 0 0 1
A-  1 0 0 0
C-  0 1 0 0
G-  0 0 1 0
T-  0 0 0 1
```

Figure 27 - Emission matrix for HMM applied to CpG islands [29].

7.2. Appendix B

In this appendix are displayed the results of the distribution of the stop codons distances in coding and non coding regions in the three different frames. In each figure, the three graphics on the top correspond to non coding regions and the three below correspond to coding regions.

Moreover, on the left are the graphics corresponding to reading frame 1, on the center the graphics of the reading frame 2 and, finally, on the right the graphics of the reading frame 3.

Figure 28 shows the distribution for the *Bifidobacteruim asteroides:*

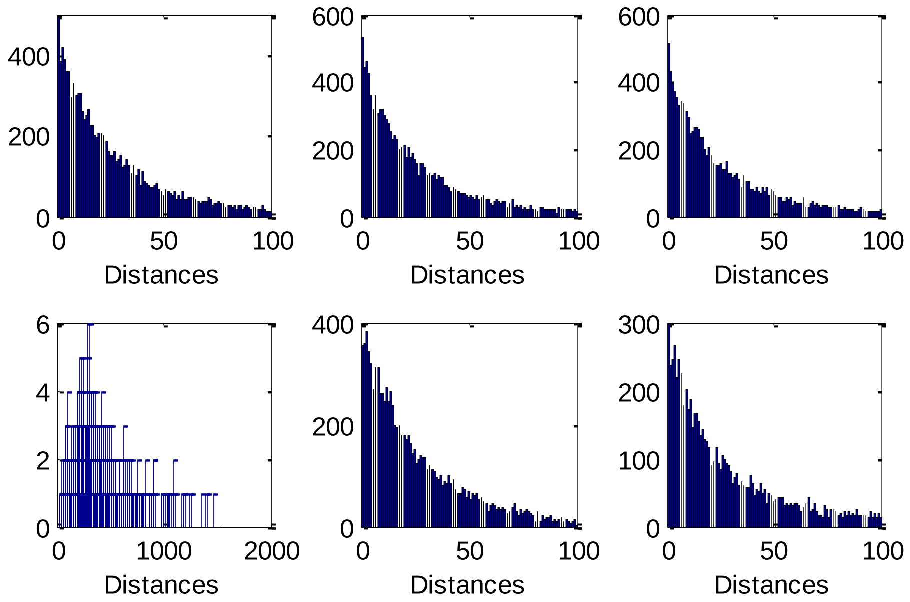


Figure 28 - Distribution of the stop codons distances in coding and non coding regions in the three different frames.

In figure 29 are the results for the *Haemophilus influenzae*:

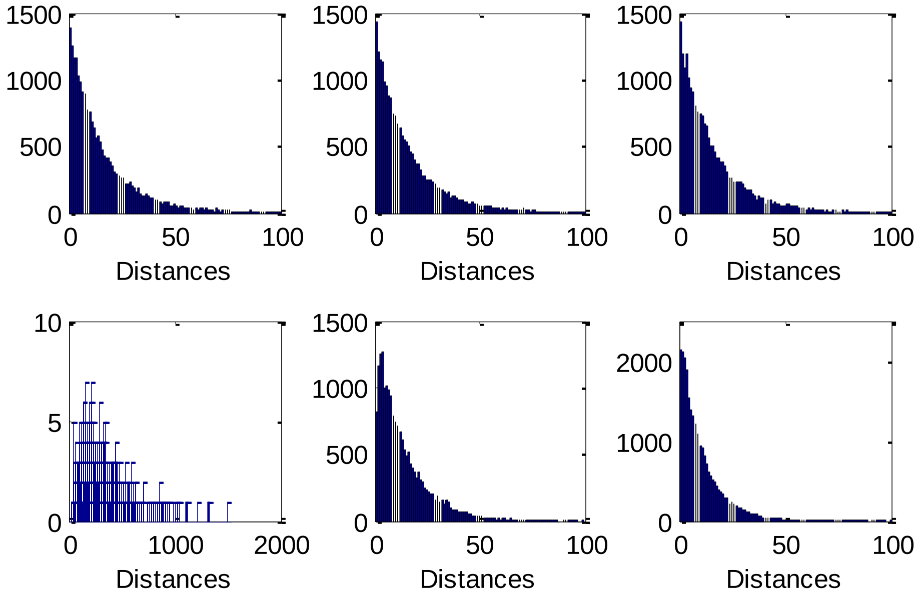


Figure 29 - Distribution of the stop codons distances in coding and non coding regions in the three different frames.

Figure 30 shows the results for the *Calliarthtron tuberculosoum*:

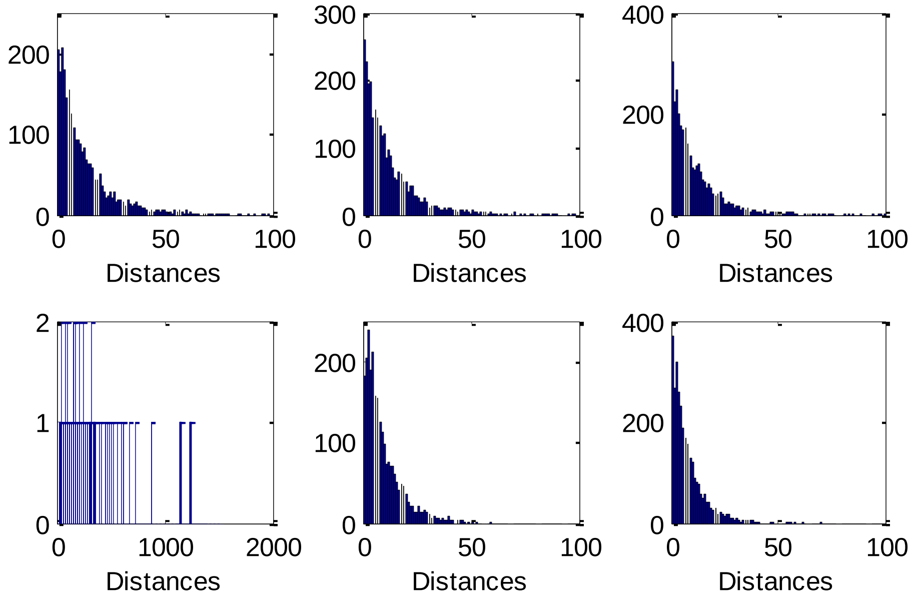


Figure 30 - Distribution of the stop codons distances in coding and non coding regions in the three different frames.

The results for the *Thermotoga maritima* are displayed in figure 31:

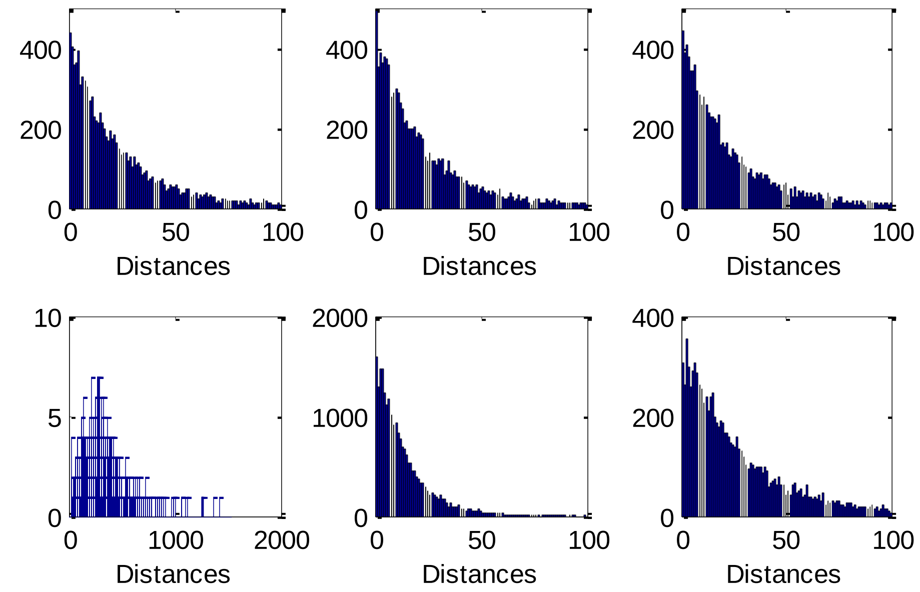


Figure 31 - Distribution of the stop codons distances in coding and non coding regions in the three different frames.

For the eukaryota *Sacccharomyces cerevisae* the results are shown in figure 32:

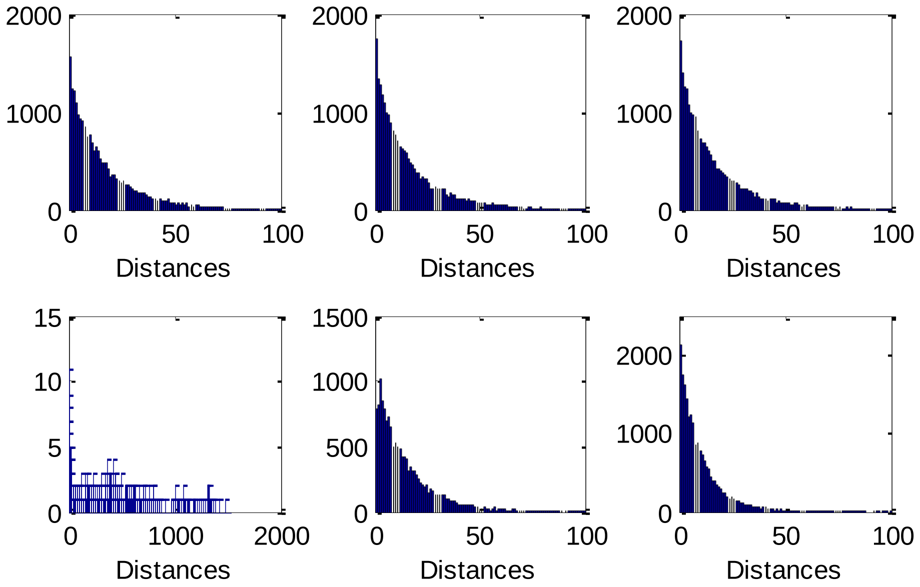


Figure 32 - Distribution of the stop codons distances in coding and non coding regions in the three different frames.

Figure 33 shows the results for the *Encephalitozoon cuniculi:*

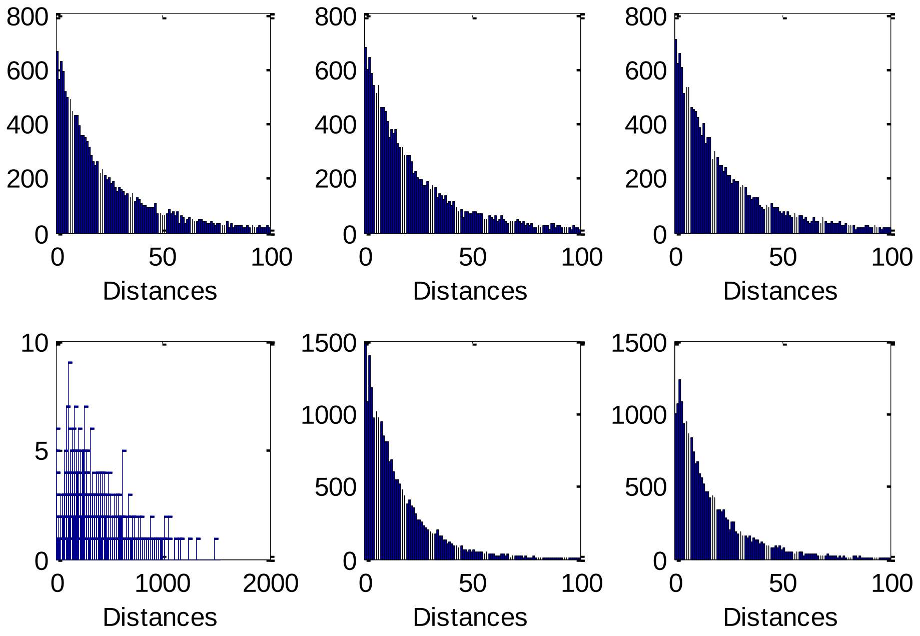


Figure 33 - Distribution of the stop codons distances in coding and non coding regions in the three different frames.

Finally, the maximum and minimum length of a gene in reading frame 1 of coding regions for each species was also determined and the results are displayed on table 12:

Species	Maximum	Minimum
Bifidobacteruim asteroides	1829	32
Haemophilus influenzae	1493	31
Calliarthtron tuberculosoum	1241	31
Thermotoga maritima	1691	31
Sacccharomyces cerevisae	2489	15
Encephalitozoon cuniculi	2410	49

Table 12 - Maximum and minimum length of a gene in reading frame 1 of coding regions for each species.

7.3. Appendix C

Figure 34 and 35 shows the distribution of the inter-CG distances for the *Saccharomyces cerevisae* genome in CpG islands and non CpG islands:

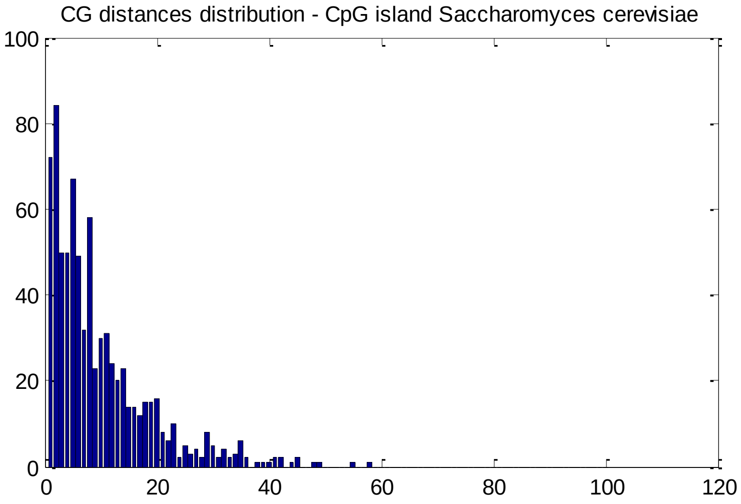


Figure 34 - Distribution of CG distances in CpG islands of the *Saccharomyces cerevisae* genome (the percentage of short distances is 58.85%, δ = 8).

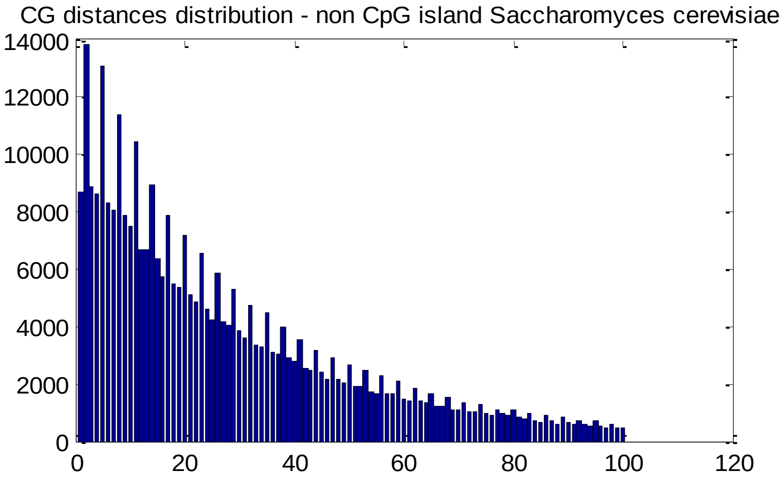


Figure 35 - Distribution of CG distances in non CpG islands of the *Saccharomyces cerevisae* genome (the percentage of short distances is 24.22%, δ = 8).

Figures 36 and 37 shows the results for the *Encephalitozoon cuniculi:*

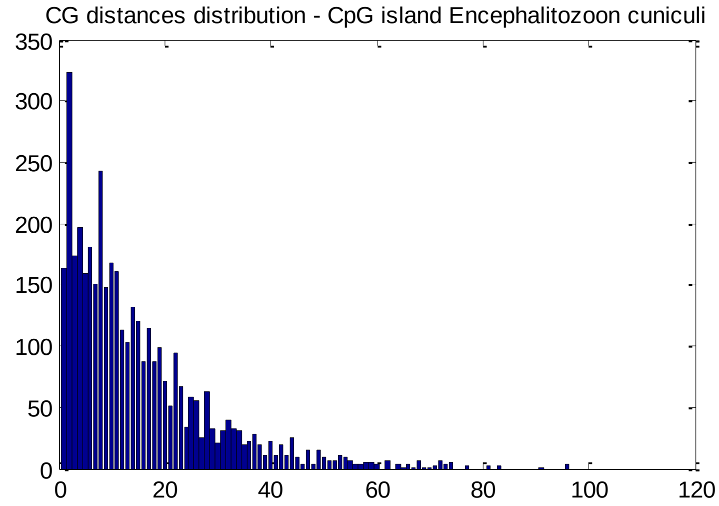


Figure 36 - Distribution of CG distances in non CpG islands of the *Encephalitozoon cuniculi* genome (the percentage of short distances is 49.93%, δ = 8).

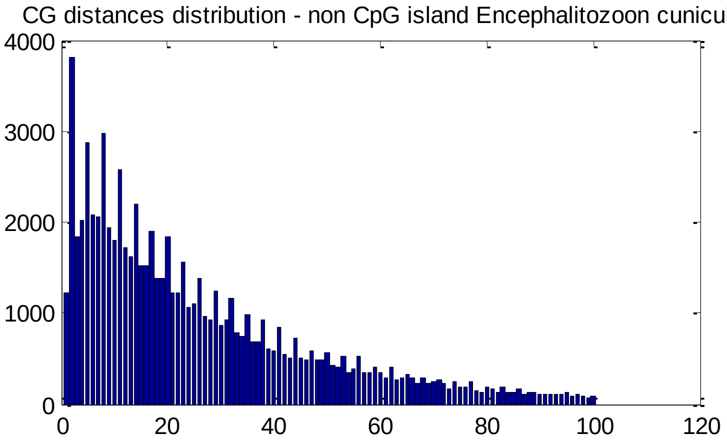


Figure 37 - Distribution of CG distances in non CpG islands of the *Encephalitozoon cuniculi* genome (the percentage of short distances is 24.45%, δ = 8).

7.4. Appendix D

The results for each chromosome are presented in the next table:

Chromosome	Accuracy (%)	Sensibility (%)	Specificity (%)
1	92,69	75,14	92,83
2	92,20	73,00	92,31
3	91,67	74,95	91,75
4	90,20	71,01	90,28
5	91,40	72,84	91,50
6	91,50	72,56	91,61
7	92,51	71,26	92,65
8	92,20	74,55	92,30
9	92,95	74,97	93,08
10	93,57	72,55	93,71
11	92,40	74,87	92,52
12	92,36	69,54	92,50
13	91,19	70,48	91,29
14	92,42	75,77	92,53
15	94,23	78,19	94,35
16	95,75	72,47	96,02
17	95,80	74,62	96,13
18	92,54	70,96	92,66
19	96,07	69,79	96,74
20	94,95	77,23	95,11
21	93,39	70,07	93,55
22	97,10	76,44	97,39
X	91,04	68,34	91,13
Y	88,62	52,31	88,69

Table 13 - Performance of the first approach of the distances algorithm for each chromosome of the *Homo Sapiens.*

8. Glossary

CHROMOSSOME - organized structure of DNA and protein found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences.

GENE - the functional units of chromosomes, corresponding to DNA fragments each formed by a specific sequence of nitrogenous bases and with a specific mission: encoding the information required for the synthesis of a protein.

NUCLEOTIDES - are biological molecules that form the building blocks of nucleic acids (DNA and RNA) constituted by a pentose, a phosphate group and a nitrogenous base (adenine (A), guanine (G), thymine (T), cytosine (C), uracil (U)).

PURINE - two fused rings of carbon and nitrogen atoms, one ring has six members and the other has five, each with two nitrogen. A group of important elements are derived from purines, including adenine and guanine.

PYRIMIDINE - a crystalline organic base that is the parent substance of various biologically important derivatives like cytosine, thymine and uracil, having a single six-member ring in which the first and third atoms are nitrogen and the rest are carbon.

AMINO ACID - the building block of proteins, containing an acid functional group and an amine functional group on adjacent carbon atoms in which each is coded for by a codon and linked together through peptide bonds (bond between the carboxyl group of one amino acid to the amino group of the other amino acid).

CODONS - A sequence of three adjacent nucleotides constituting the genetic code that determines the insertion of a specific amino acid in a polypeptide chain during protein synthesis or the signal to stop/start protein synthesis. Possible stop codons in ADN are "TGA", "TAA" and "TAG". The most common start codon is "ATG".

GENOME - the complete set of hereditary information present in an organism where all the information for its construction and operation is contained.

JUNK DNA/NONCODING DNA – describes components of an organisms DNA sequences that do not encode proteins. Much of this DNA has no known biological function, however many such sequences serve to regulate transcription of protein coding sequences.

EXONS AND INTRONS - The genes contain regions which encode proteins called exons. These regions are interrupted, in some genomes, by sequences which are not used for coding, introns.

Exons are constituted by a sequence that is not translated (Untranslated Region) and one that contains the code for a particular amino acid (Coding DNA Sequence).