16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

DETECTION OF STATISTICAL PERIODICITIES IN DNA BY CONFLICT AND ENTROPY MINIMIZATION METHODS

Nuno Monteiro, João Gomes and João Xavier

Instituto Superior Técnico − Instituto de Sistemas e Robótica, Technical University of Lisbon
Av. Rovisco Pais, 1049-001 Lisbon, Portugal
pirlruc@gmail.com, {jpg, jxavier}@isr.ist.utl.pt

ABSTRACT

Some commonly used methods for statistical analysis of DNA sequences start by mapping each nucleotide to a real or complex number and, subsequently, applying classical methods of spectral analysis to the numerical sequences obtained. Such mappings may introduce inexistent spectral peaks or attenuate others whose presence is known. Consequently, recent approaches try to reduce the dependence of DNA spectral estimation on the chosen mapping through concepts of optimization and information theory. In this paper we examine the advantages of using complex mappings combined with convex optimization algorithms (Conflict and Entropy Minimization Methods) on the selection of appropriate base values for detecting statistical periodicities in DNA. These methods are applied to real DNA sequences extracted from public databases (Genbank and Ensembl).

1. INTRODUCTION

Deoxyribonucleic acid (DNA) is responsible for differences detected among individuals of different species, or even of the same species, given that within the molecules of DNA resides the biological information for protein synthesis and regulation that defines the biochemistry of an organism, determining its characteristics.

DNA molecules consist of two complementary and antiparallel strands forming a double helix. Each strand is a sequence of monomers consisting of a pentose (deoxyribose), a phosphate group and a nitrogenous base; these monomers are designated nucleotides (or simply *bases*) and are named in compliance with their nitrogenous base. In the case of DNA, there are four possible nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T).

The genome of an organism consists of DNA molecules present in all cells (except mitochondrial and chloroplast DNA) or, in the case of some viral genomes, consists of RNA molecules. It is composed by sequences of DNA that are transcribed (genes) and intergenic regions (sequences of DNA between genes). On the other hand, genes have introns (except in prokaryotes), regions that are removed from the primary RNA transcript, and exons, regions that are present in the primary RNA transcript and mature RNA molecules. In the current paper, we will be interested in identifying exons, which can contain part of the open reading frame that codes for a specific portion of the complete protein and/or untranslated sequences. This work also aims at uncovering other periodicities in DNA, such as those that enable it to assume three-dimensional condensed forms in some organisms [1, 2].

An overview of some of the statistical methods that have been used for DNA sequence analysis is given in [7]. Our work is concerned with a particular class of methods where the sequence of nucleotides is mapped to a numerical sequence and then analyzed. Prior work in this problem includes the popular work of Anastassiou [3], where the numeric value of one of the bases was arbitrarily set to zero and the remaining ones then optimized for exon detection. In [4] mappings for DNA spectral analysis were found based on entropy minimization by exhaustive search over a grid of real base values. In the present paper we try to improve upon those methods by assuming general complex mappings without assigning *a priori* specific values to any of the nucleotides. We adopt the framework of convex optimization to efficiently find globally or locally optimal mappings without the need for exhaustive search.

2. NUMERICAL REPRESENTATION

From the standpoint of sequence analysis, DNA can be represented as a symbolic sequence of an alphabet with four letters

$$\mathbf{b} = (b_k)_{0 \le k < N},\ b_k \in \{A, C, G, T\}. \qquad (1)$$

Analysis of DNA can be carried out either at the symbolic level, or by mapping the bases to complex/real numbers and then resorting to well-established statistical tools. As mentioned, the mapping of each base to a number is important in spectral analysis for identifying intrinsic periodicities of the nucleotide sequence; however, it can also introduce some non-characteristic spectral content or attenuate other known periodicities of the sequence being studied [4, 5]. Ideally, the mathematical properties of the adopted mapping should reflect the biological properties of the polynucleotide sequence, and, simultaneously, it should be simple in order to be computationally efficient and generate results susceptible of interpretation [5].

The conversion of a nucleotide sequence of length N to a numeric sequence can be achieved by using the Voss representation based on four indicator sequences [6, 7]. These indicator sequences correspond to a binary representation, in which the number 1 identifies the presence and the number 0

This work was supported by Fundação para a Ciência e a Tecnologia through ISR/IST plurianual funding.

identifies the absence of a nucleotide *i* at the *m*-th position in a DNA fragment

$$\mathbf{u}_i(m)=\begin{cases}1, & b_m=i\\ 0, & b_m\neq i,\end{cases}\qquad i\in\{A,C,G,T\}. \tag{2}$$

The definition of indicator sequences implies that their sum is equal to one in any position *m*, therefore the four binary sequences are redundant[1]. This gives rise to another property, associated with their DFT

$$\sum_{i=\{A,C,G,T\}}\mathbf{u}_i(n)=1 \quad\Rightarrow\quad \sum_{i=\{A,C,G,T\}}\mathbf{U}_i(k)=\begin{cases}0, & k\neq 0\\ N, & k=0,\end{cases} \tag{3}$$

where

$$\mathbf{U}_i(k)=\sum_{n=0}^{N-1}\mathbf{u}_i(n)\,\mathrm{e}^{-j\frac{2\pi}{N}kn},\quad 0\le k\le N-1. \tag{4}$$

Upon adopting the Voss representation, it is possible to compute the spectral content of a given sequence without specifying any type of mapping

$$\mathbf{S}(k)=\sum_{i=\{A,C,G,T\}}\left|\mathbf{U}_i(k)\right|^2. \tag{5}$$

Nevertheless, there are other ways of representing the nucleotides of the strand being analyzed [7]. One possibility is to associate a given numeric mapping $\boldsymbol{\theta}=[\theta_A\ \theta_C\ \theta_G\ \theta_T]^{\mathrm{T}}$ to the nucleotides, obtaining

$$\mathbf{w}(n)=\theta_A\mathbf{u}_A(n)+\theta_C\mathbf{u}_C(n)+\theta_G\mathbf{u}_G(n)+\theta_T\mathbf{u}_T(n), \tag{6}$$

$$\mathbf{W}(k)=\theta_A\mathbf{U}_A(k)+\theta_C\mathbf{U}_C(k)+\theta_G\mathbf{U}_G(k)+\theta_T\mathbf{U}_T(k). \tag{7}$$

This method is prone to introducing misleading information (artifacts) in the numeric sequence, so when adopting a mapping (with the goal of finding or enhancing some intrinsic periodicity) we should resort to optimization methods. That is the rationale behind the minimum conflict and minimum entropy methods proposed below.

3. IDENTIFICATION OF CODING REGIONS

Typically, exons have a statistical periodicity of three bases that is not found in other regions of DNA molecules (in prokaryotes, this periodicity is also present in non-coding regions). This periodicity is due to the genetic code structure [5,8], which is formed by words of three consecutive nucleotides (triplets). The code can be interpreted as a correspondence between the four nucleotide alphabet of DNA and the twenty amino acid alphabet of proteins.

The nucleotides in the exonic regions can be seen as a set of triplets (codons), each encoding a given amino acid (except the STOP codon). Consequently, the concatenation of the various codons in exons comprises a sequence that encodes a specific polypeptide chain. Coding sequences are rich in C and G, and intergenic regions and introns (non-coding sequences) are rich in A and T [1, 2].

The statistical periodicity present in exonic regions translates into a peak on the spectral content at DFT index $k=L/3$ or $k=2L/3$ (DFT symmetry property), L being the length of the window on which the spectral content is computed (a recommended value of 351 is used [8]). The minimum conflict method is a training-based scheme where the mapping attempts to maximize the contribution of coding regions to the spectral content at this frequency, while minimizing that of non-coding regions. Specifically, let $\{\mathbf{x}_i\}$, $i\in\mathbf{I}$, and $\{\mathbf{y}_j\}$, $j\in\mathbf{J}$, denote two sets of 4×1 vectors containing the DFT values of the indicator sequences at index $k=L/3$.

- Vectors $\{\mathbf{x}_i\}$ are extracted from known exonic regions, **I**, of a given organism genome (in our case, chromosome XVI or XIII of *Saccharomyces cerevisiae*) that present high homology to an organism of interest. The optimal mapping is then used to detect exons in the latter organism (in our case, CG2009 of the *Drosophila melanogaster* chromosome 4, and *Caenorhabditis elegans* F56F11 cosmid).
- Vectors $\{\mathbf{y}_j\}$ are either extracted from non-coding regions (intergenic and intronic), **J**, of the same reference genome used for $\{\mathbf{x}_i\}$, or they are generated from a random (uniform) sequence of nucleotides.

The desired mapping $\boldsymbol{\theta}$ should make $\left|\mathbf{x}_i^{\mathrm{H}}\boldsymbol{\theta}\right|^2$ large and $\left|\mathbf{y}_j^{\mathrm{H}}\boldsymbol{\theta}\right|^2$ small, where the superscript $(.)^{\mathrm{H}}$ denotes complex conjugate transpose (Hermitean). To formulate the optimization problem, it is useful to work with the positive semidefinite complex matrix $\boldsymbol{\Theta}=\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}$, rather than with $\boldsymbol{\theta}$ directly, and impose a rank 1 constraint on it.

From a pattern classification perspective, $\boldsymbol{\Theta}$ defines a separating surface between the sets $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$, and it is to be chosen so that the number of violations to that surface is minimized. Ideally, we wish to have

$$\left|\mathbf{x}_i^{\mathrm{H}}\boldsymbol{\theta}\right|^2\ge 1,\quad \left|\mathbf{y}_j^{\mathrm{H}}\boldsymbol{\theta}\right|^2=0. \tag{8}$$

Deviations from the above condition are accounted for by positive coefficients u_i and v_j, whose sum is to be minimized. The optimization program is thus

$$\begin{aligned}&\min\left(\sum_{i\in\mathbf{I}}u_i+\sum_{j\in\mathbf{J}}v_j\right)\\ &\mathbf{x}_i^{\mathrm{H}}\boldsymbol{\Theta}\mathbf{x}_i\ge 1-u_i\\ &\mathbf{y}_j^{\mathrm{H}}\boldsymbol{\Theta}\mathbf{y}_j\le -1+v_j\\ &\boldsymbol{\Theta}\ge 0\\ &\operatorname{rank}\boldsymbol{\Theta}=1.\end{aligned} \tag{9}$$

To obtain a convex optimization problem, which can be efficiently solved numerically [10], the nonconvex rank constraint is dropped (relaxed). However, it was found that optimal solutions invariably exhibit near-unit rank, in which case a suitable mapping $\boldsymbol{\theta}$ is readily given by the eigenvector of $\boldsymbol{\Theta}$ associated with the largest eigenvalue. After determining the mapping $\boldsymbol{\theta}$ from each test set considered we compute the spectral content at index $k=L/3$, using a sliding window of length L = 351, and progress with a step of one nucleotide over a given DNA fragment of the organism being analyzed.

The minimum conflict method (MCM) is an improvement of Anastassiou's approach to identify coding regions of DNA, since in his work it is assumed that one of the nucleotides is mapped to zero [3], while in the method presented

[1] They can be reduced to three equivalent sequences considering that each nucleotide is associated with the vertex of a regular tetrahedron in three-dimensional space [3, 6, 7].

here the complex mapping is determined without arbitrarily fixing the value assigned to any of the nucleotides.

4. DETECTION OF PERIODICITIES IN DNA

In order to detect other periodicities in the polynucleotide chains of double-stranded DNA, besides the statistical periodicity of three bases, we adopted a non-training-based approach proposed by Galleani et al [4]. The goal is to find a mapping that minimizes the entropy associated with a power spectrum, computed in a segment of the polynucleotide sequence, which is regarded as an unnormalized probability mass function. Entropy minimization finds a mapping that leads to a "peaky" power spectrum, i.e., one where energy is concentrated around a number of significant peaks that, hopefully, reveal intrinsic properties of the symbolic sequence. We developed a formulation whereby a locally optimal complex mapping is found in four-dimensional complex space by iteratively minimizing a (non-convex) constrained cost function, thus improving on Galleani's work in which the real mapping with minimum entropy is obtained by exhaustive search in a limited bidimensional space.

We estimate the power spectrum of a given complex sequence by computing the periodogram, i.e., the squared magnitude of the Fourier transform (7)

$$\mathbf{P}(k,\boldsymbol{\theta}) = \left|\mathbf{W}(k)\right|^2, \quad \mathbf{P}(\boldsymbol{\theta}) = \left[\mathbf{P}(0,\boldsymbol{\theta}) \ldots \mathbf{P}(N-1,\boldsymbol{\theta})\right], \quad (10)$$

where the dependence on the mapping $\boldsymbol{\theta}$ is explicitly noted. We adopt the standard definition of entropy (in bits) for a set of N nonnegative values $p_0 \ldots p_{N-1}$ that sum to unity [4, 9]

$$H(p_0 \ldots p_{N-1}) = \sum_{i=0}^{N-1} H(p_i), \quad H(p_i) = -p_i \log_2 p_i . \quad (11)$$

Accordingly, we consider a normalized version of the power spectrum (unitary l_1 norm, or unit signal energy) in the optimization method, and impose an additional constraint at index $i = 0$ to avoid the trivial solution for $\mathbf{P}(\boldsymbol{\theta})$ given by a single impulse that follows from a constant map. The desired $\boldsymbol{\theta}$ then satisfies

$$\min\left(\sum_{i=0}^{N-1} H\left(\frac{\mathbf{P}(i,\boldsymbol{\theta})}{\mathbf{1}^{\mathrm{T}}\mathbf{P}(\boldsymbol{\theta})}\right)\right) \quad (12)$$
$$\mathbf{P}(0,\boldsymbol{\theta}) = 0 .$$

Above, $\mathbf{1}$ denotes an $N\times 1$ vector of 1s. According to (10), the power spectrum at index i is given by

$$\mathbf{P}(i,\boldsymbol{\theta}) = \left|\mathbf{a}_i^{\mathrm{H}}\boldsymbol{\theta}\right|^2 = \mathrm{tr}\left(\mathbf{a}_i\mathbf{a}_i^{\mathrm{H}}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}\right) = \mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}\right), \quad (13)$$

where $\mathbf{a}_i$ is a 4×1 vector containing the DFT values of the indicator sequences at index i and $\mathbf{A}_i = \mathbf{a}_i\mathbf{a}_i^{\mathrm{H}}$. We can use a positive semidefinite complex Hermitean matrix $\boldsymbol{\Theta} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{H}}$ in order to eliminate the quadratic dependence on $\boldsymbol{\theta}$. Similarly, the denominator in the evaluation function (12) becomes $\mathbf{1}^{\mathrm{T}}\mathbf{P}(\boldsymbol{\theta}) = \mathrm{tr}(\mathbf{D}\boldsymbol{\Theta})$, where $\mathbf{D} = \sum_{i=0}^{N-1}\mathbf{A}_i$ is approximately a diagonal matrix. We now eliminate the denominator of the cost function (12) by converting it into a new constraint, yielding the equivalent program

$$\begin{aligned} &\min\left(\sum_{i=0}^{N-1} H\left(\mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}\right)\right)\right) \\ &\mathrm{tr}\left(\mathbf{A}_0\boldsymbol{\Theta}\right) = 0 \\ &\mathrm{tr}\left(\mathbf{D}\,\boldsymbol{\Theta}\right) = 1 \\ &\boldsymbol{\Theta} \geq 0 \\ &\mathrm{rank}\,\boldsymbol{\Theta} = 1 . \end{aligned} \quad (14)$$

As in Section 3, we relax the optimization problem (14) by dropping the rank constraint. Still, it was found that solutions $\boldsymbol{\Theta}$ exhibit near-unit rank, so that a mapping $\boldsymbol{\theta}$ is readily obtained from an eigendecomposition of any such $\boldsymbol{\Theta}$.

The optimization problem (14) is still not in a suitable form, as the cost function is nonconvex. However, given the concave nature of entropy we can resort to a common iterative method where, at iteration k, the cost function is upper-bounded by a linear function, and the resulting convex evaluation function is then minimized to obtain $\boldsymbol{\Theta}^{k+1}$ [10]. Specifically,

$$H\left(\mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}\right)\right) \leq \alpha_i^k + \beta_i^k\left(\mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}\right) - \mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}^k\right)\right), \quad (15)$$

with $\alpha_i^k = H\left(\mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}^k\right)\right)$ and

$$\beta_i^k = \frac{d}{dx}H(x)\Big|_{x=\mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}^k\right)} = -\frac{1+\log\mathrm{tr}\left(\mathbf{A}_i\boldsymbol{\Theta}^k\right)}{\log 2} . \quad (16)$$

The terms that do not depend on $\boldsymbol{\Theta}$ are irrelevant for finding the minimizer of the linearized cost function, and are therefore neglected. In light of this linearization, at iteration k the cost function in (14) converts into

$$\min \mathrm{tr}\left(\mathbf{B}^k\,\boldsymbol{\Theta}\right), \quad \mathbf{B}^k = \sum_{i=0}^{N-1}\beta_i^k\mathbf{A}_i . \quad (17)$$

The problem is now a semidefinite program (SDP) [10], and can be solved by standard algorithms in several available optimization toolboxes. The algorithm, listed in Figure **1**, will converge to a local minimum of (14) in a few iterations. The optimization algorithm is run for several different initializations of $\boldsymbol{\Theta}$ to find the best minimizer, from which the final mapping is derived.

Minimum Entropy Method

1. Randomly compute an initial mapping $\boldsymbol{\theta}$ and use it to compute the initial power spectrum with (13).
2. Obtain the initial entropy linearization with (16) and set $k = 1$.
3. If k is equal to NITER:
   - i. Return matrix $\boldsymbol{\Theta}^k$ and verify that it has approximately rank one.

   Else:
   - ii. Use an optimization toolbox to find a matrix $\boldsymbol{\Theta}^k$ that minimizes the constrained (linearized) evaluation function.
   - iii. Use the mapping matrix $\boldsymbol{\Theta}^k$ to obtain the power spectrum
     $$\mathbf{P}\left(i,\boldsymbol{\Theta}^k\right) = \mathrm{tr}(\mathbf{A}_i\boldsymbol{\Theta}^k)$$
   - iv. Determine the linearization associated with (16).
   - v. Increment k by one and repeat point 3.

Figure 1 – Minimum entropy method pseudo-code.

5. RESULTS

5.1 Minimum Conflict Method

Mappings were determined by the method described in Section 3 and summarized in Tables **1** and **2** for the following test sets. **Coding region I:** Chromosomes XVI or XIII of *Saccharomyces cerevisiae*; **Non-coding region J:** Either a random nucleotide sequence with uniform distribution, or non-coding regions of the respective chromosomes. Once a mapping is available, we apply a sliding window (length L = 351, one-nucleotide step) to a DNA fragment of interest and compute the contribution to the peak at frequency $1/3$ (DFT index $L/3$) of its spectral content. The sequences analyzed were the *C. elegans* F56F11 cosmid (between positions 7021 and 15021) and the *D. melanogaster* CG2009 gene.

Table 1 – Mappings obtained using the chromosome XVI coding regions of *S. cerevisiae*.

Nucleo-tides	Random DNA		Non-Coding Regions	
	Mapping	Modulus	Mapping	Modulus
A	-2.30E-04 – 2.44E-21 j	2.30E-04	-6.61E-04 – 5.21E-20 j	6.61E-04
C	7.14E-05 – 2.29E-04 j	2.40E-04	-8.29E-04 + 3.98E-04 j	9.19E-04
G	-1.09E-04 + 2.64E-04 j	2.86E-04	-8.05E-04 – 9.88E-04 j	1.27E-03
T	2.39E-04 – 4.63E-05 j	2.43E-04	-1.30E-03 – 7.67E-05 j	1.31E-03

Table 2 – Mappings obtained using the chromosome XIII coding regions of *S. cerevisiae*.

Nucleo-tides	Random DNA		Non-Coding Regions	
	Mapping	Modulus	Mapping	Modulus
A	-1.80E-04 + 1.58E-21 j	1.80E-04	-9.47E-04 – 4.01E-20 j	9.47E-04
C	7.15E-05 – 1.89E-04 j	2.02E-04	-6.85E-04 + 5.99E-04 j	9.10E-04
G	-9.30E-05 + 2.43E-04 j	2.60E-04	-1.95E-03 – 2.00E-04 j	1.96E-03
T	1.97E-04 – 2.55E-05 j	1.99E-04	-1.01E-03 + 8.01E-04 j	1.29E-03

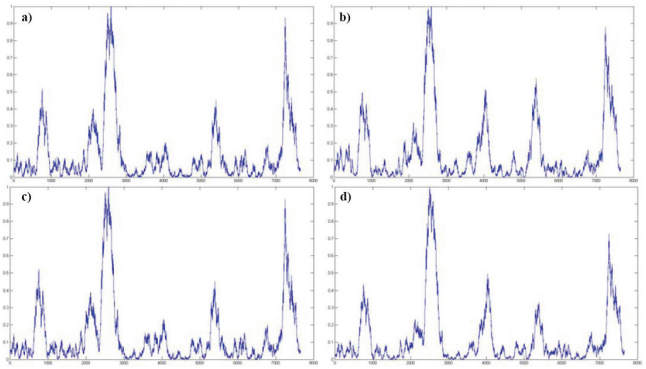

Figure 2 – Exonic regions forecast of *C. elegans*. Mappings were obtained by the minimum conflict method from *S. cerevisiae* coding regions of chromosomes XVI (top) and XIII (bottom) and **a), c)** nucleotide random sequences or **b), d)** non-coding regions of the chromossomes. Normalized spectral content is represented on the vertical axis and sequence position on the horizontal axis.

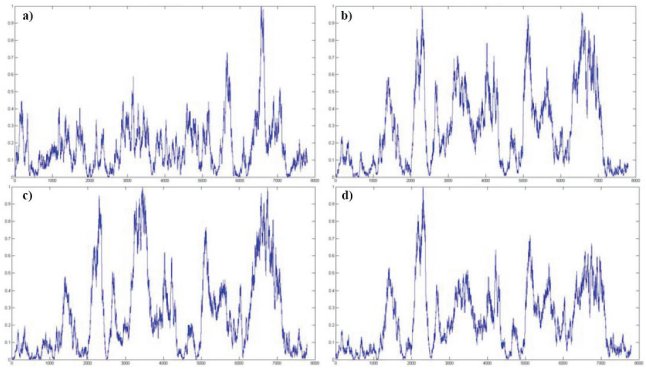

Figure 3 – Exonic regions forecast of CG2009. Mappings are the same as in Figure 2.

5.2 Minimum Entropy Method

To apply the minimum entropy method, we used *Yalmip* and *Sedumi* as our SDP solver, set NITER = 6, and performed 250 random initializations of the algorithm. Remark that the entire polynucleotide sequence of the chromosome is used (*D. melanogaster* or *C. elegans*) to obtain appropriate mappings (Table **3**). Also, to evaluate if the minimum entropy mapping can identify the coding regions of the sequences being analyzed, we computed the contribution to the peaks at indices $1/3$ and $2/3$ (since, in this optimization method, the characteristic peaks of coding regions can appear in either, or both, frequencies) using a sliding window of length L = 351 (Figure **5**).

Table 3 – Minimum entropy mapping obtained using the *C. elegans* F56F11 cosmid and the *D. melanogaster* CG2009 gene.

Nucleo-tides	*C. elegans*		*CG2009*	
	DFT	Modu-lus	DFT	Modu-lus
A	-1.64 + 1.46E-16 j	1.64	-2.21 + 8.41E-18 j	2.21
C	2.47 + 1.60 j	2.94	1.37 + 1.69 j	2.18
G	1.85 + 1.70 j	2.51	2.39 – 2.68 j	3.59
T	-0.65 – 1.73 j	1.85	0.18 + 0.53 j	0.56
Entropy	**12.183**		**12.340**	

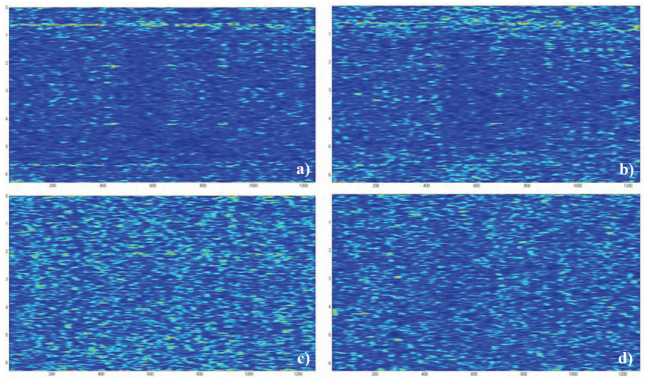

Figure 4 – Spectrograms for *C. elegans* (top) and CG2009 gene (bottom) using **a), c)** minimum entropy mapping and **b), d)** Anastassiou's mapping. Frequencies are represented on the vertical axis and sequence position on the horizontal axis.

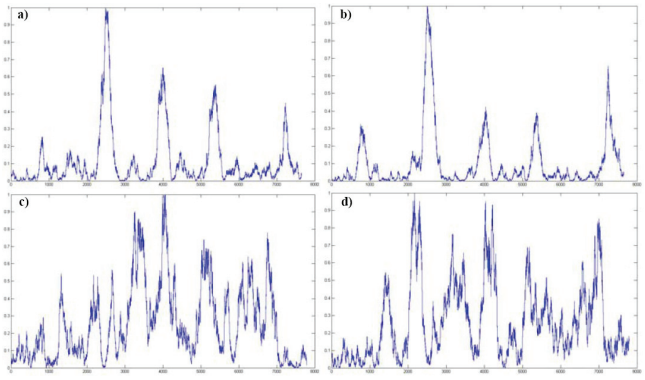

Figure 5 – Exonic regions forecast of *C. elegans* (top) and CG2009 gene (bottom) using **a), c)** minimum entropy mapping and **b), d)** Anastassiou's mapping.

6. DISCUSSION

In Figures **2** and **3** we can observe that the mappings derived from the various test sets successfully detect several coding regions of each organism (Table **4** lists the ground truth obtained from public databases). Better results are obtained when non-coding regions of *S. cerevisiae* chromosomes are used, rather than random DNA sequences, which may be partly due to the fact that the latter have a higher probability of being similar to coding regions. Note that in *D. melanogaster* the first and last exons cannot be identified.

Analysing the values in Tables **1** and **2**, we see that C+G has a higher modulus than A+T (e.g., 5.26 versus 4.73 in chromosome XVI and 4.62 versus 3.79 in chromosome XIII for random DNA), reflecting the fact that coding regions are C/G-rich (Section 3). These observations also apply to exonic regions forecast using minimum entropy mapping (MEM). The same exonic regions are detected by Anastassiou's mapping (AM) [3].

Table 4 – Exon position on the sequences analyzed.

ID	*F56F11*		*CG2009*	
	START	STOP	START	STOP
1	929	1135	601	924
2	2528	2857	2110	2628
3	4114	4377	2685	4370
4	5465	5644	4924	5037
5	7255	7605	5097	5920
6			6378	6711
7			6769	7222
8			7284	7543

In Figure **4**, we compare the spectrograms obtained using MEM and AM for both organisms studied. Generally, MEM tends to provide stronger peaks (see the discussion below regarding periods 3 and 11) and lower background noise. Also, the entropy attained by these mappings is lower for MEM (12.183 for *C. elegans* and 12.340 for *D. melanogaster*, while with AM these are 12.267 and 12.363, respectively).

Observing Figure **4**, we can identify a faint periodicity at index $L/3$ for *C. elegans* and *D. melanogaster* (due to statistical periodicity of three nucleotides in coding regions) and at $L/11$ for *C. elegans*. The peaks at lower frequencies could be due to the structural characteristics of DNA molecules. For instance, inside the cells these molecules assume a B-DNA conformation that has 10.4 base pairs (bp) per turn of the helix; however, DNA is a dynamic molecule assuming different three-dimensional structures under different conditions, such has A-DNA (forms when DNA is dehydrated and has 11bp per turn) and Z-DNA (occurs in C/G-rich regions). However, it is not clear that the lower frequency peaks appear due to the properties explained before [8].

7. CONCLUSION

This paper addressed the problem of detecting periodicities (MEM) and/or identifying exonic regions (MCM) in DNA. This is achieved through an optimization approach whereby the nucleotide sequence is mapped into a discrete numeric signal. For exonic region identification we can conclude that MCM has a similar performance to Anastassiou's mapping, whereas MEM has somewhat lower performance (although, the same exons are recognized). However, MEM is able to identify periodicities that are not detectable by Anastassiou's method or MCM, as it gives rise to full spectrograms with lower background noise and sharper peaks.

MEM has advantages over MCM, as it does not require training data (knowledge of exonic or/and non-coding regions) to detect exons and, simultaneously, enhance the strongest peaks in the power spectrum. The latter fact also presents drawbacks, as peaks surrounded by high levels of noise will be overly enhanced.

In the future, we can consider other evaluation functions for measuring entropy. MEM can also be applied to subsets of periodogram frequencies, or to fragments of a genome to determine locally-adapted mappings that better track variations in spectral content.

REFERENCES

[1] H. R. Horton et al., *Principles of Biochemistry*, Prentice-Hall, 3rd Edition, 2002.

[2] B. Alberts et al., *Molecular Biology of the Cell*, Taylor & Francis Books, 2002.

[3] D. Anastassiou, "Genomic Signal Processing", *IEEE Signal Processing Magazine*, vol. 18, pp. 8-20, 2001.

[4] L. Galleani, R. Garello, "Spectral Analysis of DNA Sequences by Entropy Minimization", in *Proc. EUSIPCO 2006*, Florence, Italy, September 4-8, 2006.

[5] P. Cristea, "Representation and Analysis of DNA Sequences", *Genomic Signal Processing and Statistics*, vol. 2, pp. 15-65, 2005.

[6] M. Akhtar, J. Epps, E. Ambikairajah, "On DNA Numerical Representation for Period-3 Based Exon Prediction", in *IEEE Gensips 2007*, Tuusula, Finland, June 10-12, 2007.

[7] V. Afreixo, P. Ferreira, D. Santos, "Fourier Analysis of Symbolic Data: A Brief Review", *Digital Signal Processing*, vol. 14, no.6, pp. 523-530, 2004.

[8] P. Vaidyanathan, B. Yoon, "The Role of Signal-Processing Concepts in Genomics and Proteomics", *Journal of the Franklin Institute*, vol. 341, pp. 111-135, 2004.

[9] D. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[10] S. Boyd, L. Vandenberghe, "Convex Optimization", *Cambridge University Press*, 2004.