14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPECTRAL ANALYSIS OF DNA SEQUENCES BY ENTROPY MINIMIZATION

Lorenzo Galleani and Roberto Garello

Dipartimento di Elettronica, Politecnico di Torino
Corso Duca degli Abruzzi, 24 - 10129 Torino, Italy
email: galleani@polito.it, garello@polito.it

ABSTRACT

Spectral analysis can be applied to study base-base correlation in DNA sequences. A key role is played by the mapping between nucleotides and real/complex numbers. In this paper, we present a new approach where the mapping is not kept fixed: it is allowed to vary aiming to minimize the spectrum entropy, thus detecting the main hidden periodicities. The new technique is first introduced and discussed through a number of case studies, then extended to encompass time-frequency analysis.

1. INTRODUCTION

Given a DNA sequence (typically taken from the genome of an organism) the study of nucleotide correlation both at short and long range is an important research issue in genomics (see [1], [2], and reference therein). Applications include characterization of random and non-random behavior, gene region prediction, detection of recurrent strings or motifs and related rules, construction of models for representing DNA structure, and so on. Spectral analysis can be a useful tool to investigate the correlation behavior, and many papers in genomic signal processing literature have been devoted to this issue (see [3] and its references). Anyway, when DNA spectral analysis is considered, a number of key issues arise, including the choice of a proper mapping between nucleotides and numbers, the usefulness of a unique quantity for representing the whole correlation properties, the need for taking into account sequence heterogeneity, and the necessity of detecting periodicities in strong noise. To cope with these issues a new technique, based on adaptive mapping and spectrum entropy minimization, is introduced and discussed in this paper.

2. CORRELATION ANALYSIS

Let us denote the four nucleotides (basis) alphabet by: $\mathbf{B} = \{A, C, G, T\}$, and a DNA sequence of length N as

$$\mathbf{s} = (b[n])_{n=0}^{N-1} \qquad b[n] \in \mathbf{B}.$$

For this sequence $\mathbf{s}$, let us introduce P_α as the occurrence probability of the nucleotide $\alpha \in \mathbf{B}$, and $P_{\alpha,\beta}[d]$ at distance $d = 1, ..., N-1$ as the joint probability of having a nucleotide α followed by a nucleotide β after d positions. For simplicity, we will suppose the DNA sequence $\mathbf{s}$ to be periodic with period N, an assumption that does not significantly alter the correlation properties of long sequences.

The two probabilities can be estimated by counting the nucleotide occurrences. In the following, we will focus on simple frequency count estimators:

$$P_\alpha = \frac{N_\alpha}{N} \qquad P_{\alpha,\beta}[d] = \frac{N_{\alpha,\beta}[d]}{N}, \quad d = 0, 1, \dots, N-1, \quad (1)$$

where N_α is the number of nucleotides α in $\mathbf{s}$ and $N_{\alpha,\beta}$ is the number of pairs (α, β) at distance d in $\mathbf{s}$ (supposed periodic).

A random sequence is composed by statistically independent symbols, then $P_{\alpha,\beta}[d] = P_\alpha P_\beta$ for each pair (α, β) and every distance d. As a consequence we can introduce as a proper measure of correlation, the sixteen *correlation functions*:

$$\Gamma_{\alpha,\beta}[d] = P_{\alpha,\beta}[d] - P_\alpha P_\beta \quad (\alpha, \beta) \in \mathbf{B}^2 \;\; d = 1, ..., N-1. \quad (2)$$

Then, the sequence $\mathbf{s}$ is random iff $\Gamma_{(\alpha,\beta)}[d] = 0$ for each (α, β) and every d, otherwise some correlation exists. The sixteen correlation functions can be further arranged in a matrix:

$$\begin{bmatrix} \Gamma_{AA}[d] & \Gamma_{AC}[d] & \Gamma_{AG}[d] & \Gamma_{AT}[d] \\ \Gamma_{CA}[d] & \Gamma_{CC}[d] & \Gamma_{CG}[d] & \Gamma_{CT}[d] \\ \Gamma_{GA}[d] & \Gamma_{GC}[d] & \Gamma_{GG}[d] & \Gamma_{GT}[d] \\ \Gamma_{TA}[d] & \Gamma_{TC}[d] & \Gamma_{TG}[d] & \Gamma_{TT}[d] \end{bmatrix} \quad d = 1, ..., N-1$$

Not all the 16 functions are independent: we certainly have $\sum_\beta \Gamma_{\alpha\beta}[d] = 0, \sum_\alpha \Gamma_{\alpha\beta}[d] = 0$, and $\Gamma_{\beta\alpha}[d] = \Gamma_{\alpha\beta}[N-d]$; non-exact, heuristic symmetries can also be invoked [1].

To study the behavior of the functions $\Gamma_{\alpha\beta}[d]$ and detect their periodicities, spectral analysis can be applied. For each correlation function $\Gamma_{\alpha,\beta}[d]$ we can compute the Discrete Fourier Transform (DFT), that will be called *correlation spectrum* in the following:

$$X_{\alpha,\beta}[k] = \sum_{d=0}^{N-1} \Gamma_{\alpha,\beta}[d] e^{-j\left(2\pi \frac{k}{N} d\right)} \quad k = 0, 1, \dots N-1. \quad (3)$$

The 16 correlation spectra $X_{\alpha\beta}$ can then be used for studying the correlation properties. Anyway, despite of this simple approach, some problems arise.

First problem: multiple representation. First of all, there are 16 correlation functions $\Gamma_{\alpha\beta}$ (and corresponding spectra $X_{\alpha\beta}$): it would be better, especially for automatic computations devoted for example to gene region prediction, to merge their information and produce a single quantity that measures the global sequence correlation properties. As an example, the 16 spectra could be added up, to form the average quantity $X^{ave}[k] = \sum_{\alpha\beta} \omega_{\alpha,\beta} X_{\alpha\beta}[k]$, where $\omega_{\alpha,\beta}$ are some proper weights.

Second problem: detection of periodicities. Real genome sequences contain a lot of “noise”, which is the consequence of genome evolution and its process of duplication, mutation, and so on. Therefore, we often have to detect periodicities buried in strong background noise.

Third problem: sequence heterogeneity. Usually, correlation functions and the corresponding spectra are computed on an entire DNA sequence: they represent an average on the sequence and they do not show where the correlation really exists. Many sequences, in fact, show different behaviors in different regions, that is they are nonstationary. Gene regions represent a typical example, since they usually show stronger correlation properties than non-coding regions.

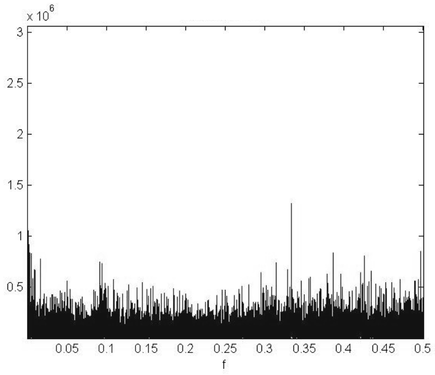

Figure 1: Classical frequency spectrum obtained from the Caenova sequence with the 4-PSK mapping.

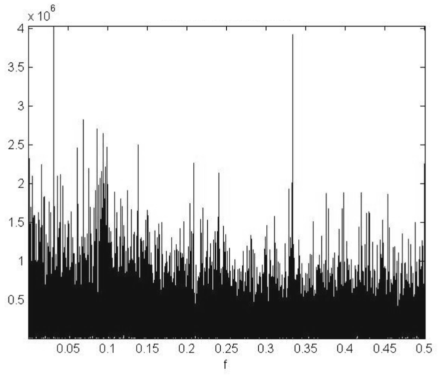

Figure 2: Classical frequency spectrum obtained from the Caenova sequence with the 4-PAM mapping.

3. MAPPING NUCLEOTIDES INTO NUMBERS

Instead of working on the 16 correlation functions and spectra, an effective approach consists in viewing the DNA sequence as a symbolic sequence over a quaternary alphabet and in analyzing it as a whole, first by translating it into a proper numeric sequence, and then by applying spectral analysis. A key point for DNA spectral analysis is then the numeric representation of the four nucleotides, which induces the mapping of DNA sequences into numeric ones.

A number of mathematical representations have been proposed in the literature and fundamental work on this subject was made by Anastassiou ([3]) and Cristea ([4]). The most popular representations are:

1. Binary indicator sequences (A=1000, C=0100, G=0010, T=0001) [1].
2. 3-D tetrahedron representation [3], [4].
3. 4-PSK complex assignment ($A = 1+i$, $G = -1+i$, $C = -1-i$, $T = 1-i$) [4].
4. 4-PAM real assignment ($A = -3$, $G = -1$, $C = +1$, $T = +3$), or similar shifted versions ($A = 0$, $G = 1$, $C = 2$, $T = 3$) [5].

Binary indicator sequences map a DNA sequence into four binary sequences. One can show that spectral analysis performed on these four sequences is essentially identical to the approach presented in Section 2, leading to the sixteen cross-spectra $X_{\alpha\beta}[k]$.

The 3-D tetrahedron has certainly a number of excellent properties, but it requires the mapping to be vectorial, forcing frequency analysis of the resulting discrete-time signal to be multidimensional (a strong complication). So, let us concentrate on the one-to-one mapping of the four nucleotides into four real or complex numbers, as done in representations 3 and 4 above.

4. THE ROLE OF LABELING

In this paper, a *labeling* l will be a one-to-one mapping between the nucleotide alphabet and the set of complex numbers:

$$l: \begin{array}{ccc} \mathbf{B} & \rightarrow & \mathbf{C} \\ \alpha & \rightarrow & \overline{\alpha} \end{array} \tag{4}$$

Somewhere, we will also use the symbol l_α to denote $\overline{\alpha} = l(\alpha)$. Given the mapping l, the DNA sequence $\mathbf{s} = (b[n])_{n=0}^{N-1}$ is translated into a unique numeric signal

$$x[n] = \overline{b}[0]\delta[0] + \overline{b}[1]\delta[n-1] + \ldots + \overline{b}[N-1]\delta[N-1], \tag{5}$$

whose spectral properties can be directly investigated. Given $x[n]$, we can compute its DFT by:

$$X[k] = \sum_{n=0}^{N-1} \overline{b}[n] e^{-j\left(2\pi \frac{k}{N} n\right)}, \quad k = 0, 1, \ldots N-1$$

and its power spectrum: $P_x[k] = |X[k]|^2$. It is well know that $P_x[k]$ is equal to

$$P_x[k] = \sum_{m=0}^{N-1} r[m] e^{-j2\pi \frac{k}{N} m}, \tag{6}$$

i.e., it is the DFT of the *autocorrelation function*:

$$r[m] = \sum_n \overline{b}[n] \overline{b}^*[n-m].$$

Clearly, $r[m]$ and $P_x[k]$ strongly depend on the labeling l. It is also important to recognize that l establishes a bridge between $P_x[k]$ and the 16 correlation spectra $X_{\alpha\beta}[k]$. In fact, by invoking the definition of $N_{\alpha\beta}$ we note that the autocorrelation function can also be computed as

$$r[m] = \sum_{\alpha\beta} l_\alpha l_\beta^* N_{\alpha\beta}[m].$$

Then

$$P_x[k] = \sum_{\alpha\beta} l_\alpha l_\beta^* \sum_m N_{\alpha\beta}[m] e^{-j2\pi \frac{k}{N} m}. \tag{7}$$

Now, by using the definition (3) of $X_{\alpha\beta}[k]$ and its connection with the pair number $N_{\alpha\beta}[k]$ established by (1) and (2), we obtain that

$$P_x[k] = c_1 \sum_{\alpha\beta} l_\alpha l_\beta^* X_{\alpha\beta}[k] + c_2\delta[k] \quad c_1, c_2 \in \mathbb{R}. \tag{8}$$

As a consequence, apart from a multiplicative constant and an additive term at zero-frequency, the spectrum $P_x[k]$ is a weighted sum of the 16 correlation spectra $X_{\alpha\beta}[k]$ via the four values l_A, l_C, l_G, l_T. This is a very simple but basic result, (i) clarifying that the symbolic sequence approach effectively solves the first problem discussed before, providing a unique representation and (ii) making evidence to the fundamental role played by the mapping.

As an example, in Fig. 1 and Fig. 2 we show the frequency spectra of a DNA sequence corresponding to a gene area of the Caenova, obtained with 4-PSK and 4-PAM mapping, respectively. The two spectra differs in a dramatic way, pointing out the strong dependence on the chosen mapping highlighted by Eq. (8).

4.1 Problems with classical mapping

A key question naturally arises: how do we choose the labeling l? In the literature, this problem is faced by considering two issues [3], [4]:

1. The mapping should respect the physical properties of the four basic molecules, A, C, G and T. That is, we would like the mapping to reflect the possible bounds of the molecules, and in general any physical property that we think can be of interest.
2. The mapping should not privilege any basis. This means it should ideally be symmetric so that the geometric distance between the transformed nucleotides be the same.

As an example, the aforementioned 4-PSK mapping [4] can be seen as a projection on the complex plane of a tetrahedron whose vertices are the four nucleotides, and is certainly a very good candidate for symbolic DNA spectral analysis. Anyway, due to its key role highlighted by Eq. (8), the mapping can produce misleading effects. As an example, we now introduce and discuss a few examples, showing some critical aspects that must be taken into account when studying DNA sequences (and any other symbolic sequence in general). The first two cases come from ad hoc built sequences, while the third is a real DNA sequence.

Critical case 1 - Two periodicities corrupted by noise. Let us consider the following discrete-time signal

$$x[n] = \sin\left(\frac{2\pi}{3}n\right) + \sin\left(\frac{2\pi}{6}n\right). \qquad (9)$$

This is a periodic signal, with a period of 6 samples, that takes three values only: $\sqrt{3}, 0, -\sqrt{3}$. Its spectrum consequently shows two peaks at frequencies 1/3 and 1/6 with equal amplitude. Now, let us suppose to build a DNA sequence $\mathbf{s}$ by associating the nucleotide symbols to the discrete-time signal. The sequence $\mathbf{s}$ will be periodic with the same period of $x[n]$, that is $N = 6$ symbols. Such operation can be considered as an inverse mapping, since it links the discrete-time signal to the sequence. As an example, we choose the following association: $A = \sqrt{3}, C = 0, G = -\sqrt{3}$. Therefore the DNA sequence $\mathbf{s}$ is given by $\mathbf{s}$ = CACCCG.... Let us now suppose to perform a spectral analysis of this DNA sequence $\mathbf{s}$ by using the 4-PSK mapping. The corresponding spectrum $P_x[k]$ is reported in Fig. 3. We notice that the two expected peaks are represented, but the one at $f_1 = 1/6$ has a very small amplitude. Also, apart from the zero frequency peak, there is an extra peak at $f = 1/2$.

Since noise is widely present in DNA sequences, let us now investigate the effect of its presence. Given $\mathbf{s}$, we produce a noisy version in this way:

1. For every nucleotide we generate a random number z uniformly distributed in the range $0 \leq z \leq 1$. If $z \leq h$, where h is a given threshold, we go to step 2, otherwise we do nothing and proceed to the next nucleotide.
2. We replace the selected nucleotide with a random nucleotide extracted with a uniform probability $P = 1/4$.

Hence, on the average hN nucleotides will be totally random in the noisy sequence. This procedure has been applied to the sequence $\mathbf{s}$ by using a threshold $h = 0.2$.

The spectrum obtained for the noisy DNA sequence by using the 4-PSK mapping is reported in Fig. 4. Unfortunately we spot one peak only! We in fact see the peak at f_2, while the peak at f_1 has disappeared. The reason is that the 4-PSK mapping produces a frequency spectrum that has a very low peak at f_1, that easily sinks in the background noise once we add it. Also, extra peaks at $f = 0$ and $f = 1/2$ carry a misleading and useless information.

Critical case 2 - Sinusoid in the sequence domain. Consider this DNA sequence, with period $N_p = 20$ nucleotides:

$$\mathbf{s} = \text{CCCCCAAAAACCCCCGGGGG...}$$

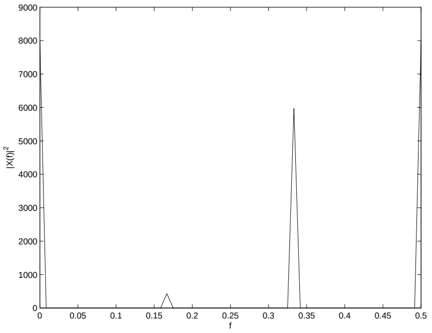

Figure 3: 4-PSK spectrum of the sequence $\mathbf{s}$ discussed in Critical case 1.

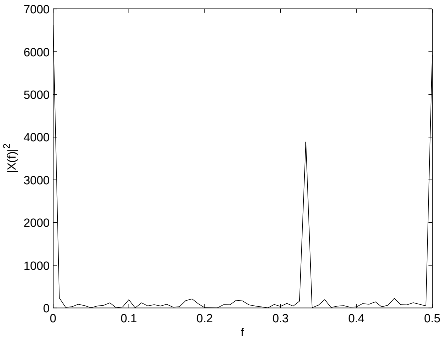

Figure 4: 4-PSK spectrum of the noised sequence $\mathbf{s}$ discussed in Critical case 1.

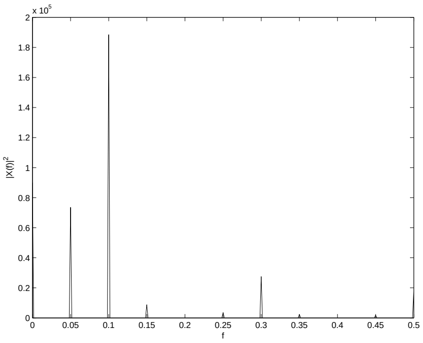

Figure 5: 4-PSK spectrum of the sequence $\mathbf{s}$ discussed in Critical case 2.

As an example, this sequence can be obtained by sampling and quantizing a simple real sinusoid. We clearly expect the spectrum of the sequence to identify the main periodicity at $f = 1/N_p = 1/20$. Beside this strong frequency component, when we map the sequence to the discrete-time signal we do not get the original real sinusoid, and therefore we expect to have low amplitude harmonics, located at frequencies that are multiples of $f = 1/20$. In Fig. 5 we show the frequency spectrum of $\mathbf{s}$ obtained via the 4-PSK mapping. As it can be seen the component at $f = 1/20$ is not the strongest one! The frequency with the taller peak is located at $f = 1/10$. There is also a DC component at $f = 0$, and some harmonics at $f > 1/10$. Anybody looking at this spectrum without knowing the original sequence s, would not be able to catch the key fact that it represents a slowly varying periodic sequence of period f_p. The conclusion that we draw from this case is that the 4-PSK mapping (and in general a classical mapping) may not represent the frequency spectrum suggested by intuition. On the contrary it may generate a representation that is misleading.

5. MINIMUM ENTROPY MAPPING SPECTRUM

Given these examples, what mapping should we use? What is the "best" labeling? This question is better reformulated in the following way: what is the frequency spectrum of a symbolic sequence? In transforming it into a numeric discrete-time sequence we have to pay attention not to alter its frequency content. As seen, a wrong mapping can hide peaks that corresponds to real periodicities of the DNA sequence. Also, if we find out that we are adding extra peaks to the spectrum because of the chosen mapping, then we are generating useless and possibly dangerous information about the sequence itself. Is there a mapping that uses the *minimum amount of information* needed, thus eliminating the extra information added by standard mappings? To answer this question, we propose a method that chooses the mapping minimizing the entropy of the frequency spectrum.

5.1 Algorithm description

For simplicity, we will consider real mappings only. Let l_A, l_C, l_G, l_T be the four real numbers corresponding to the image of the mapping l. Given the spectrum $P_x[k;l]$ (where the dependence on the mapping l has been highlighted) let us introduce its Shannon entropy H, given by

$$H\left[P_x[k;l]\right] = -\sum_{k=0}^{N/2} P_x[k;l]\log\left[P_x[k;l]\right]. \qquad (10)$$

We now define the *Minimum Entropy Mapping (MEM)* spectrum of a DNA sequence $\mathbf{s}$ as the spectrum obtained when the mapping l satisfies the minimum conditions

$$\overline{l_A}, \overline{l_C}, \overline{l_G}, \overline{l_T} = \arg \min_{l_A,l_C,l_G,l_T \in \mathbb{R}} H\left[P_x[k;l]\right]. \qquad (11)$$

To solve the minimization, we impose two constraints. Since the DC component does not contain useful information, we impose the condition

$$P_x[0;\overline{l_A}, \overline{l_C}, \overline{l_G}, \overline{l_T}] = 0.$$

Similarly, we impose the same constraint at $N/2$:

$$P_x[N/2;\overline{l_A}, \overline{l_C}, \overline{l_G}, \overline{l_T}] = 0.$$

Moreover, we scale the obtained mapping l so that the signal energy is one.

Even though the Shannon entropy in Eq. (11) is convex with respect to the energy spectrum P_x, this does not imply that it is convex with respect to the four variables l_A, l_C, l_G and l_T. This means that in general the minimization problem of Eq. (11) must be solved in a concave space.

However, the four degrees of freedom in the minimization problem, are reduced to two, thanks to the imposed constraints. Therefore, for the remaining two free variables we can easily use an exhaustive search algorithm. Since we normalize the spectrum to have unit energy, we can span the free mapping variables, say for example l_A and l_C, on a bounded space. This means that scaled versions of the same spectrum will have the same entropy H. Given this fact, we can search for the solution in a limited interval of the free variables, because we are actually searching all the possible scaled versions of the corresponding discrete-time signal, which in our formulation are hence equivalent. As an example, we have decided to limit the variables l_A and l_C to the interval$-\frac{1}{2} \leq l_A, l_C \leq \frac{1}{2}$.

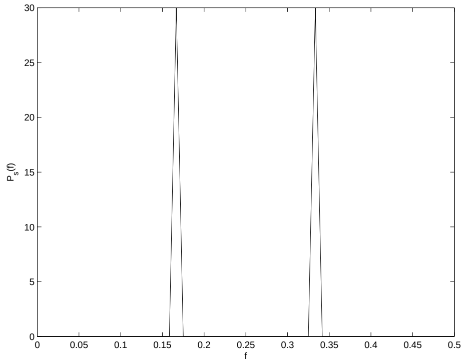

Figure 6: MEM spectrum of the sequence $\mathbf{s}$ discussed in Critical case 1.

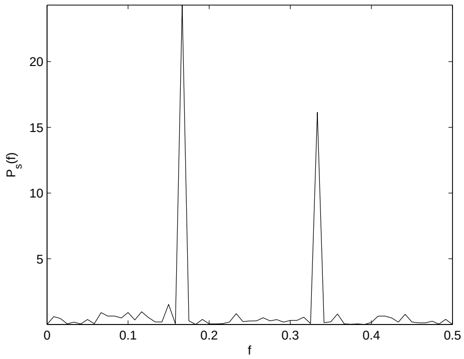

Figure 7: MEM spectrum of the noised sequence $\mathbf{s}$ discussed in Critical case 1.

5.2 Results

The new method has been applied to both ad hoc created and experimental data, and the results prove the effectiveness of the new approach. The application to ad hoc case studies is fundamental, since it allows to directly verify that the MEM spectrum can extract the information that was artificially put in the data. The application of the method to experimental data allows to generate DNA spectra that are less noisy and in general more reliable than with a standard mapping. Let us review the results of the new technique for the critical cases discussed before.

Critical case 1. In Fig. 6 we report the MEM spectrum for the sequence $\mathbf{s}$ of Critical case 1, when no noise is added. We see that it correctly represents the two known periodicities, and that no other information is present. Also, both peaks have the same height. The MEM spectrum for the noised version of the same sequence is shown in Fig. 7. We immediately notice that the spectrum shows the two peaks at $f_1 = 1/6$ and $f_2 = 1/3$, that correctly correspond to the periodicities of $N_1 = 6$ and $N_2 = 3$ nucleotides used to design the sequence. Also, some background uniform noise is present. The advantages with respect to the 4-PSK spectrum, previously shown in Fig. 3 and Fig. 4, are evident.

Critical case 2 - Sinusoid in the sequence domain. The MEM spectrum for the DNA sequence of Critical case 2 is reported in Fig. 8. We see that the new technique detects the fundamental frequency f_p, that is granted the strongest intensity. Also the higher harmonics are represented, but with a much lower amplitude, in accordance with the expected results.

Experimental data: Caenova sequence. In Fig. 9 we show the MEM spectrum of the Caenova sequence previously consid-

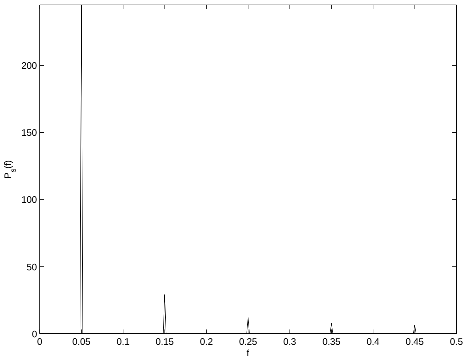

Figure 8: MEM spectrum of the noised sequence **s** discussed in Critical case 2.

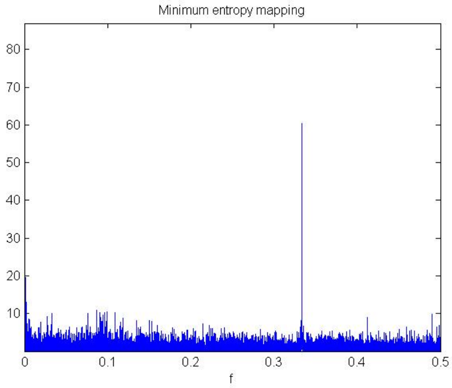

Figure 9: MEM spectrum of the Caenova sequence.

ered in Section 4. It can be seen that the MEM spectrum clearly identifies the peak at $f = 1/3$, that as known is typical of a number of gene regions [8]. A low frequency content is also present, that could be due to long term correlations in the data, and deserve future investigation. By comparing the MEM spectrum against the 4-PSK and 4-PAM mapping spectra shown in Fig. 1 and Fig. 2, we conclude that the MEM spectrum clearly identifies peaks in experimental data, proving to be very robust with respect to noise.

6. TIME-FREQUENCY MINIMUM ENTROPY SPECTRUM

DNA sequences are inherently nonstationary, as pointed out in Section 2, Problem 3. While some regions look like noise, others exhibit frequency peaks that represent periodicities. This means that the frequency content of a DNA sequence changes with position. Spectral analysis does not allow to detect such frequency variations, that are indeed very interesting and could be fundamental to understand the local meaning of the correlation structure. To cope with sequence heterogeneity and represent frequency variations we use the ideas developed in time-frequency analysis [6], a field that has been applied recently to the investigation of the properties of DNA sequences [7]. The MEM spectrum can easily be extended to time-frequency analysis. To build a time-frequency spectrum of a sequence we slide our MEM spectrum on the data, ending up with a spectrum that is a function of the analysis position n *and* and of the frequency k. So, the mapping minimization is performed also as a function of the position n. A number of case studies (both ad hoc sequences and experimental data) can be found in [9].

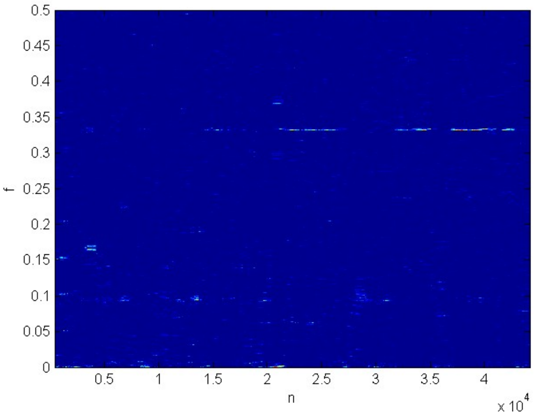

Figure 10: Time-frequency MEM spectrum of the Caenova sequence.

As an example, in Fig. 10 we show the MEM position frequency spectrum for the Caenova sequence previously considered, obtained with a window of 1000 bases. We see that the local MEM spectrum shows the presence of a peak located at $f = 1/3$, that has an intensity that changes with position. The presence of other patterns in the position-frequency plane is also brought to light.

7. CONCLUSIONS

A new technique for spectral analysis and time-frequency analysis of DNA sequences, based on an adaptive mapping, has been introduced. The spectrum entropy has been chosen as the cost function, to detect the main periodicities. Case studies have been discussed to show problems arising with classical, fixed mapping and the advantage of the new method. Future activity will focus on a wide application of the new technique to experimental DNA sequences, for investigating correlation both at short and long range. Extension of the MEM technique to larger entities is also under study.

REFERENCES

[1] Wentian Li, "The study of correlation structures of DNA sequences: a critical review," *Computers & Chemistry*, vol. 21(4), pp. 257-272, 1997.

[2] Wentian Li, on-line web page, "Bibliography on DNA correlation", http://www.nslij-genetics.org/dnacorr/

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

[4] P.D. Cristea, "Representation and analysis of DNA sequences, in " *Genomic Signal Processing and Statistics*, Hindawi Publishing Corporation, 2005.

[5] G. Rosen, "Examining coding structure and redundancy in DNA, " *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 1, pp. 62-68, January 2006.

[6] L. Cohen, *Time-Frequency Analysis*, Prentice-Hall, 1995.

[7] D. Sussillo, A. Kundaje, D. Anastassiou, "Spectrogram Analysis of Genomes, " *Eurasip Journal on Applied Signal Processing*, pp. 29-42, 2004.

[8] H. Herzel, E.N. Trifonov, O. Weiss, I. Groβe, "Interpreting correlations in biosequences," *Physica A*, vol. 249, pp. 449-459, 1998.

[9] L. Galleani and R. Garello, "Study of DNA correlation by spectrum entropy minimization," *To be submitted to IEEE Transactions on Signal Processing*, 2006.