Maximizing the SNR of DNA Spectrum for Coding Sequence Identification

Milena Marinho Arruda, Andresso da Silva and Francisco Marcos de Assis

***Abstract*— By properly mapping a DNA sequence into one or more signals, the energy spectrum of DNA sequences can reveal standard periodicities, in particular, a periodicity of three bases, to characterize coding and regions. In this sense, we propose a new method for spectral analysis for coding sequence discrimination. The method consists in defining a mapping for a given DNA sequence whose SNR of its spectrum is maximized subject to the spectral entropy constraint being greater than zero. Finally, we show that our method not only improves the coding sequence discrimination rate but also improves the SNR and spectral entropy even for those coding sequences that there are no periodicity of three bases.**

***Keywords*— coding sequence, DNA, entropy, SNR, spectral analysis.**

I. Introduction

The growth of biological databases and the need to understand how many components present in a living cell are working together to perform cellular functions are reasons that justify the interest in mathematical, statistical, and computational tools to the analysis of genomic data. The Genomic Signal Processing (GSP) refers to the use of these multidisciplinary approaches to understand the properties and structure of DNA.

Although the genetic information of an organism is encoded in DNA molecules by means of units called bases, such as: adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`), the current GSP methods require the association of a DNA sequence with a discrete-time signal by an operation called mapping. A simple and commonly used mapping scheme is the Voss representation [1]. From properly mapping a character sequence into one or more signals, digital signal processing can provide a set of useful tools for interpreting genomic information.

For instance, the energy spectrum of DNA sequences can reveal standard periodicities to characterize coding and non-coding regions, as well as intronic and exonic regions. The need for a way to discriminate coding from non-coding regions arises mainly when a gene location is only approximately known [2]. In eukaryotic cells, the DNA is divided into genes and intergenic regions. The genes are further divided into exon and intron, which is shown in Figure 1. Finally, the coding sequence (CDS) is then the portion of a gene which codes for a protein, that is, its exons.

Milena Marinho Arruda, Department of Electrical Engineering, UFCG, Campina Grande-PB, e-mail: milena.arruda@ee.ufcg.edu.br; Andresso da Silva, Department of Electrical Engineering, UFCG, Campina Grande-PB, e-mail: andresso.silva@ee.ufcg.edu.br; Francisco Marcos de Assis, Department of Electrical Engineering, UFCG, Campina Grande-PB, e-mail: fmarcos@dee.ufcg.edu.br; This work was partially supported by CNPq.

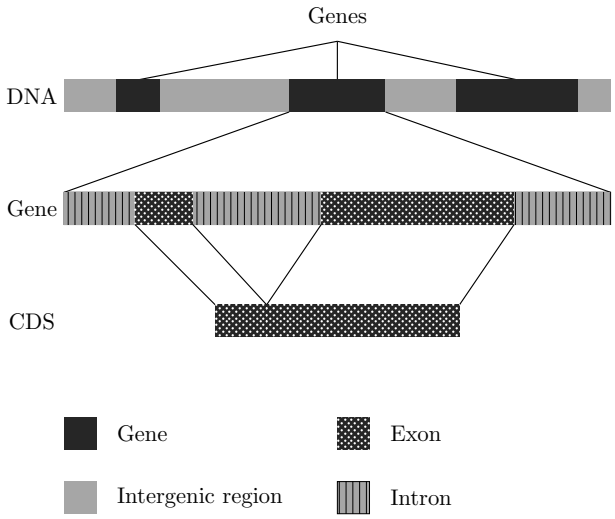


Fig. 1: The eukaryotic DNA consists of genes and intergenic regions. Moreover, the genes are composed of two regions called exon and intron which are interleaved with each other.

In this sense, Trifonov and Sussman [3] were the first ones to observe the existence of periodicities in DNA sequences from the analysis of the autocorrelation function; and, Tsonis *et al.* [4] found that whereas intronic regions show a rather random pattern, coding sequences show periodicities and in particular a periodicity of three base. Although, intuitively, the hypothesis for the origin of this periodicity derives from the triple nature of the codon and from the codon bias involved in the translation process, it proved to be insufficient. This periodic phenomenon has intrigued many biologists who seek to understand and explain such periodicity [5], [6], [7].

Since the three-base periodicity in CDS is a classical discriminatory frequency, it is often used in gene finding. The problem, however, is how to define an energy spectrum for a DNA sequence. In spite of Silverman and Linsker [8] were the first ones to define and analyze the energy spectrum of a DNA sequence, the classical approach was proposed by Voss [1]. The difference is essentially the number of discrete-time signals assigned in each case. That is, Silverman and Linsker [8] associated each of the four bases to a vertex of a regular tetrahedron resulting in three signals and, Voss [1] associated each of the four bases to binary indicator signals. Next, for the spectral analysis, the Fourier transform of each signal are evaluated and, the sum of their squared magnitude represents the energy spectrum of the DNA sequence. Coward [9] shown that these two methods yield essentially the same result.

Besides this classical approach, numerous techniques have

been proposed to improve these estimates and reduce background noise. For instance, Tiwari *et al.* [10] proposed a method in which it is sufficient to evaluate the energy density at frequency $1/3$ in a window of M samples, sliding it through the entire DNA sequence; Anastassiou [11] proposed to optimize the mapping by maximizing an objective function whose parameters are mean and standard deviation of Fourier transform; Vaidyanathan *et al.* [12] proposed the use of the antinotch filter; Galleani and Garello [13] proposed the Minimum Entropy Mapping (MEM) Spectrum, in which the spectrum of a DNA sequence should be calculated considering a real mapping for which the spectral entropy is minimized; Sahu *et al.* [14] proposed the use of the S transform; and Roy *et al.* [15] propose a minimum standard estimator.

Therefore, this paper investigates the SNR and spectral entropy as a measure to improve the DNA coding sequence identification. Although the MEM Spectrum [13] have already discussed how to choose the mapping of a DNA sequence that minimizes its spectral entropy to reveal its periodicities, the algorithm proposed by them sometimes returns a worse spectrum regarding to the signal noise ratio (SNR) analyses when compared to others representations in the literature such as Voss. Additionally, the computational time is another problem in their method since an exhaustive search algorithm was used for optimization.

We propose an optimization process that maximizes the SNR of the DNA spectrum, and, hence, minimizes the spectral entropy and the background noise. We also discuss the optimization process proposed by us in detail and present reasons to support it. Finally, we tested the method on synthetic and real DNA sequences, whose properties are known, and the results showed to be effective for the coding sequences identification. The algorithm was implemented in Python, a free open-source language.

This paper is organized as follows: Section II provides notations and definitions. In Section III, we define our method and discuss its most important features. The results are then presented in Section IV and, finally, the conclusions are elaborated in Section V.

II. Preliminaries

In this section, we describe some notations and definitions that are essentially important to the analysis in this paper, for more details we recommend [16], [17].

A. Notations and Definitions

Let s be a given DNA sequence of length N. The mapping $\mathcal{M}$ is defined as the association between the four DNA basis characters and four distinct complex numbers. That is,

$$\mathcal{M} : \mathtt{A} \mapsto a,\ \mathtt{C} \mapsto c,\ \mathtt{G} \mapsto g,\ \mathtt{T} \mapsto t, \tag{1}$$

such that,

$$a,\ c,\ g,\ t \in \mathbb{C}. \tag{2}$$

Supposing that the first four nucleotides of a given DNA sequence are $s = \mathtt{ACGT}\cdots$. From the mapping $\mathcal{M}$ we can, therefore, associate the following discrete-time signal to s,

$$x[n] = a\delta[n] + c\delta[n-1] + g\delta[n-2] + t\delta[n-3] + \cdots, \tag{3}$$

where $\delta[n]$ is the unit impulse function. Moreover, an alternative way to (3) is

$$x[n] = ax_{\mathtt{A}}[n] + cx_{\mathtt{C}}[n] + gx_{\mathtt{G}}[n] + tx_{\mathtt{T}}[n], \tag{4}$$

where $x_\alpha[n]$ is the binary indicator function that assume 1 when the n-th symbol in s is the basis $\alpha \in \{\mathtt{A},\ \mathtt{C},\ \mathtt{G},\ \mathtt{T}\}$, and 0 otherwise. Notice that the set of these four binary indicator sequences is redundant since they add to 1 for all n thus, any three indicator sequences are sufficient to determine the DNA sequence.

Example 1: The four binary indicator functions for the sequence $s = \mathtt{AACTGT}$ are: $x_{\mathtt{A}}[n] = \delta[n] + \delta[n-1]$, $x_{\mathtt{C}}[n] = \delta[n-2]$, $x_{\mathtt{G}}[n] = \delta[n-4]$, $x_{\mathtt{T}}[n] = \delta[n-3] + \delta[n-5]$.

B. Spectral Analysis

To discover underlying periodicities in genomic sequences, the spectral analysis is then performed on signal (4), whose energy spectrum is also a function of the mapping $\mathcal{M}$, that is,

$$S[k;\mathcal{M}] = \left|aX_{\mathtt{A}}[k] + cX_{\mathtt{C}}[k] + gX_{\mathtt{G}}[k] + tX_{\mathtt{T}}[k]\right|^2, \tag{5}$$

where $X_\alpha[k]$ with $\alpha \in \{\mathtt{A},\ \mathtt{C},\ \mathtt{G},\ \mathtt{T}\}$ is the Fourier transform of the respective binary indicators defined by

$$X_\alpha[k] = \sum_{n=0}^{N-1} x_\alpha[n] e^{-j\frac{2\pi}{N}nk}, \quad k = 0, 1, \ldots, N-1. \tag{6}$$

In particular, the classical approach to energy spectrum analysis was proposed by Voss [1] and it is the addition of the energy contribution of all four binary indicators of s as follows:

$$S_{\text{Voss}}[k] = \left|X_{\mathtt{A}}[k]\right|^2 + \left|X_{\mathtt{C}}[k]\right|^2 + \left|X_{\mathtt{G}}[k]\right|^2 + \left|X_{\mathtt{T}}[k]\right|^2. \tag{7}$$

Another successfully technique was proposed by Nair and Sreenadhan [18], in which, it has used the electron-ion interaction pseudopotentials (EIIP) indicator to map the DNA to a signal. This mapping is as follows: $\mathtt{A} \mapsto 0.1260$, $\mathtt{C} \mapsto 0.1340$, $\mathtt{G} \mapsto 0.0806$ and $\mathtt{T} \mapsto 0.1335$; and the energy spectrum is given by (5).

Furthermore, Galleani and Garello [13] proposed a new definition of spectrum for DNA sequences based on an entropy minimization criterion, the MEM Spectrum. In this case, the spectrum should be calculated considering a real mapping in which the spectral entropy is minimized, that is, the mapping is chosen such that

$$\overline{\mathcal{M}} = \arg \min_{a,c,g,t \in \mathbb{R}} H(\mathcal{M}), \tag{8}$$

where $H(\mathcal{M})$ is the entropy of (5) defined by

$$H(\mathcal{M}) = -\sum_{k=0}^{\lfloor N/2 \rfloor} p[k] \log p[k], \tag{9}$$

where,

$$p[k] = \frac{S[k]}{\sum_{k=0}^{\lfloor N/2 \rfloor} S[k]}. \tag{10}$$

The spectral entropy is a measure of the uniformity of energy distribution in the frequency domain. Notice that the minimum value of $H(\mathcal{M})$ is zero and occurs when $p[k] = 1$

for some k, and the maximum value occurs when the energy distribution is uniform. In this case, $H(\mathcal{M}) = \log({}^{N}/_{2} + 1)$ nats. The natural logarithm was used in (9), and, therefore, the entropy spectrum was given in nats.

The energy spectrum of DNA sequences can be slightly different when we compare the three previous methods of spectral analysis. For instance, the normalized energy spectrum for the CDS of some genes are showed in Fig. 2, 3, 4 and 5. In these figures, the horizontal axis is the normalized frequency and the vertical axis is the normalized energy spectrum. These differences show that a proper choice of the mapping can enhance the hidden information for further analysis of DNA sequences. For this reason, this paper proposes a new mapping for a given DNA sequence whose SNR of its spectrum is maximized subject to the spectral entropy constraint being greater than zero, which is discussed in Section III.

A quantitative analysis of these spectrum was given by the SNR. The SNR, defined as the ratio of signal power to the noise power, was computed by estimating the signal power, that is, the fundamental frequency, as the highest spectral component in the signal, and the computation of noise power or the background noise excluding the fundamental frequency and the DC value.

III. Materials and Methods

A. Experimental Data

The data is available at nucleotide database from National Center for Biotechnology Information (NCBI) that provides open access to biomedical and genomic information [19]. All DNA sequences have an identifier, the accession number, a simple series of digits processed by NCBI for which the sequence is referenced. In this paper, we isolated all genes for which there were no introns from chromosome XVI of *Saccharomyces cerevisiae* (accession number NC_001148). When we analyze a specific gene, we refer to it by its locus tag. Therefore, our database has 443 CDS whose average length is 1463.33 bases with standard deviation of 1008.56.

B. Maximum SNR Mapping

We define the spectrum of a DNA sequence by choosing the mapping that maximizes the SNR of its energy spectrum. Therefore, the new mapping is a complex map for which

$$\overline{\mathcal{M}} = \arg \max_{a,c,g,t \in \mathbb{C}} \mathrm{SNR}(S[k; \mathcal{M}]), \tag{11}$$

subject to the constraint

$$H(\mathcal{M}) > 0, \tag{12}$$

whose importance is discussed in Section III-C. The corresponding energy spectrum is, therefore, given by (5) using the mapping $\overline{\mathcal{M}}$.

C. Constraint Discussion

A DNA sequence is completely determined by any three of the four binary indicator sequences since they add to 1 for all n, and, therefore, the four corresponding Fourier transforms are also a redundant set in which

$$X_{\mathtt{A}}[k] + X_{\mathtt{C}}[k] + X_{\mathtt{G}}[k] + X_{\mathtt{T}}[k] = \begin{cases} 0, & k \neq 0 \\ N, & k = 0 \end{cases}. \tag{13}$$

Thus, in the search space, there are mappings whose spectrum has a stronger peak at $k = 0$ plus some minimal fluctuations at $k \neq 0$. These cases are the trivial solutions of the optimization problem.

Notice that the spectral entropy of the trivial solutions is approximately zero. Therefore, in order to avoid trivial solutions, we have to choose those mappings whose spectral entropy is greater than zero.

D. Optimization Details

In general, the SNR is not convex with respect to the mapping $\mathcal{M}$. This fact means that the optimization problem must be solved in a concave space, and, therefore, it is sensible with respect to the initial condition. Therefore, we suggest using the EIIP mapping as the initial condition of the optimization problem defined in (11) due its biological significance. Furthermore, with respect to the solver used for optimization, the Sequential Least Squares Programming (SLSQP) algorithm was preferred due its ability to search over the design space whose bounds constraints were delimited by

$$-1 < \mathrm{Re}\{a,\ c,\ g,\ t\} < 1, \tag{14}$$

and

$$-1 < \mathrm{Im}\{a,\ c,\ g,\ t\} < 1. \tag{15}$$

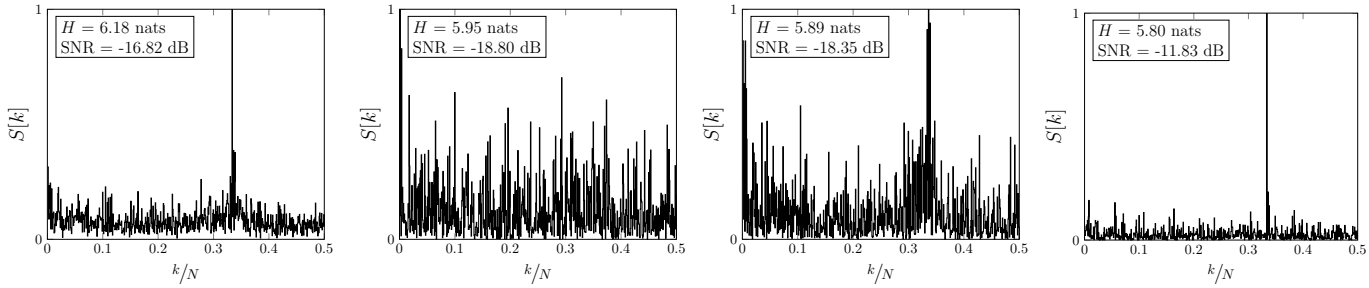
IV. Results

For the purpose of comparison, the energy spectrum of CDS in database was evaluated using the proposed method and the three methods already discussed in this paper: Voss [1], EIIP [18], MEM Spectrum [13]. The Table I summarizes the number of CDS whose algorithms identified the largest spectral peak occurring between the frequencies 0.31 and 0.35rad/sample in the spectrum of the sequences. Whether the largest peak occurs in this range, the algorithm was considered to have correctly identified the sequence. Otherwise, it has identified it incorrectly.

TABLE I: CDS identification rate by spectral analysis.

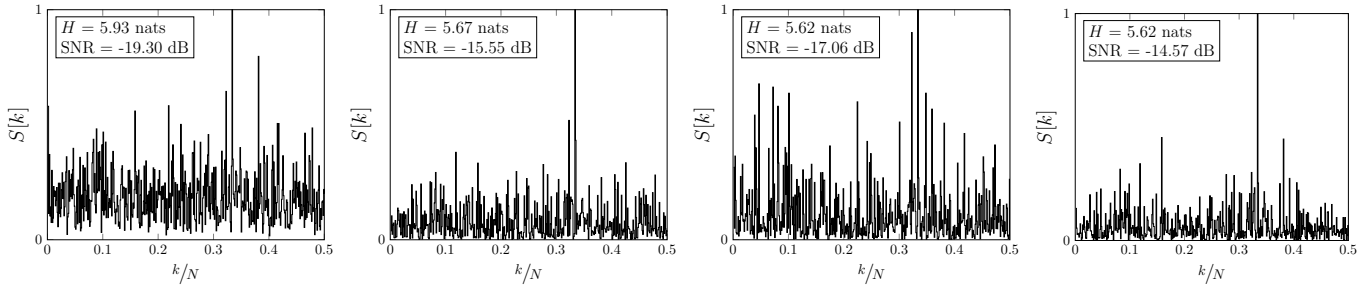
Status	Correct	Incorrect
Voss [1]	353 (79.7%)	90 (20.3%)
EIIP [18]	354 (79.9%)	89 (20.1%)
MEM Spectrum [13]	254 (57.3%)	189 (42.7%)
Proposed	374 (84.4%)	68 (15.6%)

Notice that the methods do not discriminate the three-base periodicity for all genes. There are several reasons for that. First, the way the energy spectrum is defined for a DNA sequence can influence the coding sequence identification process. For instance, the energy spectrum as defined by Voss and by us has a largest peak at frequency 0.33rad/sample for the gene YPL230W; however, this discriminatory frequency is vanishing when the EIIP method is used and the background noise increases when MEM Spectrum is evaluated (see Fig. 2).



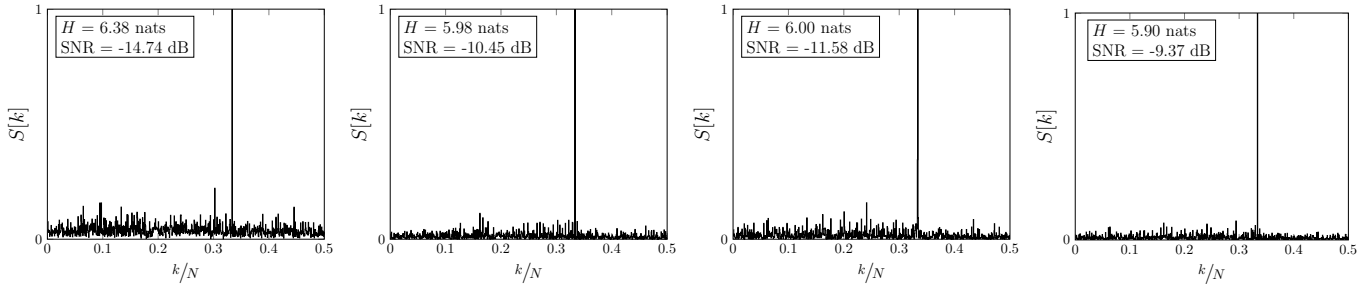
(a) Voss. (b) EIIP. (c) MEM Spectrum. (d) Proposed.

Fig. 2: Spectral analysis for the CDS of gene YPL230W from chromosome XVI of *Saccharomyces cerevisiae*.



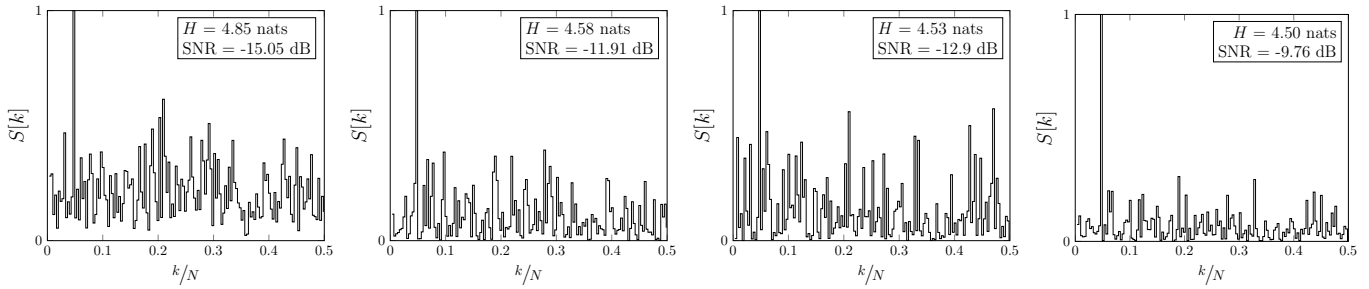
(a) Voss. (b) EIIP. (c) MEM Spectrum. (d) Proposed.

Fig. 3: Spectral analysis for the CDS of gene YPL064C from chromosome XVI of *Saccharomyces cerevisiae*.



(a) Voss. (b) EIIP. (c) MEM Spectrum. (d) Proposed.

Fig. 4: Spectral analysis for the CDS of gene YPL061W from chromosome XVI of *Saccharomyces cerevisiae*.



(a) Voss. (b) EIIP. (c) MEM Spectrum. (d) Proposed.

Fig. 5: Spectral analysis for the CDS of gene YPL278C from chromosome XVI of *Saccharomyces cerevisiae*.

A similar result is obtained for the gene YPL064C, in which the energy spectrum defined by EIIP and by us has a largest peak at frequency 0.33rad/sample; however, the background noise increases when Voss and MEM Spectrum are evaluated (see Fig. 3).

Another reason is that although the existence of three-base periodicity in CDS be a classical discriminatory frequency in the biological context, some CDS do not seem to be distinguished by it. For instance, all methods indicate that the largest peak occurs at the frequency 0.33rad/sample for the gene YPL061W (see Fig. 4). On the other hand, all methods indicate that the largest peak occurs at the frequency 0.04rad/sample for the gene YPL278C (see Fig. 5).

Although there are such limitations in discrimination of three-base periodicity for a CDS, the results from the Table I shown that our proposed method has the higher correct rate for CDS identification and the MEM Spectrum was the method with the lowest correct rate. Besides that, MEM Spectrum requires more computational effort, being impractical when we compare with the other three methods for spectral analysis.

From Fig. 2d, 3d, 4d, and 5d, notice that by maximizing the SNR, our method improves not only the graphical analysis but also the SNR and spectral entropy. Hence, the background noise of DNA spectrum is reduced and the CDS identification is improved. The most of the incorrect status returned by our proposed method (12.9%) occurs when there is no largest energy peak at frequency 0.33rad/sample also for the Voss and EIIP methods. In the remains of the incorrect classification (2.7%) of the proposed method, either the Voss or EIIP method also returns an incorrect classification. However, even for those CDS whose discriminatory frequency is not 0.33rad/sample for both Voss and EIIP, our proposed method improves the SNR and spectral entropy and (see Fig. 5d). Moreover, among the correct classification returned by our proposed method, in 11.5% of them the Voss and EIIP methods return an incorrect classification.

V. Conclusions

The new method for CDS classification of genes proposed in this paper was implemented in Python, a free open-source language, and was evaluated for CDS of genes from chromosome XVI of *Saccharomyces cerevisiae*. The method consists in defining a mapping for a given DNA sequence whose SNR of its spectrum is maximized subject to the spectral entropy constraint being greater than zero.

It is important to point out that the resulting signal from the numerical representation of a given DNA sequence is the linear combination of four indicator binary functions and there is a spectrum for each mapping $\mathcal{M}$. The SNR estimator then computes the signal power as the energy of the highest spectral component that occurs at an unknown frequency. For this reason, it is not evident and needs to be verified whether an approach that uses SNR also improves CDS classification.

In this sense, we compare our proposed method with the other three in the literature: Voss [1], EIIP [18] and MEM Spectrum [13]. Our method not only improves the CDS identification rate but also improves the SNR and spectral entropy even for those CDS whose discriminatory frequency is not in range 0.31 and 0.35rad/sample.

References

[1] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in dna base sequences," *Physical review letters*, vol. 68, pp. 3805–3808, Jun 1992.

[2] J. W. Fickett, "Recognition of protein coding regions in dna sequences," *Nucleic acids research*, vol. 10, no. 17, pp. 5303–5318, 1982.

[3] E. N. Trifonov and J. L. Sussman, "The pitch of chromatin dna is reflected in its nucleotide sequence," *Proceedings of the National Academy of Sciences*, vol. 77, no. 7, pp. 3816–3820, 1980.

[4] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, "Periodicity in dna coding sequences: implications in gene evolution," *Journal of theoretical biology*, vol. 151, no. 3, pp. 323–331, 1991.

[5] J. C. Shepherd, "Periodic correlations in dna sequences and evidence suggesting their evolutionary origin in a comma-less genetic code," *Journal of molecular evolution*, vol. 17, no. 2, pp. 94–102, 1981.

[6] J. Sánchez and I. Lopez-Villasenor, "A simple model to explain three-base periodicity in coding dna," *FEBS letters*, vol. 580, no. 27, pp. 6413–6422, 2006.

[7] E. D. Howe and J. S. Song, "Categorical spectral analysis of periodicity in human and viral genomes," *Nucleic acids research*, vol. 41, no. 3, pp. 1395–1405, 2013.

[8] B. Silverman and R. Linsker, "A measure of dna periodicity.," *Journal of theoretical biology*, vol. 118, no. 3, p. 295, 1986.

[9] E. Coward, "Equivalence of two fourier methods for biological sequences," *Journal of Mathematical Biology*, vol. 36, no. 1, pp. 64–70, 1997.

[10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by fourier analysis of genomic sequences," *Bioinformatics*, vol. 13, no. 3, pp. 263–270, 1997.

[11] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[12] P. Vaidyanathan and B.-J. Yoon, "Digital filters for gene prediction applications," in *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.*, vol. 1, pp. 306–310, IEEE, 2002.

[13] L. Galleani and R. Garello, "The minimum entropy mapping spectrum of a dna sequence," *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 771–783, 2010.

[14] S. S. Sahu and G. Panda, "Identification of protein-coding regions in dna sequences using a time-frequency filtering approach," *Genomics, proteomics & bioinformatics*, vol. 9, no. 1-2, pp. 45–55, 2011.

[15] M. Roy and S. Barman, "Effective gene prediction by high resolution frequency estimator based on least-norm solution technique," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2014, no. 1, p. 2, 2014.

[16] A. Oppenheim, A. Willsky, S. Nawab, W. Hamid, and I. Young, *Signals & Systems*. Prentice-Hall signal processing series, Prentice Hall, 1997.

[17] P. D. Cristea, "Representation and analysis of dna sequences," in *Genomic Signal Processing and Statistics* (E. R. Dougherty, I. Shmulevich, J. Chen, and Z. J. Wang, eds.), vol. 2, pp. 15–66, Hindawi, 2005.

[18] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (eiip)," *Bioinformation*, vol. 1, no. 6, p. 197, 2006.

[19] NCBI, "Nucleotide[internet].." Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 1988.