Identification of Coding Regions in Prokaryotic DNA Sequences Using Bayesian Classification

Mohammad Al Bataineh[1][0000-1111-2222-3333]

[1] Yarmouk University, Irbid 21163, Jordan
mohamadfa@yu.edu.jo

Abstract. The identification of protein-coding regions in genomic DNA sequences is a well-known problem in computational genomics. Various computational algorithms can be employed to achieve the identification process. The rapid advances in this field have motivated the development of innovative engineering methods that allow for further analysis and modeling of many processes in molecular biology. The proposed algorithm utilizes well-known concepts in communications theory, such as correlation, the maximal ratio combining (MRC) algorithm, and filtering techniques to create a signal whose maxima and minima indicate coding and noncoding regions, respectively. The proposed algorithm investigates several prokaryotic genome sequences. Two Bayesian classifiers are designed to test and evaluate the performance of the proposed algorithm. The obtained simulation results prove that the algorithm can efficiently and accurately detect protein-coding regions, which is being demonstrated by the obtained sensitivity and specificity values that are comparable to well-known gene detection methods in prokaryotes. The obtained results further verify the correctness and the biological relevance of using communications theory concepts for genomic sequence analysis.

Keywords: Gene identification; correlation; maximal ratio combining; period-3 filter; Bayesian classification.

1 Introduction

The rapid pace of advancement in computational genomics and bioinformatics has led to several innovative engineering methods for data acquisition, interpretation, and analysis. Techniques from the information theory [1]–[3], communications [4]–[11], coding theory [12]–[15], signal processing [16]–[18], machine learning [19] and various statistical methods [20], [21] have been actively researched for use in gene detection, genomic sequence analysis and alignment.

Genes are the segments of DNA that contain the coding information required for protein synthesis. A considerable target of genomic research is to understand the nature and role of the coding and noncoding information embedded in the DNA sequence structure. A crucial step in attaining this target is the detection of the gene locations in the entire DNA sequence. Several diverse methods have been proposed in the literature for gene detection in prokaryotes. For

example, probabilistic methods [22], [23], statistical methods [24]–[26], and other computational techniques including: machine learning [27], free energy calculations [28], support vector machine [29], Bayesian methods [16] information theory [30], hidden Markov model such GeneMark [20], [21], [31]–[34], and interpolated Markov model such as GLIMMER [35].

The design of a general gene identification algorithm is a compelling research problem. The gene identification method presented in this work utilizes a particular property of the DNA protein-coding regions, namely the period-3 property [36], [37] in a new novel approach using concepts from communications theory. The period-3 pattern generally gives a strong indication of the existence of coding regions. By mapping the DNA sequences to digital signals, standard digital signal processing (DSP) techniques can be implemented. The discrete Fourier transformation (DFT) [36], digital filtering [38], [39], wavelet transformations [40], Markov modeling [41] and Infinite Impulse Response (IIR) filtering [42] have shown fair performance in the detection of this period-3 behavior and, therefore, in identifying coding regions. In [17], we have proposed a novel algorithm for identifying protein-coding regions in the DNA sequences based on the period-3 property. The proposed algorithm in [17] identifies protein-coding regions by applying a digital correlating and filtering process to the entire genomic sequence under study. However, our proposed algorithm in this paper is both an enhanced and a generalized version of the work in [17] in terms of methodology, performance, and experimental validation.

This paper proposes a novel application of principles and techniques from communications theory and digital signal processing for the detection and identification of protein-coding regions in prokaryotic genomes. The proposed algorithm employs polyphase complex mapping to provide a numerical representation of the genomic sequences involved in the analysis and then uses basic concepts from communications theory and digital signal processing such as correlation, maximal ratio combining (MRC) algorithms and filtering to generate a signal whose peaks signify locations of coding regions and whose troughs signify locations of noncoding regions. The proposed gene detection algorithm investigates the complete genome sequences of several prokaryotes (e.g., MG1655 and O157H7 E. coli bacterial strains). Moreover, two Bayesian classifiers are designed to evaluate the performance of the proposed gene detection algorithm and compare it to well-known gene detection methods in prokaryotes. The conducted simulation show that the proposed algorithm can efficiently and accurately identify protein-coding regions, which is being demonstrated by the obtained sensitivity and specificity values that are comparable to the ones obtained by GLIMMER and GeneMark. The obtained results further verify the correctness and the biological relevance of using communications theory concepts for genomic sequence analysis.

The paper organization is as follows. Section 2 highlights the so-called period-3 behavior and how it can be detected using digital signal processing techniques (like the DFT) to locate protein-coding regions in the genomic structure. Section 3 presents the mathematical description of the proposed gene detection algorithm. **Algorithm I** describes the method used for maxima (corresponding to coding regions) and minima (corresponding to noncoding regions) detection. It also describes two period-3 based Bayesian classifiers designed to asses its. **Algorithm II** describes the period-3 based classification system proposed. Section 4 presents the simulation. Finally, Section 5 concludes the paper.

2 The proposed gene detection algorithm

Fig. 1 shows a schematic system-like representation of the proposed algorithm. The input parameter of the proposed detection algorithm is the genomic sequence under study, **g**, of length L_x. The output parameters are three sequences: $f[n]$, $p[n]$ and $t[n]$ whose lengths are the same as the length of the input test sequence L_x. The sequence $f[n]$ represents the correlation of the genomic sequence **g** with twenty-four hypothetically generated period-3 based subsequences after passing through a maximal ratio combining module. The latter two sequences, $p[n]$ and $t[n]$, correspond to the detected peaks and troughs in the sequence $f[n]$, respectively. Detected peaks signify coding regions while detected troughs signify noncoding regions**.**

2.1 Mathematical Description

The following procedural steps describe the proposed gene detection algorithm:

1. Convert the input genome sequence **g** (of length L_x nucleobases) to a numerical representation $\boldsymbol{x}[\boldsymbol{n}]$ using polyphase complex mapping ($A = 1,\ C = +j,\ G = -1,\ T = -j$) [43], [44]. Such mapping will allow for performing signal processing operations, such as correlation in the following steps.
2. Generate all possible hypothetical sequences ($\boldsymbol{r_i}$) that exhibit a clear period-3 pattern. Each sequence is a periodic repetition of three different nucleobases out of the four possible genetic code alphabet letters $\{A, C, G, T\}$ [45]. This approach will result in twenty-four possible sequences ($\boldsymbol{r_1}, \boldsymbol{r_2}, \dots, \boldsymbol{r_{24}}$) where each sequence is of a predefined length of ($3L_r$) nucleobases, where L_r is the number of repetitions used for each sequence.
3. Convert the hypothetical sequences ($\boldsymbol{r_i}$) obtained in step 2 to their corresponding numerical representations, $s_i[n]$, using the polyphase complex mapping described in step 1.
4. Correlate the signal, $x[n]$, with each one of the 24 hypothetical sequences, $s_i[n]$. The correlation of two signals measures how similar they are. Hence, this step will help detect the portions of the input genomic sequence that have a period-3 pattern like anyone of the twenty-four sequences, $s_i[n]$. The twenty-four corresponding correlation outputs denoted as $y_i[n]$, are the convolution of the input sequence, $x[n]$, with the time-reversed version of each of the twenty-four sequences, $s_i[n]$, as

$$y_i[n] = \sum_{m=0}^{3L_r-1} x[m]s_i[m-n], \tag{1}$$

   where each signal $y_i[n]$ is of length $(L_x + 3L_r - 1)$.
5. Truncate the first $\lfloor (3L_r - 1)/2 \rfloor$ and the last $\lceil (3L_r - 1)/2 \rceil$ elements of each sequence, $y_i[n]$. Such a step will force the latter sequences to be of a length equal to the length of the input genomic sequence, $x[n]$, which is L_x. The integer numbers $\lfloor (3L_r - 1)/2 \rfloor$ and $\lceil (3L_r - 1)/2 \rceil$ are equal if $(3L_r)$ is odd and different by one if $(3L_r)$ is even. Here, the symbol $\lfloor \cdot \rfloor$ is the largest integer not greater than the argument, while $\lceil \cdot \rceil$ is the smallest integer not less than the argument.
6. Pass the sequences, $y_i[n]$, obtained in step 5 into a maximal ratio combining (MRC) module. This combining technique multiplies each of the twenty-four signal branches by a weight factor, (α_i), that is proportional to the signal amplitude in each branch. That is to say; the combining technique further amplifies the branches with strong signals (high

amplitude), while attenuating weak signals (low amplitude). The resulting signal at the output of the MRC module is given by

$$z[n] = \sum_{i=1}^{24} \alpha_i y_i[n], \tag{2}$$

where the factors (α_i), are defined by

$$\alpha_i = \frac{|y_i[n]|}{\sum_{j=1}^{24} |y_j[n]|}. \tag{3}$$

7. Pass the resulting sequence, $z[n]$, obtained in (6) through a period-3 bandpass filter of length N with about 13 dB minimum stopband attenuation like the one described in (3). The passband center is $(\omega_0 = 2\pi/3)$. Hence, the response of the period-3 filter is given by

$$f[n] = \sum_{m=0}^{N-1} w[m] z[n-m], \tag{4}$$

whose length is $(L_x + N - 1)$.

8. Truncate the first $\lfloor (N-1)/2 \rfloor$ and the last $\lceil (N-1)/2 \rceil$ elements of $f[n]$ to force its length to be L_x elements.

The peaks of the sequence $|f[n]|$ correspond to a high correlation between $z[n]$ and $w[n]$, while troughs correspond to a low correlation. A high correlation in this context corresponds to the occurrence of a period-3 structure, and vice versa. Hence, peaks of $f[n]$ signify coding regions in both forward and reverse strands of the genomic sequence, **g**, while troughs signify noncoding regions. **Algorithm I** describes the detection method of maxima and minima. The detection of maxima (peaks) and minima (troughs) utilizes a sliding a window of length L_p for maxima and L_t for minima through the whole correlation sequence, $f[n]$, obtained in (8). At each alignment instant, maxima and minima are kept while setting all other values to zero. This step continues to generate the detected peak and trough sequences, $p[n]$ and $t[n]$, respectively.

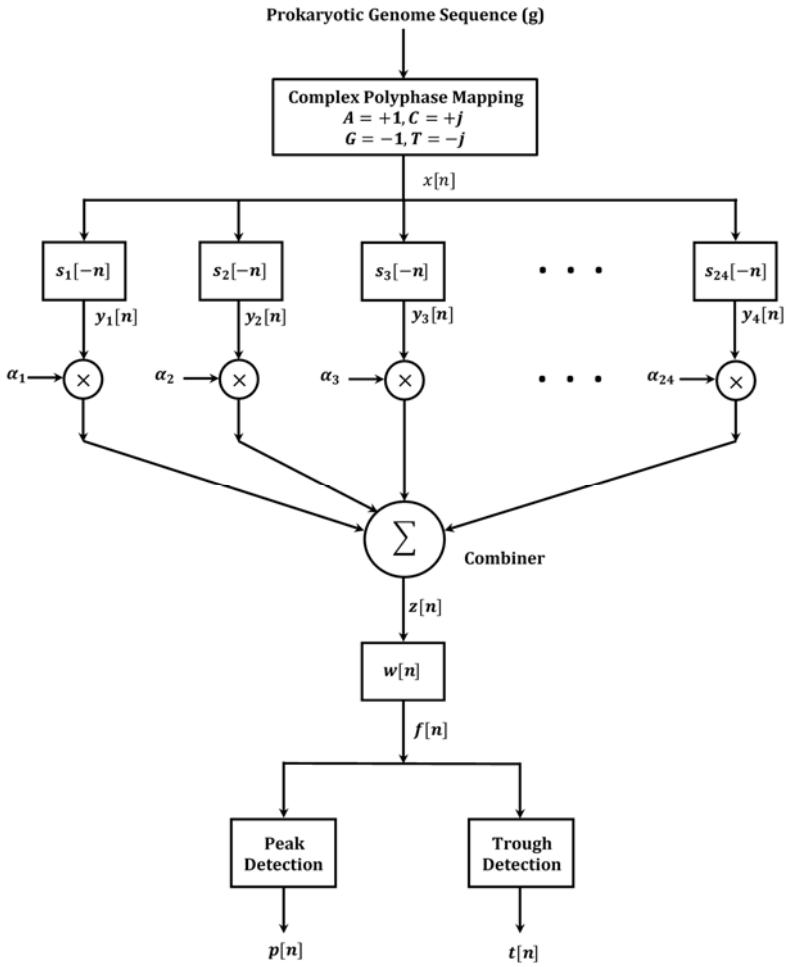

Fig. 1. The Proposed Gene Detection Algorithm

Algorithm I: Peaks and Troughs Detection Algorithm

Input: The period-3 filter output sequence,$f[n]$, obtained in (4), whose length is L_x(after being truncated); the peaks detection window length, L_p; and the troughs detection window length, L_t.

Output: The detected peaks sequence, $p[n]$, and the detected troughs sequence, $t[n]$.

Initialization: $p[n]^{(0)} = \{0,0,\dots,0\}$, $t[n]^{(0)} = \{0,0,\dots,0\}$ where $0 \le n \le L_x$.

Peaks Detection Algorithm:

For $i = 1, 2, \dots, L_x - 2L_p$, **do**

  If $\left(i > L_p\right)$ and $(i < L_x - L_p)$ **then**

- Set max to the maximum of the subsequence $f[i - L_p, i - L_p + 1, \dots, i + L_p]$

  If $f[i] \neq max$ **then**

  - Set $p[i]$ to zero,

  Else (i.e. if $f[i] = max$)

  - Set $p[i]$ to max,

Troughs Detection Algorithm:

For $i = 1, 2, \dots, L_x - 2L_t$, **do**

If $(i > L_t)$ and $(i < L_x - L_t)$ **then**

- Set min to the minimum of the subsequence $f[i - L_t, i - L_t + 1, \dots, i + L_t]$

  If $f[i] \neq min$ **then**

  - Set $t[i]$ to zero,

  Else (i.e. if $f[i] = min$)

  - Set $t[i]$ to min.

3 Simulation Results and Analysis

To demonstrate the fidelity and biological significance of the proposed gene detection algorithm, several prokaryotic genome sequences are investigated. For example, the complete genome sequence of Escherichia coli bacterial strains MG1655 and O157:H7 are used as input test sequences. Such sequences are available at the NCBI [46].

The length of each one of the 24-hypothetical period-3 based subsequences, $s_i[n]$, is selected to be 1950 (i.e., the number of repetitions, L_r, is 650). The length of the period-3 filter, $w[n]$, is selected as $N = 221$. The peaks detection window length, L_p, and the troughs detections window, L_t, are selected as 300 and 800, respectively.

Fig. 2 and Fig. 3 show the simulation result obtained by applying the proposed gene detection algorithm to test sequences extracted from the complete genome sequences of the E. coli MG1655 and O157:H7 bacterial strains, respectively. In both figures, the red-colored regions (second row of rectangles) correspond to the $5' - 3'$genome sequence (forward strand), while the green-colored regions (the upper group of rectangles) correspond to the coding regions of the $3' - 5'$ genome sequence (reverse strand). The blank regions in the latter two sequences correspond to the noncoding regions in both forward and reverse strands. It can be noticed that the proposed algorithm using the peaks and troughs detection method, described by **Algorithm I**, is able to detect the coding regions (signified by the blue-colored lines with closed circles on top) and the noncoding regions (signified by the green-colored lines with cross signs on top) in both forward and reverse strands. The detected peaks/troughs indicate the existence of a coding/noncoding region in either the forward or reverse strands, respectively.

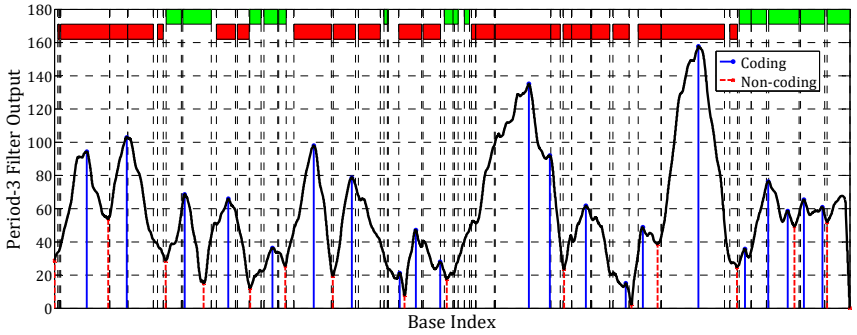

Fig. 2. Period-3 Filter Output with Peaks and Troughs identified (Applied

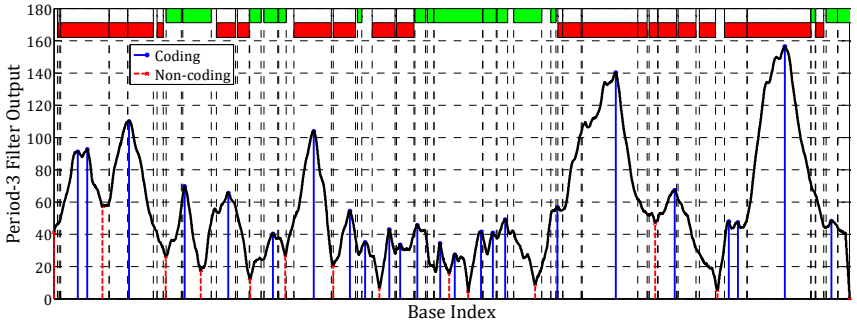

Fig. 3. Period-3 Filter Output with Peaks and Troughs identified (Applied to O157:H7).

A closer look at Fig. 2 and Fig. 3 shows that there are some peaks detected in the noncoding regions of the $5' - 3'$genome sequence (forward strand) that do not correspond to coding regions in the latter strand. This can be justified by the fact that there are coding regions (that also possess a period-3 structure) in the complementary $3' - 5'$ sequence (reverse strand), that happen to coincide with noncoding regions in the $5' - 3'$ sequence (forward strand). This remark is also being confirmed by the fact that the forward and reverse strands are totally complementary and statistically symmetric [47]. Therefore, the existence of a period-3 structure in the forward strand would imply the existence of a period-3 structure in the reverse strand as well, and vice versa. However, it can be observed that such ambiguous detected peaks have smaller amplitudes than the ones that do not coincide with coding regions in the complementary strand. In other words, the peaks with larger amplitudes most likely correspond to coding regions in the genomic sequence under study (and that is the forward strand here).

The obtained results of simulation verify that the algorithm proposed here is successful in detecting the period-3 structure embedded in the prokaryotic genomic sequence under study. However, due to the ambiguity of whether the detected peaks (with small amplitudes) correspond to coding regions in either the forward or reverse strands, the proposed gene detection method can be integrated with other detection methods, so that the areas between neighboring red lines

can be utilized to identify the coding regions. This may also be confirmed with the blue lines that point out the locations of the coding regions. The proposed algorithm will significantly increase the efficiency of coding region detection.

For the prokaryotic genome sequences (MG1655 and O157:H7), the length of the peaks detection window, (L_p), can be selected by compromising the correctness and resolution. A smaller L_p will result in more peaks and hence, better resolution. However, decreasing L_p will also produce more fake peaks that are not real maximization points that correspond to real coding regions. In a similar way, L_t can be selected as well.

It is evident that the previous simulation results are basically visual and do not always provide a sharp decision of where coding and noncoding regions precisely lie in the entire genome sequence. Due to this uncertainty in detection, we have designed two Bayesian classifiers described in detail in Section 3.2. These classifiers will allow for assessing the performance of the proposed gene detection algorithm and comparing it to well-known gene detection algorithms. Therefore, a set of all possible ORFs in the genome sequence understudy is generated. An ORF is selected if (i) it starts with a valid initiation codon (ATG, GTG or TTG), (ii) it terminates with a valid termination codon (TAG, TAA or TGA) and (iii) is at least 99 nucleobases long. This latter data set is then divided in half to form the training set and the testing set for classification. The statistical models for the two Bayesian classifiers can be constructed by training the proposed classification algorithm using the training set. Subsequently, the classification algorithm is tested using the testing data set to verify its performance in detection. Table 1 shows the obtained results of the two Bayesian classifiers when applied to the E. coli MG1655 bacterial strain being compared to the GLIMMER and GeneMark gene-finding software. The performance of the two Bayesian classifiers is assessed using the True Positive Rate (TPR, also referred to as sensitivity), the False Positive Rate (FPR, also known as fall-out), the True Negative Rate (TNR), also referred to as specificity, and the False Negative Rate (FNR). The four performance rates are defined by [48].

$$TPR = \frac{TP}{TP + FN} \times 100\%, \qquad (5) \qquad FPR = \frac{FP}{FP + TN} \times 100\%, \qquad (6)$$

$$FNR = \frac{FN}{TP + FN} \times 100\%, \qquad 7) \qquad TNR = \frac{TN}{FP + TN} \times 100\%. \qquad (8)$$

where TP, FP, FN, TN correspond to True positives, False Positives, False Negatives, and True Negatives, respectively.

The sensitivity or the TPR measure is the proportion of coding nucleotides that have been correctly predicted as coding. Additionally, the specificity (or the TNR) measure is the proportion of noncoding nucleotides that have been correctly predicted as noncoding. Both sensitivity and specificity range independently over [0,1]. However, neither sensitivity no specificity alone constitutes good measures of global accuracy. Alternatively, in the gene structure prediction literature, the preferred measure of global accuracy has traditionally been the Correlation Coefficient (CC) defined as

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}. \qquad (9)$$

Another performance measure is the Approximate Correlation Coefficient (AC), which approximates the behavior of the Correlation Coefficient (CC). It has been observed that $|AC| \geq |CC|$. In consequence, the AC measures the association between prediction and reality

appropriately and can thus be used as an alternative to the CC. Unlike the CC, the AC has a probabilistic interpretation, and it can be computed in any circumstance. The AC, introduced in [48], is defined as

$$AC = (ACP - 0.5) \times 2, \tag{10}$$

where ACP is the Average Conditional Probability defined as:

$$ACP = \frac{1}{4}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right). \tag{11}$$

Since at least two of the conditional probabilities in (20) are always defined, ACP can always be calculated as the average of the one defined. CC and AC range over [-1, 1] and usually are close to each other whenever CC is defined.

Of the two Bayesian classifiers, Classifier 1 with equal prior probabilities seems to perform the best in terms of sensitivity (or TPR) and FNR. Hence, Classifier 1 outperforms both GLIMMER and GeneMark in both TPR and FNR. However, GLIMMER and GeneMark provide better performance in terms of FPR and TNR. Classifier 2 performs better than Classifier 1 in terms of specificity. Moreover, about both CC and AC, Classifier 1 slightly performs better than Classifier 2. Overall, Classifier 1 is performing better than Classifier 2.

As claimed earlier, the proposed gene detection algorithm can efficiently extract the period-3 component and hence effectively identify protein-coding regions in the whole genomic sequences of prokaryotes. Also, it can effectively suppress the background $1/f$ noise with no added computational complexity. For a computer with an Intel(R) Core (TM) i7-4770 CPU @ 3.40GHz and 3.40GHz, on a genomic test sequence of 40,386 nucleotides long, the proposed gene detection algorithm takes about 9.7 seconds computational time to obtain the results.

Table 1. Performance evaluation of Classifier 1 and Classifier 2 compared to both GLIMMER and GeneMark with equal/unequal prior probabilities.

Classifier	(P_{w_1}, P_{w_2})	TP (TPR)	FP (FPR)	FN (FNR)	TN (TNR)	CC	AC
GLIMMER	–	3561 (85.99%)	915 (0.21%)	580 (14.01%)	434618 (99.79%)	0.8254	0.8260
GeneMark	–	3683 (88.94%)	694 (0.16%)	458 (11.06%)	434839 (99.84%)	0.8638	0.8641
Classifier 1	(0.5, 0.5)	3858 (93.17%)	110387 (25.35%)	283 (6.83%)	325146 (74.65%)	0.1494	0.3556
	(0.43, 0.57)	3712 (89.64%)	97264 (22.33%)	429 (10.36%)	338269 (77.67%)	0.1546	0.3543
Classifier 2	(0.5, 0.5)	3384 (81.72%)	101803 (23.37%)	757 (18.28%)	333730 (76.63%)	0.1321	0.3067
	(0.43, 0.57)	3100 (74.86%)	73446 (16.86%)	1041 (25.14%)	362087 (83.14%)	0.1477	0.3088

4 Conclusions

This paper proposes a novel application of principles and concepts from communications theory and digital signal processing for the detection of protein-coding regions in prokaryotic genomes. The proposed gene detection algorithm employs polyphase complex mapping to provide a

numerical representation of the genomic sequences involved in the analysis, and then uses basic concepts from communications theory and digital signal processing as correlation, maximal ratio combining (MRC) algorithms and filtering to generate a signal whose peaks signify locations of coding regions and whose troughs signify locations of noncoding regions. The proposed gene detection algorithm is applied to the complete genome sequences of several prokaryotes (e.g., MG1655 and O157H7 E. coli bacterial strains). Two Bayesian classifiers are designed for the performance evaluation of the proposed gene detection algorithm. Especially, the performance of the proposed algorithm is compared to the well-known ab initio gene detection methods: GLIMMER and GeneMark. The results verify that the algorithm can accurately and efficiently identify protein-coding regions with sensitivity and specificity values that sometimes outperform GLIMMER and GeneMark gene detection methods. The gene detection algorithm does not require any prior information about the coding regions, as does the DFT method described in [36]. Moreover, the proposed algorithm outperforms traditional methods as it puts off the need to apply a variable-length discrete Fourier transform.

References

1. G. Atkins, *Information theory and molecular biology*, vol. 327, no. 1. New York, NY: Cambridge University Press, 1993.
2. G. Battail, "Information theory and error-correcting codes in genetics and biological evolution," *Introd. to Biosemiotics New Biol. Synth.*, pp. 299–345, 2007.
3. J. Weindl, P. Hanus, Z. Dawy, J. Zech, J. Hagenauer, and J. C. Mueller, "Modeling DNA-binding of Escherichia coli sigma(70) exhibits a characteristic energy landscape around strong promoters," *Nucleic Acids Res.*, vol. 35, no. 20, pp. 7003–7010, 2007.
4. M. Al Bataineh and Z. Al-qudah, "Cognitive interference channel: achievable rate region and power allocation," *IET Commun.*, vol. 9, no. 2, pp. 249–257, 2015.
5. M. Al Bataineh, L. Huang, and G. Atkin, "TFBS detection algorithm using distance metrics based on center of mass and polyphase mapping," *2012 7th Int. Symp. Heal. Informatics Bioinforma.*, no. 1, pp. 37–40, 2012.
6. M. Al Bataineh, "Analysis of Genomic Translation Using a Communications Theory Approach," Illinois Institute of Technology, Chicago, 2010.
7. M. Al Bataineh, M. Alonso, S. Wang, W. Zhang, and G. Atkin, "Ribosome Binding Model Using a Codebook and Exponential Metric," *2007 IEEE Int. Conf. Electro/Information Technol.*, pp. 438–442, 2007.
8. M. Al Bataineh, L. Huang, I. Muhamed, N. Menhart, and G. E. Atkin, "Gene Expression Analysis using Communications, Coding and Information Theory Based Models," *BIOCOMP'09 - 2009 Int. Conf. Bioinforma. Comput. Biol.*, pp. 181–185, 2009.
9. M. Al Bataineh, L. Huang, M. Alonso, N. Menhart, and G. E. Atkin, "Analysis of gene translation using a communications theory approach," in *Advances in Experimental Medicine and Biology*, 2010, vol. 680, pp. 387–397.
10. L. Huang *et al.*, "Identification of Transcription Factor Binding Sites Based on the Chi-Square (X^2) distance of a Probabilistic Vector Model," *2009 Int. Conf. Futur. Biomed. Inf. Eng. (FBIE 2009)*, pp. 73–76, 2009.
11. J. Weindl and J. Hagenauer, "Applying Techniques from Frame Synchronization for Biological Sequence Analysis," *IEEE Int. Conf. Commun.*, pp. 833–838, 2007.
12. D. J. Reiss and B. Schwikowski, "Predicting protein-peptide interactions via a network-based motif sampler," *Bioinformatics*, vol. 20, no. SUPPL. 1, 2004.
13. Z. Dawy, P. Hanus, J. Weindl, J. Dingel, and F. Morcos, "On genomic coding theory," *Eur. Trans. Telecommun.*, vol. 18, no. 8, pp. 873–879, 2007.

14. G. L. Rosen and J. D. Moore, “Investigation of coding structure in DNA,” *2003 IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP ’03).*, vol. 2, pp. 361–364, 2003.
15. D. A. MacDonaill, “Digital parity and the composition of the nucleotide alphabet. Shaping the alphabet with error coding,” *IEEE Eng Med Biol Mag*, vol. 25, no. 1, pp. 54–61, 2006.
16. E. M. Crowley, “A Bayesian method for finding regulatory segments in DNA,” *Biopolymers*, vol. 58, no. 2, pp. 165–174, 2001.
17. L. Huang, M. A. Bataineh, G. E. Atkin, S. Wang, and W. Zhang, “A Novel gene detection method based on period-3 property,” *Conf Proc IEEE Eng Med Biol Soc*, vol. 2009, pp. 3857–3860, 2009.
18. R. Kakumani, V. Devabhaktuni, and M. O. Ahmad, “Prediction of protein-coding regions in DNA sequences using a model-based approach,” *ISCAS 2008*, vol. 18, no. 21, pp. 1918–1921, 2008.
19. E. C. Uberbacher and R. J. Mural, “Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach,” *Proc Natl Acad Sci U S A*, vol. 88, no. 24, pp. 11261–11265, 1991.
20. J. Henderson, S. Salzberg, and K. H. Fasman, “Finding genes in DNA with a Hidden Markov Model,” *J. Comput. Biol.*, vol. 4, no. 2, pp. 127–141, 1997.
21. S. R. Eddy, “Hidden Markov models and genome sequence analysis,” *Faseb J.*, vol. 12, no. 8, pp. A1327–A1327, 1998.
22. T. Yada, Y. Totoki, T. Takagi, and K. Nakai, “A novel bacterial gene-finding system with improved accuracy in locating start codons,” *DNA Res.*, vol. 8, no. 3, pp. 97–106, 2001.
23. J. Besemer, A. Lomsadze, and M. Borodovsky, “GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions,” *Nucleic Acids Res.*, vol. 29, no. 12, pp. 2607–2618, 2001.
24. M. Walker, V. Pavlovic, and S. Kasif, “A comparative genomic method for computational identification of prokaryotic translation initiation sites,” *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3181–3191, 2002.
25. S. S. Hannenhalli, W. S. Hayes, A. G. Hatzigeorgiou, and J. W. Fickett, “Bacterial start site prediction,” *Nucleic Acids Res.*, vol. 27, no. 17, pp. 3577–3582, 1999.
26. T. Nishi, T. Ikemura, and S. Kanaya, “GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences,” *Gene*, vol. 346, pp. 115–125, 2005.
27. W. S. Hayes and M. Borodovsky, “How to interpret an anonymous bacterial genome: machine learning approach to gene identification,” *Genome Res.*, vol. 8, no. 11, pp. 1154–1171, 1998.
28. Y. Osada, R. Saito, and M. Tomita, “Analysis of base-pairing potentials between 16S rRNA and 5’ UTR for translation initiation in various prokaryotes,” *Bioinformatics*, vol. 15, no. 7, pp. 578–581, 1999.
29. A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller, “Engineering support vector machine kernels that recognize translation initiation sites,” *Bioinformatics*, vol. 16, no. 9, pp. 799–807, 2000.
30. T. D. Schneider, “Measuring molecular information,” *J. Theor. Biol.*, vol. 201, no. 1, pp. 87–92, 1999.
31. J. Besemer and M. Borodovsky, “GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses,” *Nucleic Acids Res.*, vol. 33, no. suppl 2, pp. W451–W454, 2005.
32. R. Raman and G. C. Overton, “Application of hidden markov modeling in the characterization of transcription factor binding sites,” *Proc. Twenty-Seventh Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 5, pp. 275–283, 1994.

33. A. Krogh, I. S. Mian, and D. Haussler, “A Hidden Markov Model That Finds Genes in Escherichia-Coli DNA,” *Nucleic Acids Res.*, vol. 22, no. 22, pp. 4768–4778, 1994.
34. S. R. Eddy, “Hidden Markov models,” *Curr. Opin. Struct. Biol.*, vol. 6, no. 3, pp. 361–365, 1996.
35. A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg, “Identifying bacterial genes and endosymbiont DNA with Glimmer,” *Bioinformatics*, vol. 23, no. 6, pp. 673–679, 2007.
36. P. P. Vaidyanathan, “Genomics and proteomics: a signal processor’s tour,” *Circuits Syst. Mag. IEEE*, vol. 4, no. 4, pp. 6–29, 2004.
37. M. Al Bataineh and Z. Al-qudah, “A novel gene identification algorithm with Bayesian classification,” *Biomed. Signal Process. Control*, vol. 31, pp. 6–15, 2017.
38. R. Guan and J. Tuqan, “IIR filter design for gene identification,” *Gensips Proc.* Baltimore, Maryland, 2004.
39. P. Vaidyanathan and B. Yoon, “Gene and exon prediction using allpass-based filters,” *Work. Genomic Signal …*, vol. 3, 2002.
40. K.B. Murray, D. Gorse, and J. M. Thornton, “Wavelet Transforms for the Characterization and Detection of Repeating Motifs,” *J. Mol. Biol.*, vol. 316, pp. 341–363, 2002.
41. M. Borodovsky and S. Ekisheva, *Problems and solutions in biological sequence analysis*. Cambridge ; New York: Cambridge University Press, 2006.
42. P. P. Vaidyanathan and B. Yoon, “Digital Filters for Gene Prediction Applications,” *Proc. 36th Asilomar Conference on Signals, Systems, and Computers*. Monterey, CA, 2002.
43. S. D. Sharma, K. Shakya, and S. N. Sharma, “Evaluation of DNA mapping schemes for exon detection,” *2011 Int. Conf. Comput. Commun. Electr. Technol. ICCCET 2011*, pp. 71–74, 2011.
44. D. Anastassiou, “Genomic signal processing,” *IEEE Signal Proc. Mag.*, pp. 8–20, 2001.
45. P. Rangel and J. Giovannetti, *Genomes and Databases on the Internet: A practical Guide to Functions and Applications*. Horizon Scientific Press, 2002.
46. K. D. Pruitt, T. Tatusova, and D. R. Maglott, “NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, 2007.
47. P. F. Baisnee, S. Hampson, and P. Baldi, “Why are complementary DNA strands symmetric?,” *Bioinformatics*, vol. 18, no. 8, pp. 1021–1033, 2002.
48. M. Burset and R. Guigó, “Evaluation of gene structure prediction programs,” *Genomics*, vol. 34, no. 3, pp. 353–67, 1996.