Biomedical Signal Processing and Control 31 (2017) 6–15

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A novel gene identification algorithm with Bayesian classification

Mohammad Al Bataineh[a,*], Zouhair Al-qudah[b]

[a] Telecomunications Engineering Department, Yarmouk University, Irbid, Jordan
[b] Communication Engineering Department, Al-Hussein bin Talal University, Ma'an, Jordan

ARTICLE INFO

Article history:
Received 24 January 2016
Received in revised form 24 May 2016
Accepted 4 July 2016
Available online 18 July 2016

Keywords:
Gene detection
Correlation
Maximal ratio combining
Period-3 filter
Bayesian classification

ABSTRACT

The rapid advances in the field of computational genomics and bioinformatics have motivated the development of innovative engineering methods for data acquisition, interpretation, and analysis. With the help of the later methods, many processes in molecular biology can be modeled and further analyzed. Identification and discovery of the coding regions in the genomic structure using computational algorithms is a clear example of such processes. This work proposes a novel application of well-known principles and concepts from communications theory and digital signal processing for the detection of protein coding regions in prokaryotic genomes. The proposed algorithm employs a polyphase complex mapping scheme to provide a numerical representation of the genomic sequences involved in the analysis. It then utilizes concepts in communications theory such as correlation, the maximal ratio combining (MRC) algorithm, and filtering to generate a signal whose peaks and troughs signify coding and noncoding regions, respectively. The proposed algorithm is applied to several prokaryotic genome sequences. Two Bayesian classifiers are designed to evaluate the performance of the proposed algorithm. The obtained simulation results show that the algorithm is able to efficiently and accurately identify protein coding regions with sensitivity and specificity values comparable to well-known gene detection methods in prokaryotes such as GLIMMER and GeneMark. This further proves the relevance of using communications theory concepts for genomic sequence analysis.

1. Introduction

The rapid advances in the field computational genomics and bioinformatics have motivated the development of innovative engineering methods for data acquisition, interpretation, and analysis. Techniques from the information theory [1–7], communications [7–18], coding theory [5,19–30], signal processing [31–36], machine learning [37] and various statistical methods [38–44] have been actively researched for use in gene detection, genomic sequence analysis and alignment. The developed analyses made possible by the use of the latter methods allow the testing of different biological aspects related to the process of gene expression. For example, then can help in determining whether certain regions of a given genome are protein-coding sequences (i.e. gene detection). These methods promote new interdisciplinary collaborations in research and education, integrating biomedical engineering, electrical engineering and life sciences. The knowledge gain can help address fundamentally important issues that cannot be explored systematically and quantitatively by experimentation alone. Moreover, it can reduce the consumption of laboratory resources, minimize time-consuming laboratory experimentations and lead to a better understanding of the complex genetic processes.

A DNA sequence can be divided into two types of regions: genic and intergenic spaces. 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 literature for gene detection in prokaryotes. For example, probabilistic methods [45,46], statistical methods [47–49], and other computational techniques including: machine learning [50], free energy calculations [51], support vector machine [52], Bayesian methods [33] information theory [53], hidden Markov model such GeneMark [38–40,43,44,54], and interpolated Markov model such as GLIMMER [55].

[*] Corresponding author.
E-mail addresses: mohamadfa@yu.edu.jo, mffbataineh@yahoo.com (M. Al Bataineh), qudahz@ahu.edu.jo (Z. Al-qudah).

http://dx.doi.org/10.1016/j.bspc.2016.07.002
1746-8094/© 2016 Elsevier Ltd. All rights reserved.

The algorithm employed in GLIMMER has been described in detail in Ref. [56]. The method employs an interpolated Markov model to identify coding regions. Specifically, the algorithm identifies open reading frames (ORFs) of sufficient length which are most likely to be coding to give an initial model of the coding regions of the organism. This information is subsequently used in a Markov chain in order to locate all other coding sequences. The GLIMMER system is a widely used and highly accurate gene finder for prokaryotes. On the other hand, GeneMark is an HMM-like algorithm. The algorithm introduces inhomogeneous three-periodic Markov chain models of protein-coding DNA sequence that became standard in gene prediction as well as Bayesian approach to gene prediction in the two DNA strands simultaneously. The major step of the algorithm computes for a given DNA fragment posterior probabilities of either being "protein-coding" (carrying genetic code) in each of six possible reading frames (including three frames in the complementary DNA strand) or being "non-coding".

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, specifically the period-3 property [57–63] in a new novel approach using concepts from communications theory. The period-3 pattern is generally considered as a strong indication of coding regions. By mapping the DNA sequences to digital signals, standard digital signal processing (DSP) techniques the discrete Fourier transformation (DFT) [57,58,62], digital filtering [64,65], wavelet transformations [66], Markov modeling [67] and IIR filtering [63] have shown good performance in the detection of this period-3 behavior and, therefore, in identifying coding regions. However, the efficiency of the latter methods in detecting period-3 property and suppressing the background noise is obtained at the expense of increased computational complexity. In our previous work [35], 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 Ref. [35] 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 work is both an enhanced and a generalized version of the work in Ref. [35] 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 is applied to 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 obtained simulation results show that the proposed algorithm can efficiently and accurately identify protein coding regions with sensitivity and specificity values comparable to well-known gene detection methods in prokaryotes such as GLIMMER and GeneMark. This further proves the relevance of using communications theory concepts for genomic sequence analysis. Moreover, the proposed algorithm does not entail any prior information about the coding regions as does the DFT method described in Ref. [57]. It can sharply extract the period-3 component and hence effectively identify protein coding regions in the whole genomic sequences of prokaryotes. In addition, it can effectively suppress the background $1/f$ noise with no added computational complexity.

The rest of this paper is organized as follows. Section 2 highlights the so-called period-3 behavior and how it can be detected by the use of digital signal processing techniques (like the DFT) to locate protein-coding regions in the genomic structure. Section 3 gives a detailed mathematical description of the proposed gene detection algorithm. The method used for peaks (corresponding to coding regions) and troughs (corresponding to noncoding regions) detection is described in Algorithm I. It also describes two period-3 based Bayesian classifiers that are designed to evaluate the performance of the proposed algorithm. The period-3 based classification system is described in Algorithm II. Simulation results are shown and discussed in Section 4. Finally, our paper is concluded in Section 5.

2. Protein coding region identification using the period-3 property

It has been emphasized by many articles that the coding regions of the DNA possess a period-3 property caused by codon biases in the translation of codons into amino acids. This fundamental characteristic is not detected outside the coding regions, and hence can be utilized to locate the coding regions in the entire genomic structure [58,62]. This observation can be traced back to the work of Trifonov and Sussman [68].

The period-3 property implies a clear short-range correlation behavior in the coding regions of DNA sequences. However, there exists a long-range correlation behavior embedded in both the genic and intergenic regions as well. This observation was first introduced by a paper that appeared in Nature in 1992 [69], which basically addressed a new concept named the DNA Walk. Many other subsequent articles investigated correlations over longer genomic sequences that encompassed several genes. Long-range correlations have been identified both in coding and noncoding regions [70]. Conforming with the Fourier transform theory, long-range correlations entail that the Fourier transform exhibits a 1/f -behavior in the low frequency regions [71].

Indicator sequences for nucleobases in the DNA have been used in literature to facilitate the application of digital signal processing techniques for the identification of the period-3 property. For example, the indicator sequence for the nucleobase A, denoted as $x_A[n]$, is a binary sequence with 1 indicating the presence of an A and 0 indicating its absence. Following the same mapping strategy, the indicator sequences for the other nucleobases (i.e. $x_T[n]$, $x_C[n]$ and $x_G[n]$) are defined in a similar way. In particular, the discrete Fourier transform of an N-nucleobases block of the binary sequence $x_A[n]$, denoted as $X_A(k)$, is defined as

$$X_A(k) = \sum_{n=0}^{N-1} x_A[n]\, e^{-j\frac{2\pi}{N}kn}, 0 \le k \le N-1. \tag{1}$$

The DFTs $X_T(k)$, $X_C(k)$, and $X_G(k)$ are defined similarly.

When taking N a multiple of 3 and defining $S(k)$ as

$$S(k) \triangleq |X_A(k)|^2 + |X_T(k)|^2 + |X_C(k)|^2 + |X_G(k)|^2, \tag{2}$$

a plot of $S(k)$ should then show a peak at the sample value $k = N/3$ that corresponds to $2\pi/3$ center frequency. Given a long sequence of nucleobases, $S\left(N/3\right)$ can be calculated for short windows of the data, and then slid all over the sequence. Thus, a picture of how $S\left(N/3\right)$ evolves throughout the entire DNA sequence is obtained. The window size N is required to be sufficiently large ranging from a few hundreds to a few thousands. Consequently, the periodicity effect controls the background $1/f$ spectrum [57,71,72]. However, a large window size N increases the computational complexity, and

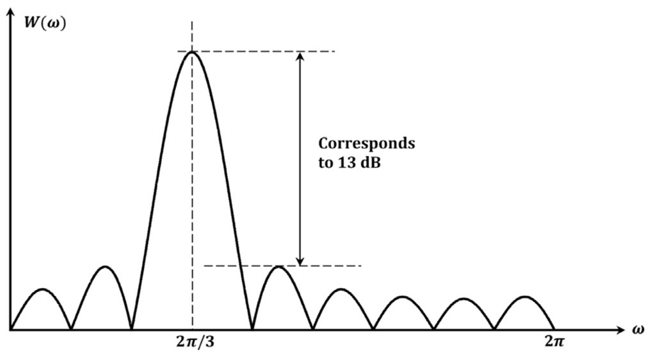


Fig. 1. Period-3 Filter Response.

also compromises the base-domain resolution in identifying the coding regions location.

Computation of the DFT with a sliding window is equivalent to applying a bandpass filter followed by downsampling at a rate determined by the separation between contiguous positions of the window [73,74]. The impulse response of the period-3 filter is represented by

$$w[n] = \begin{cases} e^{jw_0 n} & , \quad 0 \le n \le N-1 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{3}$$

The corresponding frequency response of the period-3 filter is given by

$$W(\omega) = \frac{\sin\left[\frac{(\omega-\omega_0)N}{2}\right]}{\sin\left[\frac{(\omega-\omega_0)}{2}\right]} e^{-j(\omega-\omega_0)\frac{(N-1)}{2}}, \tag{4}$$

which represents a bandpass filter whose passband is centered at $\omega_0 = 2\pi/3$ with about 13 dB minimum stopband attenuation. Fig. 1 shows the magnitude frequency response of the corresponding period-3 filter. The period-3 behavior can be effectively isolated from the background information, such as the $1/f$ noise. This can be achieved if careful attention is directed towards the design and implementation of the digital filter, and hence reducing the computational complexity.

In the previous section, a brief background is presented about using DFT for period-3 detection. However, the proposed algorithm does not use indicator sequences and DFT as described before. It uses polyphase mapping and utilizes time-domain analysis rather than frequency-domain analysis. The next section provides a detailed mathematical description of the proposed gene detection algorithm. It also describes two Bayesian classifiers designed for the purpose of the performance evaluation of the proposed algorithm. Then, some simulation results are shown to validate our theoretical study.

3. The proposed gene detection algorithm

Fig. 3 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 $\mathbf{g}$ 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 non-coding regions.

3.1. *Mathematical description of the algorithm*

The proposed gene detection algorithm can be described by the following procedural steps:

1 Convert the input genome sequence $\mathbf{g}$ (of length L_x nucleobases) to a numerical representation $\boldsymbol{x}[\boldsymbol{n}]$ using polyphase complex mapping ($A = 1, C = +j, G = -1, T = -j$) [58,75]. Such mapping will allow for performing signal processing operations such as correlation in the following steps. In this mapping, the purines (A and G) are kept in the real axis whe the pyrimidines (T and C) are kept in the imaginary axis.

2 Generate all possible hypothetical sequences ($\boldsymbol{r_i}$) that exhibit a clear period-3 pattern. Each sequence is generated as a periodic repetition of three different nucleobases out of the four possible genetic code alphabet letters $\left\{A, C, G, T\right\}$[76]. This approach will result in twenty-four possible sequences ($\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{24}$) where each sequence is of a predefined length of ($3L_r$) nucleobases, where L_r is the number of repetitions used for each sequence. For example, the sequence $\boldsymbol{r}_1 = \left\{ACGACGACG\right\}$ corresponds to ($L_r = 3$) repetitions of the genomic string AGC and hence of length ($3L_r = 9$) nucleobases. The reason that only twenty-four sequences are considered is due to the fact that a single sequence like $\boldsymbol{r}_1 = \left\{ACGACGACG\right\}$ can be looked at as a periodic repetition of either ACG or CGA or GAC. Hence, the 64 possible period-3 sequences will reduce to only 24 sequences. In other words, the genetic code includes all the possible $4^3 = 64$ distinct nucleotide triplets (codons). A periodic repetition of any of the 64 codons (say XYZ) in a sequence of a pre-specified length would result of a sequence of the form {XYZXYZXYZ$\cdots$XYZ}. This latter period-3 sequence can be considered a periodic repetition of either XYZ or YZX or ZXY codons as well. Applying this periodic structure on all the 64 possible codons of the genetic code would result in only 24 distinct period-3 sequences. Fig. 2 shows the 24 codons (marked in shade) that were used to construct the period-3 sequences in this work.

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 to each other. Hence, this step will help detect the portions of the input genomic sequence that have a period-3 pattern similar to anyone of the twenty-four sequences, $s_i[n]$. The twenty-four corresponding correlation outputs, denoted as $y_i[n]$, can be computed as the convolution of the input sequence, $x[n]$, with the time-reversed version of each of the twenty-four sequences, $s_i[n]$, as

$$y_i[n] = \sum_{m=0}^{3L_r-1} x[m]\, s_i[m-n], \tag{5}$$

where each signal $y_i[n]$ is of length $(L_x + 3L_r - 1)$.

Truncate the first $(3L_r - 1)/2$ and the last $(3L_r - 1)/2$ elements of each sequence, $y_i[n]$. This 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 $(3L_r - 1)/2$ and $(3L_r - 1)/2$ are equal if $(3L_r)$ is odd and different by one if $(3L_r)$ is even. Here, the symbol $[\cdot]$ is the largest integer not greater than the argument, while $[\cdot]$ 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 is used such

TTT	TTC	TTA	TTG	CTT	CTC	CTA	CTG	ATT	ATC	ATA	ATG	GTT	GTC	GTA	GTG
TCT	TCC	TCA	TCG	CCT	CCC	CCA	CCG	ACT	ACC	ACA	**ACG**	GCT	GCC	GCA	GCG
TAT	TAC	TAA	TAG	CAT	CAC	CAA	CAG	AAT	AAC	AAA	AAG	GAT	**GAC**	GAA	GAG
TGT	TGC	TGA	TGG	CGT	CGC	**CGA**	CGG	AGT	AGC	AGA	AGG	GGT	GGC	GGA	GGG

Fig. 2. The 24 codons used to construct the period-3 sequences.

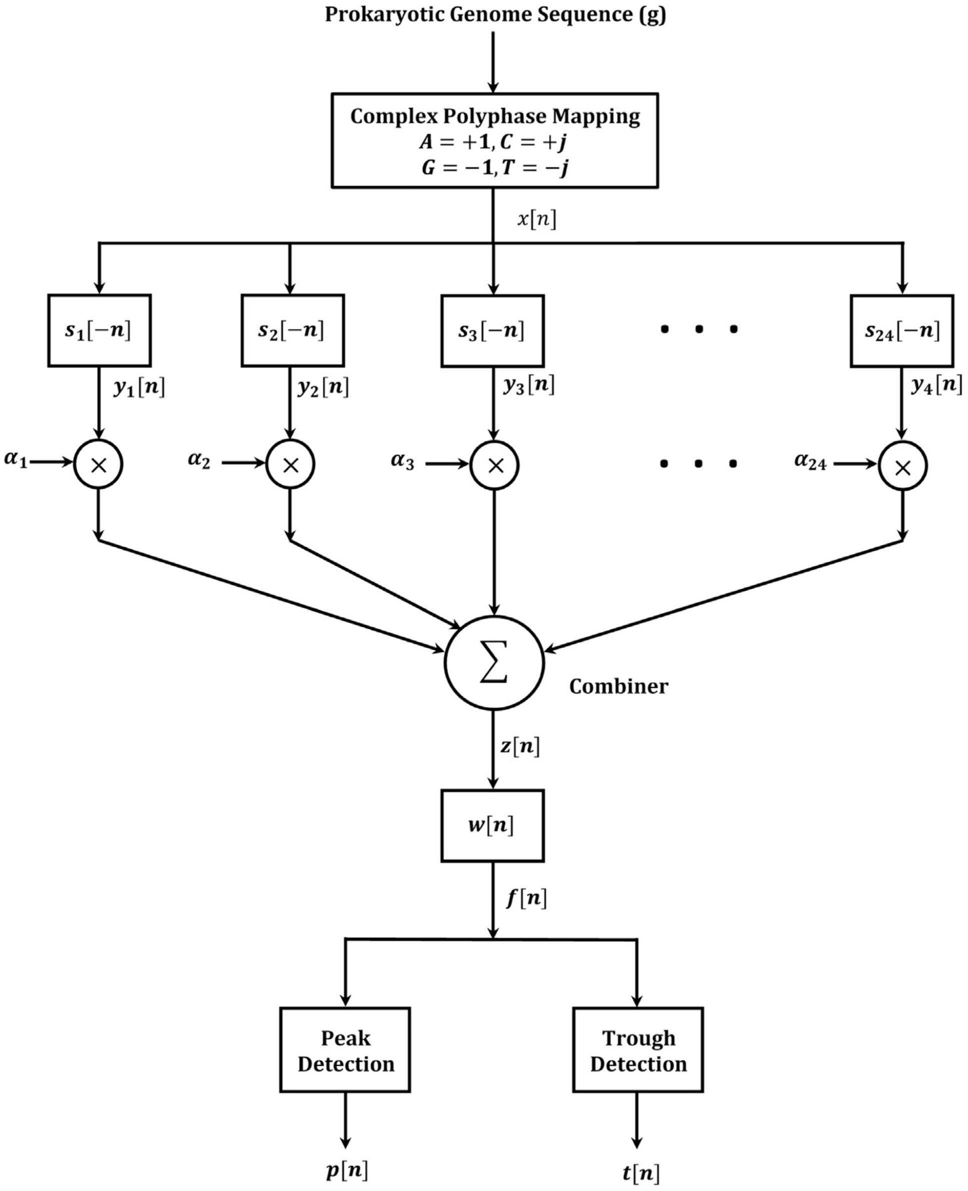


Fig. 3. The Proposed Gene Detection Algorithm.

that each of the twenty-four signal branches is multiplied by a weight factor, (α_i), that is proportional to the signal amplitude in each branch. That is, branches with strong signals (high amplitude) are further amplified, while weak signals (low amplitude) are further attenuated. The resulting signal at the output of the MRC module is given by

$$z[n] = \sum_{i=1}^{24} \alpha_i y_i [n], \tag{6}$$

where the factors, (α_i), are defined by

$$\alpha_i = \frac{|y_i[n]|}{\sum_{j=1}^{24} |y_j[n]|}. \tag{7}$$

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

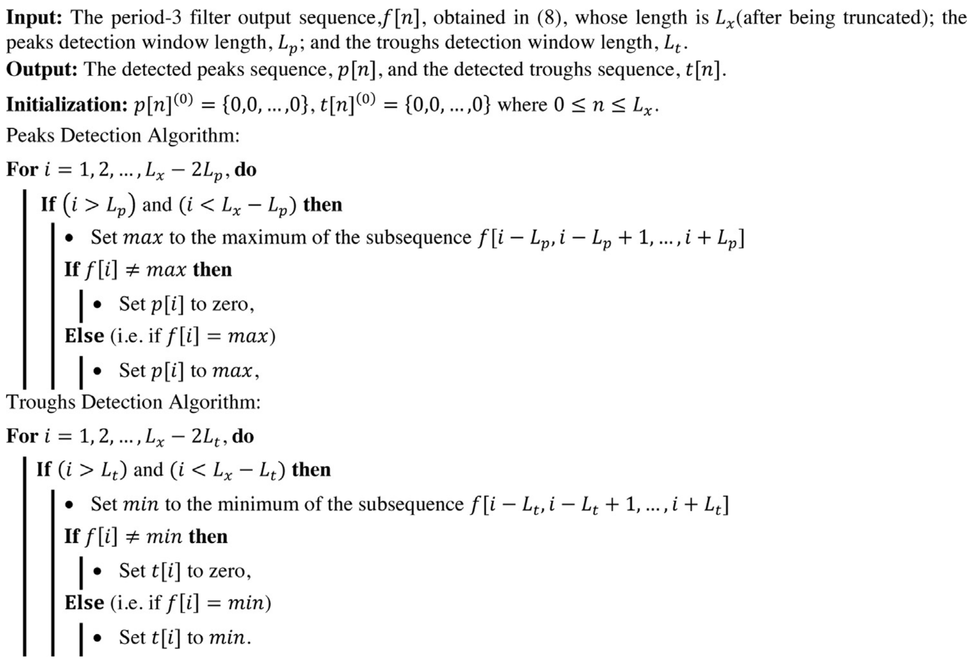
$$f[n] = \sum_{m=0}^{N-1} w[m]\, z[n-m], \tag{8}$$

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

8 Truncate the first $(N-1)/2$ and the last $(N-1)/2$ 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. The peaks and troughs detection algorithm is described in Algorithm I. The detection of peaks and troughs is achieved through sliding a window of length (L_p for peaks and L_t for troughs) 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 continues to generate the detected peak and trough sequences, $p[n]$ and $t[n]$, respectively.

Algorithm I. Peaks and Troughs Detection Algorithm



```
Input: The period-3 filter output sequence, $f[n]$, obtained in (8), 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: $p[n]^{(0)} = \{0,0,\ldots,0\}$, $t[n]^{(0)} = \{0,0,\ldots,0\}$ where $0 \le n \le L_x$.
Peaks Detection Algorithm:
For $i = 1, 2, \ldots, L_x - 2L_p$, do
    If $(i > L_p)$ and $(i < L_x - L_p)$ then
        • Set $max$ to the maximum of the subsequence $f[i - L_p, i - L_p + 1, \ldots, i + L_p]$
        If $f[i] \ne max$ then
            • Set $p[i]$ to zero,
        Else (i.e. if $f[i] = max$)
            • Set $p[i]$ to $max$,
Troughs Detection Algorithm:
For $i = 1, 2, \ldots, L_x - 2L_t$, do
    If $(i > L_t)$ and $(i < L_x - L_t)$ then
        • Set $min$ to the minimum of the subsequence $f[i - L_t, i - L_t + 1, \ldots, i + L_t]$
        If $f[i] \ne min$ then
            • Set $t[i]$ to zero,
        Else (i.e. if $f[i] = min$)
            • Set $t[i]$ to $min$.
```

3.2. Period-3 based Bayesian classification

In order to evaluate the performance of the proposed gene detection algorithm described in Section 3.1, a need arises to design a Bayesian classification system. This system can classify all possible open reading frames (ORF's) of a given genome sequence as coding or noncoding regions based on whether they possess a period-3 structure or not. The following three subsections detail the design process of the proposed period-3 based Bayesian classifier, definition of the classification variables, definition of the statistical model and finally incorporating prior knowledge of the classification classes considered.

3.2.1. Period-3 based Bayesian classifier

The elements of the Period-3 based Bayesian classifier are: the classification variable, the conditional probability of a particular value of the classification variable given a particular classification group, and the probability of the occurrence of each classification group. These elements can be denoted as: s_i, $P(s_i|w_j)$, and $P\left(w_j\right)$, respectively.

By applying the Bayesian classifier, the classification groups considered can be discriminated by the function

$$P\ w_j|s_i = \frac{P\ s_i|w_j\ P\ w_j}{P\ s_i}, \tag{9}$$

where i designates the classification variable considered (two variables are designed in this work), and j designates the classification groups investigated where $j = (Coding, Noncoding)$. The probability of a given value of the classification variable s_i, $P(s_i)$, can be calculated using the total probability theorem as

$$P(s_i) = \sum_{j=i}^{M} P\left(s_i|w_j\right) P\left(w_j\right)\cdot \tag{10}$$

The constant, M, represents the number of the classification groups considered; $M = 2$ for this work. The denominator of hand-side expression of (9) is the same for all classification groups. Therefore, and for the sake of classification purposes, we can simplify the discrimination function, (9), to

$$P\left(w_j|s_i\right) \approx \sim P\left(s_i|w_j\right) P\left(w_j\right)\cdot \tag{11}$$

3.2.2. Defining the classification variables

In the proposed Period-3 Based Bayesian Classification, two Bayesian classifiers are designed and investigated. Classifier 1 has the classification variable, s_1, as the maximum peak value that results from Algorithm I for each particular input ORF. Hence, s_1 is defined as

$$s_1 = \max\left[p\left[n\right]\right], \tag{12}$$

while Classifier 2 has the classification variable, s_2, as the maximum peak value normalized to the length of the ORF in nucleobases. Hence, s_2 is defined as

$$s_2 = \max\left[\frac{p\left[n\right]}{L_{ORF}}\right], \tag{13}$$

where L_{ORF} is the length of each particular ORF in nucleobases. This process for calculating the classification variables, s_1 and s_2, is used for every ORF in the training and test sets. We compiled our data set using GenBank sequences and annotations. All open reading frames on the forward and reverse strands, which are not valid genes (i.e. not listed at GenBank), were categorized as noncoding ORFs. On the other hand, all open reading frames on both forward and reverse strands that are listed at GenBank were categorized as coding ORFs. In order to form both of the testing and training sets, the entire data (coding and noncoding) was split in half. Algorithm II provides a detailed description of how both Bayesian classifiers (Classifier 1 and Classifier 2) work.

Algorithm II. Period-3 Based Bayesian Classification Algorithm

```
Input: The period-3 filter length, $N$; the peaks detection window length, $L_p$; the hypothetical period-3 sequence
repetition length, $L_r$; and an array of all possible open reading frames (ORFs) in the input genome sequence (in the
testing set), ORFGL, whose length is $L_{ORFGL}$.
Output: A decision that $w_i = w_1$ (i.e. the ORF is a gene) or $w_i = w_2$ (i.e. the ORF is NOT a gene)
For $i = 1, 2, \ldots, L_{ORFGL}$, do
    • Evaluate the detected peaks sequence, $p[n]$, for each input ORF by Algorithm I,
    • Evaluate the first classification variable, $s_1 = \max\left[p[n]\right]$
    • Evaluate the second classification variable, $s_2 = \max\left[p[n]\right]/L_{ORF}$, ($L_{ORF}$ is the ORF length)
    The First Bayesian Classifier (Classifier 1):
        If $P(w_1|s_1)P(w_1) > P(w_2|s_1)P(w_2)$, then
            • Select $w_i = w_1$, (i.e. make a decision that the input ORF is a gene)
        Else (i.e. if $P(w_1|s_1) < P(w_2|s_1)$)
            • Select $w_i = w_2$, (i.e. make a decision that the input ORF is NOT a gene)
    The Second Bayesian Classifier (Classifier 2):
        If $P(w_1|s_2)P(w_1) > P(w_2|s_2)P(w_2)$, then
            • Select $w_i = w_1$, (i.e. make a decision that the input ORF is a gene)
        Else (i.e. if $P(w_1|s_2)P(w_1) < P(w_2|s_2)P(w_2)$)
            • Select $w_i = w_2$, (i.e. make a decision that the input ORF is NOT a gene)
```

3.2.3. Defining the Statistical Model, $\boldsymbol{P(s_i|w_j)}$

In order to calculate $P\left(s_i|w_j\right)$ in (11), to be used in Algorithm II, the corresponding probability density functions (PDFs) for the classification variables, s_1 and s_2, of the training set are formed. $P(s_i|w_j)$ is the likelihood of a specific value of s_i given the classification group w_j. Fig. 4(a) represents $P\left(s_1|w_1\right)$, while Fig. 4(b) represents $P\left(s_1|w_2\right)$. Fig. 5(a) represents $P\left(s_2|w_1\right)$, while Fig. 5(b) represents $P(s_2|w_2)$. The latter four figures are obtained for the E. coli MG1655 bacterial genome. In Fig. 4, the horizontal axes are the classification variable values,s_1, defined in (12), and the vertical axes represent the probability of the s_1 value occurring for (a) the coding ORFs (valid genes) and (b) the noncoding ORFs (invalid genes) training set models. In Fig. 5, the horizontal axes represent the classification variable values, s_2, defined in (13), and the vertical axes are the probability of occurrence of the s_2 value for (a) the coding ORFs (valid genes) and (b) the noncoding ORFs (invalid genes) training sets. To verify that the probability density functions corresponding of the training sets considered for classification are not identical, a Wilcoxon Rank-Sum test was applied. The PDF model is the probability of occurrence of a specific value of the classification variable in a given classification group, $P(S_i = s_i|w_j)$.

3.2.4. Incorporating Prior Knowledge, $\boldsymbol{P\left(w_i\right)}$

In this work, two approaches for defining the prior probabilities,$P\left(w_1\right)$ and $P\left(w_2\right)$, used in Algorithm II are examined. One approach defines $P\left(w_i\right)$ by determining the ratio of coding and noncoding regions to the entire genomic sequence investigated in terms of sequence lengths in nucleobases, respectively. For example, based on the annotated genome sequence available at the National Center for Biotechnology Information (NCBI) [77], the prior probabilities $P\left(w_1\right)$ and $P\left(w_2\right)$ for MG1655 can be defined as 0.43 and 0.57, respectively. The second approach for defining $P\left(w_i\right)$ assumes no prior knowledge. Consequently, each classification group can be considered as equally probable. Hence, the prior probabilities,$P\left(w_1\right)$ and $P\left(w_2\right)$, are both 0.5.

4. Simulation results and analysis

In order to demonstrate the fidelity and biological significance of the proposed gene detection algorithm, it is applied to several prokaryotic genome sequences. 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 [77].

The length of each one of the twenty-four hypothetical period-3 based subsequences, $s_i\left[n\right]$, is selected to be 1950 (i.e. the number of repetitions, L_r, is 650). The length of the period-3 filter, $w\left[n\right]$, defined by (3) 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.

Figs. 6 and 7 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 (upper group of rectangles) correspond to the

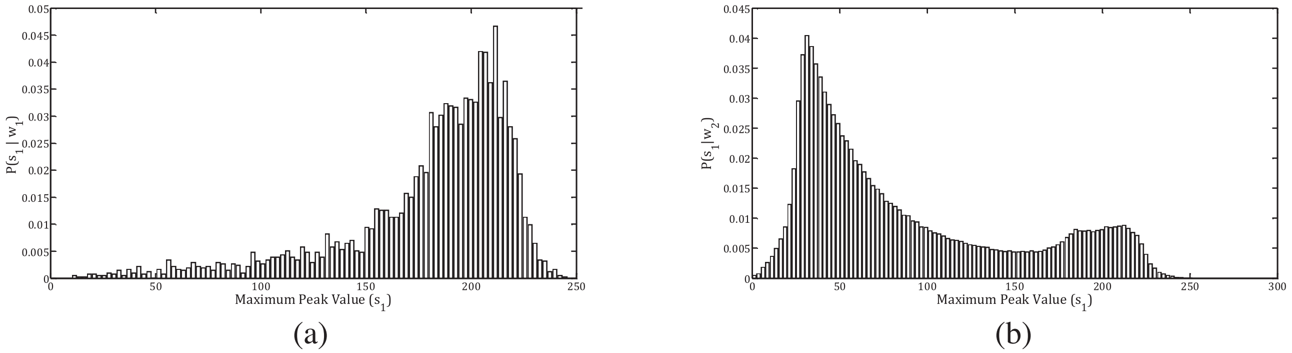


Fig. 4. The probability density functions of s_1 for (a) coding ORFs, and (b) noncoding ORFs.

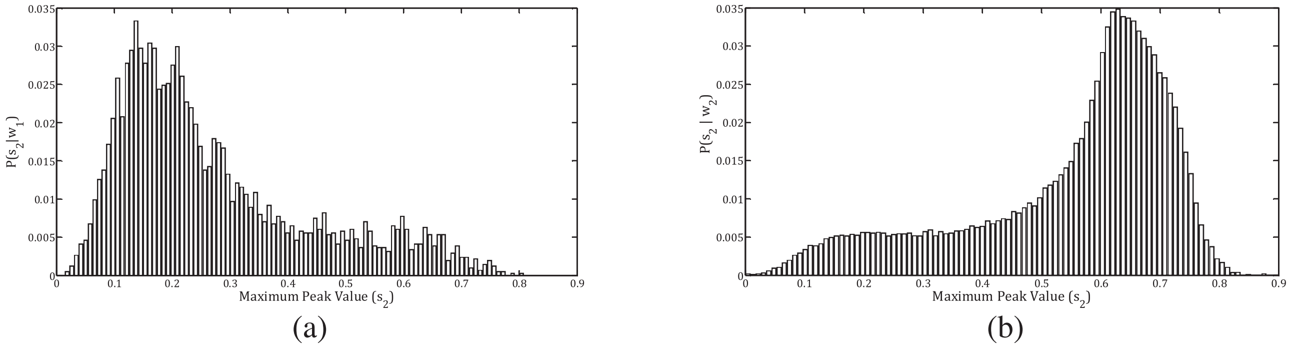


Fig. 5. The probability density functions of s_2 for (a) coding ORFs, and (b) noncoding ORFs.

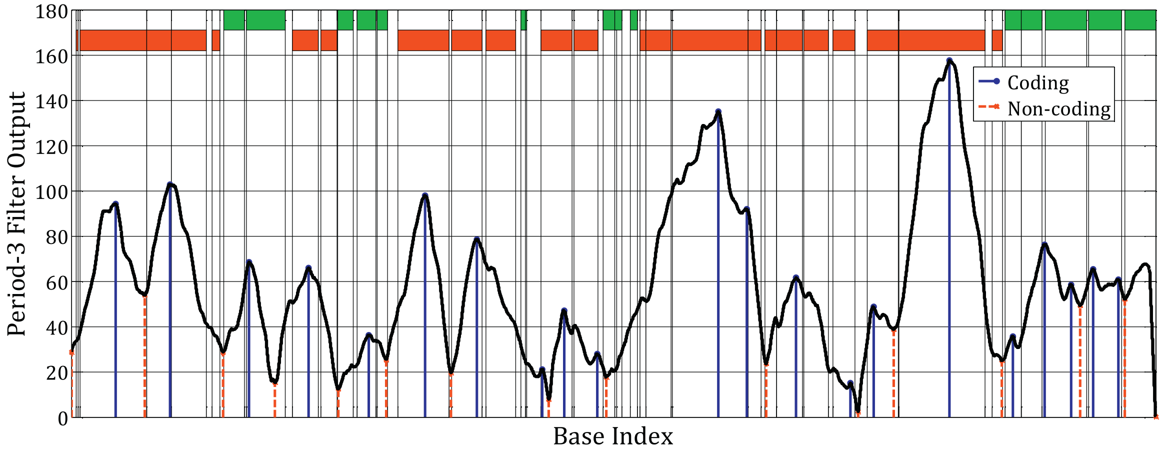


Fig. 6. Period-3 Filter Output with Peaks and Troughs identified (Applied to MG1655).

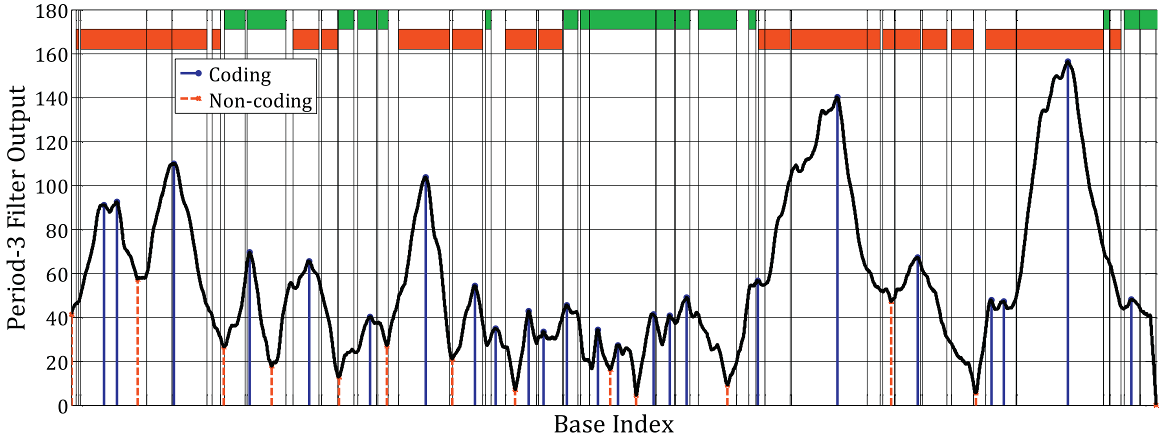


Fig. 7. Period-3 Filter Output with Peaks and Troughs identified (Applied to O157:H7).

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.

A closer look at Figs. 6 and 7 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 explained 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 is also confirmed by the fact that the forward and reverse strands are totally complementary and statistically symmetric [34]. 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. Apparently, the proposed algorithm will significantly increase the efficiency of coding regions 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 obvious that the previous simulation results are basically visual and do not always provide a sharp decision of where coding and noncoding regions exactly 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. This will allow for assessing the performance of the proposed gene detection algorithm and comparing it to well-known gene detection algorithms. To do this, a set of all possible ORFs in the genome sequence under study 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 the purpose of classification. The statistical models for the two Bayesian classifiers can be constructed by training the proposed classification algorithm using of 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 through the use of the True Positive Rate (TPR, also referred to as sensitivity), the False Positive Rate (FPR, also known as fall-out), the False Negative Rate (FNR) and the True Negative Rate (TNR, also referred to as specificity). The four performance rates are defined by Burset [78].

$$TPR = \frac{TP}{TP + FN} \times 100\%, \tag{14}$$

$$FPR = \frac{FP}{FP + TN} \times 100\%, \tag{15}$$

$$FNR = \frac{FN}{TP + FN} \times 100\%, \tag{16}$$

$$TNR = \frac{TN}{FP + TN} \times 100\%. \tag{17}$$

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 constitute 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

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}. \tag{18}$$

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 Ref. [78], is defined as

$$AC = (ACP - 0.5) \times 2, \tag{19}$$

where ACP is the Average Conditional Probability defined as:

$$ACP = \frac{1}{4}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right). \tag{20}$$

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, with regard to 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. In addition, 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 @ 3.40 GHz and 3.40 GHz, on a genomic test sequence of 40,386 nucleotides long, the proposed gene detection algorithm takes about 9.7 s computational time to obtain the final results.

Table 1
Performance evaluation of Classifier 1 and Classifier 2 compared to both GLIMMER and GeneMark with equal/unequal prior probabilities (explained in Section 3.3.4).

Classifier	(P_W, P_{W2})	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

5. Conclusions

This work 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 serval prokaryotes (e.g. MG1655 and O157H7 E. coli bacterial strains). Two Bayesian classifiers are designed for the purpose of the performance evaluation of the proposed gene detection algorithm. Specifically, the performance of the proposed algorithm is compared to the well-known ab initio gene detection methods: GLIMMER and GeneMark. The obtained simulation results show 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 Ref. [57]. Moreover, the proposed algorithm outperforms traditional methods as it puts off the need to apply variable-length discrete Fourier transform.

References

[1] G. Atkins, Information Theory and Molecular Biology, Cambridge University Press, New York, NY, 1993.

[2] R. Román-Roldán, P. Bernaola-Galván, J.L. Oliver, Application of information theory to DNA sequence analysis: a review, Pattern Recognit. 29 (1996) 1187–1194.

[3] I. Rojdestvenski, M.G. Cottam, Mapping of statistical physics to information theory with application to biological systems, J. Theor. Biol. 202 (2000) 43–54.

[4] C. Adami, Information theory in molecular biology, Phys. Life Rev. 1 (2004) 3–22.

[5] G. Battail, Information theory and error-correcting codes in genetics and biological evolution, in: Introd. to Biosemiotics New Biol. Synth, Springer, Netherlands, 2007, pp. 299–345.

[6] L. Gong, N. Bouaynaya, D. Schonfeld, Information-theoretic bounds of evolutionary processes modeled as a protein communication system, IEEE Work. Stat. Signal Process. Proc. (2007) 1–5.

[7] P. Hanus, B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, J.C. Mueller, Information and communication theory in molecular biology, Electr. Eng. 90 (2007) 161–173.

[8] M. Al Bataineh, L. Huang, G.E. Atkin, Transcription Factor Binding Site Detection Algorithm Using Distance Metrics Based on a Position Frequency Matrix Concept, IWBBIO 2014 (2nd Int. Work. Bioinforma. Biomed. Eng. (2014) 715–726.

[9] M. Al Bataineh, L. Huang, G. Atkin, TFBS detection algorithm using distance metrics based on center of mass and polyphase mapping, 2012, 7th Int. Symp. Heal. Informatics Bioinforma (2012) 37–40.

[10] M. Al Bataineh, Analysis of Genomic Translation Using a Communications Theory Approach, Illinois Institute of Technology, 2010.

[11] E.E. May, Communication theory and molecular biology at the crossroads, IEEE Eng. Med. Biol. Mag. 25 (2006) 28–29.

[12] M. Al Bataineh, M. Alonso, S. Wang, W. Zhang, G. Atkin, Ribosome binding model using a codebook and exponential metric, 2007, IEEE Int. Conf. Electro/Inform. Technol. (2007) 438–442.

[13] M. Al Bataineh, M. Alonso, S. Wang, G.E. Atkin, W. Zhang, An Optimized Ribosome Binding Model Using Communication Theory Concepts, Proc. 2007 Int. Conf. Bioinforma. Comput. Biol. (n.d.) 345–348.

[14] M. Al Bataineh, M. Alonso, L. Huang, G.E. Atkin, N. Menhart, Effect of mutations on the detection of translational signals based on a communications theory approach, Conf. Proc. IEEE Eng. Med. Biol. Soc. (2009) 3853–3856.

[15] M. Al Bataineh, L. Huang, M. Alonso, N. Menhart, G.E. Atkin, Analysis of gene translation using a communications theory approach, Adv. Exp. Med. Biol. (2010) 387–397.

[16] L. Huang, M. Al Bataineh, G.E. Atkin, M. Parra, M. del Mar Perez, I. Mohammed, W. Zhang, M. Parra, M. Perez, Identification of Transcription Factor Binding Sites Based on the Chi-Square $(X\hat{2})$ distance of a Probabilistic Vector Model, 2009 Int. Conf. Futur. Biomed. Inf. Eng. (FBIE 2009). (2009) 73–76.

[17] M. Al Bataineh, L. Huang, I. Muhamed, N. Menhart, G.E. Atkin, Gene expression analysis using communications, coding and information theory based models, BIOCOMP'09–2009, Int. Conf. Bioinf. Comput. Biol. (2009) 181–185.

[18] J. Weindl, J. Hagenauer, Applying techniques from frame synchronization for biological sequence analysis, IEEE Int. Conf. Commun. (2007) 833–838.

[19] E.E. May, Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms, North Carolina State University, 2002.

[20] E.E. May, Comparative Analysis of Information Based Models for Initiating Protein Translation in Escherichia coli K-12, NCSU, n.d.

[21] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick, An error-correcting code framework for genetic sequence analysis, J. Franklin Inst.-Eng. Appl. Math. 341 (2004) 89–109.

[22] Z. Dawy, P. Hanus, J. Weindl, J. Dingel, F. Morcos, On genomic coding theory, Eur. Trans. Telecommun. 18 (2007) 873–879.

[23] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick, Coding theory based models for protein translation initiation in prokaryotic organisms, Biosystems 76 (2004) 249–260.

[24] G.L. Rosen, Examining coding structure and redundancy in DNA, IEEE Eng. Med. Biol. Mag. 25 (2005) 62–68.

[25] D.A. Mac Donaill, Why nature chose A, C, G and U/T: an error-coding perspective of nucleotide alphabet composition, Orig. Life Evol. Biosph. 33 (2003) 433–455.

[26] M.K. GUPTA, The quest for error correction in biology, IEEE Eng. Med. Biol. Mag. 25 (2006) 46–53.

[27] L.S. Liebovitch, Y. Tao, A.T. Todorov, L. Levine, Is there an error correcting code in the base sequence in DNA? Biophys. J. 71 (1996) 1539–1544.

[28] G.L. Rosen, J.D. Moore, Investigation of coding structure in DNA, 2003 IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP '03). 2 (2003) 361–364.

[29] N. Stambuk, On circular coding properties of gene and protein sequences, Croat. Chem. Acta 72 (1999) 999–1008.

[30] D.A. MacDonaill, Digital parity and the composition of the nucleotide alphabet. Shaping the alphabet with error coding, IEEE Eng. Med. Biol. Mag. 25 (2006) 54–61.

[31] V. Veljkovic, I. Cosic, B. Dimitrijevic, D. Lalovic, Is it possible to analyze DNA and protein sequences by the methods of digital signal-processing, IEEE Trans. Biomed. Eng. 32 (1985) 337–341.

[32] A. Arneodo, Y. DAubentonCarafa, E. Bacry, P.V. Graves, J.F. Muzy, C. Thermes, Wavelet based fractal analysis of DNA sequences, Phys. D-Nonlinear Phenom. 96 (1996) 291–320.

[33] E.M. Crowley, A Bayesian method for finding regulatory segments in DNA, Biopolymers 58 (2001) 165–174.

[34] P.F. Baisnee, S. Hampson, P. Baldi, Why are complementary DNA strands symmetric? Bioinformatics 18 (2002) 1021–1033.

[35] L. Huang, M.A. Bataineh, G.E. Atkin, S. Wang, W. Zhang, A Novel gene detection method based on period-3 property, Conf. Proc. IEEE Eng. Med. Biol. Soc. 2009 (2009) 3857–3860.

[36] R. Kakumani, V. Devabhaktuni, M.O. Ahmad, Prediction of protein-coding regions in DNA sequences using a model-based approach, ISCAS 18 (2008) (2008) 1918–1921.

[37] E.C. Uberbacher, 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. 88 (1991) 11261–11265.

[38] J. Henderson, S. Salzberg, K.H. Fasman, Finding genes in DNA with a Hidden Markov Model, J. Comput. Biol. 4 (1997) 127–141.

[39] R. Raman, 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. 5 (1994) 275–283.

[40] A. Krogh, I.S. Mian, D. Haussler, A Hidden Markov Model that finds genes in Escherichia-Coli DNA, Nucleic Acids Res. 22 (1994) 4768–4778.
[41] W.N. Grundy, T.L. Bailey, C.P. Elkan, M.E. Baker, Meta-MEME: motif-based hidden Markov models of protein families, Comput. Appl. Biosci. 13 (1997) 397–406.
[42] M. Borodovsky, J. McIninch, GENEMARK: parallel gene recognition for both DNA strands, Comput. Chem. 17 (1993) 123–133.
[43] S.R. Eddy, Hidden Markov models, Curr. Opin. Struct. Biol. 6 (1996) 361–365.
[44] S.R. Eddy, Hidden Markov models and genome sequence analysis, FASEB J. 12 (1998) A1327.
[45] T. Yada, Y. Totoki, T. Takagi, K. Nakai, A novel bacterial gene-finding system with improved accuracy in locating start codons, DNA Res. 8 (2001) 97–106.
[46] J. Besemer, A. Lomsadze, 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. 29 (2001) 2607–2618.
[47] M. Walker, V. Pavlovic, S. Kasif, A comparative genomic method for computational identification of prokaryotic translation initiation sites, Nucleic Acids Res. 30 (2002) 3181–3191.
[48] S.S. Hannenhalli, W.S. Hayes, A.G. Hatzigeorgiou, J.W. Fickett, Bacterial start site prediction, Nucleic Acids Res. 27 (1999) 3577–3582.
[49] T. Nishi, T. Ikemura, S. Kanaya, GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences, Gene 346 (2005) 115–125.
[50] W.S. Hayes, M. Borodovsky, How to interpret an anonymous bacterial genome: machine learning approach to gene identification, Genome Res. 8 (1998) 1154–1171.
[51] Y. Osada, R. Saito, M. Tomita, Analysis of base-pairing potentials between 16 S rRNA and 5′ UTR for translation initiation in various prokaryotes, Bioinformatics 15 (1999) 578–581.
[52] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, K.-R. Müller, Engineering support vector machine kernels that recognize translation initiation sites, Bioinformatics 16 (2000) 799–807.
[53] T.D. Schneider, Measuring molecular information, J. Theor. Biol. 201 (1999) 87–92.
[54] J. Besemer, M. Borodovsky, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses, Nucleic Acids Res. 33 (2005) W451–W454.
[55] A.L. Delcher, K.A. Bratke, E.C. Powers, S.L. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23 (2007) 673–679.
[56] A.L. Delcher, D. Harmon, S. Kasif, O. White, S.L. Salzberg, Improved microbial gene identification with GLIMMER, Nucleic Acids Res. 27 (1999) 4636–4641, http://dx.doi.org/10.1093/nar/27.23.4636.
[57] P.P. Vaidyanathan, Genomics and proteomics: a signal processor's tour, Circuits Syst. Mag. IEEE. 4 (2004) 6–29.
[58] D. Anastassiou, Genomic signal processing, IEEE Signal Proc. Mag. (2001) 8–20.
[59] V.R. Chechetkin, A.Y. Turygin, Sizedependence of three-periodicity and long-range correlations in DNA sequences, Phys. Lett. A 199 (1995) 75–80.
[60] J.W. Fickett, The gene identification problem: an overview for developers, Comput. Chem. 20 (1996) 103–118.
[61] J.W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Res. 10 (1982) 5303–5318.
[62] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by fourier analysis of genomic sequences, CABIOS 13 (1997) 263–270.
[63] P.P. Vaidyanathan, B. Yoon, Digital filters for gene prediction applications, Proc. 36th Asilomar Conf. Signals, Syst. Comput. 1 (2002) 306–310.
[64] R. Guan, J. Tuqan, IIR filter design for gene identification, Gensips Proc. (2004).
[65] P. Vaidyanathan, B. Yoon, Gene and exon prediction using allpass-based filters, Work. Genomic Signal. 3 (2002).
[66] K.B. Murray, D. Gorse, J.M. Thornton, Wavelet transforms for the characterization and detection of repeating motifs, J. Mol. Biol. 316 (2002) 341–363.
[67] M. Borodovsky, S. Ekisheva, Problems and Solutions in Biological Sequence Analysis, Cambridge University Press, Cambridge; New York, 2006.
[68] E.N. Trifonov, J.L. Sussman, The pitch of chromatin DNA is reflected in its nucleotide sequence, Proc. Natl. Acad. Sci. U. S. A. 77 (1980) 3816–3820.
[69] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long-range correlations in nucleotide sequences, Nature 356 (1992) 168–170.
[70] Z.G. Yu, V.V. Anh, B. Wang, Correlation property of length sequences based on global structure of the complete genome, Phys. Rev. E—Stat. Nonlinear Soft Matter Phys. 63 (2001) 1–8.
[71] A. Papoulis, Systems and transforms with applications in optics, Ser. Syst. Sci. Malabar Krieger 1968 (1) (1968).
[72] A. Lohmann, Systems and transforms with applications in optics, J. Franklin Inst. 288 (1969) 328.
[73] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Pearson Education India, 1993.
[74] T. Saramäki, R. Bregovic, Multirate systems and filter banks, Multirate Syst. Des. Appl. 2 (2001) 27–85.
[75] S.D. Sharma, K. Shakya, S.N. Sharma, Evaluation of DNA mapping schemes for exon detection, 2011, Int. Conf. Comput. Commun. Electr. Technol. ICCCET 2011 (2011) 71–74.
[76] P. Rangel, J. Giovannetti, Genomes and Databases on the Internet: A Practical Guide to Functions and Applications, Horizon Scientific Press, 2002.
[77] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 35 (2007).
[78] M. Burset, R. Guigó, Evaluation of gene structure prediction programs, Genomics 34 (1996) 353–367.

Mohammad F. Al Bataineh was born in Irbid, Jordan in 1979. He received his B.S. degree in Telecommunications Engineering with high honors from Yarmouk University, Jordan, in 2003. He received his M.S. and PhD degrees in Electrical Engineering with excellent distinction from Illinois Institute of Technology (IIT) in 2006 and 2010, respectively. His research interests are focused in the application of communications, coding theory, and information theory concepts to the interpretation and understanding of information flow in biological systems such as gene expression. Since September 2010, Mohammad Al Bataineh has been with the Telecommunications Engineering Department at Yarmouk University, Jordan, where he is currently an assistant professor. He teaches undergraduate courses in Signals and Systems, Analog Communications, Digital Communications, Probability and Random Processes, Digital Signal Processing for the graduate level, and Information Theory and Coding for the graduate level.

Zouhair Al-qudah was born in Irbid, Jordan in 1979. He received his B.S. degree in Telecommunications Engineering from Yarmouk University, Jordan, in 2002. He received his M.S. degree in Electrical Engineering, with emphasis on digital communications and signal processing for wireless communication, from Kalmar University College, Sweden in 2006. He received his PhD degree in Electrical Engineering from Southern Methodist University at Dallas, Texas in 2013. Since August 2013, he has been with Al-Hussein Bin Talal University at Ma'an, Jordan, where he is currently an Assistant Professor. His research interest span various aspects of multipath fading channels, including Multiuser information theory, interference cancellation techniques, and practical coding techniques for Dirty Paper problem