Bayesian Classification of Ribosome Binding Sites in Prokaryotic Genome Sequences: A Communications Theory Approach

Article *in* International Journal of Bioscience Biochemistry and Bioinformatics · January 2017

DOI: 10.17706/ijbbb.2017.7.3.133-142

CITATION
1

READS
23

2 authors:

Mohammad Fayez Al Bataineh
Yarmouk University
49 PUBLICATIONS **121** CITATIONS

SEE PROFILE

Zouhair Al-qudah
Al-Hussein Bin Talal University
47 PUBLICATIONS **306** CITATIONS

SEE PROFILE

Bayesian Classification of Ribosome Binding Sites in Prokaryotic Genome Sequences: A Communications Theory Approach

Mohammad F. Al Bataineh[1*], Zouhair J. Al-Qudah[2]

[1] Telecommunications Engineering Department, Yarmouk University, Irbid, Jordan.
[2] Department of Communication Engineering, Al-Hussein bin Talal University, Ma'an, Jordan.

* Corresponding author. Tel.: +962795159463; email: mohamadfa@yu.edu.jo
Manuscript submitted January 15, 2017; accepted April 4, 2017.
doi: 10.17706/ijbbb.2017.7.3.133-142

Abstract: Dramatic advances in genomics and computational biology have resulted in large amounts of data and have encouraged the development of computational algorithms for the identification and analysis of coding regions. This paper proposes a novel application of fundamental principles and concepts from communications theory for the identification of exact translation initiation sites in prokaryotic genomes. It employs several Bayesian classifiers to assess the performance of the ribosome binding sites detection algorithms investigated in this work. The proposed classification algorithms utilize well-known principles in communications theory such as cross correlation and Euclidean distance based metrics to make precise real-time decisions of weather a given open reading frame (ORF) is a valid protein coding region or not. The simulation results confirm that the proposed Bayesian classification algorithms can provide a efficient and accurate gene identification with sensitivity and specificity values comparable to the ones obtained by the well-known prokaryotic gene detection methods such as GLIMMER and GeneMark. This further confirms the significance of applying communications theory concepts to genomic sequence analysis.

Key words: Gene detection, cross correlation, Euclidean distance, Bayesian classification.

1. Introduction

Researchers are increasingly interested in studying the relevance of using communications theory concepts and tools to model and understand the information flow in biological systems. Particularly, by analyzing the gene expression process, various analogies with the field of digital data transmission can be clearly noticed. Principles from communications, coding, information theory, detection theory, and pattern recognition can be utilized to reveal further similarities between the latter fields [1]–[6].

A DNA sequence can be divided into two types of regions: genes and intergenic spaces. Genes are the segments of DNA that contain the coding information required for the synthesis of protein. A considerable target of genomic research is to realize the nature and the role of the coding and non-coding information embedded in the DNA sequence structure. A crucial step to achieve this target is to identify gene locations in the genomic sequence under study. Several methods have been proposed in literature for gene detection in prokaryotes. For instance, probabilistic methods such as RBSFinder and GeneHacker Plus [7], and GeneMarkS [8], Statistical methods as in [9]. Some gene detection methods incorporate certain biological factors in an attempt to quantify valid translational start sites [10]. Such factors include the free energy that

results from the binding of the ribosome to its binding site (RBS), and the distance the separates the RBS from the initiation codon. Other computational gene detection methods employ machine learning, Bayesian methods [11], information theory, hidden Markov models such as GeneMark [12], and interpolated Markov models such as GLIMMER [13]. Despite their increased overall accuracy, GLIMMER and GeneMark usually necessitate longer test sequences for reliable gene detection.

The Bayesian classification algorithms proposed in this work for gene detection utilize two main properties of genomic sequences: i) the fact that the last thirteen-bases sequence of the $3' - \text{end}$ of the $16S\ rRNA$ molecule play an important role in the identification of coding and non-coding regions in the entire genomic structure, and ii) the so-called period-3 property [14] of protein coding regions. This work is proposing a novel application of principles and practices from communications theory and digital signal processing for the detection and identification of coding regions in prokaryotes. The proposed algorithms employ several numerical representations of the genomic sequences involved in the conducted analyses. Moreover, the proposed algorithms utilize basic concepts from communications theory, coding theory, and digital signal processing as cross-correlation, Euclidean distance, matched filter, and other distance metrics to design several classification algorithms that can efficiently identify coding and non-coding regions. The proposed classification algorithms are applied to the complete genomic sequences of different prokaryotes (e.g. MG1655 and O157H7 E. coli bacterial strains). 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 confirms the significance of applying communications theory concepts to genomic sequence analysis.

The rest of this paper is organized as follows. Section II highlights the biological significance of the last thirteen bases of the 16S ribosomal RNA (16S rRNA) molecule and its role in gene identification. Section III provides a clear and detailed description of the proposed Bayesian classification algorithms that are designed for gene identification. The design of the process of Bayesian classification along with the definition of the classification variables and the statistical model are provided in Section IV. Simulation results are shown and discussed in Section V. Finally, the paper is concluded in Section VI. Procedure for Paper Submission.

2. The Last Thirteen Bases of 16S ribosomal RNA

The ribosome recognition of the initiation codon is made possible when the $Shine\text{-}Dalgarno$ (SD) sequence is detected [15]. A conserved structure of the SD sequence is given by AGGAGG. This consensus structure is a part of the Watson-Crick complementary sequence of the last thirteen bases of the $3'\text{-}end$ of $16S\ rRNA$ given by $\{UAAGGAGGUGAUC\}$, which corresponds to $\{TAAGGAGGTGAUC\}$ in the original genomic sequence. The $16S$ r RNA molecule is the part of the small subunit (30S) of a prokaryotic ribosome that interacts with the SD sequence via base-pairing [14]. In the light of this biological background, the four Bayesian classifiers proposed in this work are basically based on the use of the latter thirteen-bases sequence for ribosome binding site detection. In other words, the proposed classification algorithms are trying to model the biological mechanism used by the ribosome to detect valid initiation sites.

3. The Proposed Bayesian Classification Algorithms

In this paper, four Bayesian classifiers are designed to evaluate the performance of the corresponding proposed ribosome binding sites detection algorithms. The four classifiers are described in details in the (A-D) subsections. The first four classifiers assume the ribosome is using the sequence of the last

thirteen-bases as a template to decide whether a given ORF is a valid protein coding sequence or not.

3.1. The Euclidean Distance Metric Based Classifier

In this classifier, the classification variable is based on a Euclidean distance metric that is used to identify the last thirteen-bases sequence in the genomic test sequence under study. This metric is calculated at every single nucleobase of the genomic test sequence as detailed by **Algorithm A**.

The Euclidean distance that separates two different points $\boldsymbol{A}$ and $\boldsymbol{B}$ in the Euclidean n-dimensional space is given by the length of the line segment ($\overline{\boldsymbol{AB}}$) that connects between them. If the points $\boldsymbol{A}$ and $\boldsymbol{B}$ are given by the n-tuples: $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$, respectively, then the distance (d) between them is defined as:

$$d(\boldsymbol{A}, \boldsymbol{B}) = d(\boldsymbol{B}, \boldsymbol{A}) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (1)$$

Algorithm A: The Euclidean Distance Metric Based Classification Algorithm

Input: $\boldsymbol{S}_{13}$: The last thirteen-bases sequence of the 16S rRNA molecule; $\boldsymbol{S}_t$: the genomic test sequence whose design is described in Section V; **ORFGL**: and an array of all possible ORFs in the input genome sequence (in the testing set), whose length is L.
Output: The classification variable $\boldsymbol{S}_1 = [s_{11}, s_{12}, \dots, s_{1L}]$ whose elements are s_{1j} where $j = 1, 2, 3, \dots . L$. The subscript 1 corresponds to the first classification variable $\boldsymbol{S}_1$.
For $i = 1, 2, \dots, L$, **do**

- Map $\boldsymbol{S}_t$ and $\boldsymbol{S}_{13}$ sequences using the numerical quaternary representation($A = 0, C = 1, G = 2, and\ T = 3$).
- Slide $\boldsymbol{S}_{13}$ along $\boldsymbol{S}_t$ and calculate the Euclidean distance at each alignment using (1) by setting $\boldsymbol{A} = \boldsymbol{S}_t$ and $\boldsymbol{B} = \boldsymbol{S}_{13}$, and record the obtained distance as a function of base position.
- Detect the minimal point of the resulting vector in step 3, and save it as s_{1i} as the i^{th} value of the first classification variable S_1.

A minimal point in the third step of **Algorithm A** represents a perfect match of the last thirteen-bases sequence. However, partial matches of the latter sequence results in several local minima whose amplitudes increase as the number of mismatches increase. In other words, this algorithm can provide an accurate identification of the last thirteen-bases sequence, $\boldsymbol{S}_{13}$, in the genomic test sequence, $\boldsymbol{S}_t$, and can account for mismatches as well.

3.2. The Cross-Correlation Metric Based Classifier

In telecommunications, the presence of a given signal or a template can be detected in another signal using cross correlation. This can be achieved by convolving the first signal with a time-reversed version of the second signal or vice versa. In other words, to detect the last thirteen-bases sequence $\boldsymbol{S}_{13}$ in the genomic sequence under study, the input test genomic sequence $\boldsymbol{S}_t[n]$ can be looked at as the input to a matched filter whose impulse response is given by $h[n] = \boldsymbol{S}_{13}[-n]$. The output $z[n]$ is evaluated by

$$z[n] = \sum_{k=-\infty}^{\infty} S_t[k] S_{13}[n+k] \,.\, x \quad (2)$$

Algorithm B provides a description of the cross-correlation metric based classification algorithm.

Algorithm B: The Cross-Correlation Metric Based Classification Algorithm

Input: $\boldsymbol{S}_{13}$: The last thirteen-bases sequence of the 16S ribosomal RNA molecule; $\boldsymbol{S}_t$: the genomic test sequence whose design is described in Section V with a length of L_{S_t} bases; **ORFGL**: and an array of all

possible open reading frames (ORFs) in the input genome sequence (in the testing set), whose length is L.

Output: The classification variable $\boldsymbol{S}_2 = [s_{21}, s_{22}, \dots, s_{2L}]$ whose elements are s_{2j} where $j = 1, 2, 3, \dots, L$. The subscript 2 corresponds to the second classification variable $\boldsymbol{S}_2$.

For $i = 1, 2. \dots . L$, **do**

- Map $\boldsymbol{S}_t$ and $\boldsymbol{S}_{13}$ sequences to their corresponding binary representation using ($A = 00, C = 01, G = 10, and\ T = 11$).
- Replace all zeros in the previous step by (-1) for better correlation results.
- Correlate $\boldsymbol{S}_t$ and $\boldsymbol{S}_{13}$ using (2).
- Detect the maximal point of the resulting vector in step 3, and save it as s_{2i} as the i^{th} value of the first classification variable S_2.

3.3. The Exponential Detection Metric Based Classifier

In this classifier, the last thirteen-bases sequence of the 16S ribosomal RNA molecule is slid over each test sequence $\boldsymbol{S}_t$ one nucleobase at a time. An exponential metric associated with the total number of matches at each alignment is calculated as described in **Algorithm C**. The input test sequence $\boldsymbol{S}_t$ and the last thirteen-bases sequence $\boldsymbol{S}_{13}$ are mapped to their equivalent binary representation using (A=00, C=01, G=10, and T=11).

Algorithm C: The Exponential Detection Metric Based Classifier

Input: $\boldsymbol{S}_{13}$: The last thirteen-bases sequence of the 16S ribosomal RNA molecule; $\boldsymbol{S}_t$: the genomic test sequence whose design is described in Section V with a length of L_{S_t} bases; **ORFGL**: an array of all possible open reading frames (ORFs) in the input genome sequence (in the testing set), ORFGL, whose length is L.

Output: The classification variable $\boldsymbol{S}_3 = [s_{31}, s_{32}, \dots, s_{3L}]$ whose elements are s_{4j} where $j = 1, 2, 3, \dots . L$. The subscript 3 corresponds to the fourth classification variable $\boldsymbol{S}_3$.

Initialize $a = 2$ **(**The parameter a is used to control the exponential growth of the weighting function W)

For $i = 1. 2. \dots . L$, **do**

- Generate the test sequence S_t as described in Section V.

For $j = 1. 2. \dots . L_{S_t} - 13 + 1$; (13 is the length of S_{13} sequence in bases).

Extract a window (w) of length equal to thirteen bases starting at the j^{th} position and ending at the $(j + 12)^{th}$ position.

Initialize $W = 0,\ M = 0$.

For $k = 1, 2, \dots, 13.$ **do**

If the k^{th} base in the extracted window is equal to the k^{th} bases in S_{13}, **then**

- **Increment** the number of matches $M = M + 1$**.**
- **Set** $w(k) = a^M$.

Else

- **Set** $w(k) = 0$.

Set $W = W + w(k)$.

Set $weight(j) = W$.

- **Select** the classification variable s_{3i} as the maximal point of the "weight" vector whose length is $L_{S_t} - 13 + 1$ bases.

3.4. The Free Energy Metric Based Classifier

A free energy Table in [16] shows the free binding energy (in kcal/mol) that quantifies the strength of the

bond that happens between the ribosome and its binding site. Specifically, the free binding energy is given per each dinucleotide in the 16S ribosomal RNA with its complement in the genomic sequence. This free energy is utilized in this classifier to calculate a corresponding distance metric that is used as a classification variable. The distance metric is calculated at every alignment between the genomic test sequence $\boldsymbol{S}_t$ and the last thirteen-bases of the 16S ribosomal RNA molecule $\boldsymbol{S}_{13}$. For instance, a bond between the dinucleotide GU in the mRNA and CA in the $16S$ ribosomal RNA results in a free energy of $-2.1\ kcal/mol$.

Algorithm D provides a detailed description of the underlying classification algorithm.

Algorithm D: The Free Energy Metric Based Classifier

Input: $\boldsymbol{S}_{13}$: The last thirteen-bases sequence of the 16S ribosomal RNA molecule; $\boldsymbol{S}_t$: the genomic test sequence whose design is described in Section V with a length of L_{S_t} bases; **ORFGL**: an array of all possible ORF in the input genome sequence (in the testing set), ORFGL, whose length is L.
Output: The classification variable $\boldsymbol{S}_4 = [s_{41}, s_{42}, \dots, s_{4L}]$ whose elements are s_{4j} where $j = 1, 2, 3, \dots, L$. The subscript 4 corresponds to the fourth classification variable $\boldsymbol{S}_4$.
For $i = 1, 2, \dots, L$, **do**

- **Generate** the test sequence S_t as described in Section V.

For $j = 1, 2, \dots, L_{S_t} - 13 + 1$; (13 is the length of S_{13} sequence in bases).

Extract a window (w) of length equal to thirteen bases that starting at the j^{th} position and ending at the $(j+12)^{th}$ position.
Initialize the free energy $E = 0$.
For $k = 1, 2, \dots, 12$, **do**

- **If** the k^{th} base doublet in the extracted window $\{w_k w_{k+1}\}$ is equal to the k^{th} base doublet in S_{13} $\{S_{13_k} S_{13_{k+1}}\}$, **then**
  - **Set** $FED = \{w_k w_{k+1}\}$**,** where FED: Free Energy Doublet.
  - **Set** $E = E + energy(FED)$. where the function $energy$ returns the free energy associated with the base doublet FED as given in Table 1.

Else

- **Keep** E unchanged.

- **Set** $FreeEnergy(j) = E$.

- **Select** the classification variable $s_{4i} = \min(FreeEnergy)$ at the i^{th} alignment.

Table 1. Energy Doublets

Pairs of bases energy			
AA -0.9	GA	AG	GG
AU -0.9	GU	AC	GC
UA -1.1	CA	UG	CG
UU -0.9	CU	UC	CC

4. Bayesian Classification

The performance of the proposed ribosome binding site detection algorithms is evaluated using Bayesian classification described by **Algorithm E**. A clear description of the classification variables S_1, S_2, S_3, and S_4 is provided in **Algorithms A**, **B**, **C**, and **D**, respectively.

Algorithm E: Bayesian Classification Algorithm

Input: The classification variable vector S_i whose values are s_{ij}; and an array of all possible ORF in the input genome sequence (in the testing set), ORFGL, whose length is L.

Output: A decision that $W_k = W_1$ (i.e. the ORF is a gene) or $W_k = W_2$ (i.e. the ORF is NOT a gene)

For $i = 1, 2, \ldots, L_{ORFGL}$, **do**

- Obtain the classification vector values from either one of the proposed classification criteria,

  If $P(W_1|s_{ij})P(W_1) > P(W_2|s_{ij})P(W_2)$, **then**

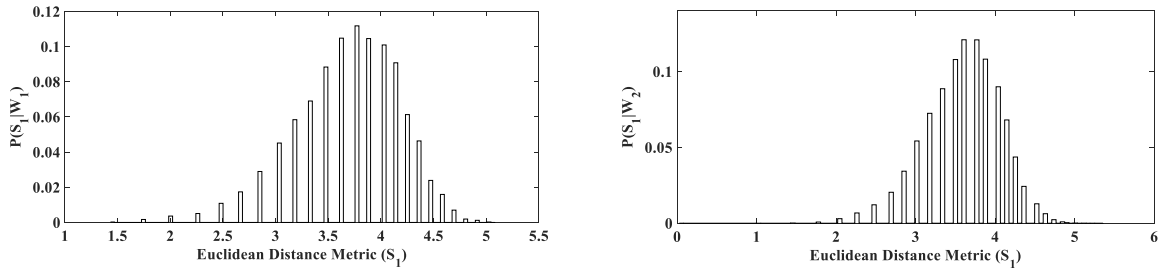
  - Select $W_k = W_1$, (i.e. classify the input ORF as a gene)

  Else (i.e. if $P(W_1|s_{ij}) < P(W_2|s_{ij})$)

  - Select $W_k = W_2$, (i.e. classify the input ORF as not a gene)

4.1. Defining the Statistical Model, $P(S_i|W_j)$

To determine $P(s_{ij}|W_k)$ in (5) to be used in **Algorithm E**, the corresponding probability density functions (PDFs) for the classification variables s_{ij} of the training set are formed. $P(S_i|W_k)$ is the conditional probability of S_i given the class W_k. Fig. 1.a corresponds to $P(S_1|W_1)$, while Fig. 1.b corresponds to $P(S_1|W_2)$ where S_1 is the Euclidean Distance Metric Classifier described by **Algorithm A** before. The two figures are obtained for E. coli MG1655 bacterial genome. The horizontal axes represent the classification variable values of S_1, and the vertical axes represent their corresponding probabilities in the coding ORFs (valid genes) and in non-coding ORFs (invalid genes) training set models, respectively. The probability density function model is basically the probability of a given values of the classification variable that occurs in each classification class, $P(S_1 = s_{ij}|W_j)$.



(a) (b)

Fig. 1. The probability density functions of S_1 for (a) coding ORFs, (b) for noncoding ORFs.

4.2. Incorporating Prior Knowledge, $P(W_i)$

In this work, the prior probabilities $P(W_1)$ and $P(W_2)$ used in **Algorithm E** are assumed equal (i.e. no prior information).

5. Simulation Results and Analysis

In order to demonstrate the fidelity and biological significance of the proposed ribosome binding site classification algorithms, the proposed algorithms are applied to the complete genome sequence of Escherichia coli bacterial strains MG1655 available at the NCBI [17].

The test sequence used for analysis in this paper has the following structure: (i) the 29 bases of the noncoding region that precede the initiation codon of a possible open reading frame (ORF), (ii) the initiation codon (ATG or GTG or TTG)), and (iii) twenty-eight bases of the coding region that immediately follow the initiation codon. Accordingly, the test sequence structure is given by:

$$[b_{-29} \ldots b_{-1}\ A\ U\ G\ b_{+3} \ldots b_{+29}\ b_{+30}].$$

The ribosome spans approximately 30 nucleobases of the mRNA sequence at a time [18]. Consequently, a 60-base test sequence is sufficient to represent the region of the mRNA sequence that binds to the 16S ribosomal RNA during the initiation process.

To evaluate the performance of the proposed ribosome binding site detection algorithms, 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 split in half to get a training set and a testing set for classification. The statistical models for the four Bayesian classifiers can be constructed by training the proposed classification algorithms using of the training datasets. Subsequently, the classification algorithms are tested using the testing datasets to verify its performance in detection. Table 2 shows the obtained results of the four Bayesian classifiers when applied to the E. coli MG1655 bacterial strain being compared to the GLIMMER, GeneMark gene finding software, and to the period-3 gene detection algorithm proposed in [5]. The performance of the four 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 False Negative Rate (FNR) and the True Negative Rate (TNR, also referred to as specificity). The four performance rates are defined by

$$TPR = \frac{TP}{TP+FN} \times 100\%, \tag{3}$$

$$FPR = \frac{FP}{FP+TN} \times 100\%, \tag{4}$$

$$FNR = \frac{FN}{TP+FN} \times 100\%, \tag{5}$$

$$TNR = \frac{TN}{FP+TN} \times 100\%. \tag{6}$$

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

The Correlation Coefficient (CC) is considered a preferred measure of global accuracy and is defined as

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}. \tag{7}$$

Another performance measure is the Approximate Correlation Coefficient (AC) given by

$$AC = (ACP - 0.5) \times 2, \tag{8}$$

where ACP is the Average Conditional Probability defined as:

$$ACP = \frac{1}{4}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right). \tag{9}$$

The four Bayesian classifiers are also applied to all possible ORFs. The corresponding simulation results are shown in Table 2 to compare with the performance obtained by the 60-bases sequence described before.

Table 2 shows the simulation result obtained by applying the proposed ribosome binding site detection algorithms to test sequences extracted from the complete genome sequences of the Escherichia coli MG1655 bacterial strains. The obtained results of simulation verify that the algorithms proposed here are

successful in detecting ribosome binding sites in the prokaryotic genomic sequence under study with sensitivity and specificity values comparable to well-known gene detection methods in prokaryotes such as GLIMMER and GeneMark.

Table 2. Performance Evaluation of Classifiers A, B, C, and D

Classifier	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
MPMBC (Maximum Peak Metric Based Classifier); $S_t = ORF$ [5]	3858 (93.17%)	110387 (25.35%)	283 (6.83%)	325146 (74.65%)	0.1494	0.3556
NMPMBC (Normalized Maximum Peak Metric Based Classifier); $S_t = ORF$ [5]	3384 (81.72%)	101803 (23.37%)	757 (18.28%)	333730 (76.63%)	0.1321	0.3067
(A) EDMBC (Euclidean Distance Metric Based Classifier); $S_t = \{b_{-29} \dots b_{+30}\}$	1472 (35.55%)	110161 (25.29%)	2669 (64.45%)	325372 (74.71%)	0.0228	0.0538
(B) CCMBC (Correlation Coefficient Metric Based Classifier); $S_t = \{b_{-29} \dots b_{+30}\}$	3357 (81.07%)	321961 (73.92%)	784 (18.93%)	113572 (26.08%)	0.0157	0.0375
(C) ExDMBC (The Exponential Detection Metric Based Classifier); $S_t = \{b_{-29} \dots b_{+30}\}$	2801 (67.64%)	226117 (51.92%)	1340 (32.36%)	209416 (48.08%)	0.0304	0.0816
(D) FEMBC (Free Energy Metric Based Classifier) $S_t = \{b_{-29} \dots b_{+30}\}$	2769 (66.87%)	126212 (28.98%)	1372 (33.13%)	309321 (71.02%)	0.0804	0.1980
(A) EDMBC (Euclidean Distance Metric Based Classifier); $S_t = ORF$	2444 (59.02%)	224764 (51.61%)	1697 (40.98%)	210769 (48.39%)	0.0143	0.0384
(B) CCMBC (Correlation Coefficient Metric Based Classifier); $S_t = ORF$	3872 (93.50%)	405295 (93.06%)	269 (6.50%)	30238 (6.94%)	0.0017	0.0026
(C) ExDMBC (The Exponential Detection Metric Based Classifier); $S_t = ORF$	2180 (52.64%)	184468 (42.35%)	1961 (47.36%)	251065 (57.65%)	0.0201	0.0534
(D) FEMBC (Free Energy Metric Based Classifier); $S_t = ORF$	2760 (66.65%)	243959 (56.01%)	1381 (33.35%)	191574 (43.99%)	0.0207	0.0552

Out of the four Bayesian classifiers, Classifier B with equal prior probabilities seems to perform the best in terms of sensitivity (or TPR) and FNR. Hence, Classifier E outperforms both GLIMMER and GeneMark in both TPR and FNR. However, GLIMMER and GeneMark provide better performance in terms of FPR and TNR. Classifier D outperforms the other three classifiers in terms of specificity.

6. Conclusions

This work is proposing a novel application of principles and concepts from communications theory and digital signal processing for the gene detection in prokaryotic genomes. The proposed gene detection algorithm employs several mapping schemes 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, matched filter, Euclidean distance, and other distance metrics based on the use of the last thirteen-bases sequence of the 16S ribosomal RNA molecule to identify coding and noncoding regions of the whole genomic sequence under study. The proposed gene detection algorithms are applied to the complete genome sequences of serval prokaryotes (e.g. MG1655 E. coli bacterial strain).

Four Bayesian classifiers are designed for the performance evaluation of the proposed ribosome binding site detection algorithms compared to well-known gene detection methods such as 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 comparable to GLIMMER and GeneMark gene detection methods.

References

[1] May, E. E., Vouk, M. A., Bitzer, D. L., & Rosnick, D. I. (2004). Coding theory based models for protein translation initiation in prokaryotic organisms. *Biosystems*, 76(1–3), 249–260.

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

[3] Dawy, Z., Gonzalez, F., Hagenauer, J., & Mueller, J. C. (2005). Modeling and analysis of gene expression mechanisms: a communication theory approach. *IEEE Int. Conf. Commun.*, *2*, 815–819.

[4] Bataineh, M. A., Huang, L., Alonso, M., Menhart, N., & Atkin, G. E. (2010). Analysis of gene translation using a communications theory approach. *Advances in Experimental Medicine and Biology*, *680*, 387–397.

[5] Bataineh, M. A., & Al-Qudah Z. (2017). A novel gene identification algorithm with Bayesian classification. *Biomed. Signal Process. Control*, *31*, 6–15.

[6] Bataineh, M. A. (2010). Analysis of genomic translation using a communications theory approach. Illinois Institute of Technology, Chicago.

[7] Yada, T., Totoki, Y., Takagi, T., & Nakai, K. (2001). A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res.*, *8(3)*, 97–106.

[8] Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). 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(12)*, 2607–2618.

[9] Walker, M., Pavlovic, V., & Kasif, S. (2002). A comparative genomic method for computational identification of prokaryotic translation initiation sites. *Nucleic Acids Res.*, *30(14)*, 3181–3191.

[10] Hannenhalli, S. S., Hayes, W. S., Hatzigeorgiou, A. G., & Fickett, J. W. (1999). Bacterial start site prediction. *Nucleic Acids Res.*, *27(17)*, 3577–3582.

[11] Crowley, E. M. (2001). A Bayesian method for finding regulatory segments in DNA. *Biopolymers*, *58(2)*, 165–174.

[12] Besemer J., & Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.*, *33(2)*, W451–W454.

[13] Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, *23(6)*, 673–679.

[14] Steitz J. A., & Jakes, K. (1975). How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proceedings of Natl Acad Sci U S A*, *72(12)*, 4734–4738.

[15] Shine J., & Dalgarno, L. (1974). The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proceedings of Natl Acad Sci U S A*, *71(4),* 1342–1346.

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

[17] Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, *35(1)*.

[18] Lewin, B. (1995). *Genes V*. Oxford Univ. Press. New York, NY.

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.

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.