EURASIP Journal on Applied Signal Processing 2004:1, 81–91

Segmentation of DNA into Coding and Noncoding Regions Based on Recursive Entropic Segmentation and Stop-Codon Statistics

Daniel Nicorici

Tampere International Center for Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere FIN-33101, Finland
Email: daniel.nicorici@tut.fi

Jaakko Astola

Tampere International Center for Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere FIN-33101, Finland
Email: jaakko.astola@tut.fi

Received 28 February 2003; Revised 15 September 2003

Heterogeneous DNA sequences can be partitioned into homogeneous domains that are comprised of the four nucleotides A, C, G, and T and the stop codons. Recursively, we apply a new entropic segmentation method on DNA sequences using Jensen-Shannon and Jensen-Rényi divergences in order to find the borders between coding and noncoding DNA regions. We have chosen 12- and 18-symbol alphabets that capture (i) the differential nucleotide composition in codons and (ii) the differential stop-codon composition along all the three phases in both strands of the DNA. The new segmentation method is based on the Jensen-Rényi divergence measure, nucleotide statistics, and stop-codon statistics in both DNA strands. The recursive segmentation process requires no prior training on known datasets. Consequently, for three entire genomes of bacteria, we find that the use of nucleotide composition, stop-codon composition, and Jensen-Rényi divergence improve the accuracy of finding the borders between coding and noncoding regions in DNA sequences.

Keywords and phrases: recursive segmentation, DNA sequence, information divergence measures, statistics of stop codons, Bayesian information criterion.

1. INTRODUCTION

The computational identification of genes and coding regions in DNA sequences is a major goal and a long-lasting topic for molecular biology, especially for the human genome project [1, 2]. One of the main goals of the human genome project is to provide a complete list of annotated genes that will be used in the biomedical research. Also, methods for reliable identification of genes in anonymous sequences of DNA can speed the process. A number of such methods exist but their predictive performance for finding genes is still not satisfactory [3]. There are two basic problems in gene finding: detection of protein-binding sites of the genes and detection of regions that code for proteins. These problems still are not satisfactorily solved, and the reliable detection of genes and coding regions in DNA sequences is critical for the success of the computational gene discovery from annotated genome sequences [4]. We address in this study the problem of finding the coding regions in DNA sequences that code for proteins.

Almost everything in the organism of living beings is made of proteins. According to the central dogma that forms the backbone of molecular biology, the DNA codes for the production of messenger RNA (mRNA) during the transcription process. The ribosomes "read" this information and use it for protein synthesis during the translation process.

The main genetic material in the prokaryote and the eukaryote cells is represented by the nucleic DNA molecules that have a well-studied structure. There are four kinds of nucleotides that differ by their nitrogenous bases: adenine (A), cytosine (C), thymine (T), and guanine (G). Along two strands of DNA double helix, a pyrimidine in one chain always faces a purine in the other and only the complementary base pairs T-A and G-C exist. A pyrimidine contains bases T and G, and purine contains bases A and C. Also, there is a large redundancy of the protein-coding regions in DNA that is distributed unevenly. There are $4^3 = 64$ codons to specify only 21 outputs, where 20 are amino acids and one output (stop codon) signals the end of the translation process.

One generic feature of DNA sequences is that their statistical properties are not homogeneously distributed along the sequence [5]. There is evidence of long-range correlations in genomic DNA, and it has been attributed to the presence of complex heterogeneities in the DNA sequences [6, 7, 8]. However, the current biological knowledge about coding regions in DNA is still limited to the structure of the codon and functional sites of the genes. The fact that the composition of the nucleotides for positions inside the codon (periodicity of three nucleotides) is different for the coding regions than the noncoding ones provides a strong signal for detection [9, 10].

Many algorithms have been developed for gene recognition based on three-base periodicity [11, 12, 13, 14], codon-usage measure [2], dicodon-usage measure [15], and position-weight matrix [16]. Fickett [17, 18] presents several algorithms for recognizing complete genes and one algorithm for recognizing coding regions. The accuracy of these algorithms for the complete gene recognition is generally high when they are tested on Guigo's dataset [3], but is not so good for the recently completed genomes of different organisms.

Segmentation methods are computational methods used to identify the homogeneous regions based on entropy measures. They are important for DNA-sequence analysis when identifying the borders between coding and noncoding regions [5, 7, 19, 20]. Also, recursive segmentation of DNA sequences has been used for detecting the existence of the isochores, and CpG islands, detecting replication origin and terminus, and complex patterns such as telomeres, and evaluating the genomic complexity [5, 6]. The Jensen-Shannon divergence is one of the most widely used methods for segmenting DNA sequences [5, 6, 7, 19, 20, 21], and is used for recursively separating DNA sequences in homogenous regions with respect to its neighbors. The criterion for continuing the recursive segmentation process can be based on (i) statistical significance [19, 20, 22], or (ii) Bayesian information criterion (BIC) [5, 6, 7, 21].

In this study, we analyze the recursive entropic segmentation for DNA sequences from different bacteria, but this can be easily extended to other DNA sequences of other organisms. All the bacteria's genomes referred to in this study are available on the site of European Bioinformatics Institute (http://www.ebi.ac.uk/genomes/). In [19], Bernaola-Galvan et al. use a 12-symbol alphabet and Jensen-Shannon divergence for finding the borders between coding and noncoding regions in DNA. The 12-symbol alphabet is based on nucleotide statistics inside codons. It is well known that the coding regions contain stop codons within maximum two phases and noncoding regions contain usually stop codons within all three phases [23]. In order to take into account these statistical properties of coding regions, we use the recursive segmentation algorithm proposed by Bernaola-Galvan et al. [19], a new 18-symbol alphabet that takes into account the nonuniform distribution of stop codons within all three phases, Jensen-Rényi divergence, and a new stopping criterion. The stopping criterion based on BIC for recursive segmentation was proposed by Li [5, 7]. Our approach uses only general statistical properties of coding regions. In this way, the prior training on data sets is avoided and furthermore, the search for additional biological information such as splice and promoter regions may also be avoided. It is noted that such additional information could be incorporated in a more concrete implementation of the algorithm [19]. Consequently, for three entire genomes of bacteria, we find that the use of nucleotide and stop-codon composition, and Jensen-Rényi divergence improve the accuracy of finding the borders between coding and noncoding regions in DNA sequences.

2. STOP-CODON STATISTICS

The distribution of stop codons in DNA coding regions is different than in the noncoding regions. Also, it is well known that the stop codons are strong signals in DNA sequences. In coding regions, the stop codons are usually distributed along two phases (reading frames) with the exception of the stop codon that is in a reading frame and signals the end of a gene. This knowledge is employed implicitly by hidden Markov models used in different gene-finding algorithms [4, 24, 25]. Explicitly, for the first time, the stop-codon statistics is used for recognizing coding regions in studies of Wang et al. [23] and Carpena et al. [26].

Different DNA sequences from different organisms are studied in order to show the distribution of stop codons along all three phases in coding and noncoding regions. There are extracted DNA sequences of different lengths—40, 80, 120, and 160 base pairs (bp)—from the following three randomly chosen prokaryote organisms: *Methanococcus jannaschii* (GenBank acc. L77117), *Chlamydia muridarum* (GenBank acc. AE002160), and *Chlamydophila pneumoniae* (GenBank acc. BA000008). The DNA sequences are taken randomly from coding and noncoding regions of the previous bacteria, and they are not overlapping on the same DNA strand.

Table 1 shows the counts of DNA sequences that have stop codons in one, two, and three phases, and no stop codons in neither of the three phases. There is no DNA coding region with stop codons within all three phases, as is shown in Table 1. We take advantage of this by introducing a new alphabet that considers also the stop-codon statistics and Jensen-Rényi divergence.

Also, in Figure 1, it is shown that the counts of stop-codons along all three phases are increasing rapidly with the length of noncoding regions, and in Figure 2, the counts of the stop codons along three phases are decreasing rapidly with the length of coding regions. Similar observations as in Figures 1, 2, and Table 1 have been used before for the introduction of the stop-codon statistics into the gene-finding field [23].

Figures 3 and 4 show the histograms of the lengths of noncoding and coding DNA regions from bacteria *Methanococcus jannaschii*, *Chlamydia muridarum*, and *Chlamydophila pneumoniae*; none of the coding regions of the three chosen bacteria have the length less than 50 bp, but there exist very short noncoding regions.

Table 1: Distribution of stop codons along phases for coding and noncoding DNA regions.

DNA sequence	Sequence length [bp]	Number of sequences	No stop codons [%]	Stop codons in one phase(s) [%]		
				one	two	three
Coding	40	8000	8.21	44.64	47.15	0
Noncoding	40	8000	5.32	31.08	46.36	17.24
Coding	80	4000	1.23	18.15	80.62	0
Noncoding	80	4000	0.45	6.30	37.80	55.35
Coding	120	2000	0.10	6.85	93.05	0
Noncoding	120	2000	0.30	1.85	22.60	75.25
Coding	160	1400	0	3.36	96.64	0
Noncoding	160	1400	0	0.70	13.20	86.20

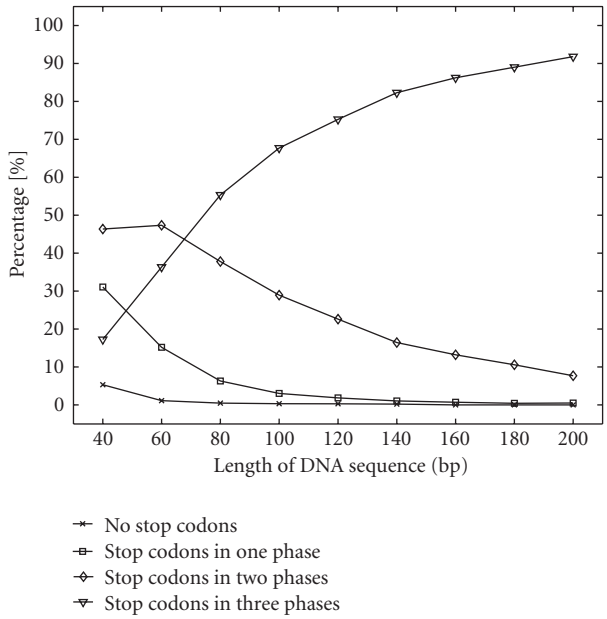


Figure 1: Distribution of stop codons along three phases in noncoding DNA regions.

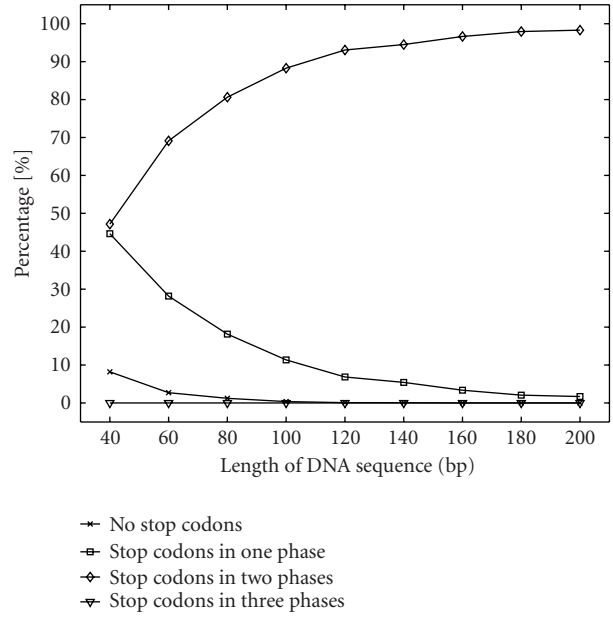


Figure 2: Distribution of stop codons along three phases in coding DNA regions.

The segmentation method based on nucleotide statistics [7] detects the coding regions even when they are on the opposite DNA strand. Thus, the stop-codon statistics along all three phases should also be considered on both DNA strands. As is shown in Figure 5, the stop codons on the reverse DNA strand appear in the given DNA strand, where the stop codons TAA, TAG, and TGA are situated as TCA, CTA, and TTA. When the codon CTA is met on a given DNA strand, it is known that it represents the stop codon TAG on the opposite DNA strand. In this way, the stop-codon statistics in both DNA strands is the same with the statistics of the six codons TAA, TAG, TGA, TCA, CTA, and TAA along a single DNA strand.

3. THE JENSEN-SHANNON DIVERGENCE

The Jensen-Shannon divergence quantifies the difference between two or more probability distributions and is widely used for DNA segmentation [5, 7, 19, 20, 21]. The Jensen-Shannon divergence D_{JS} between m probability distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ with the corresponding weights is defined as

$$D_{JS}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] = H\left[\sum_{j=1}^{m} \pi^{(j)} \cdot \mathbf{p}^{(j)}\right] - \sum_{j=1}^{m} \pi^{(j)} \cdot H[\mathbf{p}^{(j)}], \quad (1)$$

where $\mathbf{p}^{(j)} \equiv (p_1^{(j)}, p_2^{(j)}, \ldots, p_k^{(j)})$ are probability distributions satisfying the usual constraints $\sum_{i=1}^{k} p_i^{(j)} = 1$ and $0 \le p_i^{(j)} \le 1$, for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, m$; and $\pi^{(j)}$ are the weights of the distributions $\mathbf{p}^{(j)}$, satisfying the constraints $\sum_{j=1}^{m} \pi^{(j)} = 1$ and $0 \le \pi^{(j)} \le 1$. The Shannon entropy of the probability distribution $\mathbf{p}$ used in (1) is defined as

$$H[\mathbf{p}] = -\sum_{i=1}^{k} p_i \cdot \log_2 p_i. \quad (2)$$

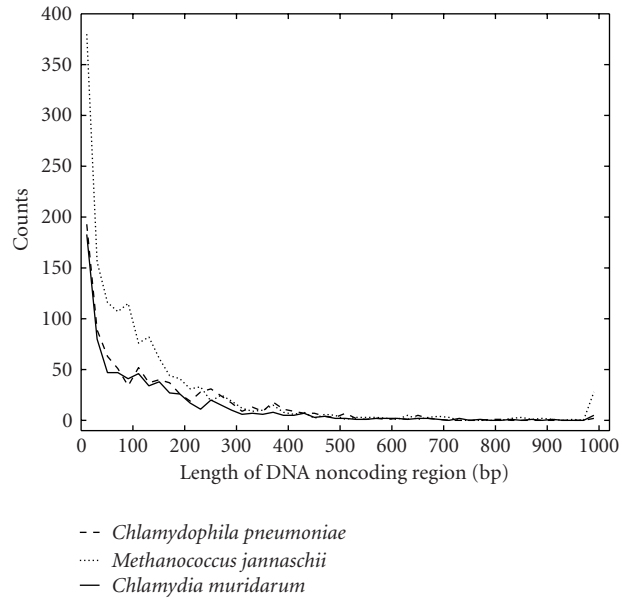


Figure 3: Histograms of the lengths of noncoding DNA regions.

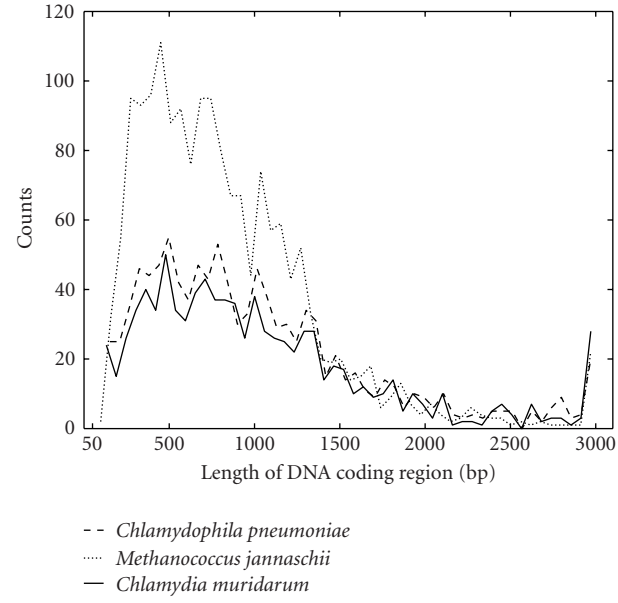


Figure 4: Histograms of the lengths of coding DNA regions.

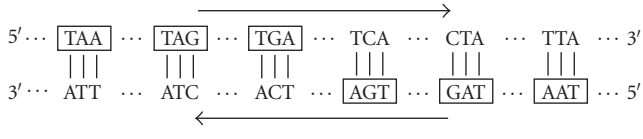


Figure 5: Stop codons in both strands of DNA.

Figure 6 illustrates the three-dimensional representation of the Jensen-Shannon divergence with equal weights for two Bernoulli probability distributions. Some mathematical

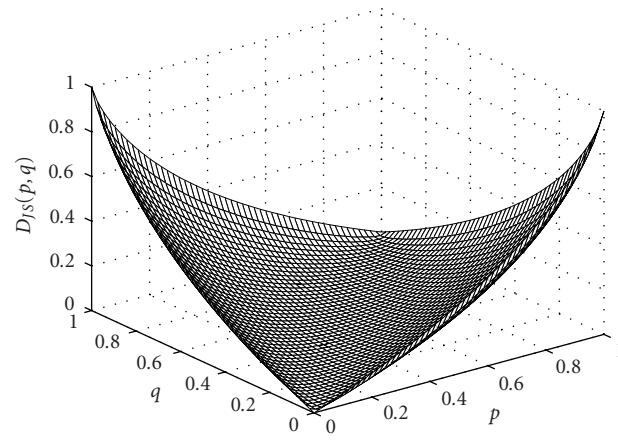


Figure 6: Three-dimensional representation of Jensen-Shannon divergence $D_{JS}(\mathbf{p},\mathbf{q})$, where $\mathbf{p} = (p, 1-p)$, $\mathbf{q} = (q, 1-q)$, and $\pi = (0.5, 0.5)$.

properties for the m-ary case that are important for its application as a divergence measure are the following:

(i) the use of Jensen inequality implies

$$D_{JS}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] \geq 0, \tag{3}$$

where $D_{JS}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] = 0$ if and only if $\mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \cdots = \mathbf{p}^{(m)}$;

(ii) the divergence D_{JS} is symmetric in its arguments $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$, that is, is invariant for any permutation of its arguments;

(iii) the divergence D_{JS} is well defined even if $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ are not absolutely continuous.

4. THE JENSEN-RÉNYI DIVERGENCE

The Jensen-Rényi divergence, as Jensen-Shannon divergence, is defined as a similarity measure between two or more probability distributions, and is used in image registration [27]. The Jensen-Rényi divergence D_{JR_α} between m probability distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ with the corresponding weights is defined as

$$D_{JR_\alpha}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] = R_\alpha\left[\sum_{j=1}^{m} \pi^{(j)} \cdot \mathbf{p}^{(j)}\right] - \sum_{j=1}^{m} \pi^{(j)} \cdot R_\alpha[\mathbf{p}^{(j)}]. \tag{4}$$

The Rényi entropy of the probability distribution $\mathbf{p}$ referred to in (4) is defined as

$$R_\alpha[\mathbf{p}] = \frac{1}{1-\alpha} \cdot \log_2 \sum_{i=1}^{k} p_i^\alpha, \tag{5}$$

where $\alpha > 0$ and $\alpha \neq 1$. For $\alpha > 1$, the Rényi entropy is neither concave nor convex [27]. For $\alpha \in (0, 1)$, the Rényi

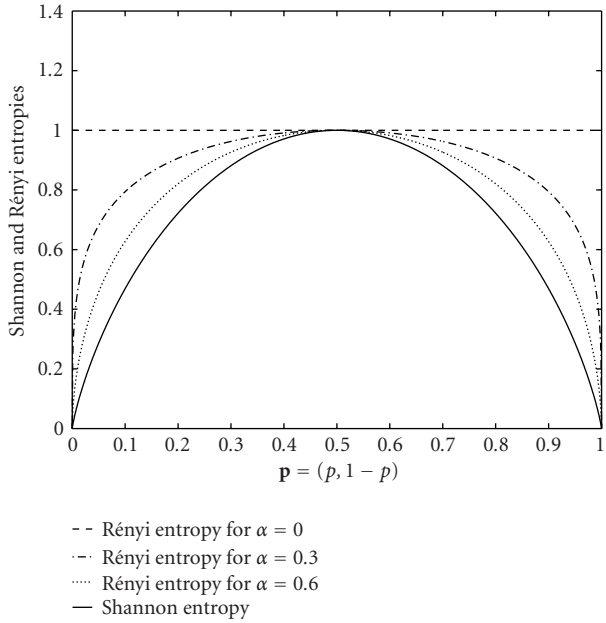


Figure 7: Shannon and Rényi entropies of Bernoulli distribution $\mathbf{p} = (p, 1-p)$ for different values of α.

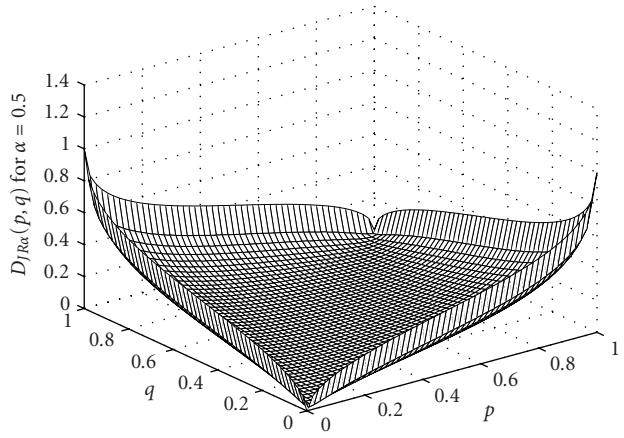


Figure 8: Three-dimensional representation of Jensen-Rényi divergence $D_{JR_\alpha}(\mathbf{p}, \mathbf{q})$, where $\mathbf{p} = (p, 1-p)$, $\mathbf{q} = (q, 1-q)$, $\pi = (0.5, 0.5)$, and $\alpha = 0.5$.

entropy is concave and tends to Shannon entropy $H[\mathbf{p}]$ as $\alpha \to 1$ [27]. The Rényi entropy is a nonincreasing function of α, and thus $R_\alpha[\mathbf{p}] \geq H[\mathbf{p}]$, for all $\alpha \in (0, 1)$. We restrict in this study $\alpha \in (0, 1)$, unless otherwise is specified. As shown in Figure 7, the measure of uncertainty is at a minimum when Shannon entropy is used and it increases as α decreases. The Rényi entropy attains a maximum uncertainty when α is equal to zero [27].

Figure 8 illustrates the three-dimensional representation of the Jensen-Rényi divergence for two Bernoulli probability distributions. Some mathematical properties for the m-ary case, for all $\alpha \in (0, 1)$, that are important for its application as a divergence measure [27] are the following:

(i) the use of Jensen inequality implies

$$D_{JR_\alpha}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] \geq 0, \tag{6}$$

where $D_{JR_\alpha}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}] = 0$ if and only if $\mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \cdots = \mathbf{p}^{(m)}$;

(ii) the divergence D_{JR_α} is symmetric in its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$, that is, is invariant for any permutation of its arguments;

(iii) the divergence D_{JR_α} is well defined even if $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(m)}$ are not absolutely continuous.

5. DETECTION OF BORDERS BETWEEN CODING AND NONCODING REGIONS USING RECURSIVE SEGMENTATION

We use the approach proposed by Bernaola-Galvan et al. [19, 20] and Li [5, 7] for segmentation of DNA sequences in homogeneous regions that are coding and noncoding. The recursive segmentation of a DNA sequence is as follows. First, the DNA sequence of length N_T is converted into a sequence of symbols with length N using a k-symbol alphabet. We sweep through the symbol sequence, and compute at every position i, where $i = 1, \ldots, N$, that divides the sequence into a left and a right sequence, the entropy of the whole, left, and right sequences. The position where the divergence reaches its maximum is accepted as a cutting point. Further, we recursively apply the segmentation to the left and to the right sequences until the maximized divergence measure is above a certain threshold. For the Jensen-Shannon divergence, the threshold is based on BIC. If the maximized divergence measure is above the threshold, the sequence is segmented, and if not, the segmentation is stopped for the respective sequence.

The Jensen-Shannon divergence D_{JS} is as follows:

$$D_{JS} = \max_i D_{JS}(i) = \left[H - \frac{i}{N} H_l - \frac{N-i}{N} H_r \right], \tag{7}$$

where H, H_l, and H_r are the Shannon entropies (2) of the whole, left, and right sequences, respectively [5, 7, 19, 20]. The weights are i/N and $(N-i)/N$ for the left and right sequences, respectively, where i is the point that divides the sequences into two sequences. In his study, Grosse et al. [22] shows that Jensen-Shannon divergence, as introduced previously, can be interpreted as the mutual information in the framework of information theory.

The Jensen-Rényi divergence D_{JR_α} is as follows:

$$D_{JR_\alpha} = \max_i D_{JR_\alpha}(i) = \left[R_\alpha - \frac{i}{N} R_{\alpha,l} - \frac{N-i}{N} R_{\alpha,r} \right], \tag{8}$$

where R_α, $R_{\alpha,l}$, and $R_{\alpha,r}$ are the Rényi entropies (5) of the whole, left, and right sequences, respectively.

Bernaola-Galvan et al. [19] introduces a 12-symbol alphabet in order to take into account the differential nucleotide composition in codons. The phase of the nucleotide, for this alphabet, is defined as $m = (n \bmod 3) + 1$, where $m \in \{1, 2, 3\}$, and n is the position of the nucleotide in the

Table 2: Symbol mapping for 12-symbol alphabet.

Nucleotide	Phase	Symbol
A	1	A_1
	2	A_2
	3	A_3
C	1	C_1
	2	C_2
	3	C_3
G	1	G_1
	2	G_2
	3	G_3
T	1	T_1
	2	T_2
	3	T_3

Table 3: Stop-codon mapping for 18-symbol alphabet.

Triplets of nucleotides (codons)	Phase	Symbol
TGA, TAG, or TAA	1	S_1
	2	S_2
	3	S_3
TCA, CTA, or TTA	1	S'_1
	2	S'_2
	3	S'_3

DNA sequence. Each nucleotide of the DNA sequence is substituted by the symbols from $\mathcal{A}_{12} = \{A_1, A_2, A_3, C_1, C_2, C_3, G_1, G_2, G_3, T_1, T_2, T_3\}$, as is also shown in Table 2.

We introduce in this study an 18-symbol alphabet that takes into account also the nonuniform distribution of stop-codons in both DNA strands, along all three phases [23]. Thus, the nucleotides and the stop codons are substituted by the symbols from $\mathcal{A}_{18} = \{A_1, A_2, A_3, C_1, C_2, C_3, G_1, G_2, G_3, T_1, T_2, T_3, S_1, S_2, S_3, S'_1, S'_2, S'_3\}$, where the symbols for nucleotides are as for $\mathcal{A}_{12}$ alphabet (Table 2). The symbols S_1, S_2, and S_3 are the stop codons TAA, TAG, and TGA in the given DNA strand, and S'_1, S'_2, and S'_3 are the stop codons AGT, GAT, and AAT on the opposite DNA strand, as shown in Table 3. The phase of a stop codon is defined the same as for a nucleotide with the exception that n represents the position of the first nucleotide of the given codon. For example, the DNA sequence ACTTAA is converted using the 18-symbol alphabet as $A_1C_2S'_3T_3S_1T_1A_2A_3$.

These two alphabets, together with the two divergence measures, are used for finding the borders between coding and noncoding regions in different DNA sequences from bacterium *Rickettsia prowazekii*, as shown in Figures 9 and 10.

In Figure 9, we plot the D_{JR_α} ($\alpha = 0.5$) and D_{JS} with $\mathcal{A}_{12}$ and $\mathcal{A}_{18}$ alphabets along a DNA sequence. The DNA sequence is composed of two randomly chosen regions from bacterium *Rickettsia prowazekii*. The first region of 1016 bp belongs to a coding region and the second one of 1151 bp belongs to a noncoding region. Figure 9 shows that using both divergences and both alphabets, we are able to find the border between the coding and noncoding region. Using $\mathcal{A}_{12}$ alphabet with both divergences, the cut is found at 11 bp to the right of the real border, and using $\mathcal{A}_{18}$ alphabet with both divergences, the cut is found at 4 bp to the left of the real border. In Figure 10, we plot both divergences using the both alphabets along a DNA sequence that contains a coding region of 810 bp from gene **RP172** followed by the original noncoding region of 1477 bp as it appears in the chromosome of bacterium *Rickettsia prowazekii.*

In Table 4, we analyze the same DNA sequence as in Figure 10 and it can be seen that using the alphabet $\mathcal{A}_{18}$ and the Jensen-Rényi divergence, we get the closest cut to the real border between the coding and noncoding regions. When the segmentation is applied on a single continuous DNA sequence followed by the "original" noncoding region as in Figure 10, using the alphabet $\mathcal{A}_{12}$ is not anymore possible to detect with a reasonable accuracy the border between the two regions, because the coding region "leaks," for a small portion, into the noncoding region. The region where the leaking phenomena happens has the same nucleotide composition as a coding region even though it is a noncoding region. This region does not have the same stop-codons composition as a coding region and because of this, using $\mathcal{A}_{18}$ alphabet, we are able to find a much closer border to the real one. The "leaking" regions appear usually in vicinity of the coding regions and they are removed in the cases when two randomly chosen, coding and noncoding, regions are joined arbitrary together, as in Figure 9. The Jensen-Rényi divergence takes better advantage of the $\mathcal{A}_{18}$ alphabet than Jensen-Shannon divergence because the counts of the stop-codons are much less than the counts of the nucleotides. The Jensen-Rényi divergence emphasizes better the difference between the regions with different stop-codon statistics. Thus, using the $\mathcal{A}_{18}$ alphabet and Jensen-Rényi divergence, we are able to detect better the border due to the introduction of the biological knowledge in the segmentation method.

6. STOPPING CRITERION FOR RECURSIVE SEGMENTATION

The stopping criterion in the case of Jensen-Shannon divergence can be considered from the point of view of the hypothesis testing and the model selection framework. For the hypothesis testing framework, the probability that the value of D_{JS} can be obtained by chance is computed by the null hypothesis that the sequence is homogeneous. The exact form of the null distribution is difficult to find [5, 28] but Grosse et al. [9, 22] suggest an empirical form of the null distribution based on numerical simulation.

In this study, the stopping criterion for segmentation using Jensen-Shannon divergence is based on model selection that has been introduced by Li in his studies [5, 7]. The model is judged by how well it fits the data and how complex it is. Thus the stopping criterion tests if a two-random-subsequence model is better than the one-random-sequence

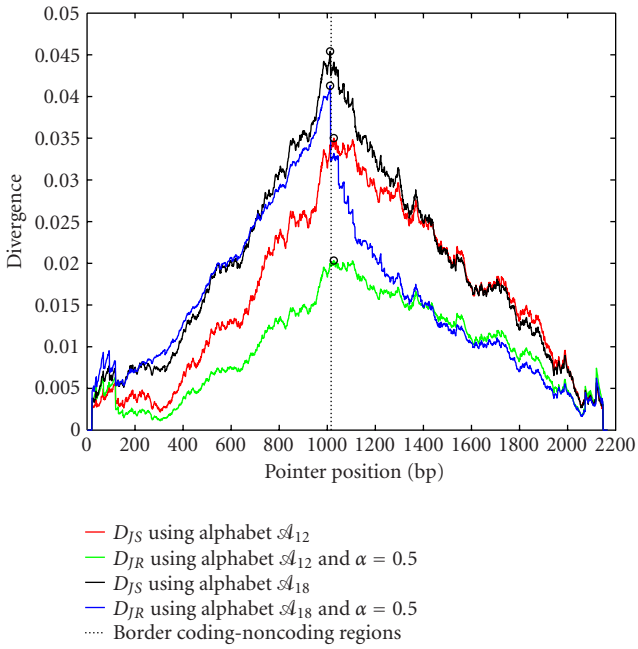


Figure 9: Jensen-Shannon divergence and Jensen-Rényi divergence versus cutting position for a DNA sequence containing a randomly chosen coding region and a randomly chosen noncoding region. The maximum values for the divergences are circled on the graph.

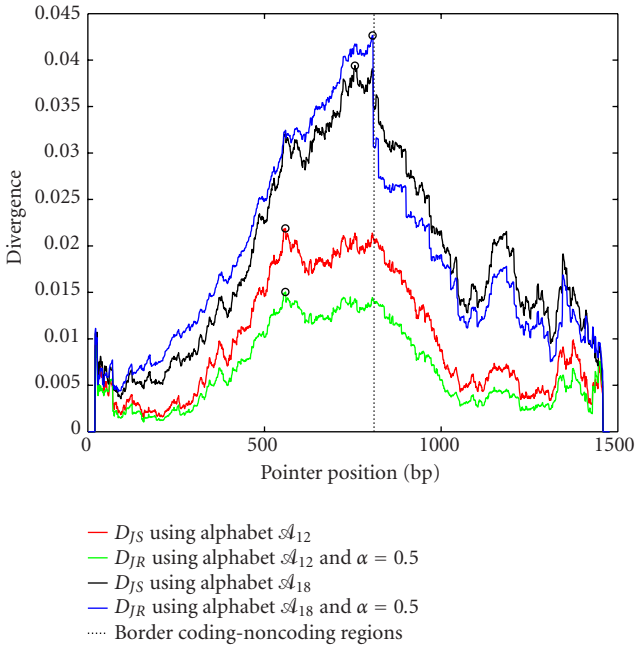


Figure 10: Jensen-Shannon divergence and Jensen-Rényi divergence versus cutting position for a DNA sequence containing a coding region followed by a noncoding region. The maximum values for the divergences are circled on the graph.

model. If the two-random-subsequence model is better, then the cut will be accepted, otherwise it is not. For balancing the goodness-of-fit of the model to the data with the number of parameters, the BIC is used as follows:

$$\Delta\text{BIC} = -2 \cdot \log L + K \cdot \log_2 N, \tag{9}$$

where $L = L_2/L_1$, L_1 and L_2 are the maximum likelihood of the models before and after the cut is made, respectively; $K = K_2 - K_1$, K_1 and K_2 are the number of free parameters before and after the cut is made, respectively; and N is the length of the sequence [5, 7]. In order to continue the recursive segmentation procedure and to decide if a cut is significant or not, the BIC should be reduced, that is, $\Delta\text{BIC} < 0$. This leads to

$$2 \cdot N \cdot D_{JS} > K \cdot \log_2 N. \tag{10}$$

In order to decide when the segmentation algorithm using D_{JS} has to be stopped, Li [5, 7] introduced, as a measure, the segmentation strength as

$$s = \frac{2 \cdot N \cdot D_{JS} - K \cdot \log_2 N}{K \cdot \log_2 N}. \tag{11}$$

The BIC stopping criterion is introduced here only for Jensen-Shannon divergence. In order to decide when the segmentation algorithm using D_{JR_α} has to be stopped, we introduce a new segmentation strength, derived empirically, as

$$s = \frac{2 \cdot N \cdot D_{JR_\alpha} - K \cdot \log_2 N}{K \cdot \log_2 N}. \tag{12}$$

The recursive segmentation continues, or a cut is accepted as significant as long as $s \geq s_0$, where s_0 can be set by the user. By setting the s_0, one affects the threshold used to make the decision if a cut is significant or not. For the $\mathcal{A}_{12}$ and $\mathcal{A}_{18}$ alphabets, the segmentation strength is defined by (11) or (12), where $K = 10$ and $K = 16$, respectively. The segmentation strengths for D_{JS} and D_{JR_α} have a closely related expression. Special cases of Jensen-Rényi divergence are obtained for $\alpha = 1/2$ for which one obtains the log Hellinger distance squared and for $\alpha = 1$ for which one obtains the Kullback-Liebler divergence [29]. For $\alpha = 1$, one obtains $D_{JR_\alpha} = D_{JS}$.

In this study, the standard stopping criterion is the stopping criterion where a cut is accepted as significant as long as $s \geq s_0$, where s is the segmentation strength in (11) and (12). A DNA sequence that does not have stop codons along all three phases has a very high probability (Figures 1 and 2) to be a coding region, and in this case it does not need to be

Table 4: Cuts obtained using different methods for segmentation for the same DNA sequence as in Figure 10.

Segmentation method	Distance from border
D_{JS} with $\mathcal{A}_{12}$ alphabet	251 bp (left)
D_{JR} ($\alpha = 0.5$) with $\mathcal{A}_{12}$ alphabet	251 bp (left)
D_{JS} with $\mathcal{A}_{18}$ alphabet	54 bp (left)
D_{JR} ($\alpha = 0.5$) with $\mathcal{A}_{18}$ alphabet	4 bp (left)

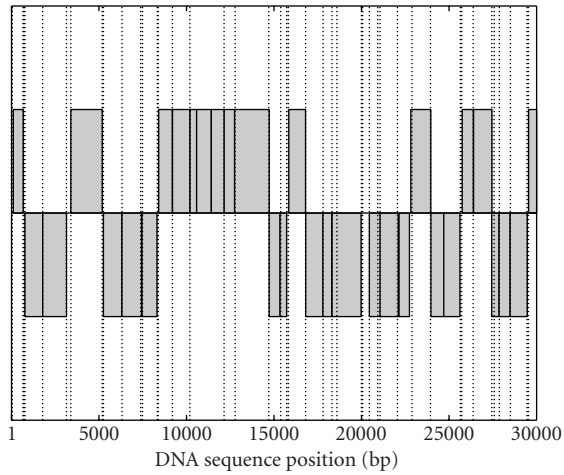


Figure 11: Comparison between the known coding regions (gray regions with solid lines as borders) of a DNA sequence from bacteria *Borrelia burgdorferi* and the borders (vertical dashed lines) obtained through recursive segmentation using Jensen-Rényi divergence ($\alpha = 0.5$), $\mathcal{A}_{18}$ alphabet, and standard stopping criterion. The coding regions oriented downwards are situated on the opposite DNA strand.

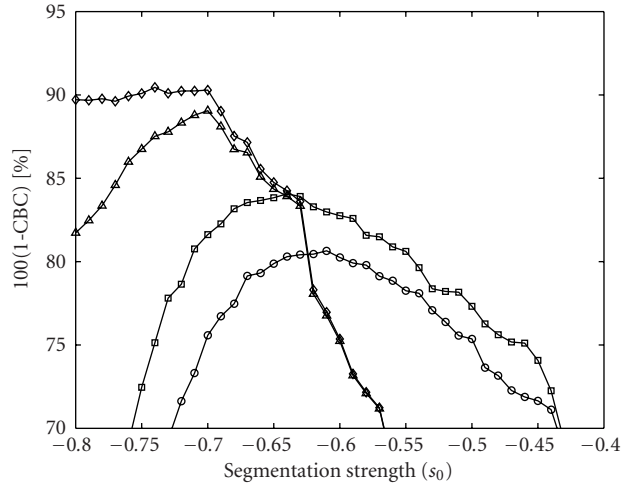


Figure 12: Accuracies of recursive segmentation for different thresholds of segmentation strength using Jensen-Shannon and Jensen-Rényi divergences with $\mathcal{A}_{12}$ and $\mathcal{A}_{18}$ alphabets and two stopping criterions for the genome of bacterium *Rickettsia prowazekii*.

segmented further. Thus, we introduce a new stopping criterion as follows. A cut is accepted as significant if $s \geq s_0$ and the segmented sequence has stop codons in all three phases. Hence, a DNA sequence is not segemented further if it has stop codons only in two phases.

In this study, the DNA sequences smaller than 40 bp in length are not segmented further in the recursive segmentation process because we consider that it is not statistically enough to separate them into two subsequences with a high confidence and the stop-codon statistics is not anymore relevant for such small sequences, as shown in Table 1.

7. EXPERIMENTAL RESULTS

In order to quantify the coincidence between cuts (CBC) obtained using the recursive segmentation algorithm and known borders between coding and noncoding regions, we use the following measure, introduced by Bernaola-Galvan et al. [19]:

$$\mathrm{CBC} = \frac{1}{2}\left[\sum_i \frac{\min_j |b_i - c_j|}{N_T} + \sum_j \frac{\min_i |b_i - c_j|}{N_T}\right], \quad (13)$$

where $\{b_i\}$ is the set of all borders between coding and noncoding regions, $\{c_j\}$ is the set of all cuts produced by the segmentation, and N_T represents the total length of the DNA sequence. The measure CBC is the average of the error in the determination of the correct boundaries between coding and noncoding regions, so the value $(1 - \mathrm{CBC})$ is a reasonable measure of the accuracy of the borders detected between coding and noncoding regions [19].

In Figure 11, a comparison is shown between the known regions of a DNA sequence containing the first 30000 bp from the beginning of the genome of bacterium *Borrelia burgdorferi* and the predicted borders obtained through recursive entropic segmentation using Jensen-Rényi divergence with the $\mathcal{A}_{18}$ alphabet and standard stopping criterion. The threshold of the segmentation strength is $s_0 = -0.55$ where the parameter CBC achieves its overall minimum. The borders between coding and noncoding regions are detected very close to the real ones as shown in Figure 11.

We show in Figures 12, 13, and 14 the results of the recursive segmentation for different values of the segmentation strength—using Jensen-Shannon and Jensen-Rényi divergences with alphabets $\mathcal{A}_{12}$ and $\mathcal{A}_{18}$, and two stopping criterions—of the whole genomes of the bacteria *Rickettsia prowazekii* (GenBank acc. AJ235269, length 1111523 bp), *Borrelia burgdorferi* (GenBank acc. AE000783, length 910724 bp), and *Methanococcus jannaschii* (GenBank acc. L77117, length 1664970 bp). For recursive segmentation of all three genomes with Jensen-Rényi divergence and $\mathcal{A}_{18}$ alphabet, we use $\alpha = 0.5$. This value has been found by segmenting the whole genome of bacterium *Rickettsia prowazekii*, using standard stopping criterion, for $\alpha = 0, 0.1, 0.2, \ldots, 0.9, 1$ and choosing the value for α, where the maximum of segmentation accuracy occurs. The recursive segmentation, using Jensen-Rényi divergence with $\mathcal{A}_{12}$ alphabet, achieves the maximum of the accuracy for $\alpha = 1$ that is the same as Jensen-Shannon divergence. Hence, the Jensen-Rényi divergence takes better advantage of the introduction of the stop-codon statistics than the Jensen-Shannon divergence does.

The recursive segmentation using the Jensen-Rényi divergence with $\mathcal{A}_{18}$ alphabet and new segmentation criterion

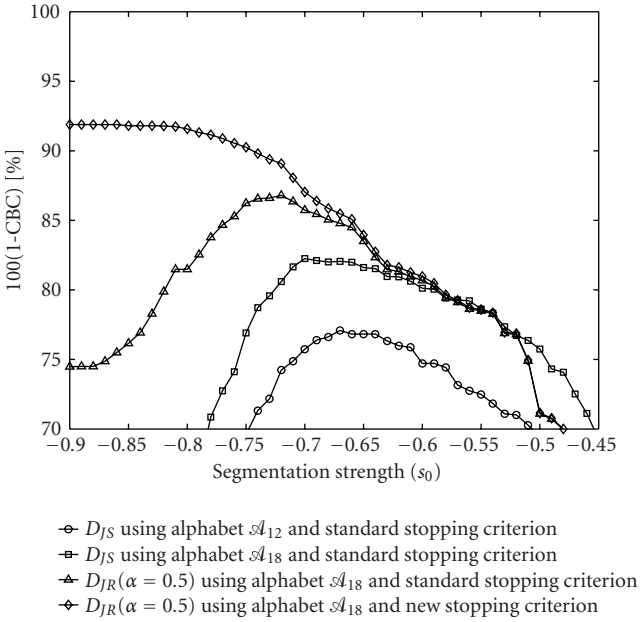


Figure 13: Accuracies of recursive segmentation for different thresholds of segmentation strength using Jensen-Shannon and Jensen-Rényi divergences with $\mathcal{A}_{12}$ and $\mathcal{A}_{18}$ alphabets and two stopping criterions for the genome of bacterium *Borrelia burgdorferi*.

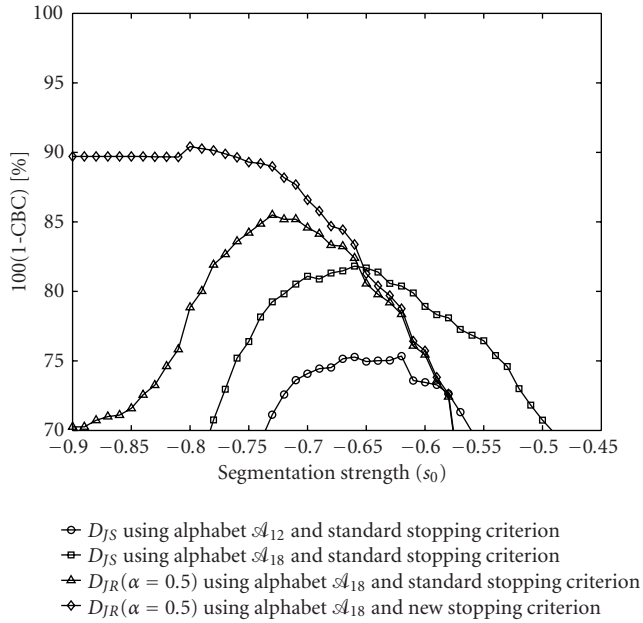


Figure 14: Accuracies of recursive segmentation for different thresholds of segmentation strength using Jensen-Shannon and Jensen-Rényi divergences with $\mathcal{A}_{12}$ and $\mathcal{A}_{18}$ alphabets and two stopping criterions for the genome of bacterium *Methanococcus jannaschii*.

achieves the best overall maximum accuracies for the whole genome of the three bacteria. Bernaola-Galvan et al. [19] achieves the maximum of accuracy in detecting the borders of 80% compared with our 80% with the same Jensen-Shannon divergence and same $\mathcal{A}_{12}$ alphabet. We use the standard stopping criterion based on BIC, compared with the statistical significance used by Bernaola-Galvan et al. [19]. Our newly introduced segmentation method that uses Jensen-Rényi divergence with $\mathcal{A}_{18}$ alphabet and the new stopping criterion gives an accuracy of 90% for $s_0 = -0.74$, that is, higher than 80% reported by Bernaola-Galvan et al. [19]. Also the accuracies for bacteria *Borrelia burgdorferi* and *Methanococcus jannaschiie* are improved from 77% and 75% with Jensen-Shannon divergence using $\mathcal{A}_{12}$ alphabet and standard stopping criterion to 91% and 89% with Jensen-Rényi divergence using $\mathcal{A}_{18}$ alphabet and new stopping criterion, respectively. The improvement in accuracy is explained by the use of Jensen-Rényi divergence that takes better advantage of the stop-codon statistics than Jensen-Shannon divergence does. Also, the introduction of the new stopping criterion in this study improves the accuracies of the segmentation. From Figures 12, 13, and 14, a good value of the threshold for the segmentation strength is $s_0 = -0.75$ for segmenting other genomes of bacteria with Jensen-Rényi divergence ($\alpha = 0.5$) using $\mathcal{A}_{18}$ alphabet and new stopping criterion. Even though, for $s_0 > -0.75$, higher accuracies can be achieved in some situations, this is not always true due to the scattering of coding regions in genome.

Consequently, our results that use the newly introduced approach, based on Jensen-Rényi divergence with the $\mathcal{A}_{18}$ alphabet and new stopping criterion, appear to be more accurate than those obtained using only Jensen-Shannon divergence with $\mathcal{A}_{12}$ alphabet and standard stopping criterion, in finding the borders between coding and noncoding regions.

8. DISCUSSION

In this study, we introduce a new segmentation method based on Jensen-Rényi divergence, an 18-symbol alphabet, and a new stopping criterion for finding the borders between coding and noncoding regions. The new segmentation method applied to three bacteria genome improves the accuracies of the border detection compared to the standard segmentation procedures previously reported. We employ the composition of stop codons over all three phases along the DNA sequence in the 18-symbol alphabet and in the new stopping criterion for improving the accuracy of finding the borders between coding and the noncoding DNA regions.

The assumptions built in other gene-finding systems as GENMARK, VEIL [25], and MORGAN [30] have a number of shortcomings [30] that do not affect the recursive entropic segmentation in finding the borders between coding and noncoding regions. A direct comparison between gene-finding and recursive segmentation for finding the borders between coding and noncoding regions is difficult to make because the gene-finding systems perform very well on small DNA sequence that contains only one gene or very few coding regions. The recursive entropic segmentation performs better on long DNA sequences with a large number of genes, in order to gain statistics. The present segmentation

algorithms [5, 7, 19] rely heavily on statistical properties for finding the coding, noncoding, and other regions of interests in DNA, but the gene-finding systems [4, 25, 30] use biological knowledge regarding functional sites, together with statistics for finding genes. Also, the recursive segmentation needs no prior training compared with gene-finding systems that require extensive training on known datasets. In eukaryotes are much more short coding-regions that are more "scattered" than in prokaryotes and thus it is more difficult to find their borders-based statistical properties as in [5]. The genomes analyzed in this study belong only to prokaryotes that have the coding regions much more compact than in eukaryotes.

9. CONCLUSION

There is an increasing need to develop new algorithms for finding coding regions in DNA sequences. In this study, we introduce a new segmentation method based on Jensen-Rényi divergence with an 18-symbol alphabet and new stopping criterion for finding the borders between coding and noncoding regions in prokaryotes. We use recursive segmentation along with a stopping criterion based on Bayesian information criterion (BIC). Together, they offer a novel method to view the compositional heterogeneity of a DNA sequence. The success comes from the utilization of the stop-codon statistics in all three phases along the DNA sequence and use of Jensen-Rényi divergence. For three entire genomes of bacteria, we found that the use of Jensen-Rényi divergence, nucleotide composition, and stop-codon composition improves the accuracy of finding the borders between coding and noncoding regions in DNA sequences, compared to the standard segmentation procedures previously reported.

REFERENCES

[1] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, no. 17, pp. 5303–5318, 1982.

[2] R. Staden and A. D. McLachlan, "Codon preference and its use in identifying protein coding regions in long DNA sequences," *Nucleic Acids Research*, vol. 10, pp. 141–156, 1982.

[3] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, no. 3, pp. 353–367, 1996.

[4] D. Nicorici, J. Astola, and I. Tabus, "Computational identification of exons in DNA with a hidden Markov model," in *Workshop on Genomic Signal Processing and Statistics*, Raleigh, NC, USA, October 2002.

[5] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse, "Applications of recursive segmentation to the analysis of DNA sequences," *Computers and Chemistry*, vol. 26, no. 5, pp. 491–510, 2002.

[6] R. K. Azad, J. S. Rao, W. Li, and R. Ramaswamy, "Simplifying the mosaic description of DNA sequences," *Phys. Rev. E*, vol. 66, no. 031913, pp. 1–6, 2002.

[7] W. Li, "New stopping criteria for segmenting DNA sequences," *Phys. Rev. Lett.*, vol. 86, no. 25, pp. 5815–5818, 2001.

[8] P. D. Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871–888, 2003.

[9] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, "Species independence of mutual information in coding and noncoding DNA," *Phys. Rev. E*, vol. 61, no. 5, pp. 5624–5629, 2000.

[10] W. Li, G. Stolovitzky, P. Bernaola-Galvan, and J. L. Oliver, "Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes," *Genome Research*, vol. 8, no. 9, pp. 916–928, 1998.

[11] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, "Periodicity in DNA coding sequences: Implications in gene evolution," *Journal of Theoretical Biology*, vol. 151, no. 3, pp. 323–331, 1991.

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

[13] D. Anastassiou, "DSP in Genomics," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1053–1056, Salt Lake City, Utah, USA, May 2001.

[14] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using allpass-based filters," in *Workshop on Genomic Signal Processing and Statistics*, Raleigh, NC, USA, October 2002.

[15] R. Farber, A. Lapedes, and K. Sirotkin, "Determination of eukaryotic protein coding regions using neural networks and information theory," *J. Mol. Biol.*, vol. 226, pp. 471–479, 1992.

[16] R. Staden, "Computer methods to locate signals in nucleic acid sequences," *Nucleic Acids Research*, vol. 12, no. 1, pp. 505–519, 1984.

[17] J. W. Fickett, "Finding genes by computer: the state of the art," *Trends in Genetics*, vol. 12, no. 8, pp. 316–320, 1996.

[18] J. W. Fickett, "The gene identification problem: an overview for developers," *Computer and Chemistry*, vol. 20, no. 1, pp. 103–118, 1996.

[19] P. Bernaola-Galvan, I. Grosse, P. Carpena, J. L. Oliver, R. Roman-Roldan, and H. E. Stanley, "Finding borders between coding and noncoding DNA regions by an entropic segmentation method," *Phys. Rev. Lett.*, vol. 85, no. 6, pp. 1342–1345, 2000.

[20] P. Bernaola-Galvan, R. Roman-Roldan, and J. L. Oliver, "Compositional segmentation and long-range fractal correlations in DNA sequences," *Phys. Rev. E*, vol. 53, no. 5, pp. 5181–5189, 1996.

[21] D. Nicorici, J. A. Berger, J. Astola, and S. K. Mitra, "Finding borders between coding and noncoding DNA regions using recursive segmentation and statistics of stop codons," in *Proceedings of the 2003 Finnish Signal Processing Symposium*, pp. 231–235, Tampere, Finland, May 2003.

[22] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. L. Oliver, and H. E. Stanley, "Analysis of symbolic sequences using the Jensen-Shannon divergence," *Phys. Rev. E*, vol. 65, no. 041905, pp. 1–16, 2002.

[23] Y. Wang, C. T. Zhang, and P. Dong, "Recognizing shorter coding regions of human genes based on the statistics of stop codons," *BioPolymers*, vol. 63, no. 3, pp. 207–216, 2002.

[24] M. Borodovsky and J. McIninch, "GENMARK: parallel gene recognition for both DNA strands," *Computer and Chemistry*, vol. 17, no. 2, pp. 123–134, 1993.

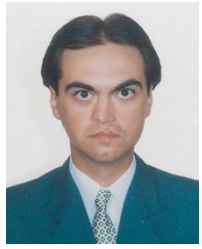
[25] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in DNA with a hidden Markov model," *Journal of Computational Biology*, vol. 4, no. 2, pp. 127–141, 1997.

[26] P. Carpena, P. Bernaola-Galvan, R. Roman-Roldan, and J. L. Oliver, "A simple and species-independent coding measure," *Gene*, vol. 300, no. 1–2, pp. 97–104, 2002.

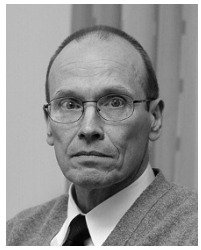
[27] Y. He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1211–1220, 2003.

[28] A. N. Pettitt, "A simple cumulative sum type statistic for the change-point problem with zero-one variables," *Biometrika*, vol. 67, no. 1, pp. 79–84, 1980.

[29] A. O. Hero and O. J. J. Michel, "Rényi information divergence via measure transformations on minimal spanning trees," in *Proc. IEEE 2000 International Symposium on Information Theory*, p. 414, Sorrento, Italy, June 2000.
[30] S. Salzberg, A. Delcher, K. Fasman, and J. Henderson, "A decision tree system for finding genes in DNA," *Journal of Computational Biology*, vol. 5, no. 4, pp. 667–680, 1998.



Daniel Nicorici received his B.S. and M.S. degrees in electrical engineering from Technical University of Cluj-Napoca, Romania, in 1999 and 2000, respectively. Since 2001, he has been with Tampere University of Technology, Finland, as a Researcher. He is currently pursuing his Ph.D. at Tampere International Center for Signal Processing. His research interest focuses on genomic signal processing.



Jaakko Astola (IEEE Fellow) received his B.S., M.S., Licentiate, and Ph.D. degrees in mathematics (specialising in error-correcting codes) from Turku University, Finland, in 1972, 1973, 1975, and 1978, respectively. From 1976 to 1977, he was with the Research Institute for Mathematical Sciences of Kyoto University, Kyoto, Japan. Between 1979 and 1987, he was with the Department of Information Technology, Lappeenranta University of Technology, Lappeenranta, Finland, holding various teaching positions in mathematics, applied mathematics, and computer science. In 1984, he worked as a Visiting Scientist in Eindhoven University of Technology, the Netherlands. From 1987 to 1992, he was an Associate Professor in applied mathematics at Tampere University, Tampere, Finland. Since 1993, he has been a Professor of signal processing and Director of Tampere International Center for Signal Processing, leading a group of about 60 scientists. From 2001 to 2006, he was nominated Academy Professor by Academy of Finland. His research interests include signal processing, coding theory, spectral techniques, and statistics.