

DIVERGENCE MEASURES FOR DNA SEGMENTATION

Daniel Nicorici*, Jaakko Astola
Tampere International Center for Signal Processing,
Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND
Email: Daniel.Nicorici@tut.fi, Jaakko.Astola@tut.fi

Abstract

Entropy-based divergence measures have shown promising results in many areas of engineering and image processing. In this study, we use the Jensen-Shannon and Jensen-Rényi divergence measures for DNA segmentation. Based on these information theoretic measures and protein shape coded in DNA, we propose a new approach to the problem of finding the borders between coding and noncoding DNA regions. We find that Jensen-Rényi divergence together with an alphabet based on protein shape characteristics brings improvements in accuracy for DNA segmentation. Our results demonstrate that the proposed approach is highly accurate, in finding the borders between coding and noncoding regions, and requires no “prior training” on known data sets.

Keywords: Jensen-Rényi Divergence, Jensen-Shannon Divergence, Protein Coding DNA Regions.

1. INTRODUCTION

The computational identification of the genes and the protein-coding regions in the 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 the annotated genes that will be used in the biomedical research. Also, methods that can reliably identify the genes in anonymous sequences of DNA, generated by the genome project, can speed the process. A number of such methods exist but their predictive performance for finding

* Corresponding author

genes is still not satisfactory [3]. There are two basic problems in gene finding: detection of the functional sites of the genes, and detection of the regions that code for proteins. These are not satisfactorily solved and the reliable detection of genes and coding regions in the 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 borders between coding regions in the DNA sequences, that code for proteins, and noncoding regions.

Almost everything in the organism of the living beings is made of, or by proteins. According to the central dogma, that forms the backbone of molecular biology, the DNA codes for the production of messenger RNA (mRNA) during transcription process, and the ribosomes “read” the information that is coded in the mRNA, and use it for protein synthesis during the translation process. Also, the spatial shape of the proteins is critical to their function.

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), guanine (G), thymine (T), and cytosine (C). Along the two strands of the DNA double-helix, a pyrimidine (bases T and C) in one chain always faces a purine (bases A and G) in the other and only the complementary base pairs T-A and C-G exists. Also, there is a large redundancy of the protein-coding regions in DNA that is distributed unevenly, as there are $4^3=64$ codons (sequence of three nucleotides) to specify only 21 outputs that are 20 aminoacids and one output that signals the end of the translation process.

One generic feature of the 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 of DNA is still limited to the structure of the codon and the 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 coding regions than the noncoding ones, provides a strong signal for detection [9,10,28].

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

The segmentation methods are computational methods used to identify the homogeneous regions and they are important for DNA sequence analysis for identifying the borders between the coding and the noncoding regions [5,7,18,19]. The Jensen-Shannon divergence is one of the most widely used methods for segmenting the DNA sequences [5,6,7,18,19,20,21] and it is used for separating recursively the DNA sequences in homogenous regions with respect to its neighbours.

In this study, we introduce a new approach for finding the borders between coding and noncoding DNA regions, without *a priori* training, based on usage of Jensen-Rényi divergence and the shape of the proteins that are coded in the DNA. Our approach uses only statistical general properties of the coding DNA and proteins. In this way, the prior training on data sets is avoided and furthermore the search for

additional biological information (splice sites, promoter regions) can be also avoided; however such additional information could be easily incorporated and exploited in a concrete implementation of the algorithm. Our approach can be used for very long DNA sequences or very short that contains only partial coding DNA regions.

2. THE JENSEN-SHANNON DIVERGENCE

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

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

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

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

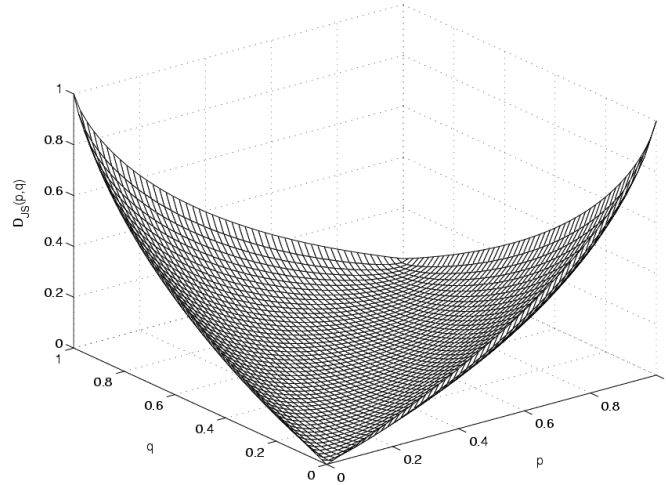


Figure 1 – Representation in three dimensions of Jensen-Shannon divergence $D_{JS}(\mathbf{p}, \mathbf{q})$ where $\mathbf{p} = (p, 1-p)$, $\mathbf{q} = (q, 1-q)$, $\boldsymbol{\pi} = (0.5, 0.5)$

Figure 1 illustrates the three dimensional representation of the Jensen-Shannon divergence with equal weights for two Bernoulli probability distributions.

Some mathematical properties of D_{JS} for the m -ary case that are important for its application as a divergence measure are:

1. the use of Jensen inequality implies that

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

with $D_{JS}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}] = 0$ if and only if $\mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \dots = \mathbf{p}^{(m)}$.

2. divergence D_{JS} is symmetric in its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$, i.e. D_{JS} is invariant for any permutation of its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$.
3. divergence D_{JS} is well defined even if $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$ are not absolutely continuous.

3. 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 it has been introduced by Yun He [22] and it has been used only in image registration [22]. The Jensen-Rényi divergence (D_{JR_α}) between m probability distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$ with the corresponding weights $\pi^{(1)}, \pi^{(2)}, \dots, \pi^{(m)}$ is defined as

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

where $\mathbf{p}^{(j)} \equiv (p_1^{(j)}, p_2^{(j)}, \dots, p_k^{(j)})$ are probability distributions satisfying the usual constraints $\sum_{i=1}^k p_i^{(j)} = 1$ and $0 \leq p_i^{(j)} \leq 1$ for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, m$, and $\pi^{(j)}$ are the weights of the distributions $\mathbf{p}^{(j)}$, satisfying the constraints $\sum_{j=1}^m \pi^{(j)} = 1$ and $0 \leq \pi^{(j)} \leq 1$. The Rényi entropy of the probability distribution \mathbf{p} used in Equation (4) is defined as

$$R_\alpha[\mathbf{p}] = \frac{1}{1-\alpha} \cdot \log_2 \sum_{i=1}^k p_i^\alpha, \quad \alpha > 0 \text{ and } \alpha \neq 1. \quad (5)$$

For $\alpha > 1$, the Rényi entropy is neither concave nor convex [22]. For $\alpha \in (0, 1)$ the Rényi entropy is concave and tends to Shannon entropy $H(\mathbf{p})$ as $\alpha \rightarrow 1$ [23]. The Rényi entropy is a non-increasing function of α , and thus

$$R_\alpha(\mathbf{p}) \geq H(\mathbf{p}), \quad \forall \alpha \in (0, 1). \quad (6)$$

We restrict in this study $\alpha \in (0, 1)$, unless otherwise is specified. As shown in the Figure 2, 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 [22].

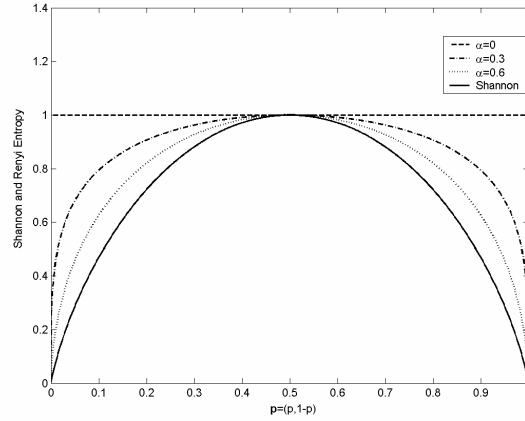


Figure 2 – Shannon and Rényi entropy of Bernoulli distribution $\mathbf{p} = (p, 1-p)$ for different values of α

Figure 3 illustrates the three dimensional representation of the Jensen-Rényi divergence for two Bernoulli probability distributions.

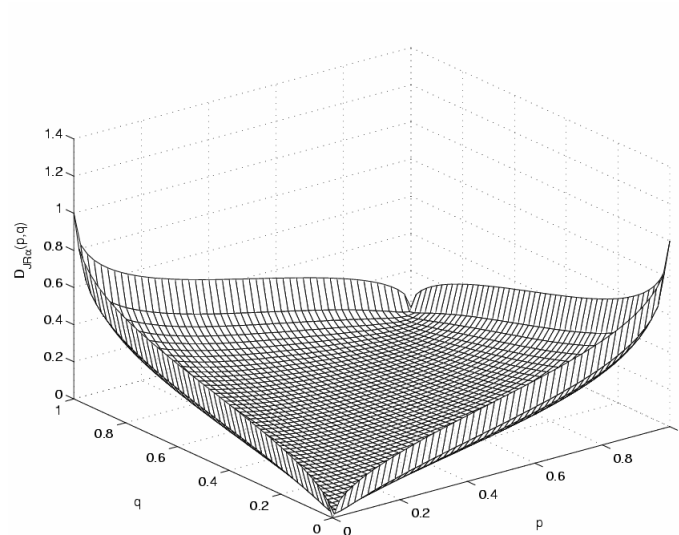


Figure 3 – Representation in three dimensions of Jensen-Rényi divergence $D_{JR_\alpha}(\mathbf{p}, \mathbf{q})$ where $\mathbf{p} = (p, 1-p)$, $\mathbf{q} = (q, 1-q)$, $\boldsymbol{\pi} = (0.5, 0.5)$, $\alpha = 0.3$

Some mathematical properties of D_{JR_α} for the m -ary case that are important for its application as a divergence measure are:

1. the use of Jensen inequality implies that

$$D_{JR_\alpha}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}] \geq 0, \quad (7)$$

with $D_{JR_\alpha}[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}] = 0$ if and only if $\mathbf{p}^{(1)} = \mathbf{p}^{(2)} = \dots = \mathbf{p}^{(m)}$.

2. divergence D_{JR_α} is symmetric in its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$, i.e. D_{JR_α} is invariant for any permutation of its arguments $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$.
3. divergence D_{JR_α} is well defined even if $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(m)}$ are not absolutely continuous.

4. PROTEIN CODING REGIONS OF DNA

Amino acids are the basic building blocks of proteins. The shape of the proteins is critical to their function, and their shape is largely a result of the bonds that form between the side chains of amino acids that make the protein. Thus it is concluded that a primary purpose of the side chains in amino acids is to give proteins their shape that dictates their function [24]. There are twenty amino acids that are used to form proteins, and they are presented in Table 1 with their corresponding triplets of DNA nucleotides [24]. For example the codons GGG, GGA, GGC, and GGT code for the same amino acid, called Glycine (Gly).

Table 1 – Genetic code mapping codons (DNA) to amino acids

First position (5' end)	Second position				Third position (3' end)
	G	A	C	T	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	T
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	T
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	T
T	Trp	Stop codon	Ser	Leu	G
	Stop codon	Stop codon	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	T

The approach used by Bernaola-Galvan [18,19], Li [5,7] for finding the coding and noncoding DNA regions with entropic segmentation, separates the DNA sequences in homogeneous regions based on differential nucleotide composition of the neighboring DNA regions. In this study, we introduce a new approach for finding the borders between coding and noncoding DNA regions based on differential composition of the four types of codons within three reading frames in order to take into account the shape of the protein that is coded in DNA. The Jensen-Shannon and Jensen-Rényi divergences are used to measure the differential compositions.

Table 2 – Classification of amino acids based on the characteristics of their R groups (side chains)

Classification	Amino acids
Nonpolar	Gly, Ala, Val, Leu, Ile, Met, Pro, Phe, Trp
Charged polar	Ser, Thr, Asn, Cys, Tyr, Gln
Uncharged polar	Lys, Arg, His, Asp, Glu

Timberlake [25], classifies the amino acids based on characteristics of their R groups (side chains), identifying amino acids as nonpolar, charged polar and uncharged polar, as presented in Table 2. Thus using this classification of amino acids we are taking into consideration the protein shape that is coded in DNA.

There are 64 possible codons and they can be divided into four groups: three groups that are based on the chemical characteristics of the amino acids that the

codons code for (Table 2), and one more group that contains the stop codons (signal the end of the translation process) and do not code for any amino acid.

4. DETECTION OF BORDERS BETWEEN CODING AND NONCODING DNA REGIONS

We use in this study the approach proposed by Bernaola-Galvan [18,19] and Li [5,7] for segmentation of the DNA sequences in homogeneous regions, that are coding and noncoding. The segmentation algorithm sweeps through the given DNA sequence and computes for every position i ($0 < i < N$), that divides the sequence into two subsequences, the divergence measure using an alphabet of k -symbols. At the position where the divergence reaches its maximum is accepted as cutting point.

The Jensen-Shannon divergence (D_{JS}) for DNA segmentation [5,18,20] is given (derived from Equation 1) as

$$D_{JS} = \max_i D_{JS}(i) = \max_i \left[H - \frac{i}{N} H_L - \frac{N-i}{N} H_R \right] \quad (8)$$

where:

- H is the Shannon entropy of the whole sequence;
- H_L Shannon entropy of the subsequence on the left side of the partition point;
- H_R Shannon entropy of the subsequence on the right side of the partition point;
- N number of symbols in whole sequence.

In his study Grosse [21] shows that Jensen-Shannon divergence D_{JS} , as is introduced previously, can be interpreted as the mutual information, in the framework of information theory.

We use also the Jensen-Rényi divergence that has been used introduced by Yun He for image registration [22], for DNA segmentation. The Jensen-Rényi divergence (D_{JR_α}) for DNA segmentation is given (derived from Equation 4) as

$$D_{JR_\alpha} = \max_i D_{JR_\alpha}(i) = \max_i \left[R_\alpha - \frac{i}{N} R_{\alpha,L} - \frac{N-i}{N} R_{\alpha,R} \right] \quad (9)$$

where:

- R_α is the Rényi entropy of the whole sequence;
- $R_{\alpha,L}$ Rényi entropy of the subsequence on the left side of the partition point;
- $R_{\alpha,R}$ Rényi entropy of the subsequence on the right side of the partition point;
- N number of symbols in whole sequence.

The Shannon and Rényi entropies are computed using two different alphabets. In his study Bernaola-Galvan [18] introduces a 12-symbol “nucleotide alphabet”, in order to take into account the differential nucleotide composition within codons. The phase of nucleotides (position of the nucleotides inside the codons) for this alphabet is taken into account. The phase is defined as $m = ((n-1) \bmod 3) + 1$ where $m \in \{1, 2, 3\}$ and

n is the position of the nucleotide inside the DNA sequence. Each nucleotide from the DNA sequence is substituted by the following symbols from the nucleotide alphabet $\mathcal{A}_{12} = \{A_1, A_2, A_3, C_1, C_2, C_3, G_1, G_2, G_3, T_1, T_2, T_3\}$ where, for example, A_2 is the nucleotide A within phase 2 [18]. For example the DNA sequence ACTGACTGCCAT is translated using the nucleotide alphabet \mathcal{A}_{12} as $A_1C_2T_3G_1A_2C_3T_1G_2C_3C_1A_2T_3$.

We introduce in this study a new 12-symbol protein alphabet in order to take into account the protein shape encoded into DNA, and it is defined as $\mathcal{A}_{12}^* = \{NP_1, NP_2, NP_3, CP_1, CP_2, CP_3, UP_1, UP_2, UP_3, SC_1, SC_2, SC_3\}$. The symbols NP, CP, UP, and SC represents the four groups of codons that are built based on chemical properties of amino acids that they code for, and they are defined in Table 3, and Tables 1 and 2. The four groups of codons are described into section 4.

Table 3 - Classification of codons based on the characteristics of protein characteristics

Corresponding Codons	Alphabet symbol
Nonpolar aminoacids	NP
Charged polar aminoacids	CP
Uncharged polar aminoacids	UP
Stop codon	SC

The reading frame of the codons for protein alphabet \mathcal{A}_{12}^* is taken into account and it is defined as $m = ((n-1) \bmod 3) + 1$ where $m \in \{1, 2, 3\}$ and n is position of the codon inside the DNA sequence. Each codon of the DNA sequence is substituted by the symbols from the protein alphabet \mathcal{A}_{12}^* . For example the DNA sequence ACTGACTGCCAT from Figure 4 is translated using the protein alphabet \mathcal{A}_{12}^* (Tables 1, 2 and 3) as $CP_1NP_2SC_3UP_1CP_2NP_3CP_1SC_2NP_3UP_1$.

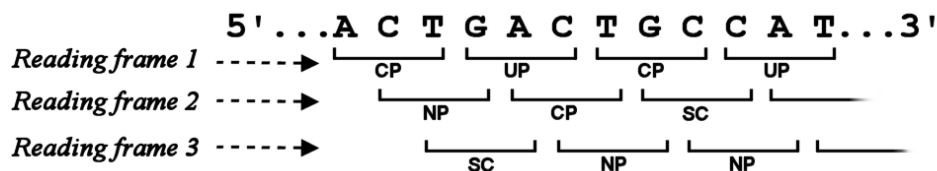


Figure 4 – Reading frames in a DNA sequence and the mapping into the protein alphabet \mathcal{A}_{12}^*

The reading frames for codons and the phases for nucleotides are considered from the beginning of the given DNA sequence. In Figure 5, we show the Jensen-Shannon divergence, using the nucleotide alphabet \mathcal{A}_{12} , and Jensen-Rényi divergences, using the nucleotide alphabet \mathcal{A}_{12} and protein alphabet \mathcal{A}_{12}^* , along two DNA sequences from bacterium *Rickettsia prowazekii* (GenBank acc. AJ235269). In Figure 5.a. is used for segmentation a DNA sequence of length 2288 bp obtained by joining an arbitrary-chosen coding region (length 1016 bp) with a arbitrary-chosen noncoding region (length 1272 bp). In Figure 5.b. is used for segmentation a single continuous DNA sequence of length 4380 bp that contains a coding region (last 2560 bp of the gene RP856) followed by the “original” noncoding region (1280 bp in length).

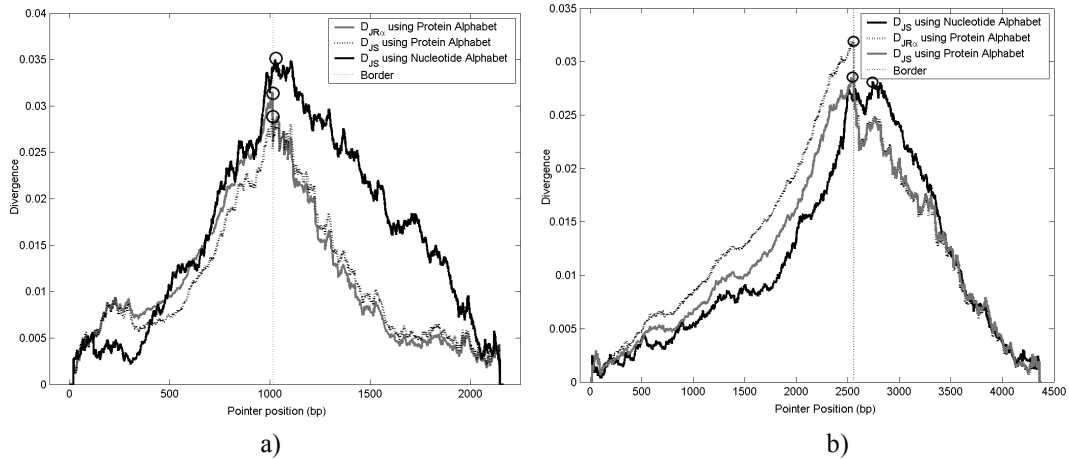


Figure 5 - Jensen-Shannon divergence (D_{JS}) and Jensen-Rényi divergence ($D_{JR\alpha}$) for $\alpha=0.8$ vs. cutting position for a DNA sequence (the maximum values are circled on the graph):
a) obtained by joining an arbitrary coding region and an arbitrary noncoding region;
b) containing in the beginning a coding region followed by a noncoding region;

The divergences are computed and the cuts are made where the divergences achieve their maximum. The cuts obtained for the DNA sequences from Figure 5.a. and Figure 5.b are presented in Tables 4 and 5, respectively.

Table 4 - Cuts obtained using different methods for segmentation of DNA sequence from Figure 5.a

Segmentation method	Distance from real border
Jensen-Shannon divergence using nucleotide alphabet \mathcal{A}_{12}	11 bp (right)
Jensen-Shannon divergence using protein alphabet \mathcal{A}_{12}^*	5 bp (left)
Jensen-Rényi divergence using protein alphabet \mathcal{A}_{12}^* , ($\alpha=0.8$)	5 bp (left)

Table 5 - Cuts obtained using different methods for segmentation of DNA sequence from Figure 5.b

Segmentation method	Distance from real border
Jensen-Shannon divergence using nucleotide alphabet \mathcal{A}_{12}	184 bp (right)
Jensen-Shannon divergence using protein alphabet \mathcal{A}_{12}^*	26 bp (left)
Jensen-Rényi divergence using protein alphabet \mathcal{A}_{12}^* , ($\alpha=0.8$)	4 bp (left)

For a DNA sequence created through joining an arbitrary-chosen coding region with an arbitrary-chosen noncoding region the cuts are slightly more precisely using the protein alphabet (Figure 5.a and Table 4). But, when the segmentation is applied on a single continuous DNA sequence (Figure 5.b and Table 5) that contains a coding region followed by the “original” noncoding region, the use of protein alphabet \mathcal{A}_{12}^* and Jensen-Rényi divergence brings a major improvement in detecting the border between coding and noncoding region.

5. RESULTS

In order to quantify the coincidence of the cut (CC) obtained using the segmentation algorithm and the known border between coding and noncoding region, we use the measure that is a variation of the measure introduced by Bernaola-Galvan [18]

$$CC = \frac{|b-c|}{N} \quad (10)$$

where b is the border between coding and noncoding region, and c is the cut produced by the segmentation, and N the total length of the DNA sequence. The measure CC is the error in the determination of the correct boundary between coding and noncoding region, so the value $(1-CC)$ is a reasonable measure of the accuracy of the border detected between coding and noncoding region [18].

We apply the segmentation using the Jensen-Shannon and Jensen-Rényi divergences with the nucleotide and protein alphabets on a dataset of DNA sequences and the results are presented in Table 6. The dataset consist of 1046 DNA sequences (orientation 5' to 3') randomly chosen from bacteria *Chlamydophila pneumoniae* (GenBank Acc. BA000008), *Borrelia burgdorferi* (GenBank Acc. AE000783), *Rickettsia prowazekii* (GenBank Acc. AJ235269) and every sequence contains only one coding region (at least 100 bp in length) and one noncoding region (at least 100 bp in length). The Jensen-Rényi divergence for $\alpha \rightarrow 0$ emphasizes more the DNA regions that are noncoding and we find that $\alpha \in [0.7, 0.8]$ represents a good choice for DNA segmentation.

Table 6 – Estimated accuracies of the cuts obtained using different segmentation methods

Segmentation method	100(1-CC) [%]
Jensen-Shannon divergence using nucleotide alphabet \mathcal{A}_{12}	83.95
Jensen-Rényi divergence using nucleotide alphabet \mathcal{A}_{12} , ($\alpha=0.7$)	81.55
Jensen-Shannon divergence using protein alphabet \mathcal{A}_{12}^*	88.82
Jensen-Rényi divergence using protein alphabet \mathcal{A}_{12}^* , ($\alpha=0.7$)	92.87

The entropic segmentation method that is presented in this study can be used also without any modification in detecting the borders between the coding and the noncoding regions in the eukaryotes, but further tests are needed. The Jensen-Rényi divergence together with the new protein alphabet achieves the best accuracy but the Jensen-Rényi divergence fails to bring any improvement for the nucleotide alphabet. Thus, the Jensen-Rényi divergence takes advantage of the newly introduced protein alphabet. The accuracy of detecting the borders between coding and noncoding regions is higher when the protein alphabet is used instead of the nucleotide alphabet. Thus the introduction of the biological knowledge in the segmentation method brings improvements in the accuracy of detecting the borders between coding and noncoding DNA regions.

6. CONCLUSIONS

In this study, we introduce a new segmentation method for finding the borders between coding and noncoding DNA regions, without *a priori* training, based on usage of Jensen-Rényi divergence using a new alphabet based upon the shape of the proteins that are coded in the DNA. To our knowledge, we propose for the first time the use of the Jensen-Rényi divergence for DNA segmentation. Our approach uses only statistical general properties of coding DNA and proteins. In this way, the prior training on data sets is avoided. The segmentation method presented here can also used for improving the accuracy of the gene finding methods through providing additional information. Also, we have applied the segmentation method to a dataset of

DNA sequences from bacteria *Chlamydomophila pneumoniae*, *Borrelia burgdorferi*, *Rickettsia prowazekii*. For this dataset of DNA sequences, it is shown that the accuracy of the border detection using the Jensen-Rényi and the new protein alphabet is improved compared to the standard segmentation method that uses Jensen-Shannon divergence with nucleotide alphabet, previously reported. The success seems to come from the utilization of the Jensen-Rényi divergence together with the alphabet based on protein shape characteristics. There is a constant need to study new and better algorithms for finding DNA coding regions in bacteria and humans.

7. REFERENCES

- [1] J.W. Fickett, "Recognition of protein coding regions in DNA sequences", *Nucleic Acids Research*, vol. 10, 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, R. Guigo, "Evaluation of gene structure prediction programs", *Genomics*, Vol. 34, pp. 353-357, 1996.
- [4] D. Nicorici, J. Astola, I. Tabus, "Computational identification of exons in DNA with a Hidden Markov Model", *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, IEEE, CP2-06, Oct. 2002.
- [5] W. Li, P. Bernaola-Galvan, F. Haghghi, I. Grosse, "Applications of recursive segmentation to the analysis of DNA sequences", *Computers and Chemistry*, Elsevier, Vol. 26, pp. 491-510, 2002.
- [6] R.K. Azad, J.S. Rao, W. Li, R. Ramaswamy, "Simplifying the mosaic description of DNA sequences", *Physical Review E*, vol. 66(031913), pp. 1-6, 2002.
- [7] W. Li, "New stopping criteria for segmenting DNA sequences", *Physical Review E*, Vol. 86(25), pp. 5815-5818, June 2001.
- [8] P.D. Cristea, "Large Scale features in DNA genomic signals", *Signal Processing*, Elsevier, vol. 83, pp. 871-888, 2003.
- [9] I. Grosse, H. Herzel, S.V. Buldyrev, H.E. Stanley, "Species independence of mutual information in coding and noncoding DNA", *Physical Review E*, vol. 61, pp. 5624-5629, 2000.
- [10] W. Li, G. Stolovitzky, P. Bernaola-Galvan, J.L. Oliver, "Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes", *Genome Research*, vol. 8, pp. 916-928, 1998.
- [11] A.A. Tsonis, J.B. Elsner, & P.A. Tsonis, "Periodicity in DNA coding sequences: Implications in gene evolution", *Journal of Theoretical Biology*, vol. 151, pp. 323-331, 1991.
- [12] S. Tiwari, S. Ramachandran, S. Bhattacharya, A. Bhattacharya, R. Ramaswamy, "Prediction of probable genes by Fourier analysis of Genomic sequences", *CABIOS*, vol. 13, pp. 263-270, 1997.
- [13] R. Farber, A. Lapedes, K. Sirotkin, "Determination of eukaryotic protein coding regions using neural networks and information theory", *J. Mol. Biol.*, vol. 226, pp. 471-479, 1992.
- [14] R. Staden, "Computer methods to locate signals in nucleic acid sequences", *Nucleic Acids Research*, vol. 12, pp. 505-519, 1984.
- [15] J.W. Fickett, "Finding genes by computer: the state of the art", *Trends in Genetics*, vol. 12, pp. 316-320, 1996.

- [16] J.W. Fickett, "The gene identification problem: an overview for developers", *Computer & Chemistry*, vol. 20, pp. 103-118, 1996.
- [17] M. Burset, R. Guigo, "Evaluation of gene structure prediction programs", *Genomics*, Vol. 34, pp. 353-357, 1996.
- [18] P. Bernaola-Galvan, I. Grosse, P. Carpena, J.L. Oliver, R. Roman-Roldan, H.E. Stanley, "Finding borders between coding and noncoding DNA regions by an entropic segmentation method", *Physical Review E*, vol. 85(6), pp. 1342-1345, 2000.
- [19] P. Bernaola-Galvan, R. Roman-Roldan, J.L. Oliver, "Compositional segmentation and long-range fractal correlations in DNA sequences", *Physical Review E*, vol. 53(5), pp. 5181-5189, 1996.
- [20] D. Nicorici, J.A. Berger, J. Astola, S.K. Mitra, "Finding borders between coding and noncoding DNA regions using recursive segmentation and statistics of stop codons", *FINSIG 2003*, (to appear), April 2003.
- [21] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman Roldan, J.L. Oliver, H.E. Stanley, "Analysis of symbolic sequences using Jensen-Shannon divergences", *Physical Review E*, vol. 65(041905), pp. 1-16, 2002.
- [22] Yun He, A. Ben Hamza, Hamid Krim, "A Generalized Divergence Measure for Robust Image Registration", *IEEE Transactions on Signal Processing*, vol. 51, no. 5, (to appear), May 2003.
- [23] A. Rényi, "On Measures of Entropy and Information", *Selected papers of Alfred Rényi*, vol. 2, pp. 525-580, 1976.
- [24] M.K. Campbell, "Biochemistry – 2nd Edition", Saunders College Publish, 1995.
- [25] K.C. Timberlake, "Chemistry – 5th Edition", Haper-Collins Publishers Inc, NY, 1992.