Romanian Reports in Physics, Vol. 66, No. 4, P. 1078–1087, 2014

APPLICATIONS OF ENTROPIC DIVERGENCE MEASURES FOR DNA SEGMENTATION INTO HIGH VARIABLE REGIONS OF CRYPTOSPORIDIUM SPP. GP60 GENE

IONELA MIRELA NEAGOE[1,2], D. POPESCU[3,4,a], V.I.R. NICULESCU[5]

[1] "Carol Davila" University of Medicine and Pharmacy, Department of Parasitology, D. Gerota 19-21, Bucharest, Romania,
[2]National Institute of Research and Development for Microbiology and Immunology "Cantacuzino", Splaiul Independentei 103, Bucharest, Romania, E-mail: ineagoe_labparasit@yahoo.com
[3] Institute of Mathematical Statistics and Applied Mathematics of Romanian Academy, Department of Mathematical Modelling in Life Sciences, Calea 13 Septembrie 13, Bucharest, Romania,
[4] University of Bucharest, Faculty of Biology, Department of Physiology and Biophysics, Spl. Independentei 91-95, Bucharest, Romania, [a]E-mail: popescu1947@yahoo.com
[5] National Institute for Nuclear Physics and Engineering, P.O.Box MG-6, RO-077125 Bucharest-Magurele, Romania, Email: filo_niculescu@yahoo.com

Received August 7, 2013

Abstract. In this paper we have applied several entropic segmentation methods for obtaining preliminary borders between different regions of the coding gene gp60 subtypes of Cryptosporidium spp parasite. We have calculated Jensen-Shannon and Jensen-Renyi divergences measures for experimental and data bases *Cryptosporidium*. The statistical comparisons obtained by segmenting the DNA borders have shown that divergence Jensen-Shannon and Jensen-Renyi are accurate enough to indicate that regions with high variability: the microsatellite region and the hypervariable region, respectively.

Key words: DNA sequences, *Cryptosporidium* variability, gp60 gene, Jensen-Shannon and Jensen-Renyi divergences measures.

1. INTRODUCTION

With the advance of sequencing technology and the availability of great databases, molecular biology investigations require increasingly more to use new techniques that are similar to those developed in statistical physics in order to analyze, identify genes or coding gene regions [1] and also to characterize by comparing various gene profiles [2].

The main characteristic of DNA sequences is that their statistical properties are not homogeneously distributed along the sequence [3]. The presence of complex heterogeneities in nucleotide sequences results in mosaic patterns in DNA [4]. The heterogeneity with fluctuating density in the DNA sequences can be

observed mainly in poor or rich G+C regions, A-T strand asymmetry, the origin or terminus replication, regions of periodicity at 3 nucleotides [3, 5]. Periodical sequences are found also in the polymorphic regions as simple sequence repeats or microsatellites and minisatellites sequences [6].

The information entropy measures are very useful for identifying homogenous regions and evaluating the gene complexity [1]. Computational approach as segmentation analysis is a powerful way of examining the large-scale organization of DNA sequence [4]. The most commonly used procedure is based on maximization of the Jensen-Shannon divergence through which a given heterogeneous DNA sequence is recursively separated into compositionally homogeneous subsequences called domains (or patches). This segmentation method was proposed mainly to find the borders between coding and non-coding homogenous DNA regions but also to detect the heterogenic domains (periodicity of three-pattern, G+C regions) [3, 5, 7]. A more accurate segmentation method was introduced by Nicorici and it was based on measuring of Jensen-Renyi divergence in both DNA strands [1].

In this study we analyzed the recursive entropic segmentation for DNA sequences of several coding gp60 gene variants [12 unpublished data, 13, 15] from the genome of two different species of a protozoan parasite *Cryptosporidium hominis* and *Cyptosporidium parvum*. In order to perform a comparison between known coding variable regions and predicted borders we used for partition of DNA sequences the segmentation methods proposed by Bernaola-Galvan *et al.* [7], W. Li [3, 5] and Nicorici [1].

2. STATISTICAL PARAMETERS USED

We were defined two probability distributions:

$P = \{p_1, p_2 .....p_N\}$ and $Q = \{q_1, q_2 .....q_N\}$

$\sum_1^N p_i = 1$, $\sum_1^N q_i = 1$, $0 \le p_i, q_i \le 1$, $i = 1, 2, ...N$.

1. The informational entropy of source S1 (entropy of distribution P) is defined as [8]:

$$H[Y] = -\sum_{i=1}^{N} p_i \log_2 p_i . \quad (1)$$

2. The Renyi entropy is a mathematical generalization of Shannon entropy, and is defined [10]:

$$R_\alpha(Y) = \frac{1}{1-\alpha} log_2 \sum_{i=1}^{N} p_i^\alpha , \quad (2)$$

where $\alpha \in (0,1)$. The Renyi entropy $R_\alpha(Y)$ tends to Shannon entropy $S(Y)$ as $\alpha \to 1$ and $R_\alpha(Y) \geq S(Y)$. It attains a maximum uncertainty when $\alpha = 0$, in contrast to Shannon entropy which attains a minimum uncertainty when α decreases [11].

3. The Jensen-Shannon divergence is defined as [9]:

$$D_{JS}\left(p^{(1)}, p^{(2)},..., p^{(M)}\right) \equiv H\left(\sum_{j=1}^{M} \pi^{(j)} p^{(j)}\right) - \sum_{j=1}^{M} \pi^{(j)} H\left(p^{(j)}\right), \quad (3)$$

where $p^{(j)} = (p_1^{(j)}, p_2^{(j)}, \ldots, p_N^{(j)})$ are probability distributions satisfying the constraints $\sum_{i=1}^{N} p_i^{(j)} = 1$ and $0 \leq p^{(j)} \leq 1$ for $i = 1, 2\ldots, N$ and $j = 1,2,\ldots,M$.

The Jensen-Shannon divergence is a measure that can quantify the distance between two or M probability distributions. The compared distributions can be weighted by assigning some weights vectors $\pi^{(1)}, \pi^{(2)},\ldots, \pi^{(M)}$ which satisfy the usual constraints $\sum_{j=1}^{M} \pi^{(j)} = 1$ and $0 \leq \pi^{(j)} \leq 1$. In this way, it can take into account the subsequences with different length from which the probability distributions are computed [9].

4. The Jensen-Renyi divergence is defined as [11]:

$$D_{JR}\left(\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(M)}\right) \equiv R_\alpha\left(\sum_{j=1}^{M} \pi^{(j)} \mathbf{p}^{(j)}\right) - \sum_{j=1}^{M} \pi^{(j)} R_\alpha\left(\mathbf{p}^{(j)}\right) \quad (4)$$

where $p^{(j)} = (p_1^{(j)}, p_1^{(j)}, \ldots, p_N^{(j)})$ are probability distributions satisfying the constraints $\sum_{i=1}^{N} p_i^{(j)} = 1$ and $0 \leq p^{(j)} \leq 1$ for $i = 1, 2\ldots, N$ and $j = 1,2,\ldots,M$.

The Jensen-Renyi divergence is used as measure for determining the similarity between two or M probability distributions with some assigned weights $\pi^{(1)}, \pi^{(2)},\ldots, \pi^{(M)}$ which satisfy the usual constraints $\sum_{j=1}^{M} \pi^{(j)} = 1$ and $0 \leq \pi^{(j)} \leq 1$.

3. THE DATA SET

All allelic sets of the coding gp60 gene (which encodes glycoprotein of 60kDa involved in host enterocytes stage invasion by the parasite) investigated in this study containing two highly variable regions: microsatellite region (which consists in a variable number of nucleotide triplets repetitions) and hypervariable region that print allelic subtype [13, 15]. In addition, it has been postulated that the high polymorphism of gp60 gene might be caused to selective pressure exerted by the immune system of the host [14].

All *Cryptosporidium* genetic subtypes tested as models of probability distributions in this study were downloaded from http://www.ncbi.nlm.nih.gov:

1. Reference 1 DNA sequence (GenBank: EU052234) with 903bp belongs to *Cryptosporidium hominis,* and its complete subgenotype name is IaA13R7 [16].

2. Reference 2 DNA sequence (GenBank: HQ005735) with 803bp belongs to *Cryptosporidium parvum,* and complete subgenotype name is IIaA17G1R1 [17].

3. Sample DNA sequence (was deposited in GenBank and the accession number will be communicated us after a month) with 711bp; it was isolated from a

HIV positive patient infected with *Cryptosporidium hominis* [12], belongs to allele family *Ib* and its complete subgenotype name is IbA10G2.

4. DETECTION BORDERS BETWEEN DIFFERENT CODING VARIABLE DNA REGIONS

DNA partition was achieved by applying a segmentation algorithm based on the movement of a sliding pointer along the sequence at the border of each partition the pointer cut in two subsequences. The Jensen-Shannon divergence and Jensen-Renyi divergence measures [1, 3, 5, 7] were computed for each cutting point i (where i = 3,...N–2 and N = the sum of nucleotides {A, C, G, T} in the selected sequence) between two neighboring polymorphic regions.

The Jensen-Shannon divergence type for DNA segmentation is defined as follows [1, 5, 7]:

$$D_{JS} = max_i \left[H - \frac{i}{N} H_L - \frac{N-i}{N} H_R\right], \quad (5)$$

where H, H_L ,H_R are Shannon entropy of the whole, left and right subsequence.

The Jensen-Renyi divergence type for DNA segmentation introduced by Yun He [11] is:

$$D_{JR} = max_i \left[R_\alpha - \frac{i}{N} R_{\alpha L} - \frac{N-i}{N} R_{\alpha R}\right], \quad (6)$$

where: R_α, $R_{\alpha L}$, $R_{\alpha R}$ are Renyi entropy of the whole, left and right subsequence, and $\alpha \in (0,1)$.

5. RESULTS AND DISCUSSIONS

1. The subsequences obtained by partition of sample DNA sequence (length 711bp) with allele family Ib (subgenotype IbA10G2), are indicated in Fig. 1 and Table 1. The cuts for i = 28 are the distance from left border = 28bp and distance from right border = 683bp; for i = 64, the distance from left border = 64bp and distance from right border = 647bp.

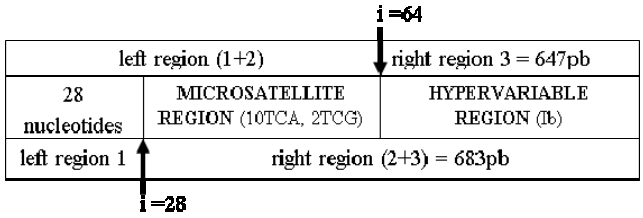

Fig. 1 – Cuts obtained for two different partition points in the structure of subgenotype IbA10G2.

Table 1

The number of each nucleotide type in whole sequence of subgenotype IbA10G2, left and right subsequence for different *i*

IbA10G2 subgenotype	A	C	G	T
Whole DNA sequence	237	154	168	152
Left region 1, i=28	7	5	7	9
Right region (2+3), i=28	230	152	161	143
Left region (1+2), i=64	17	17	9	21
Right region 3, i=64	220	140	159	131

(i) Jensen-Shannon divergence results of subgenotype IbA10G2 for two different cutting points and $S_{[IbA10G2]}$= 1.97714 were calculated below:

For i = 28 (S_{L1} = 1.97014; S_{R1} = 1.97498; $D_{JS1[IbA10G2]}$= 0.00235), where 1 = (left region 1, right region 2+3).

For i = 64 (S_{L2} = 1.94153; S_{R2} = 1.97114; $D_{JS2[IbA10G2]}$= 0.00866), where 2 = (left region 1+2, right region 3).

Table 2

Variation of Jensen-Renyi divergences of subgenotype IbA10G2 for different *i* and Renyi entropy values in relation with α Є (0-1)

α	$R_{\alpha[IbA10G2]}$	$D_{JR1\ (i=28)}$	$D_{JR2\ (i=64)}$
0.01	**- 0.08153**	**0.00009**	**0.0012**
0.05	**- 0.42480**	**0.0005**	**0.0020**
0.1	**- 0.89681**	**0.0010**	**0.0042**
0.2	**- 2.01780**	**0.0023**	**0.0094**
0.3	**- 3.45909**	**0.0040**	**0.0161**
0.4	**- 5.38081**	**0.0062**	**0.0250**
0.5	**- 8.07121**	**0.0093**	**0.0376**
0.6	**- 12.10680**	**0.0140**	**0.0563**
0.7	**- 18.83283**	**0.0217**	**0.0877**
0.8	**- 32.64093**	**0.0371**	**0.1503**
0.9	**- 72.64093**	**0.0836**	**0.3381**
0.95	**- 153.35307**	***0.1764***	**0.7140**

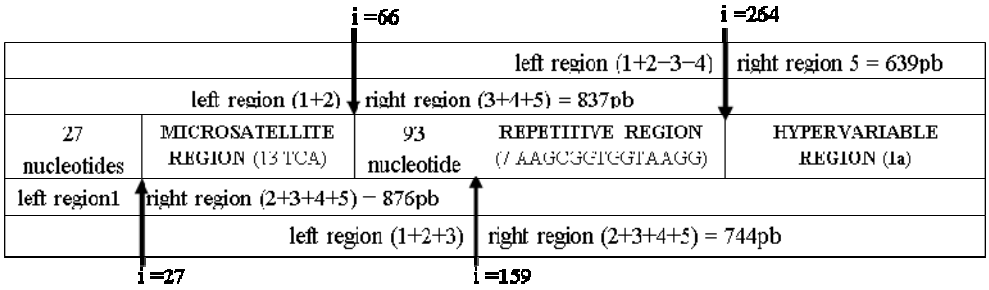

Fig. 2 – Borders obtained for four different partition points in the structure of subgenotype IaA13R7.

(ii) Jensen-Renyi divergence results (Table 2) were computed for two different cutting points of subgenotype IbA10G2 and different α values.

Table 3

The numbers of each nucleotide type in whole sequence of subgenotype IaA13R7, left and right subsequences for different *i*

IaA13R7 subgenotype	A	C	G	T
Whole DNA sequence	300	168	240	195
Left region 1, where i=27	6	3	9	9
Right region (2+3+4+5), i=27	294	165	231	186
Left region (1+2), where i=66	19	16	9	22
Right region (3+4+5), i=66	281	152	231	173
Left region (1+2+3), i=159	56	36	36	31
Right region (4+5), i=159	244	232	204	164
Left region (1+2+3+4), i=264	91	43	85	45
Right region 5, where i=264	209	125	155	150

(i) Jensen-Shannon divergence results of subgenotype IaA13R7 for different cutting points and $S_{[IaA13R7]}$= 1.96516 were calculated below:

For i = 27 (S_{L1} = 1.89106; S_{R1} = 1.96407; D_{JS1}= 0.00327), where 1 = (left region 1, right region 2+3+4+5).

For i = 66 (S_{L2} = 1.93307; S_{R2} = 1.95830; D_{JS2}= 0.00870), where 2 = (left region 1+2, right region 3+4+5)

For i = 159 (S_{L3} = 1.96051; S_{R3} = 1.96285; D_{JS3}= 0.00272), where 3 = (left region 1+2+3, right region 4+5).

For i = 264 (S_{L4} = 1.91761; S_{R4} = 1.97432; D_{JS4}= 0.00742), where 4 = (left region 1+2+3+4, right region 5).

Table 4

Variation of Jensen-Renyi divergences subgenotype IaA13R7 for four different *i* and Renyi entropy values in relation with α Є (0-1)

α	$R_{[IaA13R7]}$	$D_{JR1\ (i=27)}$	$D_{JR2\ (i=66)}$	$D_{JR3\ (i=159)}$	$D_{JR4\ (i=264)}$
0.01	**- 0.08221**	**0.0001**	**0.0004**	**0.0001**	**0.0003**
0.05	**- 0.42837**	**0.0007**	**0.0020**	**0.0006**	**0.0016**
0.1	**- 0.90433**	**0.0016**	**0.0043**	**0.0014**	**0.0036**
0.2	**- 2.03475**	**0.0036**	**0.0096**	**0.0030**	**0.0081**
0.3	**- 3.48813**	**0.0061**	**0.0164**	**0.0052**	**0.0140**
0.4	**- 5.42598**	**0.0095**	**0.0256**	**0.0082**	**0.0215**
0.5	**- 8.13897**	**0.0143**	**0.0383**	**0.0123**	**0.0323**
0.6	**- 12.20846**	**0.0214**	**0.0575**	**0.0184**	**0.0484**
0.7	**- 18.99093**	**0.0333**	**0.0894**	**0.0287**	**0.0753**
0.8	**- 32.55589**	**0.0571**	**0.1533**	**0.0492**	**0.1291**
0.9	**- 73.25075**	**0.1284**	**0.3450**	**0.1108**	**0.2906**
0.95	**- 154.64046**	**0.2710**	**0.7283**	**0.2339**	**0.6134**

(ii) Jensen-Renyi divergence results (Table 4) (obtained by joining DNA regions: 1 and 2 +3+4+5 / 1+2 and 3+4+5 / 1+2+3 and 4+5 / 1+2+3+4 and 5) were computed to four different cutting points of subgenotype IaA13R7 and different α values.

3. The subsequences obtained by partition of Reference 2 DNA sequence (length 803bp) with allele family IIa (IIaA17G1R1 subgenotype), are shown in Fig. 3 and Table 5. The cuts obtained for i = 15 are distance from left border = 15bp and distance from right border = 788bp; for i = 69 (distance from left border = 69bp and distance from right border = 735bp); for i = 75 (distance from left border = 75bp and distance from right border = 728bp).

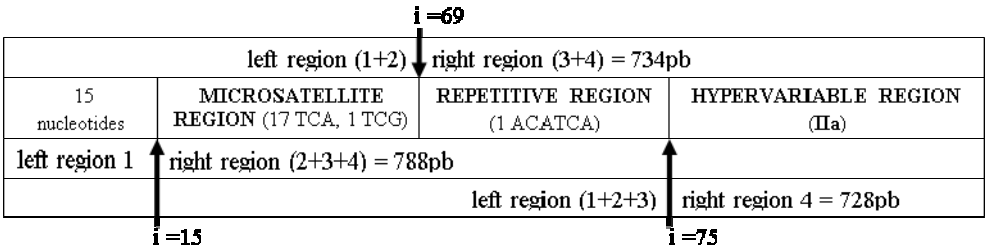

Fig. 3 – Borders obtained for three different partition points in the structure of subgenotype IIaA17G1R1

(i) Jensen-Shannon divergence results of subgenotype IIaA17G1R1 for different cutting points and $S_{[IaA13R7]}$= 1.98429 were calculated below:

For i = 15 (S_{L1} = 1.78194; S_{R1} = 1.98203; D_{JS1}= 0.00600), where 1 = (left region 1, right region 2+3+4).

For i = 69 (S_{L2} = 1.89127; S_{R2} = 1.98141; D_{JS2}= 0.01063), where 2 = (left region 1+2, right region 3+4).

For i = 75 (S_{L3} = 1.88254; S_{R3} = 1.98194; D_{JS3}= 0.01164), where 3 = (left region 1+2+3, right region 4).

Table 5

The numbers of each nucleotide type in whole sequence of subgenotype IIaA17G1R1, left and right subsequences for different i

IIaA17G1R1 subgenotype	A	C	G	T
Whole DNA sequence	253	180	181	189
Left region 1, where i=15	1	3	6	5
Right region (2+3+4), i=15	252	177	175	184
Left region (1+2), where i=69	18	21	7	23
Right region (3+4), i=69	285	159	174	166
Left region (1+2+3), i=75	21	23	7	24
Right region (4+5), i=75	232	157	174	165

Table 6

Variation of Jensen-Renyi divergences subgenotype IIaA17G1R1 for three different *i* and Renyi entropy values in relation with α Є (0-1)

α	$R_{[IIaA17G1R1]}$	$D_{JR1\ (i=15)}$	$D_{JR2\ (i=69)}$	$D_{JR3\ (i=75)}$
0.01	**- 0.08142**	**0.0003**	**0.00052**	**0.0021**
0.05	**- 0.42422**	**0.0015**	**0.00271**	**0.0110**
0.1	**- 0.89557**	**0.0031**	**0.00572**	**0.0231**
0.2	**- 2.01502**	**0.0071**	**0.01286**	**0.0521**
0.3	**- 3.45432**	**0.0122**	**0.02205**	**0.0893**
0.4	**- 5.37339**	**0.0190**	**0.03429**	**0.1390**
0.5	**- 8.06009**	**0.0284**	**0.05144**	**0.2084**
0.6	**- 12.09013**	**0.0426**	**0.07716**	**0.3127**
0.7	**- 18.80688**	**0.0662**	**0.12003**	**0.4864**
0.8	**- 32.940359**	**0.1136**	**0.20576**	**0.8340**
0.9	**- 72.54081**	**0.2556**	**0.46295**	**1.8761**
0.95	**- 153.14170**	**0.5340**	**0.97735**	**3.9607**

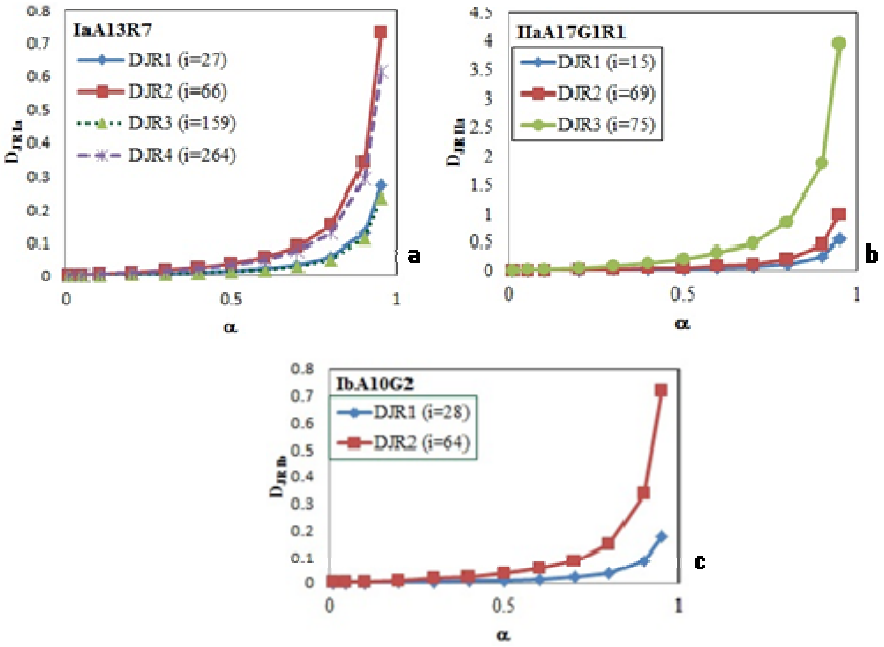

Fig. 4 – Jensen-Renye divergences for different cutting positions for a DNA sequence *versus* α values: **a)** the maximum values for divergence in the Reference1 DNA sequence of *C. hominis* is detected to $D_{JR2\ (i=66)}$, respectively between DNA subsequence containing microsatellite region (1+2) and subsequence that includes the repetitive R and hypervariable regions the (3+4+5); **b)** the maximum values for divergence in the Reference 2 of *C. parvum* is detected to $D_{JR3\ (i=69)}$, respectively between DNA subsequence that includes microsatellite region and repetitive region R (1+2+3) and subsequence that containing only the hypervariable regions the (4); **c)** the maximum values for divergence in the Sample DNA sequence of *C. hominis* is detected to $D_{JR2\ (i=64)}$, respectively between DNA subsequence that includes microsatellite region (1+2) and subsequence that containing only the hypervariable regions the (3).

(ii) Jensen-Renyi divergence results (Table 5) (obtained by joining DNA regions: 1 and 2+3+4 / 1+2 and 3+4 / 1+2+3 and 4 / 1+2+3 and 4) were computed to three different cutting points of subgenotype IIaA17G1R1 and different α values.

For visualizing the predicted borders with the highest value of divergence obtained by applying DNA segmentation method, we plotted Jensen-Renyi divergences corresponding to different *i* in relation to α adjusted from 0 to 1. As you may see in Fig. 4, Jensen-Renyi the divergences values calculated for all subsequences of each analyzed subgenotypes increases linearly with values of α. On the other side, the segmentation method based on Jensen-Renyi divergence confirms results obtained by Jensen-Shannon divergence for all three subtypes analyzed.

Using DNA segmentation methods based on Jensen-Shannon and Jensen-Renyi divergences proved efficient and accurate to underline the borders of the highest different and variable regions of the gp60 gene: microsatellite region and hypervariable region, respectively. Detection of borders predicted by these methods confirmed the variability of known polymorphic regions

6. CONCLUSIONS

This preliminary report present two segmentation methods based on JS and JR divergences for characterization and quantification of known variable regions from DNA sequences of three gp60 gene subgenotypes belong to two different species of *Cryptosporidium*. For all three *Cryptosporidium* subgenotypes analyzed, the DNA segmentation methods raised the accuracy of finding the borders between the highest known variable regions of *Cryptosporidium* gp60 gene. Although these algorithms are widely used especially in finding coding regions of DNA sequences, they could be used also to determine regions with high genetic variability in the coding gene gp60 from the genome of *Cryptosporidium* species investigated. On the other hand these DNA segmentation methods were an effective alternative to highlight the compositional heterogeneity of DNA sequence analyzed especially in the microsatellite region of gp60 gene.

REFERENCES

1. D. Nicorici, J. Astola, *Divergence measures for DNA segmentation*, ftp://ftp.dec.com/pub/toomany/paper31.pdf.
2. G. Benson, Bioinformatics, **18**, Suppl.2, S44–S53 (2002).
3. W.Li, Phys. Rev. E., **86**, *25*, 5815–5818 (2001).
4. R.K. Azad et al., Phys. Rev. E., **66**, 031913 (2002)
5. W. Li et al., Computers and Chemistry, **26**, *5*, 491–510 (2002).

6. E.V. Korotkov, D.A. Phoenix, Pac. Symp. Biocomput., 222–31 (1997).
7. P. Bernaola-Galvan et al. Phys. Rev., **85**, *6*, 1342–1345 (2000).
8. C. Shannon. W. Weaver, *Mathematical Theory of Communication*, Illinois University, Illinois Press, Urbana, 1945.
9. I. Grosse et al, Phys. Rev., **65** (041905), 1–16 (2002).
10. A. Renyi, *On Measures of Entropy and Information*, Selected papers of Alfred Renyi, **2**, 525–580, (1976).
11. Yun He et al, IEEE Transactions on Signal Processing, **51**, *5* (2003).
12. I.M. Neagoe, D. Popescu, N. Niculescu, *Alternative methods for statistical characterization and quantification of Cryptosporidium spp. gp60 gene variability*, Rom. Rep. in Phys., 66, 683–692, 2014.
13. R.M. Chalmers, C. Jackson, K. Elwin, S. Hadfield, P. Hunter, *DWI0851*: *Investigation of Genetic Variation within Cryptosporidium hominis for Epidemiological Purposes,* (http://dwi.defra.gov.uk/research/completed-research/2000todate.htm), 2007.
14. L.D. Sybley et al., *Evolution of Virulence in Eukaryotic Microbes*, Wiley-Blackwell, 2012.
15. L. Xiao, Y. Feng , FEMS Immunol Med Microbiol., **52**, *3*, 309–23 (2008).
16. G.D. Sturbaum, D.A. Schaefer, B.H. Jost, C.R. Sterling, M.W. Riggs, Molecular & Biochemical Parasitology, **159**, *2*, 138–141 (2008).
17. RM Chalmers , RP Smith , SJ Hadfield , K Elwin , M Giles, Parasitol. Res., **108**, *5*, 1321–5 (2011).