Romanian Reports in Physics, Vol. 66, No. 3, P. 683–692, 2014

ALTERNATIVE METHODS FOR STATISTICAL CHARACTERIZATION AND QUANTIFICATION OF CRYPTOSPORIDIUM SPP. GP60 GENE VARIABILITY

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.Gerota19-21, 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, 13, Calea 13 Septembrie, Bucharest, Romania
[4] University of Bucharest, Department of Physiology and Biophysics, Faculty of Biology, 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 used statistical methods for to study DNA sequences different species of *Cryptosporidium:* two reference DNA sequences from public GenBank database and one sample DNA sequence isolated and analyzed by sequencing in our laboratory, which was proposed to be deposited in GenBank. Using statistical concepts as Shannon entropy, Onicescu informational energy, different statistical distances types we ordered the three DNA sequences according to their structure variability. We have also introduced two indices: the index associated to structural information quantities Shannon entropy and Onicescu energy and the altered index associated to the structural discrepancy quantity, respectively. We have approach firstly these statistical methods for *Cryptosporidium*.

Key words: DNA sequences, *Cryptosporidium* variability, gp60 gene, statistical methods.

1. INTRODUCTION

DNA sequences can be analyzed in different ways; ones of them are statistical analysis and algebraic method [1, 2]. Among the statistical analyzes of DNA sequences, those that use measures of entropy and divergence for finite nucleotide alphabet have become increasingly used in statistical genetics studies. DNA alphabet is represented by four types of nucleotides that differ in their nitrogenous bases adenine (A), cytosine (C), guanine (G) and thymine (T). These nitrogenous bases (purines A, G, and pyrimidines C, T) are the interconnection

substrate between the two strands of the DNA double helix. In other words, base pairs between the two DNA chains are carried out only on the basis of complementarity existing between a purine on one chain and pyrimidine on the other chain or between A and T or G and C. One of the major goals of DNA sequence analysis is the understanding of the overall organization of DNA in regions as genes, promoters, repetitions, etc, and its characteristics. The genetic information of DNA can vary in a population and it already established that genetic variability is attributed to the tendency of a whole set of genes or genotypes to become different from one individual to another. The set of alleles or variants of a gene are closely linked to the expression of particular characteristics or phenotype.

Variability of DNA sequence or gene polymorphism among individuals or populations can be assessed and quantified by different mathematical functions [3, 4] able to describe the analyzed information content.

In the present work, we analyze the genetic variability of coding gp60 gene from the genome of the two dominant species of the protozoan parasite *Cryptosporidium* involved in the etiology of diarrheal syndrome with potentially opportunistic character in patients with human immunodeficiency virus (HIV). In order to quantify this variability we chose to apply some methods borrowed from physics, based on entropy and divergence. Our motivation was to understand the complexity and difference of structure between the different genetic subtypes of the two species of *Cryptosporidium* (*C. hominis* and *C. parvum*) in statistical terms. All *Cryptosporidium* DNA sequences referred to in this study are available on http://www.ncbi.nlm.nih.gov.

2. STATISTICAL PARAMETERS USED

The DNA sequences may be considered as a statistical system. Over many years ago, there are efforts for statistical characterization of the deterministic systems of apparent randomness. The system entropy was first successfully parameter. Other statistical parameters were introduced in order to describe the randomness degree, the structure and correlations or statistical complexity of a system, or the difference between two distributions. Here we define some of the statistical parameters used in this work.

Firstly, we define two probability distributions:

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

$$\sum_{1}^{N} p_i=1,\qquad \sum_{1}^{N} q_i=1,\qquad 0\le p_i,q_i\le 1,\ i=1,2,.....N\ .$$

These distributions can describe two sources of information, S1 and S2, each with N elements. In this case, we define a discrete random variable Y, where p_k and

q_k is the probability that the variable Y have the value of k element from source S1 and source S2, respectively.

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

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

The informational entropy was introduced by Shannon and is a measure of the information amount contained in the source, or the degree of randomness that characterizes the state of a physical system (or stochastic process).

2. The *square deviation from uniformity D* defined as [9]:

$$D[Y] = \sum_{i=1}^{N} \left(p_i - \frac{1}{N} \right)^2 . \tag{2}$$

The quantity, D measures the discrepancy for distribution P from uniform distribution, characterized by maximal randomness which occurs for entropy $H = \log_2 N$ and $p_i = 1/N$.

3. The *statistical complexity measure* [8]:

$$C_{LMC}[Y] = H[Y] \times D[Y]. \tag{3}$$

It was proposed this form for $C_{LMC}[Y]$, because it must be vanish for distributions which describe perfect order (the Shannon entropy vanishes, $H[Y] = 0$) and maximal randomness ($D[Y] = 0$ for uniform distribution, because $p_i = 1/N$).

4. The informational energy. The discrete informational energy of the random variable Y was defined by Onicescu [10, 11]:

$$E[Y] = \sum_{i=1}^{N} p_i^2 . \tag{4}$$

5. Statistical information indices for classification of probability distributions characteristics are:

– the Index associated to structural information quantities H and E, which was extended from [12]:

$$I[Y] = \frac{E[Y]}{H[Y]}; \tag{5}$$

– the altered index associated to the structural discrepancy quantity D:

$$T[Y] = \frac{E[Y]}{H[Y]D[Y]}. \tag{6}$$

6. *Bhattacharyya distance* is defined as:

$$\rho(P,Q) = -\ln\left(\sum_{i=1}^{N}\sqrt{p_i q_i}\right) \tag{7}$$

and measures the similarity between two probability distributions [5].

It is evidently that, $\rho(P,Q)=1$, for identical distributions and $0 \le \rho(P,Q) \le \infty$

7. *Kullback-Leibler distance* is definite as:

$$K(P,Q) = \sum_{i=1}^{N} p_i \log_2 \frac{p_i}{q_i} \tag{8}$$

and it represents a measure for how much the two distributions differ one to another [6]. It is also known as asymmetric divergence. It has units of bits, as Shannon entropy. The probability distribution P has a dominant role in the $K(P,Q)$ value. From this reason, the Kullback-Leibler distance is asymmetric.

8. In order to equalize the role of the two distributions, and to symmetrizing the Kullback-Leibler distance, D.H. Jonson and S. Sinanovic define a *new distance as the harmonic mean* of $K(P,Q)$ and $K(Q,P)$ [7]:

$$\frac{1}{R[P,Q]} = \frac{1}{K[P,Q]} + \frac{1}{K[Q,P]}. \tag{9}$$

This distance is also named the resistor-average distance from the formula of equivalent resistance for parallel connected resistors.

3. GENERAL FEATURES OF GP60 GENE AND MODELS OF PROBABILITY DISTRIBUTIONS

Gp60 gene is a variable marker which consists in microsatellite region represented by tandem repeats of serine coding trinucleotide TCA/TCG/TCT at 5' end of the gene and an extensive, hypervariable, non-repeated region located downstream [13, 14]. The last variable region provide information for determining allele family of each Cryptosporidium species (six different allele family Ia-Ig were definite for *C. hominis* and nine IIa-IIi for *C. parvum*) [14, 15]. Unlike other known allelic families, allele families Ia and IIa contain an additional different repetitive sequence (R) located between the serine amino acid coding microsatellite region and hypervariable region (Fig.1). The full subgenotype profile name is done first by the allele family, followed by the number of each repetitive triplet denoted by the last nucleotide component (*A*/*G*/*T*) and in the particular genetic subtypes the number of repetitions (*R*) are added [14, 15].

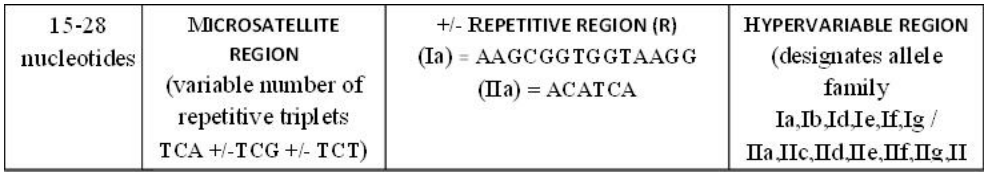

Fig. 1 – Distribution scheme of polymorphic regions (~ 900bp) in a gp60 subgenotype of *Cryptosporidium*.

On the other hand, variable gp60 gene is an important functional gene that codes for a 60kDa precursor protein involved in invasive stage of the parasite and neutralization of human immune response [14]. Thus, knowledge of parasitic subgenotype family identity might establish a possible link between some biological characteristic of the parasite and clinical presentation [15] or response to the treatment.

As models of probability distributions we used two reference DNA sequences retrieved from public database GenBank using the Entrez search engine, and a sample DNA sequence.

(*i*) Reference 1 DNA sequence (accession number EU052234) with 903bp (Table 1), belongs to *Cryptosporidium hominis,* and complete subgenotype name is IaA13R7 [16].

Table 1

The number and type of each nucleotide, number and type of repetitions for subgenotype IaA13R7

Nucleotide Type	No. of each Nucleotide type	Order of Repetitions in IaA13R7 (No. and Type)	
A	300	13 TCA	7 AAGCGGTGGTAAGG
C	168		
G	240		
T	195		

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

(*iii*) For the intra species comparisons we used a sample DNA sequence isolated from the stool of a HIV positive patient infected with *Cryptosporidium hominis*. Approximately, 850bp from Gp60 gene of isolated DNA was amplified by nested PCR [13], purified, and subsequently analyzed by sequencing. The sequence searches for comparing polymorphic regions of DNA sequence obtained, against all the sequences on GenBank database were performed using BLAST, (http://www.ncbi.nlm.nih.gov/BLAST/).

Table 2

The number and type of each nucleotide, number and type of repetitions for subgenotype IIaA17G1R1

Nucleotide Type	No. of each Nucleotide type	Order of Repetitions in IIaA17G1R1 (No. and Type)			
A	253	2 TCA	1 TCG	15 TCA	1 ACATCA
C	180				
G	181				
T	189				

Sample DNA sequence was deposited in GenBank and the accession number will be communicated us after a month; it belongs to allele family Ib and complete subgenotype name is IbA10G2 (Table 3) with 711bp.

Table 3

The number and type of each nucleotide, number and type of repetitions for subgenotype IbA10G2

Nucleotide Type	No. of each Nucleotide type	Order of Repetitions in IbA10G2 (No. and Type)				
A	237	7 TCA	1 TCG	2 TCA	1 TCG	1 TCA
C	154					
G	168					
T	152					

4. RESULTS AND DISCUSSIONS

4.1. INFORMATION AND DISCREPANCE QUANTITIES FOR SEQUENCE VARIABILITY COMPLEXITY BETWEEN SUBGENOTYPES

It is well know that the general indicators of structure and correlation for statistical complexity between some probability distributions are Shannon entropy H and square deviation from uniformity D. In this study, Shannon entropy was considered a strategy to characterize quantitatively the variability of DNA sequence between the selected subgenotypes. The substrate of total order and disorder structure is given by the fact that Shannon's entropy assumes a minim value (H_{min}=0) for different probabilities ($p_1 = 1$; $p_2 = ... p_N = 0$), and a maxim value (H= log$N$) for equal probabilities ($p_1 = p_2 = ... p_N = 1/N$), respectively [12].

In order to achieve a better characterization of structure variability between analyzed subgenotypes, we used the Onicescu informational energy E as a finer measure to quantify the nucleotide dispersion distributions than the Shannon entropy [18]. This function reaches minimum value ($E_{min} = 1/N$) for total disorder (equal probabilities, $p_1 = p_2 = ... p_N = 1/N$) and maximum value ($E_{max}$=1) for minimal disorder (total order, $p_1 = 1$; $p_2 = ... p_N = 0$) [12].

The discrete random variable *Y* runs over all allowed values of the information sources associated to the three subgenotypes mentioned above. Their elements are the four nucleotides, each of them containing one of nitrogenous bases adenine (A), cytosine (C), guanine (G) and thymine (T).

The corresponding probabilities were calculated taking into account the Tables 1–3 are used for the calculation of all statistical parameters defined in Ch. 2. The obtained results were written in Table 4.

We observed that the statistical parameters show differences in all three analyzed subgenotypes. The Shannon entropy *H* values vary inversely with the values of the other statistical parameters used (*D, C_{LMC}, E, I*, and *T*). In this way, informational energy *E* confirms Shannon entropy. Between subgenotypes belonging to two different *Cryptosporidium* species, IIaA13G2R1 of *C. parvum* has the highest Shannon entropy value. On the other hand, Shannon entropy shows a difference within the same species of *Cryptosporidium* (*C. hominis*); subgenotype IbA10G2 has a higher value than IaA13R7.

Table 4

Structural information and discrepancy quantities H, D, C_{LMC}, E, I, T, calculated for each species-references and DNA sequence sample

Subgenotype and Species	H[Y]	D[Y]	C_{LMC} [Y]	E[Y]	I [Y]	T[Y]
IaA13R7 *C. hominis*	1.96516	0.26005	0.51104	0.26226	0.13345	0.26114
IbA10G2 *C. hominis*	1.97397	0.25675	0.50682	0.25956	0.13149	0.25944
IIaA17G1R1 *C. parvums*	1.98429	0.25324	0.50249	0.25572	0.12887	0.25647

It is worthy to note that the values of these statistical parameters can be correlated with the diversity and number of repetitions of DNA sequences.

Subgenotype IIaA17G1R1 shows the highest sequence variability (Tabel 2), as demonstrated by the presence in the microsatellite region of two different repetitive triplets (TCA and TCG) which are found in variable number and interleaved distributed and also by the existence of an additional repetitive region R consisting of six nucleotides (ACATCA).

At the other extreme, subgenotype IaA13R7 has a structure more ordered than the subtype IIaA17G1R1 (Tabels 1, 3). In the microsatellite region, variability is represented only by a single type of repeating triplets (TCA) found in lower number. Despite this, it contains an additional repetitive region R of 15 nucleotides arranged in a greater number of repetitions than the subtype IIaA17G1R. In contrast, subtype Ib does not contain a supplementary repetitive regions R but the microsatellite region has two types of triples (TCA and TCG) in variable number and interleaved distribution.

Thus, "the correlations" between the microsatellite region's repetitive components could print a level of randomness and unpredictability between extremes of complexity. Such "correlations" justify the possibility of a connection between the values of statistical complexity measures used here and random repetitive triples of microsatellite region; the result is that their average growth rate is given by the Shannon entropy rate.

It remains to be appreciated better in the future if measuring the randomness and unpredictability of DNA sequence variation does not adequately reflect the correlational structure in its behavior or the existence of certain genetic subtype variations and their capacity to give chronic infection with *Cryptosporidium* in HIV positive patients.

4.2. DISTANCE MEASURES TO ASSESS THE DIVERGENCE BETWEEN SUBGENOTYPES

Here in, we depicted how much the selected probability distributions are related to each other by applying three distance measures proposed in references [5, 6, 7]. The distances Bhattacharyya ρ and Kullback-Leibler K have the additive property [7] in the statistically independent cases with structural non-identically distributed random variables as the subgenotypes in our study.

As you may see in the Tables 5, 6, and 7, in the case of Kullback-Leibler distance, because of its asymmetry, distance between p_i and q_i, $K(p_i \| q_i)$ is different from that distance between q_i and p_i $K(q_i \| p_i)$. On the other hand, the distance ρ for each group of distributions has values very close to 1 indicating a relatively high similarity between the two compared distributions. Distance ρ with the lowest value

Table 5

Distance measures *ρ, K*, and *R* calculated by comparing two different population distributions (IaA13R7and IbA10G2 subgenotypes) of the same species (*Cryptosporidium hominis*)

Subgenotype and species	**A**	**C**	**G**	**T**	**$\rho(p_{Ia}, p_{Ib})$**	**$K(p_{Ia} \| p_{Ib})$**	**$K(p_{Ib} \| p_{Ia})$**	**$R(p_{Ia}, p_{Ib})$**
IaA13R7 *C. hominis*	**300**	**168**	**240**	**195**	**0.99898**	**- 0.00584**	**- 0.00591**	**- 0.00294**
IbA10G2 *C. hominis*	**237**	**154**	**168**	**152**				

is observed between the different subtypes of the two species of *Cryptosopridium*, IaA13R7 and IIaA17G1R1, respectively. This dissimilarity may be related to different structural content of the region microsatellite and repetitive regions *R* between the two distributions. This aspect in particular that of the microsatellite region is somewhat supported by the distance *K* value indicating a high divergence between genetic subtypes of the same species of *Cryptosporidium*. However,

taking into account that the symmetric resistor-average distance *R* is more accurate than Kullback-Leibler distance *K* [7], the distance *R* confirms the Bhattacharyya measure ρ and supports the claim that the highest difference in structure is observed between IaA13R7 and IIaA17G1R1.

Table 6

Distance measures *ρ, K*, and *R* calculated by comparing two different population distributions (IaA13R7 and IIaA17G1R1 subgenotypes) of the two species of *Cryptosporidium*

Subgenotype and species	**A**	**C**	**G**	**T**	**$\rho(p_{Ia},p_{IIa})$**	**$K(p_{Ia}\|p_{IIa})$**	**$K(p_{IIa}\|p_{Ia})$**	**$R(p_{Ia},p_{IIa})$**
IaA13R7 ***C. hominis***	**300**	**168**	**240**	**195**	**0.99796**	**-0.00117**	**-0.00118**	**-0.00589**
IIaA17G1R1 ***C. parvum***	**253**	**180**	**181**	**189**				

Table 7

Distance measures *ρ, K*, and *R* calculated by comparing two different population distributions (IbA10G2 and IIaA17G1R1 subgenotypes) of the two species of *Cryptosporidium*

Subgenotype and species	**A**	**C**	**G**	**T**	**$\rho(p_{Ib},p_{IIa})$**	**$K(p_{Ib}\|p_{IIa})$**	**$K(p_{IIa}\|p_{Ib})$**	**$R(p_{Ib},p_{IIa})$**
IbA10G2 ***C. hominis***	**237**	**154**	**168**	**152**	**0.99952**	**- 0.00278**	**- 0.00281**	**- 0.00140**
IIaA17G1R1 ***C. parvum***	**253**	**180**	**181**	**189**				

5. CONCLUSIONS

In this preliminary report we present some statistical measures to characterize and quantify coding gp60 gene variability from the genome of the two dominant species of the protozoan parasite *Cryptosporidium* involved in the etiology of human gastrointestinal disease. Based on entropy theory, we evaluated the information and divergence of statistical structural complexity between genetic subtypes belonging to the two species of *Cryptosporidium* (*C. hominis* and *C. parvum*) and between those that belong to the same species. As additional statistical tools for estimating information-theoretic distance from selected data, we adopted three divergence-based measures. Together these information theoretic methods have provided a new perspective to highlight the high variability in the microsatellite region structure between different *Crytosporidium* gp60 gene variants. Thus these indicators of correlation structure may be successfully used to scale differences between species and also within species.

REFERENCES

1. P. W. Messer et al., Phys. Rev. Lett., **94**, 138103 (2005).
2. A. Ashrafi, P. Farhami, Rom. J. Phys., **57**, *3–4*, 720–725 (2012).
3. E.E. Berg and J.L. Hamrick, Can. J. For. Res. **27**, 415–424 (1997).
4. N. Ryman and O. Leimar, Molecular Ecology, **18**, 1084–1087 (2009).
5. F. Aherne et al, Kybernetika, **32**, *4*, 1–7 (1997).
6. S. Kulback, *Information Theory and Statistics*, Wiley, New York, 1959.
7. D.H. Johnson, S. Sinanovic, *Symmetrizing the Kullback-Leibler distance* (http://www.ece.rice.edu/~dhj/resistor.pdf), 2000.
8. C. Shannon. W. Weaver, *Mathematical Theory of Communication*, Illinois University, Illinois Press, Urbana, 1945.
9. A.P. Feldman, J.P. Crutchfield, Phys. Lett. A, **238**, 244–252 (1998).
10. O. Onicescu, St. Cerc. Mat., **19**, *10*, 1419–1420 (1966).
11. V. Stefanescu-Greci, *Applications of Informational Energy and Correlation*, Bucharest, 1979.
12. C. Lepadatu and E. Nitulescu, Acta Chim Slov, **50**, 539–546 (2003).
13. M. Alves, L. Xiao, I.M. Sulaiman, A.A. Lal, O. Matos, and F. Antunes, J. Clin. Microbiol., **41**, *6*, 2744–2747 (2003).
14. 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.
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. R.M Chalmers , RP Smith , SJ Hadfield , K Elwin , M Giles, Parasitol Res., **108**, *5*, 1321–5 (2011).
18. K.Ch. Chatzisavvas, C.P. Panos, S.E. Massen, *Information-Theoretic Comparison of Quantum Many-Body Systems* (http://arxiv.org/abs/quant-ph/0305106v1), 2003.