ANALYSIS OF THE *CRYPTOSPORIDIUM* SPP GP60 GENE VARIABILITY APPLYING INFORMATION THEORY

D. POPESCU[*#], *IONELA MIRELA NEAGOE*[**,***], *SUZANA E. CILIEVICI*[***,****], *DIANA R. CONSTANTIN*[*****], *V.I.R. NICULESCU*[******]

https://www.doi.org/10.59277/RJB.2024.1.01

[*]Department of Mathematical Modelling in Life Sciences, "Gheorghe Mihoc-Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, 13, Calea 13 septembrie, Bucharest-050711, Romania, [#]e-mail: dghpopescu@gmail.com
[**]Parasitology & Micology Laboratory, "Cantacuzino" National Medico-Military Institute for Research and Development, 103, Splaiul Independenței, Bucharest-050096, Romania
[***]"Carol Davila" University of Medicine and Pharmacy, Parasitology Chair, 19-21, Dimitrie Gerota street, Bucharest-020027, Romania
[****]Parasitic Diseases Clinic, Colentina Universitary Hospital, 19–21, Șoseaua Ștefan cel Mare, Bucharest, Romania
[*****]Astronomical Institute of the Romanian Academy, 5, Cuțitul de Argint street, Bucharest-040557, Romania
[******]Plasma and Radiation Physics, National Institute for Lasers, 409, Atomiștilor Street, Măgurele, Ilfov, Romania

Abstract. In this paper we used statistical methods to understand the genetic information of DNA considered as a statistical system. The alphabet of a DNA sequence is defined by the four nucleotides: adenine, cytosine, guanine, and thymine. The order of nucleotides along the DNA sequences encodes the genetic information. We have analyzed three *Cryptosporidium* DNA sequences: one DNA sequence isolated and analyzed in our laboratory and two DNA reference sequences from the public database GenBank. Each DNA sequence is considered as a statistical system and is represented by a random variable and an associate probability distribution. The Shannon entropy, Renyi entropy, Onicescu informational energy and square deviation from uniform distribution are used in order to measure the degree of randomness for the three statistical systems. The similarity and difference between the three DNA sequences of the two *Cryptosporidium* species (*Cryptosporidium hominis* and *Cryptosporidium parvum*) were assessed by calculating the statistical distance between the probability distributions associated with each pair of DNA sequences. Each of the three DNA sequences pairs with one of the other two sequences and forms three pairs of sequences. Using the associated probability distributions, the statistical distance between them can be calculated. Bhattacharyya distance measures similarity degree between the two probability distributions. The Kullback-Leiber and the resistor-average distances measure the difference between the two distributions.

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

Received: October 2023;
in final form December 2023.

ROMANIAN J. BIOPHYS., Vol. 34, No. 1, P. 1–12, BUCHAREST, 2024

INTRODUCTION

The methods of algebraic analysis and especially those of statistical investigation based on measures of entropy and divergence are increasingly approached and constitute part of the ways of characterizing the nucleotide alphabet from the composition of gene sequences [3, 19–21].

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 is 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.

The different mathematical functions capable of describing the content of the analyzed information may also be applicable in the estimation or calculation of intrapopulation genetic polymorphism [4, 25].

We apply physics methods based on notions of entropy and divergence to quantify variability at the genotypic level.

Our goal was to determine in statistical terms the structural complexity and similarity or dissimilarity between the different subtypes of the species *Cryptosporidium hominis* and *Cryptosporidium parvum.*

BASIC DNA STRUCTURE AND GENETIC INFORMATION

The main goal of modern genetics consists in the decoding and understanding of the DNA information. The basic alphabet of DNA is defined by the four types of nucleotides.

These nucleotides are of two types: pyrimidines [cytosine (C), thymine (T)], and purines [adenine (A), guanine (G)].

A five-carbon sugar molecule and a phosphate molecule are attached to each nitrogenous base to form a nucleotide that is the chemical building blocks of the DNA polymer. A phosphate group of one nucleotide links to the sugar of another nucleotide and form a DNA strand.

These bases are connected to each other between the two strands of the DNA to form basic units named base pairs (bp). 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 a pyrimidine on the other chain or between adenine and thymine or guanine and cytosine. A smaller base (a purine) is paired with a bigger base (a pyrimidine) in order to maintain a constant distance between the two DNA strands. The stability of DNA is assured by the hydrogen bonds. So,

adenine and thymine are paired and joined by two hydrogen bonds; guanine pairs with cytosine and are joined by three hydrogen bonds. The order of nucleotides along the DNA strands encodes the genetic information [20, 21].

STATISTICAL CONCEPTS AND METHODS FOR DNA ANALYSIS

A DNA sequence may be considered as a succession of symbols (nucleotides, codons, etc.) that is named alphabet. It is a source of information, and it can be assimilated to a statistical system.

Let's consider a discrete variable, $X = (x_1, x_2, \ldots x_N)$, to which we assign a probability distribution $P = (p_1, p_2, \ldots p_N)$. The probability that the random variable X has the value x_k, is equal to p_k.

We remind the probabilities properties: $\sum_{i=1}^{N} p_i = 1$ and $0 \leq p_i \leq 1$, $i = 1, 2, \ldots N$. The remainder of this section defines the statistical concepts that will be used for statistical analysis of DNA sequences.

SHANNON ENTROPY

The Shannon entropy is defined as [28]:

$$S(X) = -\sum_{i=1}^{N} p_i \log_2 p_i \quad (1)$$

It is a measure of the degree of randomness from a physical system characterized by the random variabile X.

The Shannon entropy has a minim value ($S_{min} = 0$) when the system is characterized by a perfect order ($p_1 = 1$; $p_2 = 0 \ldots p_N = 0$), and a maxim value ($S_{max} = \log_2 N$) for a complete disorder when all probabilities are equal ($p_1 = p_2 = \ldots p_N = 1/N$).

The square deviation D, defined as [10]:

$$D(X) = \sum_{i=1}^{N} \left(p_i - \frac{1}{N} \right)^2 \quad (2)$$

measures the deviation for distribution P from uniform distribution, characterized by maxim value of Shannon entropy.

RENYI ENTROPY

The Renyi entropy is a generalization of Shannon entropy. It is defined as [24]:

$$R_\alpha(X) = \frac{1}{1 - \log_2 \sum_{i=1}^{N} p_i^\alpha} \quad (3)$$

where $\alpha \geq 0$ and $\alpha \neq 1$. The Renyi entropy $R_\alpha(X)$ approaches Shannon's entropy $S(X)$ as $\alpha \to 1$ and $R_\alpha(X) \geq S(X)$.

ONICESCU ENERGY

The discrete informational energy of the random variable X was introduced by Onicescu [22, 26]:

$$E(X) = \sum_{i=1}^{N} p_i^2 \tag{4}$$

The informational energy Onicescu reaches a minimum value (E_{min}= 1/N) for uniform distribution, $p_1 = p_2 = ... \; p_N = 1/N$ (total disorder) and a maximum value ($E_{max} = 1$) for $p_1 = 1$; $p_2 = 0$; ... $p_N = 0$ (total order).

Two indices of statistical information were introduced to classify the characteristics of the probability distributions:

– the index, $I(X)$, associated to structural information quantities, Shannon entropy, $S(X)$, and Onicescu energy, $E(X)$ [16]:

$$I(X) = \frac{E(X)}{S(X)} \tag{5}$$

– the altered index, $T(X)$, associated to the square deviation from uniformity $D(X)$:

$$T(X) = \frac{E(X)}{S(X)D(X)} \tag{6}$$

SIMILARITY AND DIFFERENCE

Sometimes, it is useful to compare two or more probabilities distributions. So, we are considering random variable, $X(x_1, x_2, \ldots x_N)$ and $Y(y_1, y_2, \ldots y_N)$ having the probabilities distributions $P(p_1, p_2, \ldots p_N)$ and $Q(q_1, q_2, \ldots q_N)$, respectively. In order to measure the similarity and difference between two probabilities distributions, P and Q, we define two distances:

Bhattacharyya distance

Bhattacharyya distance, $\rho(P,Q)$, is defined as:

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

This distance measures similarity degree between the two probability distributions, P and Q, and meets the conditions: $0 \leq \rho(P,Q) < \infty$ and $\rho(P,Q) = 1$ (for identical distributions) [1].

Kullback-Leibler distance

Kullback-Leibler distance, $K(P,Q)$, is defined as:

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

The probability distribution, *P*, has a dominant role in the distance *K*(*P*,*Q*). The Kullback-Leiber distance is asymmetric and measures the difference between the two distributions, *P* and *Q* [15].

Resistor-average distance

The resistor-average distance, *R*(*P*, *Q*), is defined as:

$$\frac{1}{R(P,Q)} = \frac{1}{K(P,Q)} + \frac{1}{K(Q,P)} \quad (9)$$

This distance, *R*(*P*,*Q*) have the resistor-average name, because the formula (9) is similar to the formula of equivalent resistance for parallel connected resistors.

R(*P*,*Q*) is the harmonic mean of *K*(*P*,*Q*) and *K*(*Q*,*P*) [14]. The resistor-average distance symmetrizes the Kullback-Leibler distance and equalizes the role of the two distributions, *P* and *Q* [14].

COMMON ASPECTS OF GP60 SUBTYPING TOOL AND THE ANALYZED MATERIAL

The gp60 (also called Cpgp40/15) gene encodes a 60 kDa glycoprotein located on the surface of the apical region of the invasive parasitic stages of *Cryptosporidium*. This protein is involved in the neutralization of antibody responses in humans [7, 13, 23, 27].

On the other hand, the gp60 locus is a highly polymorphic marker in the *Cryptosporidium* genome and, for this reason, is widely used in genetic subtyping. Both species *Cryptosporidium parvum* and *Cryptosporidium hominis* are polyphyletic in the gp60 gene [24].

The gp60 target has a sequence content similar to short tandem repeats (STRs), also known as microsatellites or simple sequence repeats, at the 5′ end of the gene. It presents tandem repeats of serine that encode the trinucleotides TCA/TCG/TCT [24].

Variations in the number of trinucleotides repeats andextensive sequence differences in the non-repetitive regions (placed downstream of the microsatellite region) classify each species (*Criptosporidium parvum* and *Criptosporidium hominis*) into several subtype families also known as alleles [2, 6].

The non-repetitive regions designate the name of each allele family belonging to *Cryptosporidium* species. It is hypervariable (H) between subtypes belonging to the allele family (Ia – Ig; IIa – IIg) [17, 18, 23].

Also, allele families Ia and IIa contain a different repetitive sequence (R) located between the microsatellite region and the non-repetitive downstream region H (Table 1).

Thus, the full name of each subtype at the gp60 site begins first with the allele family type, (Ia, Ib, Id, Ie, If, Ig for *Criptosporidium hominis*, and IIa, IIb, IIc, IId, IIe, IIf, IIg for *Criptosporidium parvum*) followed by the number of repeats. In these repeats the trinucleotiedes found are represented as follows: TCA by the letter A, TCG by the letter G and TCT by the letter T.

Table 1

Distribution of polymorphic regions in a gp60 subtype of *Cryptosporidium*

Non-repeated region	**Microsatellite region (triplet repeats)**	**+/–Repetitive region (R)**	**Non-repeated region H**
≥15 bp	TCA+/-TCG+/-TCT	(Ia)5'-A(A/G)ACGGTGGTAAGG-3' (IIa)5'-ACATCA-3'	Ia-Ig/IIa-IIg

In the case of particular allele families, repeats (R) are added to the end of the subtype name [6, 20, 31].

We used DNA sequences belonging to a sample and two different references as models of probability distributions for the intraspecies comparisons. The two DNA sequence references were taken from public database GenBank [33].

1. The DNA sequence sample was isolated from the stool of HIV positive patient 3 infected with *C. hominis*, subtype IbA10G2 and has 897 bp. Accession number in European Nucleotide Archive is LT556068.

2. The DNA sequence reference 1 belongs to *C. hominis*, subtype IaA13R7 and has 895 bp [17]. Accession number in GenBank is JX088408.

3. The DNA sequence reference 2 belongs to *C. parvum*, subtype IIaA17G1R1 and has 868 bp [29]. Accession number in GenBank is KC995129.

All the three subtypes of *Cryptosporidium*, IbA10G2, IaA13R7, IIaA17G1R1, are included in Table 2. For a better appreciation of the amino acid sequence after applying of mathematical methods, the size of nucleotide sequences belonging to references has been adjusted (from 895 to 891 bp for DNA sequence reference 1, and from 868 to 867 bp DNA sequence reference 2).

Table 2

The number of nucleotides and order of repetitions contained in the selected subtypes belonging to *Cryptosporidium* spp

Subgenotype	**A**	**C**	**G**	**T**	**Order of repetitions (Number and type)**				
IbA10G2-897 bp	292	195	213	197	7 TCA	1 TCG	2 TCA	1 TCG	1 TCA
IaA13R7-891 bp	288	172	234	193	13 TCA	7 R-AAGACGGTGGTAAGG			
IIaA17G1R1-867 bp	266	196	198	206	2 TCA	1 TCG	15 TCA	1R-ACATCA	

The subtype IbA10G2 was chosen as a sample because it was the most dominant in this study and it was associated with all categories of human cryptosporidiosis. Also, according to studies carried out at European level, this subtype has an anthroponotic character and is frequently involved in waterborne epidemics [5]. References were chosen on their higher sequence alignment in BLAST with new subtypes detected in this study.

In addition, previous reports indicated that:

The cases of human infections with subtype IIaA17G1R1 were especially linked to exposure to zoonosis [25]. Although the potential for human-to-human transmission, especially through the nosocomial route, exists [13].

IaA13R7 was frequently reported in cases of HIV infection [5], in children [11] as well as in wild animals [32] despite the fact that *C. hominis* species was previously considered anthroponotic. This explains the high potential of this subtype to spread and transmit with great probability through drinking water [32].

RESULTS AND DISCUTIONS

STATISTICAL ANALYSIS OF THE SEQUENCE VARIABILITY BETWEEN *CRYPTOSPORIDIUM* SUBTYPES

The squared deviation D and the Shannon entropy S are known as parameters widely applied in identifying a pattern of structure and statistical complexity between probability distributions. Thus, in order to quantify the variability between the DNA sequences taken in the study, it was considered that the Shannon entropy has a minimum value ($S_{min} = 0$) defining the total structural order (there is only one state with probability equal to 1) and a maximum value ($S_{max} = \log_2 N$) to define total disorder (all states have equal probability).

Additionally, the information energy, E, was approached as a superior statistical parameter to quantify the dispersion distributions of the nucleotide symbols belonging to the analyzed subtypes [9]. This function evolves inversely proportional to the Shannon. The information energy, E, has a maximum value ($E_{max} = 1$) for total structural order and a minimum value ($E_{min} = 1/N$) for total structural disorder.

The four different nucleotides [adenine (A), cytosine (C), guanine (G) and thymine (T)] constitute the source of information written in DNA structure of the three subtypes of *Cryptosporidium* selected for analysis.

Based on the existing data in Table 2, the probabilities correlated to the distributions of the three DNA sequences were calculated.

The results obtained by applying statistical methods for the calculted probabilities are presented in Table 3.

Values of Shannon entropy evolve inversely proportional to the values of the other statistical parameters (*D*, *E*, *I*, *T*). *Cryptosporidium parvum*, subtype IIaA17G1R1, has the highest Shannon entropy value. Shannon entropy, *S*, for subtype IbA10G2 is higher than for subtype IaA13R7. Dissimilarities were also observed between the subtypes of the same *C. hominis* species.

The subtype IIaA17G1R1 stands out for the diversity and number of DNA sequence repeats. In the STR region of this subtype there are two different repetitive triplets (TCA and TCG) in variable numbers.

Table 3

The values of the statistical parameters calculated on the investigated DNA sequences

Subtypes and species	$S(X)$	$D(X)\cdot 10^2$	$E(X)$	$I(X)$	$T(X)$
IIaA17G1R1-867 bp *C. parvum*	1.98794	0.43588	0.25436	0.12795	29.35463
IbA10G2-897 bp *C. hominis*	1.97853	0.78482	0.25785	0.13032	16.60508
IaA13R7-891 bp *C. hominis*	1.97242	0.97385	0.25974	0.13169	13.52216

Overall, the subtype belonging to the allelic family IIa is characterized by a much more uniform structural distribution with more equal probabilities of the four DNA nucleotide symbols. The high value of the Shannon entropy and the low value of the informational energy define this structure and argue for a higher complexity of subtype IIaA17G1R1.

On the other hand, compared to subtype IIaA17G1R1, IaA13R7 presents a more uneven distribution due to the unequal probabilities of the occurrence of certain nucleotide symbols from the four known and implicitly a more ordered structure (Table 2). The region of STRs is defined by a reduced number of triple repeats (TCA) and the additional repetitive R region of 15 nucleotides is arranged in a higher number of repeats than in subtype IIaA17G1R1.

In the composition of subtype IbA10G2, the STRs region has two types of triples (TCA and TCG) in variable numbers with an interspersed distribution missing and the repetitive R region missing.

The disposition and composition of the repetitive structures, especially in the region of the STRs, justify the connection with the values of the statistical complexity measures obtained here. The average growth rate of repeating triplets is given by the Shannon entropy rate.

This means that if the informational entropy, *S*, is higher, the nucleotide distribution in a DNA sequence is less stable and implicitly more complex.

STATISTICAL DISTANCES ASSESS THE DIVERGENCE BETWEEN *CRYPTOSPORIDIUM* SUBTYPES

The Bhattacharyya distance and Kullback-Leibler distance are known for their additive property in the statistically independent cases with non-identically distributed structural random variables [1, 14, 15]. These statistical parameters and symmetric resistor-average distance *R* have been approached to appreciate the degree of relatedness between the probability distributions selected in this study.

According to the asymmetry of the Kullback-Leibler measure, the distance between p_1 and p_2, $K(p_1 \| p_2)$, is different from the distance between p_2 and p_1, $K(p_2 \| p_1)$ (Table 4). On the other hand, if the Bhattacharyya distance ρ for two distributions is closer to 1, a greater degree of similarity exists between the two compared distributions. Thus, the greatest dissimilarity was noted between the different subtypes of the two *Cryptosporidium* species (IaA13R7 and IIaA17G1R1). In this distribution group the distance ρ reached the lowest value (Table 4).

This difference can be justified by the different structural components at the level of the STRs region and the repetitive R regions in the two DNA distributions.

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, the symmetric resistor-average distance *R*, which is characterized by a higher accuracy than the Kullback-Leibler *K* distance, confirms the Bhattacharyya distance parameter ρ and supports the statement that the greatest dissimilarity in structure is observed between IaA13R7 and IIaA17G1R1 [14].

Table 4

The distances ρ, *K*, and *R* calculated for each pair of two different population distributions

First subtype	$\rho(p_1,p_2)\cdot 10^2$	$K(p_1,p_2)\cdot 10^2$	$K(p_2,p_1)\cdot 10^2$	$R(p_1,p_2)\cdot 10^2$	**Second subtype**
IIaA17G1R1 ***C. parvum***	4.4796	0.2595	0.2576	0.1293	IbA10G2 *C. hominis*
IbA10G2 ***C. hominis***	6.4977	0.3681	0.3691	0.1843	IaA13R7 *C. hominis*
IaA13R7 ***C. hominis***	15.9825	0.9213	0.9226	0.4610	IIaA17G1R1 *C. parvum*

The DNA sequence belonging to the IIaA17G1R1 subtype proved to present the greatest structural disorder and to be the most different compared to the other subtypes. This aspect could argue for a higher individuality of the IIaA17G1R1 subtype. At the same time, it could explain the spread and implicitly the ability to give more limited infections compared to the subtypes of *Cryptosporidium hominis* taken in the study. On the other extreme, IaA13R7 having a much more stable and

less complex structure could justify the higher capacity of demographic spread reported at the international level and to cause infections in both humans and animals [31, 32].

In Table 4 we find the distances ρ, *K*, and *R* calculated for comparing each pair of two different population distributions of the three species of *Cryptosporidium* (IbA10G2 with IaA13R7, IaA13R7 with IIaA17G1R1 and IIaA17G1R1 with IbA10G2).

CONCLUSIONS

In this paper we applied statistical measures of that we used as tools to evaluate the variability of the gp60 gene in the genome of the two dominant species of the protozoan parasite: *Cryptosporidium parvum* and *Cryptosporidium hominis*. Based on the notion of informational entropy, the structural information and the divergence of the statistical structural complexity were quantified at both the between species and at the intraspecies level. Additionally, by approaching the three measures based on divergences, the informational theoretical distance between the selected data was statistically estimated.

Taken together these information theoretic methods have provided a new perspectiveon the high variability in the microsatellite region structure between different *Crytosporidium* gp60 subtypes. Thus, these indicators of correlation structure may be successfully used to scale differences between species.

REFERENCES

1. AHERNE, F., N. THACKER, P. ROCKET, The Bhattacharyya metric as an absolute similarity measure for frequency coded data, *Kybernetika*, 1997, **32**, 1–7.
2. ALVES, M., L. XIAO, I.M. SULAIMAN, A.A. LAL, O. MATOS, F. ANTUNES, Subgenotype analysis of *Cryptosporidium* isolates from humans, cattle, and zoo ruminants in Portugal, *J. Clin. Microbiol*., 2003, **41**, 2744–2747.
3. ASHRAFI, A., P. FARHAMI, Characterization of 3D visualization method for DNA sequences, *Rom. J. Phys*., 2012, **57**, 720–725.
4. BERG, E.E., J.L. HAMRICK, Quantification of genetic diversity at allozyme loci, *Can. J. Res.*, 1997, **27**, 415–424.
5. CACCIÒ, S.M., R.M. CHALMERS, Human cryptosporidiosis in Europe, C*linical Microbiology and Infection*, 2016, **22**(6), 471–480.
6. CAMA, V.A., J.M. ROSS, S. CRAWFORD, V. KAWAI, R. CHAVEZ-VALDEZ, D. VARGAS, A. VIVAR, E. TICONA, M. NAVINCOPA, J. WILLIAMSON, Y. ORTEGA, R.H. GILMAN, C. BERN, L. XIAO, Differences in clinical manifestations among *Cryptosporidium* species and subtypes in HIV-infected persons, *J. Infect. Dis.* 2007, **196**, 684 –691.
7. CEVALLOS, A.M., X. ZHANG, M.K. WALDOR, S. JAISON, X. ZHOU, S. TZIPORI, M.R. NEUTRA, H.D. WARD, Molecular cloning and expression of a gene encoding *Cryptosporidium parvum* glycoproteins gp40 and gp15, *Infect. Immun*., 2000, **68**, 4108–4116.

8. CHALMERS, R.M., C. JACKSON, K. ELWIN, S. HADFIELD, P. HUNTER, DWI0851: Investigation of genetic variation within *Cryptosporidium hominis* for epidemiological purposes, 2007, *http://dwi.defra.gov.uk/research/ completed-research/2000todate.htm*.
9. CHATZISAVVAS, K.CH., C.P. PANOS, S.E. MASSEN, Information-theoretic comparison of quantum many-body systems, *Quantum Physics*, 2003, *http://arxiv.org/abs/quant-ph/0305106v1*.
10. FELDMAN, A.P., J.P. CRUTCHFIELD, Measures of statistical complexity: Why?, *Phys. Lett. A*, 1998, **238**, 244–252.
11. GALVAN-DIAZ A.L, K. BEDOYA-URREGO, A. MEDINA-LOZANO, J. URAN-VELASQUEZ, J.F. ALZATE, G. GARCIA-MONTOYA, Common occurrence of *Cryptosporidium hominis* in children attending day-care centers in Medellin, Colombia, *Parasitol. Res*., 2020, **119**(9), 2935–2942.
12. GROSSE, I., P. BERNAOLA-GALVAN, P. CARPENA, R. ROMAN-ROLAND, J. OLIVER, H.E. STANLEY, Speciesindependence of mutual information in coding and noncoding DNA, *Phys. Rev*., 2002, **65**, 1–16.
13. HATALOVA, E., T. GUMAN, V. BEDNAROVA, V.T. SIMOVA, M. LOGOIDA, M. HALANOVA, Occurrence of *Cryptosporidium parvum* IIaA17G1R1 in hospitalized hemato-oncological patients in Slovakia, *Parasitology Research*, 2022, **121**, 471–476.
14. JOHNSON, D.H., S. SINANOVIC, Symmetrizing the Kullback-Leibler distance, *IEEE Transactions on Information Theory*, 2001, *http://www.ece.rice.edu/~dhj/resistor.pdf*.
15. KULBACK, S., *Information Theory and Statistics*, Wiley, New York, 1959.
16. LEPADATU, C., E. NITULESCU, Information energy and information temperature for molecular systems, *Acta. Chim. Slov.*, 2003, **50**, 539–546.
17. LI, N., L. XIAO, V.A. CAMA, Y. ORTEGA, R.H. GILMAN, M. GUO, Y. FENG, Genetic recombination and *Cryptosporidium hominis* virulent subtype IbA10G2, *Emerg. Infect. Dis*., 2013, **19**(10), 1573–1582.
18. LI, F., Z. ZHANG, S. HU, W. ZHAO, J. ZHAO, M. KVÁČ, Y. GUO, N. LI, Y. FENG, L. XIAO, Common occurrence of divergent *Cryptosporidium* species and *Cryptosporidium parvum* subtypes in farmed bamboo rats (*Rhizomyssinensis*), *Parasites & Vectors*, 2020, **13**, 149–165.
19. MESSER, P.W., P.F. ARNDT, M. LÄSSIG, Solvable sequence evolution models and genomic correlations, *Phys. Rev. Lett*., 2005, **94**, 138103.
20. NEAGOE, I.M., D. POPESCU, V.I.R. NICULESCU, Alternative methods for statistical characterization and quantification of *Cryptosporidium* spp. Gp60 gene variability, *Rom. Report Phys*., 2014, **66**(3), 683–692.
21. NEAGOE, I.M., D. POPESCU, L. LAZAR, V.I.R. NICULESCU, S. MICLOS, Human cryptosporidiosis: species, subtypes, differences in pathogeny and clinical manifestations, and mathematical methods for DNA sequence analysis, in: *Advance in Medicine and Biology*, L.V. Berhardt ed., Nova Science Publishers Inc., New York, USA, Chapter 7, 2016, pp. 109–166.
22. ONICESCU, O., Energie informationnelle, *Comptes Rendus de l'Académie des Sciences Paris A*, 1966, **263**, 841–842.
23. PRIEST, J.W., KWON, J.P., ARROWOOD, M.J., LAMMIE, P.J., Cloning of the immunodominant 17-kDa antigen from *Cryptosporidium parvum*, *Mol. Biochem. Parasitol*., 2000, **106**, 261–271.
24. RENYI, A., On measures of entropy and information, *Papers of Alfred Renyi*, 1976, **2**, 525–580.
25. RYMAN, N., O. LEIMAR, G_{ST} is still a useful measure of genetic differentiation - a comment on Jost's *D*, *Molecular Ecology*, 2009, 18, 2084–2087.
26. STEFANESCU-GRECI, V., *Applications of Informational Energy and Correlation*, Romanian Academy House, Bucharest, 1979.
27. STRONG, W.B., R.G. NELSON, Preliminary profile of the *Cryptosporidium parvum* genome: An expressed sequence tag and genome survey sequence analysis, *Mol. Biochem. Parasitol*., 2000, **107**, 1–32.

28. SHANNON, C., W. WEAVER, *Mathematical Theory of Communication*, Illinois University, Illinois Press, Urbana, 1945.
29. TOMAZIC, M.L., J. MAIDANA, M. DOMINGUEZ, E.L. URIARTE, R. GALARZA, C. GARRO, M. FLORIN-CHRISTENSEN, L. SCHNITTGER, Molecular characterization of *Cryptosporidium* isolates from calves in Argentina, *Vet. Parasitol*., 2013, **198**(3-4), 382–386.
30. WINTER, G., A.A. GOOLEY, K.L. WILLIAMS, M.B. SLADE, Characterization of a major sporozoite surface glycoprotein of *Cryptosporidium parvum*, *Funct. Integr. Genomics*, 2000, **1**, 207–217.
31. XIAO, L., U. RYAN, Molecular Epidemiology, In: *Cryptosporidium and Cryptosporidiosis*, R. Fayer, L. Xiao, Eds, cap. 5, CRC Press, 2008, pp. 119–173, ISBN 13 978-1-4200-5226-8.
32. ZAHEDI, A, A. PAPARINI, F. JIAN, I. ROBERTSON, U. RYAN, Public health significance of zoonotic *Cryptosporidium* species in wildlife: Critical insights into better drinking water management, *International Journal for Parasitology: Parasites and Wildlife*, 2016, **5**(1), 88–109.
33. ****http://www.ncbi.nlm.nih.gov/genbank*.