Wavelet spectrogram - based DNA analysis for the assessment of *Cryptosporidium spp*. Gp60 subgenotypes variation

I. M. NEAGOE[a,b], S. MICLOS[c,*], D. POPESCU[d,e], D. SAVASTRU[c], V. I. R. NICULESCU[f], M. DAMIAN[b], L. LAZAR[a,g], S. DONTU[c], M. TAUTAN[c]

[a]*"Carol Davila" University of Medicine and Pharmacy, 19-21 D. Gerota St., Bucharest, Romania*
[b]*National Institute of R&D for Microbiology and Immunology "Cantacuzino", Spl. Independentei 103, Bucharest, Romania*
[c]*National Institute of R&D for Optoelectronics - INOE 2000, 409 Atomistilor St., Magurele, Ilfov, RO-077125, Romania*
[d]*Institute of Mathematical Statistics and Applied Mathematics of Romanian Academy, Calea 13 Septembrie 13, Bucharest, Romania*
[e]*Faculty of Biology, University of Bucharest, Spl. Independentei 91-95, Bucharest, Romania*
[f]*National Institute for Laser, Plasma and Radiation Physics, 409 Atomistilor St., P.O.Box MG-36, Bucharest-Magurele, RO-077125, Romania*
[g]*"Colentina" Clinical Hospital, 19-21 Stefan cel Mare St., Bucharest, Romania*

Some variable regions of *Cryptosporidium* gp60 gene may have a potential implication in the pathogenesis of disease caused by this parasite. DNA sequencing provides enough material to specific computational processing methods for the characterization of DNA diversity. A wavelet spectrogram method has been adopted to evaluate the degree of similarity and difference of variable DNA information between three subgenotypes of two species of *Cryptosporidium* parasite. One of the three wavelet spectrogram analyzed DNA sequences, was extracted, amplified, sequenced in our laboratories, and deposited in GenBank. Using additional a mathematically index and Multidimensional Scaling tool, we emphasized some features of analyzed *Gp60* subgenotypes in the perspective of wavelet analysis.

(Received December 4, 2013; accepted July 10, 2014)

Keywords: Wavelet spectrogram analysis, Multidimensional Scaling, DNA sequence, *Cryptosporidium* genetic diversity, *Gp60* gene

1. Introduction

Cryptosporidium spp. is a significant parasitic pathogen of humans and other animals which it is recognize as a major cause of a severe diarrheal disease mainly in young or immunocompromised organisms. Human infections are predominantly induced by species *C. hominis* and *C. parvum* [1]. Among the many markers applied in epidemiological survey of *Cryptosporidium* species, a variable fragment of the *Gp60* gene encoding a sporozoit surface glycoprotein was commonly used. The polymorphic nature of this gene is given by microsatellite region composed on serine coding tri-nucleotide repetitions and a hypervariable region based on were identified numerous alleles [2,3].

Following the widespread use of DNA sequencing for genotyping and subtyping of clinical and environmental isolates, a large amount of information on the *Cryptosporidium spp* DNA variation are available to be accessed and decoded by mathematics assessment methods [3,4]. Although generally DNA studies are useful in many other directions [5 and therein], some computational processing methods of DNA information, especially wavelet transform [6] and wavelet spectrogram analysis [7] have received more attention due to its ability to capture global and local features of DNA structure. Some types of wavelet were modulated in order to localize, identify or describe the observations of DNA information (such as SNPs, microsatellites, coding triplets, GC content, recombination rate) [6,8-10]. Taking into account the relationship between genetic content and divergence, wavelet analysis transforms a sequence of observations in a series of coefficients able to describe variation of the signal at broader scales [11]. On the other hand, on the light of patterns obtained from wavelet spectrograms (*WS*), continuous wavelet method tends to be the best for the purpose to evaluate DNA regions variation by comparing different genetic information [7, 12].

The wavelet transform add to optical spectroscopy in order to study the properties of DNA biopolymer (13-16]. From the wavelet spectrograms one can obtain the relevant information available in capturing global and local characteristics of DNA biopolymer structure [6].

Based on the fact that information is encoded by the four types of nucleotides which differ in the level of nitrogenous bases adenine (A), cytosine (C), guanine (G) and thymine (T), in this paper we evaluate the *Cryptosporidium Gp60* gene diversity in terms of *WS*. We adopted the model of Shannon wavelet transform [12] for an imaging estimation of dissimilarities between three subgenotypes of *C. parvum* and *C. hominis*. For a quantitative assessment of the differences between wavelet

patterns obtained, we applied additional a comparison index and Multidimensional Scaling (MDS) visualization tool [17].

2. Experimental

2.1 *Cryptosporidium spp. Gp60* gene data

For signal processing were chosen from the literature two reference DNA sequences of *Gp60* gene belonging to the species *C hominis* and *C. parvum*. They are designated as two different subgenotypes (IaA13R7 and IIaA17G1R1), and can be acquired from public database GenBank under accession numbers: EU052234 and HQ005735, respectively [18,19].

For signal comparing we used a sample DNA sequence analyzed by us applying other mathematical methods [3,4]. It has been isolated from an immunocompromised patient with diarrheal syndrome, amplified by PCR (Polymerase Chain Reaction) with Applied Biosystem AmpGene 2700 termocycler, and sequenced with ABI PRISM 3100 Avant sequencer in our "Cantacuzino" Institute laboratories.

The consensus DNA sample sequence was obtained by analyzing and processing of DNA sequencing chromatograms on forward (Fig. 1) and reverse strands. It was successively compared with sequences of the international database using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), designated as *C. hominis* subgenotype IbA10G2 and deposited in GenBank under accession number HG423391.

Each of the three analyzed DNA sequences (in FASTA format) received an identifier composed of letters and a number. First letter is C (from *Cryptosporidium*), the second is H or P (from the species *C. hominis* or *C. parvum*) and R for a reference sequence or S for a sample sequence. The number 1 or 2 designed the two allelic families I and II belonging to the two species of *Cryptosporidium* (*C. hominis* and *C. parvum*). The last letter A or B is assigned to allelic type (*a* or *b*).

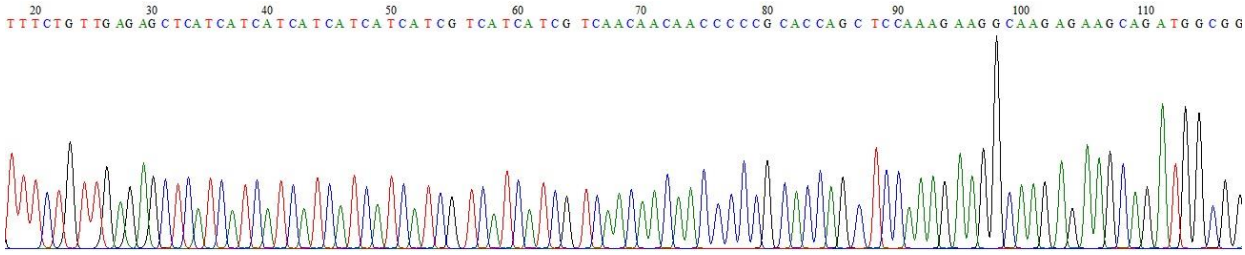


Fig. 1. DNA sequencing chromatogram segment (forward DNA strand) of the DNA sample CHS1B with the microsatellite region.

Table 1. Composition of the five analyzed DNA sequences.

	CHR1A	CHS1B	CPR2A
Adenine	300	237	253
Cytosine	168	154	180
Guanine	240	168	181
Thymine	195	152	189
Total	**903**	**711**	**803**

The first reference DNA sequence CHR1A (Fig. 2a) has *n* = 903 nucleotide length and it includes a microsatellite region with 13 TCA repeats (between nucleotide positions 28 – 66) and an independent repetitive region R with 7 AAGCGGTGGTAAGG repeats (nucleotide positions 161 – 265). The second reference DNA sequences CPR2A (Fig. 2c) has *n* = 803 nucleotides and includes a microsatellite region with 17 TCA and 1TCG repeats (between nucleotide positions 16-69) followed immediately by an independent region R with 1 ACATCA repeats. Using the BLAST tool, between the two reference sequences CHR1A and CPR2A were determined specific identities of 93% (between nucleotide positions 13-71 and 1-59 in the region containing microsatellites) and 86% (between nucleotide positions 321-903 and 222-795 at the end of the hypervariable region).

The sample sequence CHS1B (Fig. 2b) is 711 nucleotides long and contains a microsatellite motif with 10 TCA and 2 TCG repeats (between nucleotide positions 28 – 63 in FASTA format consensus DNA sequence, access number HG423391). It shows identities in BLAST with reference CHR1A of 81% (between nucleotide positions 2-139 and 1-141) and 82% (nucleotide positions 293-707 and 349-772). On the other side, the sample CHS1B presents identities with the reference CPR2A of 87% (nucleotide positions 14-70 and 1-57) and 80% (nucleotide positions 215-711 and 172-668) in accord with the BLAST determinations. Sample sequence shows similarity ends (including microsatellite region and terminal portion of the hypervariable region) with the two references. Besides nucleotide positions mentioned above, the sample CHS1B is very different of references CHR1A and CPR2A.

2.2 DNA information decoding

The symbols {A,C,G,T} corresponding to the four nucleotides types of the three DNA sequences analyzed, were converted into numerical values after the following

scheme [12]: A = 1+0*i*, C = −1+0*i*, T =0 +*i* and G = 0−*i*, where $i=\sqrt{-1}$. This translation leads to the transformation of each DNA sequence into a complex "signal" *x* (t) where t is interpreted as a "time" serial number of nucleotides along each sequence. With other words the signal *x*(t) describing the DNA sequence composition is defined as follows:

$$x(\mathbf{k})=\begin{cases}1 & \text{if the position k in the DNA sequence is occupied by the nucleotide } \mathbf{A}\\ i & \text{if the position k in the DNA sequence is occupied by the nucleotide } \mathbf{T}\\ -1 & \text{if the position k in the DNA sequence is occupied by the nucleotide } \mathbf{C}\\ -i & \text{if the position k in the DNA sequence is occupied by the nucleotide } \mathbf{G}\end{cases}$$

We will name this function which describes a DNA sequence as "genomic function".

2.3 Wavelet spectrograms analysis

The information extraction from the three converted DNA sequences was performed by analysis of spectrograms obtained by applying continuous wavelet transformation (*WT*) [13] defined as:

$$W(\tau,s)=\int_{-\infty}^{+\infty}\frac{x(t)}{\sqrt{s}}\cdot\psi^{*}\left(\frac{t-\tau}{s}\right)\cdot dt \quad (1)$$

where the parameters *s* (*s* > 0) and *τ* represent the dyadic dilation (scale parameter), respectively the dyadic position (translation parameter), and *ψ(t)* is a function called the mother wavelet. The symbol * denotes the complex conjugate. The maxim values of *s* and *τ* are equal to DNA sequence length. The best results were obtained using as mother wavelet the real Shannon wavelet [12]:

$$\psi(x)=\frac{\sin(2\pi x)-\sin(\pi x)}{\pi x} \quad (2)$$

Shannon wavelet function being real, the complex conjugate * in Eq. (1) becomes useless. The "time" *t* takes discrete values between 1 and *n* (the length of the DNA sequence), *s* and *τ* take also the same values as "time" *t*. Finally, the integral turns into a sum:

$$W(\tau,s)=\sum_{t=1}^{n}\frac{x(t)}{\sqrt{s}}\cdot\psi\left(\frac{t-\tau}{s}\right) \quad (3)$$

This *WT* were calculated for each of the three DNA sequence of a given *Cryptosporidium* subtype converted into complex-valued signal. For each sequence a spectrogram pattern was depicted.

2.4 Comparison index r_{ij}

The "distance" between two data sequences (*i* and *j*) were estimated using the measure r_{ij} [12]:

$$r_{ij}=\sqrt{2\cdot\left(\mu_i-\mu_j\right)^2+\left(\sigma_{\bar{s}i}-\sigma_{\bar{s}j}\right)^2+\left(\sigma_{\bar{\tau}i}-\sigma_{\bar{\tau}j}\right)^2} \quad (4)$$

where μ represents the average of the data sequence, $\sigma_{\bar{S}}$ − standard deviation on direction $\bar{\mathbf{s}}$, $\sigma_{\bar{\tau}}$ − standard deviation on direction $\bar{\boldsymbol{\tau}}$; *i* and *j* = 1, 2, 3 (all the analyzed DNA sequences).

2.5 MDS (Multidimensional scaling) tool

Basis on the measure r_{ij} values, a symmetrical correlation matrix R_{3x3} for comparing all three DNA sequences was constructed. The MDS (Multidimensional scaling) method was approached as alternative to represent in a lower dimensional map the set of data points whose similarities are defined in a higher dimensional space obtained by wavelet [20]. The graphical representation of two dimensional MDS map for the data points and its building based on the matrix r_{ij} elements were created in MATLAB.

3. Results and discussions

3.1. Features of the three DNA sequences obtained using the wavelet spectrogram analysis

After obtaining wavelet spectrograms for the set of the three DNA sequences, the charts were normalized for coordinates *s* and *τ* in order to compare DNA sequences with different lengths [12]. So, $\bar{\mathbf{s}}=s/s_{max}$ and $\bar{\boldsymbol{\tau}}=\tau/\tau_{max}$.

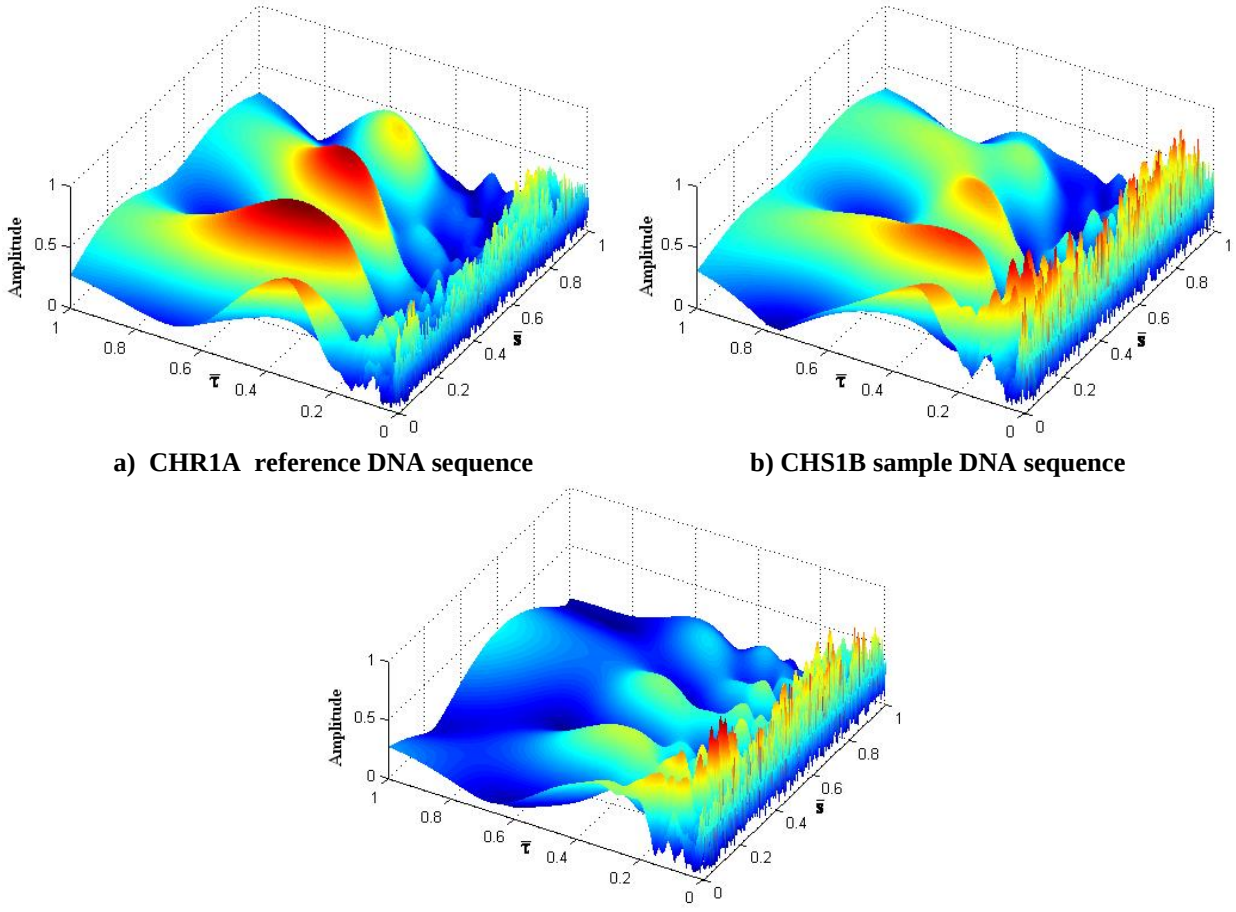
Using as landmark the DNA variability of information embedded into the three different subgenotypes of *Cryptosporidium* species, we evaluated by comparison the Shannon wavelet ability to capture and reveal the differences and similarities between different two references CHR1A, CPR2A and CHR1B sample. Fig. 2 depicts the clear distinction of wavelet imaging features for the three converted DNA sequences of Cryptosporidium. The degree of qualitative similarity between the three analyzed wavelet patterns is mainly attributed to the number and appearance of relevant items or peaks mostly spaced and placed in the region having $\bar{\mathbf{s}}\in(0, 0.5)$ and $\bar{\boldsymbol{\tau}}\in(0.6,1)$ in Fig. 2.

We remind that wavelet spectrogram has a better time (DNA length) and poorer frequency resolution at low scale parameter $\bar{\mathbf{s}}$ (high frequency) and better frequency and poor time resolution for high scale parameter $\bar{\mathbf{s}}$ (low frequency). Looking at the Fig. 2a (CHR1A) we easily see three important items with high resolution. The coordinates of their centers are placed at high values of $\bar{\boldsymbol{\tau}}$ (that is at the end of the consensus DNA sequence) and low values of $\bar{\mathbf{s}}$ (high frequency resolution). Such observation is available for Fig. 2b (CHS1B). For CPR2A appear three items (for $\bar{\mathbf{s}}<0.3$) also, but are localized at the middle of the DNA sequence. These items contain

information about a property of a DNA sequence segment. In our opinion, this information may suggest that the number of nucleotides A and T is greater than the number of nucleotides C and G along this segment, or inversely. This property may be not common for the all the three sequences. For an easier explanation, we named thymine and adenine as "positive" nucleotide, and cytosine and guanine as "negative" nucleotide, in relation with the sign of the numerical values associated above. If we thinking the DNA sequence as a succession of nucleotide triplets (named codons) which code for the amino acids, the building blocks of proteins, then some observations may be made about the proteins structure and their function. For example, the segments of DNA sequences corresponding to the three identified items contain more codons with two or three "positive" nucleotides and fewer codons with two or three "negative" nucleotides.

Another lot of items appear at the bottom half of the pictures of the fig. 2, corresponding to the first half of each DNA sequences. The small items positioned within the region $\bar{\tau} \in (0.2, 0.45)$ and $\bar{s} \in (0, 0.25)$ indicate relatively short DNA segments of these sequences, with a higher percentage in GC content (over 50%).

However visually it is difficult to estimate accurately the relationship between detected items and the positions of the variable regions with higher similarity between DNA sequences analyzed. Therefore, to assess the regional dissimilarities between two DNA sequences, the ability of the eye to appreciate the differences between wavelet patterns obtained for these sequences is not sufficient.



a) CHR1A reference DNA sequence

b) CHS1B sample DNA sequence

c) CPR2A reference DNA sequence

Fig. 2. The wavelet spectrograms for DNA sequences of Gp60 gene belonging to three species Cryptosporidium.

3.2 Quantify the differences between the three wavelet spectrograms

The time expression features of the three analyzed DNA sequences which were not distinguishable by visual inspection of Shannon wavelet charts, can be mathematically estimated by measure r_{ij}.

The index r_{ij} emphasized the great similarity in the polymorphic ends of the two reference sequences (CHR1A and CPR2A) according to BLAST. From the perspective of DNA sequence length parameter, it is observed that the

greatest differences occur between sequences that have very different length (CHR1A and CHS1B) while the largest similarities are calculated between sequences of very close lengths (CHR1A and CPR2A. Somehow, it is possible that the large differences in length between DNA sequences lead to wavelet exacerbation of structural DNA differences. On the other side, the index r_{ij} suggests that *WT* highlights cumulatively areas with highest similarity (over 80% by evaluating BLAST) without taking into account the length of these regions in analyzed DNA sequences. Depending on the sequence variability of these areas, the wavelet estimates global degree of similarity or difference between DNA sequences. The additive property is also observed in the case of Kullbach-Leibler distance [21]. Our results obtained by wavelet quantitative estimation of the *Cryptosporidium* subgenotypes variation confirm Kullbach-Leibler distance results between the same three subtypes obtained in another study of ours [3].

Table 2. Distances $r_{ij} \cdot 10^4$ between DNA sequences.

R_{3x3}	CHR1A	CHS1B	CPR2A
CHR1A	0	4.38187	2.61486
CHS1B	4.38187	0	3.00877
CPR2A	2.61486	3.00877	0

3.3 Features of the wavelet analysis of DNA sequences by MDS tool

MDS approach in the perspective of the Shannon wavelet transform and r_{ij} index was employed to map mathematical distances between analyzed DNA sequences.

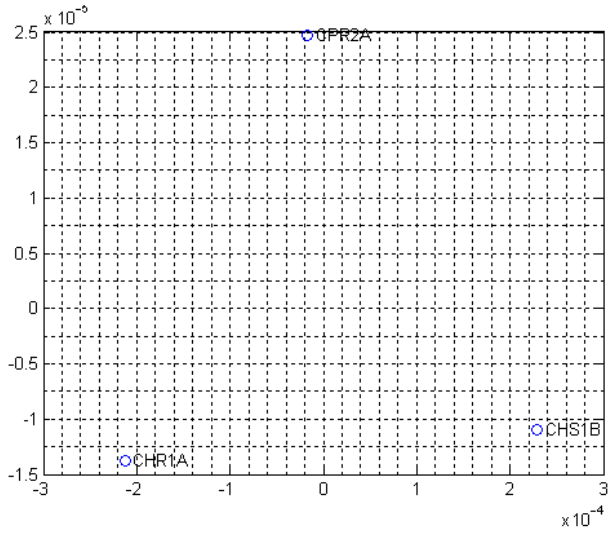


Fig. 3. MDS chart for DNA sequences of Gp60 gene belonging to three species Cryptosporidium.

Looking at the MDS map (fig. 3), it should be noted that the genetic distances built on the r_{ij} values between the three wavelet transformed sequences indicate that the largest difference occurs in the same species of *Cryptosporidium hominis* between subgenotypes CHR1A and CHS1B. The large distance between different subgenotypes of same allelic families can be attributed to the composition and position of the hypervariable region and repetitive regions of CHR1A (82% identities according to BLAST). In contrast the smallest distance is recorded on the map MDS between different species *C. hominis* and *C. parvum* with subgenotypes CHR1A and CPR2A (in accord with BLAST identities of 93% in the microsatellites area and 86% in hypervariable region). It should also be noted that for the overall assessment of genetic variation between sequences, the potential effect of sequence length on *WT* must completely remove by comparing the DNA sequences of the same length. In support of these results, we note that in addition to the potential parameter related to sequence length and nucleotide composition of the variable regions (microsatellite and hypervariable regions) there is a potential third parameter that can influence wavelet analysis. It is associated with the nucleotide position or distance of the interested variable region.

4. Conclusions

WT has the great ability to develop signal components at different resolutions and highlight some feature of DNA related mainly to the position and composition of nucleotide regions. The real Shannon wavelet model can be considered as a promising method for capturing DNA dissimilarities between different subgenotypes of the same species or different species of *Cryptosporidium* through qualitative and quantitative comparisons. We also defined a mathematical function that describes a DNA sequence by introducing the name "genomic function" and we hope this name will be adopted by the scientific world. All *Cryptosporidium* subgenotypes comparisons of normalized wavelet values obtained for the references and sample used in this study, confirm the DNA structural similarities highlighted by BLAST. The complex pattern of wavelet transformation is dependent on the sequence length and consequently there is an addiction to different maximum values of s and τ. These two wavelet parameters of the DNA sequence (DNA sequence variation and length) can amplify or attenuate each other. MDS as visualization tool creates a hierarchy of differences between sequences analyzed according to computed genetic distances between them. Based on this study, we hypothesize that in terms of comparing DNA sequences of the same length, Shannon wavelet analysis has the potential to be widely applied to evaluate the nucleotide composition and genetic variation of coding DNA regions with possible implications in pathogenicity.

Acknowledgments

We thank the Molecular Epidemiology laboratory of "Cantacuzino" Institute, Bucharest, Romania which has provided us the access to some equipment for PCR analysis and also NCBI organizations for allowing the deposit and access of *Cryptosporidium spp* DNA sequences referred to this study (http://www.ncbi.nlm.nih.gov). These wavelet results were obtained in the frame of Romanian National Authority for Scientific Research "PARTNERSHIP IN PRIORITY DOMAINS" Programme Contract nr. 184/2012 "MOIST".

References

[1] A. Gherasim, M. Lebbad, M. Insulander, V. Decraene, A. Kling, M. Hjertqvist, A. Wallensten, Euro Surveill., **15**(17), 46 (2012).
[2] G. Widmer, Y. Lee, Appl. Environ. Microbiol, **76**(19), 6639 (2010).
[3] I. M. Neagoe, D. Popescu, V. I. R. Niculescu, Romanian Reports in Physics, **66**, 3 (2014).
[4] I. M. Neagoe, D. Popescu, V. I. R. Niculescu, Romanian Reports in Physics, **66**, 4 (2014).
[5] R. Bunghez, O. Dumitrescu, E. Vasile, S. Doncea, R. M. Ion, J Optoelectron Adv Mater, **15**(11-12), 1380 (2013).
[6] A. J. Vilella, A. Blanco-Garcia, S. Hutter, J. Rozas, Bioinformatics Applications Note, **21**(11), 2791 (2005).
[7] I. M. Neagoe, S. Miclos, D. Popescu, D. Savastru, D. Steriu, S. Dontu, V. I. R. Niculescu, M. Tautan, J Optoelectron Adv Mater, **8**(3–4), 408 (2014).
[8] Y. Hur, H. Lee, BMC Bioinformatics, **12**, 146 (2011).
[9] L. Wang, L. D. Stein, BMC Bioinformatics, **11**, 550 (2010).
[10] M. Brandstrom, A. T. Bagshaw, N. J. Gemmell, H. Ellegren, Mol. Biol. Evol., **25**(12), 2579 (2008).
[11] C. C. A. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myres, B. Silverman, P. Donnelly, D. Bentley, G. McVean, PloS Genetics, **2**(9), 1375 (2006).
[12] J. A. T. Machado, A.. C. Costa, M. D. Quelhas, Genomics, **98**, 155-163 (2011).
[13] D. Baleanu, Physics and Technology, Chapter **15**, 353-371 (2012).
[14] V. I. R. Niculescu, V. Babin, M. Dan, J. Optoelectron. Adv. Mater. **4**(4), 971 (2002).
[15] I. Gruia, S. B. Yermolenko, M. I. Gruia, P. V. Ivashko, T. Stefănescu, J. Optoelectron. Adv. Mater.-Rapid Comm. **4**(4), 523 (2010).
[16] S. B. Yermolenko P. V. Ivashko, A. Prydiy, I. Gruia, J. Optoelectron. Adv. Mater.-Rapid Comm. **4**(4), 527(2010).
[17] http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html
[18] G. D. Sturbaum, D. A. Schaefer, B. H. Jost, C. R. Sterling, M. W. Riggs, Mol Biochem Parasitol, **159**(2), 138 (2008).
[19] R. M. Chalmers, R. P. Smith, S. J. Hadfield, K. Elwin, M. Giles, Parasitol Res. **108**(5), 1321 (2011).
[20] A. M. Costa, J. T. Machado, M. D. Quelhas, Bioinformatics, **27**(9), 1207 (2011).
[21] http://www.ece.rice.edu/~dhj/resistor.pdf.

*Corresponding author: miclos@inoe.ro