DNA structural information from *Giardia intestinalis* tpi gene assemblages using the wavelet spectrogram analysis

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

[a]*Department of Parasitology, "Carol Davila" University of Medicine and Pharmacy, D.Gerota 19-21, Bucharest, Romania*
[b]*National Institute of Research and Development 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]*Department of Mathematical Modelling in Life Sciences, Institute of Mathematical Statistics and Applied Mathematics of Romanian Academy, Calea 13 Septembrie 13, Bucharest, Romania*
[e]*Department of Physiology and Biophysics, 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*

Molecular investigations of *Giardia intestinalis tpi* polymorphic gene have facilitated approaching of some mathematical methods to describe the genetic information. In this study, we applied wavelet spectrogram analysis for estimating the degree of similarity and difference between different assemblages and intra-assemblages of infectious parasite *Giardia intestinalis*. Five different DNA sequences were investigated, of which two sample sequences were isolated, sequenced and analyzed in our laboratory and subsequently deposited in GenBank. Using additional a quantitative comparison index and Multidimensional Scaling as visualization method we revealed some characteristics of DNA information embedded in the analyzed *tpi* sequences.

(Received December 5, 2013; accepted March 13, 2014)

Keywords: Wavelet spectrogram analysis, Multidimensional Scaling, DNA sequence, *Giardia tpi* gene; genetic assemblages

1. Introduction

Giardia intestinalis is one of the most ubiquitous enteric parasites that may cause an infectious diarrhea syndrome mainly in humans. With the increasing use of genetic sequencing analysis in the pathogenesis and epidemiology studies of this infectious protozoan a large volume of information can be available to computational processing for DNA (deoxyribonucleic acid) research. *Tpi* gene is one of the most investigated genes of *Giardia intestinalis* genome because it is a phylogenetic marker with a high degree of polymorphism [1] which codes for glycolytic enzyme triosephosphate isomerase involved in the parasitic energetic metabolism [2].

In recent years, a number of mathematical methods for DNA study have been reported and some of them are based on wavelet transform [3,4,15,16]. Wavelets transform (WT) has received much attention in the literature due to its practical contributions [5,6,12]. These theoretical methods add to optical spectroscopy as a means of studying the properties of DNA biopolymer [12-14]. In the study of DNA, the wavelet spectrogram allows obtaining the relevant information from the analyzed data and could be valuable in capturing global and local characteristics of DNA structure [7].

DNA is a long polymer containing millions of nucleotides (the monomer). The structure of DNA of all species comprises two helical chains each coiled round the same axis.

Several wavelet models have been adapted depending on what it was intended to describe or to detect in the DNA structure [3,4,7,8]. But from the perspective of comparing similarities between different DNA structural information, real Shannon wavelet was considered by far the best wavelet function to represent the complex patterns emerging after wavelet transformation of DNA information [8]. This type of continuous wavelet transform was chosen as a strategy to reveal some important properties of the structure of DNA and to demonstrate the effective as a mathematical tool for description of DNA information [8]. Also, the statistical analysis have used for DNA study which is a complex informational biopolymer [12–14].

In the present paper we approached some wavelet descriptions to analyze the polymorphic DNA information embedding in the coding *tpi* gene from the genome of the

parasite *Giardia intestinalis*. Two genetically distinct lineages of this parasitic protozoan also designated as assemblages A and B known to infect humans were investigated applying methods proposed by Machado et al.[8]. Our motivation was to assess the intra and inter-assemblages differences of *Giardia intestinalis* at both qualitative by WT and quantitative level by using a comparative index [8] and the visualization tool Multidimensional Scaling (MDS) [9].

2. Experimental

2.1 *Giardia intestinalis tpi* gene data

For processing the *tpi* gene DNA sequences we take into account the constitution alphabet of DNA double helix which is represented by four types of nucleotides that differ in their nitrogenous bases adenine (A), cytosine (C), guanine (G) and thymine (T). Each type of base is interconnected with only one type of base between the two strands of the DNA forming the base pairing (bp) A-T and G-C.

As initial models for signal processing we accessed two reference DNA sequences of *tpi* gene belonging to the genetic assemblages A and B of *Giardia intestinalis*. They are retrieved from public database GenBank under accession numbers: AY368161 and AY368163 respectively [1]. The inter-assemblages comparisons were performed using two sample DNA sequences isolated from the stool of two patients infected with *Giardia intestinalis.* Approximately 500bp of *tpi* gene from DNA isolates was sequenced from purified nested PCR (Polymerase chain reaction) products [1], compared with published sequences on GenBank database using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) and assigned to *tpi* assemblages A and B based on sequence variation, presence and position of single nucleotide polymorphisms (SNPs) [10]. The sample sequences of *Giardia intestinalis* partial *tpi* gene used in this study were deposited under accession number HG792784 and HG792785 respectively.

For a better comparison by wavelet processing of DNA information, five DNA sequences (in FASTA format) were analyzed. Each one received an identifier composed of three letters and a number. First letter is G (from *Giardia*), the second is R for a reference sequence or S for a sample sequence. The last letter, A or B, is assigned to the genetic assemblage type of *Giardia intestinalis*. The number represents the length of the sequence. There are two reference DNA sequences: GRA532 (Fig. 1a) and GRB532 (Fig. 1b), both being complete (n = 532 nucleotides) and two samples: GSA532 (Fig. 1d) and GSB480 (Fig. 1e). The sample sequence GSA532 is 532 nucleotides long and differs of reference GRA532 by three SNPs (at nucleotide positions 27, 79 and 96 according to the GenBank accession number L02116). The sample GSB480 is a slight variation of the reference GRB532, but it is shorter and has five SNPs (at nucleotide positions 111, 216, 271, 471 and 483 according to the GenBank sequence L02116). The DNA sequence supplementary introduced for wavelet analysis is another reference, GRB480 (Fig. 1c), which is an excerpt of reference GRB532, starting with position 32 and having a length of 480 nucleotides.

Table 1. Composition of the five analyzed DNA sequences.

	GRA532	GSA532	GRB532	GRB480	GSB480
Adenine	131	129	138	128	129
Cytosine	130	131	120	105	106
Guanine	167	168	156	139	139
Thymine	104	104	118	108	106
Total	**532**	**532**	**532**	**480**	**480**

2.2 DNA encoding and data processing

DNA information of the five sequences assigned to wavelet processing was converted using a translation scheme [8] which transforms the four nitrogenous bases {A, C, G, T} into complex numbers, as it follows: A = 1+0i, C = −1+0i, T =0 +i and G = 0−i, where $i = \sqrt{-1}$.

Each DNA sequence is described, after the above mentioned conversion, by a complex "temporal" signal $x(t)$, the "time" t being the running number of the DNA sequence. Applying a continuous wavelet transform [6] to this signal, it will result:

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

Here the parameters s (s > 0) and τ represent the dyadic dilation (scale parameter), respectively the dyadic position (translation parameter), and ψ is a function called the

mother wavelet. The symbol * denotes the complex conjugate. The best results were obtained using as mother wavelet the real Shannon wavelet [8]:

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

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 are also in the same range. Finally, the integral turns into a sum:

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

After obtaining wavelet charts for the given values of n and set of five DNA sequences, the second step of our analysis is to evaluate the similarities and difference between them. The charts were normalized for coordinates s and τ in order to compare DNA sequences with different lengths. So, $\bar{s} = s / s_{\max}$ and $\bar{\tau} = \tau / \tau_{\max}$.

The "distance" between two data sequences (i and j) may be estimated using the measure r_{ij} as comparison index [8]:

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

where μ represents the average of the data sequence, $\sigma_{\bar{s}}$ – standard deviation on direction $\bar{s}$, $\sigma_{\bar{\tau}}$ – standard deviation on direction $\bar{\tau}$; i and $j = 1 \ldots m$ where $m = 5$ (all the analyzed DNA sequences).

The third step in the analysis consists in exposure embedded patterns in the correlation matrix data. The symmetrical correlation matrix R_{5x5} for comparing all five DNA sequences was constructed on the basis of the measure r_{ij} values. For this propose we approach the MDS (Multidimensional scaling) method 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 [11]. 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 five DNA sequences extracted by wavelet spectrogram analysis

The charts of the DNA sequences processed by Shannon wavelet (Fig. 1) show visually the differences and the similarities between the DNA sequences. All charts represent normalized absolute values of the wavelets for each analyzed DNA sequence with the conversion of $\bar{s}$ and $\bar{\tau}$ values into the interval [0,1].

Wavelet analysis successfully captured temporal distinctions between the two different genetic assemblages types A and B of *Giardia intestinalis*. Comparing wavelet complex patterns of the references GRA532 and GRB532 (Fig. 1a and 1b) is clearly revealed the huge difference between them. A local proof in DNA structure shows that GRA532 and GRB532 sequences differ in 103 of the 532 points. A high difference is also seen between wavelet complex patterns of the samples GSA532 and GSB480 (Fig. 1d and 1e). Although the local evidence in DNA structure reveals a smaller difference between them (98 points in 532) than references GRA532 and GRB532. In the Fig.1c and 1e, wavelet charts of the same reference GRB but with different lengths show a high resemblance. Apparently, they are slightly different in the range of normalized wavelet values [0.4 – 1] 10^{-5}. As a result of comparison between reference (GRA532 or GRB480) and sample (GSA532 or GSB480) wavelet models appropriate for each type of assembly, overall many similarities were observed. Minor differences are recorded only at higher wavelet levels [0.5-0.9] 10^{-5} according to the values of $\bar{\tau}$, (Fig. 1a with 1d and Fig. 1c with 1e). These observations are in agreement with the records of the SNPs positions in DNA structure of GRB480 and GSB480 which differs only in 5 of the 480 points while GRA532 and GSA532 differs only in 3 of the 532 points.

Overall we observe that the wavelet charts are different from each other and a relative similarity is detected only by a few small items corresponding to low values of $\bar{\tau}$. GSA532 and GRA532 patterns are characterized by 6 and 5 distinct items for the average values of $\bar{\tau}$ while GRA and GRB reveals several levels according to $\bar{\tau}$. However, we did not find any correlation between the positions of SNPs in the analyzed DNA sequences and the distinct items located in the center of the wavelet charts. From the point of view of continuous wavelet analysis, data capture and their representation in the form of complex patterns was a potential way for qualitative characterization of these sequences based on the similarities or differences between them. But, for a quantitative characterization, a visual estimation is not sufficient and some statistic processing of the data can be easily done.

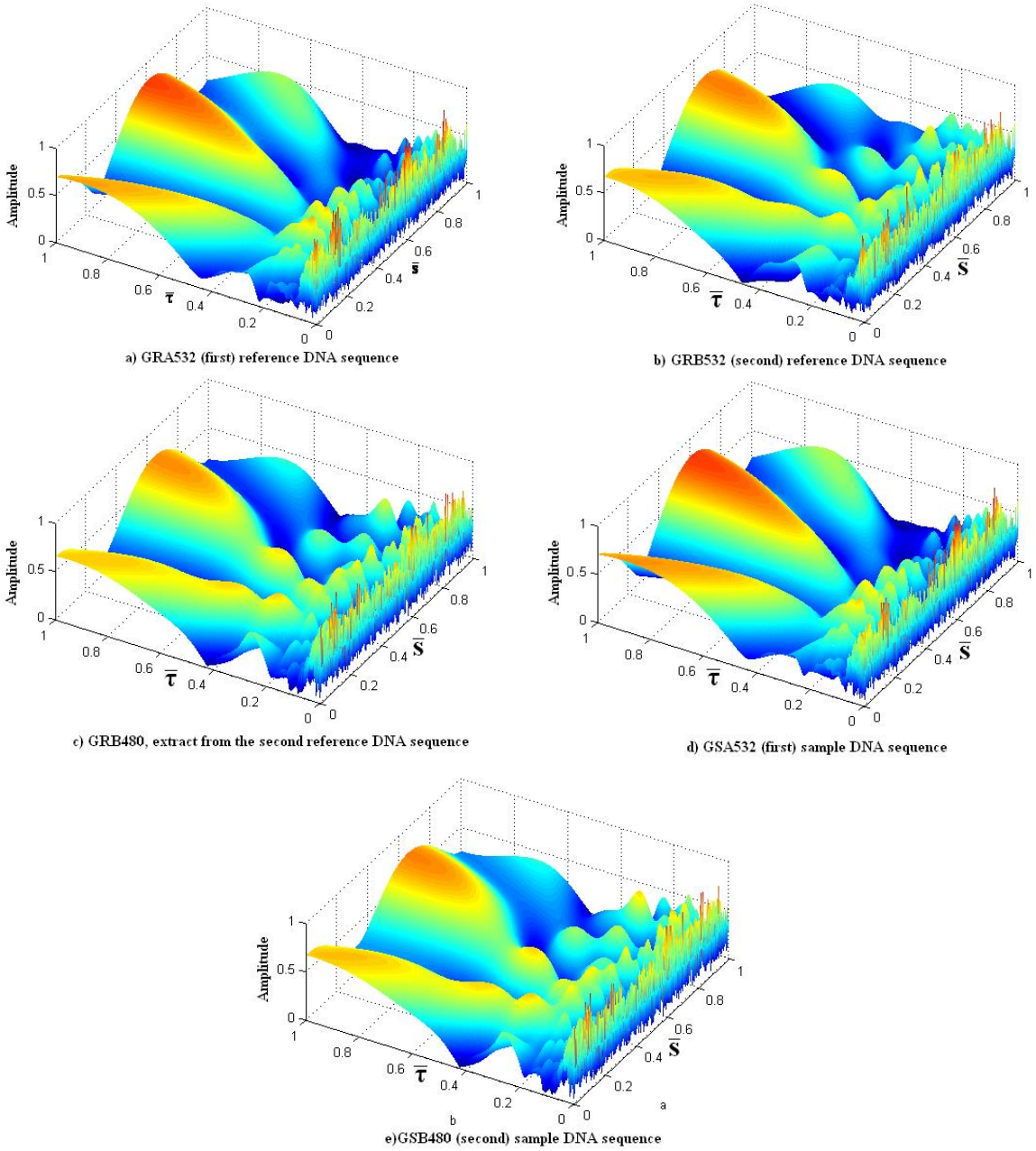

Fig. 1. The wavelet spectrograms for the five selected DNA sequences.

3.2 Mathematically estimation of distances between the five wavelet patterns using r_{ij} index

As shown in the Table 2, the symmetric matrix R_{5x5} was calculated on the r_{ij} values for all possible pairs of sequences i and j in terms of normalization WT of each DNA sequence analyzed. The measure (r_{ij}) gives a mathematical support to the visual estimation of Shannon wavelet charts. Besides the slight difference (less than 0.6 % of points) between sequences GRA532 and GSA532 ($r_{14} = r_{41} = 0.47492 \cdot 10^{-4}$) and that between sequences GRB480 and GSB480 (less than 1.1 % of points, $r_{25} = r_{52} = 0.28717 \cdot 10^{-4}$), the large difference (almost 19.5 % of points) between sequences GRA532 and GRB532 ($r_{13} = r_{31} = 2.10879 \cdot 10^{-4}$) certifies the visual estimation given by Shannon wavelet charts.

Larger differences occur when the sequences have different lengths. Sequences GRB532 and GRB480 differ only by length (GRB480 is a subset of GRB532), but this difference (9.8 %) is substantial ($r_{35} = r_{53} = 2.98293 \cdot 10^{-4}$).

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

R_{5x5}	GRA532	GRB480	GRB532	GSA532	GSB480
GRA532	0	3.20600	2.10879	0.47492	3.17285
GRB480	3.20600	0	2.98089	3.34014	0.28717
GRB532	2.10879	2.98089	0	2.30410	2.98293
GSA532	0.47492	3.34014	2.30410	0	3.26724
GSB480	3.17285	0.28717	2.98293	3.26724	0

3.3 Features of the five DNA structure mapping by MDS tool

MDS is an alternative approach in the perspective of the Shannon wavelet transform and the r_{ij} index. Fig. 2 illustrates the same similarities or differences between the five DNA sequences as those highlighted by the r_{ij} index. In this way are marked the inter-assemblages distances and intra-assemblages groups. In terms of mathematical calculations performed and visualized by MDS at intra-assemblages groups, although local in the analyzed sequences, the number of SNPs is greater between GRB480 and GSB480, overall at DNA structure the two sequences compared are more similar than sequences GSA532 and GRA532 that differ between by a smaller number of SNPs. At the level of inter-assemblages, greatest distance is observed between GSA532 and GRB480. Our results support the influence of the three important sequence parameters in this mathematical analysis by comparing: the length of the structure, the number of SNPs and especially the position or distance between SNPs of analyzed structure.

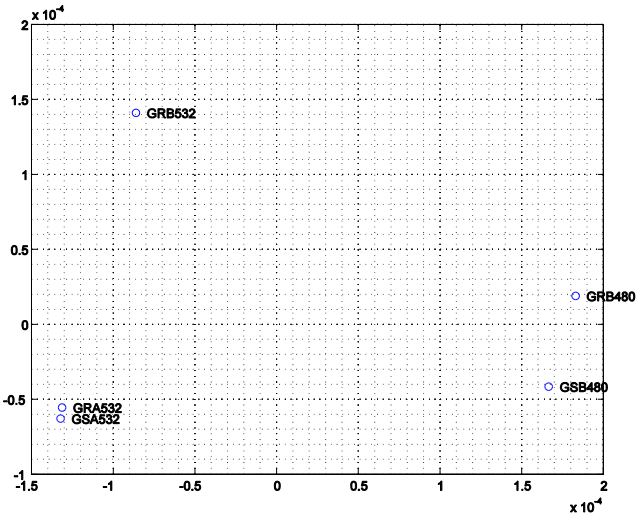

Fig. 2. MDS chart for the five considered DNA sequences.

4. Conclusions

In summary, the use of real Shannon wavelet to capture information from the five analyzed DNA sequences allows obtaining patterns that can be compared and interpreted to describe information. All the inter-assemblage comparisons of normalized wavelet values obtained for different references and samples used in this study, confirm the wide differences in the structure of DNA information between the two different genetic assemblages types A and B of *Giardia intestinalis*. The complex pattern of wavelet transformation is strongly dependent on the choice of sequence length and consequently there is an addiction to different maximum values of $\bar{s}$ and $\bar{\tau}$. The intra-assemblages comparisons between wavelet patterns belonging to the references and samples with the same length of each genetic assemblage type exhibited a much higher degree of similarity than the inter-assemblage comparisons. The length of the sequence is a factor which influences both the wavelet pattern and especially the quantitative index r_{ij}. MDS as visualization tool creates a hierarchy of differences between sequences analyzed according to distance scales between them.

On the other hand, wavelet analysis as a way of describing and association of captured DNA information facilitates new research directions for obtaining mathematical data by which we can make predictions and assessments of results in relation to other significant parameters of the parasite *Giardia intestinalis*.

Acknowledgments

We thank the Molecular Epidemiology laboratory of Cantacuzino Institute, Bucharest, Romania which has provided us some equipment for PCR analysis and also NCBI organizations for allowing the access and deposit of *Giardia intestinalis* DNA sequences referred to this study (http://www.ncbi.nlm.nih.gov). These 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] I. M. Sulaiman, J. Jiang, A. Singh, L. Xiao, Applied and Environmental Microbiology, **70**(6), 3776 (2004).
[2] G. Hernanadez-Alcantara, A. Torres-Larios, S. Enriquez-Flores, I. Garcia-Torres, A. Castillo-Villaneuva, S. T. Mendez, I. de la Mora-de la Mora, S. Gomez-Manzo, A. Torres-Arroyo,G. Lopez-Velazquez, H. Reyes-Vivas, J. Oria-Hernandez, PLOS-ONE, **8**, 7 (2013).
[3] Y. Hur, H. Lee, BMC Bioinformatics, **12**, 146 (2011).

[4] L. Wang, L. D. Stein, BMC Bioinformatics, **11**, 550 (2010).
[5] M. Kumar, S. Pandit, International Journal of Nonlinear Science, **13**, 3, 325 (2012).
[6] D. Baleanu, Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology, Chapter **15**, 353 (2012).
[7] A. J. Vilella, A. Blanco-Garcia, S. Hutter, J. Rozas, Bioinformatics Applications Note, **21**(11), 2791 (2005).
[8] J. A. T. Machado, A. C. Costa, MD Quelhas, Genomics, **98**, 155 (2011).
[9] http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html.
[10] C. M. Wielinga, R. C. A. Thompson, Parasitology, **134**, 1795 (2007).
[11] A. M. Costa, J. T. Machado, M. D. Quelhas, Bioinformaics, **27**(9), 1207 (2011).
[12] V. I. R. Niculescu, V. Babin, M. Dan, J. Optoelectron. Adv. Mater. **4**(4), 971 (2002).
[13] I. Gruia, S. B. Yermolenko, M. I. Gruia, P. V. Ivashko, T. Stefănescu, J. Optoelectron. Adv. Mater.-Rapid Comm. **4**(4), 523 (2010).
[14] S. B. Yermolenko P. V. Ivashko, A. Prydiy, I. Gruia, J. Optoelectron. Adv. Mater.-Rapid Comm. **4**(4), 527 (2010).
[15] I. M. Neagoe, D. Popescu, V. I. R. Niculescu, Alternative methods for statistical characterization and quantification *of Cryptosporidium spp*. Gp60 gene variability, Romanian Reports in Physics, **66**, 3, (2014).
[16] I. M. Neagoe, D. Popescu, V.I.R. Niculescu, Applications of entropic divergence measures for DNA segmentation into high variable regions of *Cryptosporidium spp.* Gp60 gene, Romanian Reports in Physics, **66**, 4, (2014).

*Corresponding author: miclos@inoe.ro