A SENSITIVE METHOD FOR DETECTING DINUCLEOTIDE ISLANDS AND CLUSTERS THROUGH DEPTH ANALYSIS

Paul Gagniuc [1]***, Dănuţ Cimponeriu*** [1]***, Nicolae Mircea Panduru*** [2]***, Monica Stavarachi*** [1]***,***
Mihai Toma [1]***, Constantin Ionescu-Tîrgovişte*** [3]***, Lucian Gavrilă*** [1]

[1] Human Genetics and Molecular Diagnosis Laboratory, Department of Genetics, University of Bucharest, Romania
[2] Department of Pathophysiology, “Carol Davila” University of Medicine and Pharmacy from Bucharest, Romania
[3] National Institute of Diabetes, Nutrition and Metabolic Diseases, Romania

Abstract

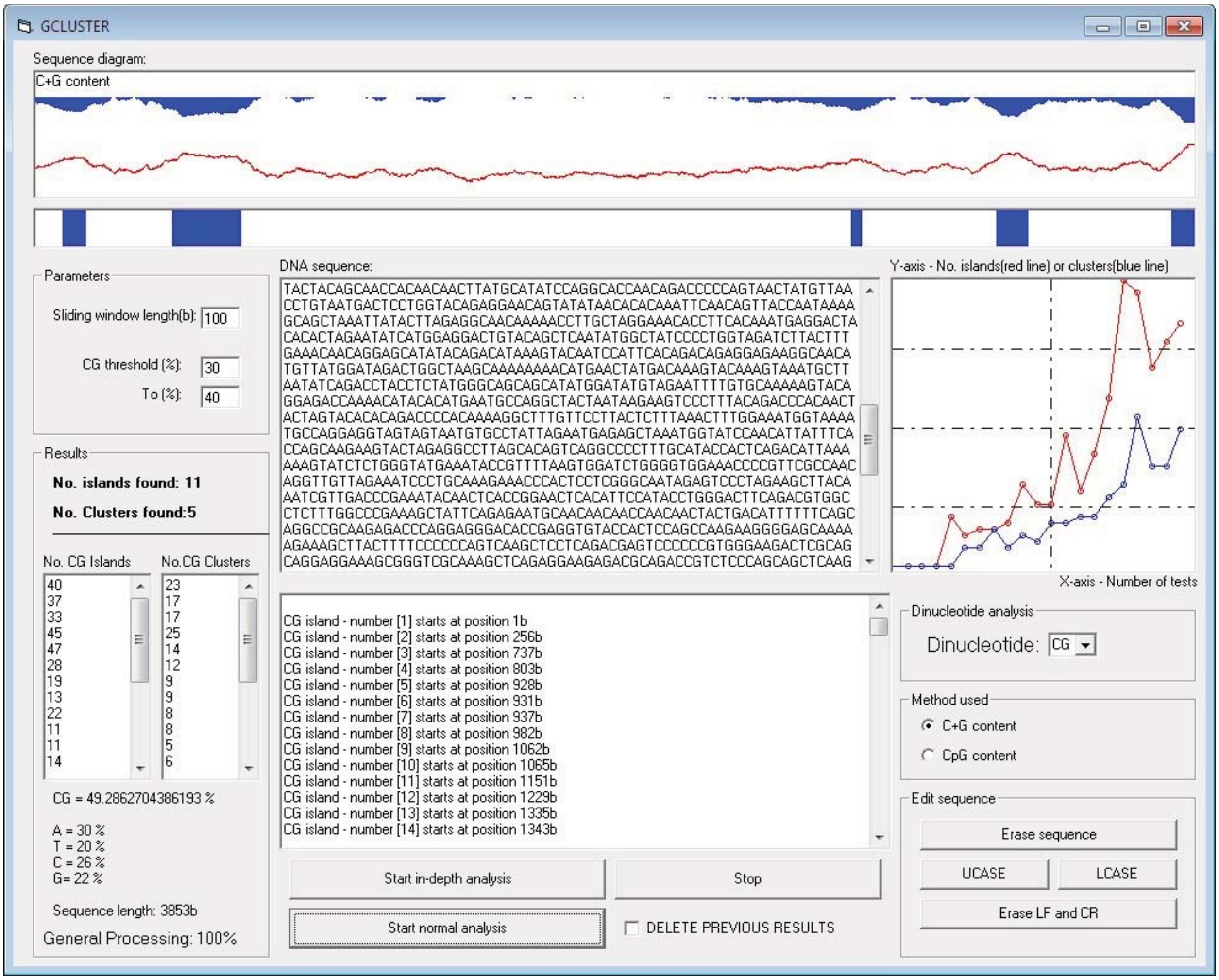
The field of bioinformatics is an essential asset for modern biology. In recent years, after the appearance of GWS (Genome Wide Scan) studies, powerful bioinformatics methods have been developed. In order to understand the genetic basis of various diseases, especially polygenic diseases (diabetes, obesity and vascular disease), we have implemented a dynamic method named "in-depth analysis" to detect and interpret CpG islands, CpG clusters and other dinucleotide structures. In-depth analysis is made through repeated tests with different dinucleotide thresholds. GCLUSTER is our design for "in-depth analysis". We tested GCLUSTER with randomly generated DNA sequences, multiple genes from Homo sapiens and several types of viral genomes.

key words*: nucleotide repeats; dinucleotide; CpG Clusters; CpG Islands; DNA Computation.*

Introduction

In recent years we have witnessed a tremendous increase of studies and algorithms in many bio-related fields. Diabetes, obesity and metabolic syndrome are complex metabolic disorders whose genetic bases are suggested by their heritability. In recent years, 36 loci were identified with type 2 diabetes and some of them with both obesity and type 2 diabetes [1, 2]. Extracting useful information from unstructured data, such as the human genome, is not an easy task. Using text processing solutions we can acquire significant information on gene-gene or gene-protein interactions and their physiological function. Software tools can answer various questions that arise when taking a step forward in the field of genetics.

Many of these studies have been conducted on DNA sequences and particularly on dinucleotide structures, namely on CpG sites (Cytosine - phosphate - Guanine) [3, 4, 5]. CpG islands are genomic regions (at least 200 bp) which contain a high frequency of CpG sites (CpG ratio observed/expected higher than 60%). CpG islands (CGIs) also known as HTF islands (HpaII tiny fragments) [6] are preferentially located near the transcription start site (TSS) in the promoter region of housekeeping genes [7, 8, 9]. CpG islands methylation is important in human cancer research [10]. CpG sites methylation within the promoters of genes can lead to their low expression or their complete silencing. Unmethylated CpG sites near promoters lead to gene expression. A study made on Xenopus genome showed that shorter CGIs are functional [11]. Functional CGIs found in Xenopus genome are not only smaller but also have a lower G+C content. This proves that a clear definition of CGIs has yet to be debated. Small genes (up to 300b in size) are still to be discovered. Very small CpG clusters (CGCs) can highlight new genes that are not currently detected through conventional methods.



```
Parameters
Sliding window length(b): 100
CG threshold (%): 30
To (%): 40

Results
No. islands found: 11
No. Clusters found:5

No. CG Islands   No.CG Clusters
40               23
37               17
33               17
45               25
47               14
28               12
19               9
13               9
22               8
11               8
11               5
14               6

CG = 49.2862704386193 %

A = 30 %
T = 20 %
C = 26 %
G= 22 %

Sequence length: 3853b
General Processing: 100%

DNA sequence:
TACTACAGCAACCACAACAACTTATGCATATCCAGGCACCAACAGACCCCCAGTAACTATGTTAA
CCTGTAATGACTCCTGGTACAGAGGAACAGTATATAACACACAAATTCAACAGTTACCAATAAAA
GCAGCTAAATTATACTTAGAGGCAACAAAAACCTTGCTAGGAAACACCTTCACAAATGAGGACTA
CACACTAGAATATCATGGAGGACTGTACAGCTCAATATGGCTATCCCCTGGTAGATCTTACTTT
GAAACAACAGGAGCATATACAGACATAAAGTACAATCCATTCACAGACAGAGGAGAAGGCAACA
TGTTATGGATAGACTGGCTAAGCAAAAAAAACATGAACTATGACAAAGTACAAAGTAAATGCTT
AATATCAGACCTACCTCTATGGGCAGCAGCATATGGATATGTAGAATTTTGTGCAAAAAGTACA
GGAGACCAAAACATACACATGAATGCCAGGCTACTAATAAGAAGTCCCTTTACAGACCCACAACT
ACTAGTACACACAGACCCCACAAAAGGCTTTGTTCCTTACTCTTTAAACTTTGGAAATGGTAAAA
TGCCAGGAGGTAGTAGTAATGTGCCTATTAGAATGAGAGCTAAATGGTATCCAACATTATTTCA
CCAGCAAGAAGTACTAGAGGCCTTAGCACAGTCAGGCCCCTTTGCATACCACTCAGACATTAAA
AAAGTATCTCTGGGTATGAAATACCGTTTTAAGTGGATCTGGGGTGGAAACCCCGTTCGCCAAC
AGGTTGTTAGAAATCCCTGCAAAGAAACCCACTCCTCGGGCAATAGAGTCCCTAGAAGCTTACA
AATCGTTGACCCGAAATACAACTCACCGGAACTCACATTCCATACCTGGGACTTCAGACGTGGC
CTCTTTGGCCCGAAAGCTATTCAGAGAATGCAACAACAACCAACAACTACTGACATTTTTTCAGC
AGGCCGCAAGAGACCCAGGAGGGACACCGAGGTGTACCACTCCAGCCAAGAAGGGGAGCAAAA
AGAAAGCTTACTTTTCCCCCCAGTCAAGCTCCTCAGACGAGTCCCCCCGTGGGAAGACTCGCAG
CAGGAGGAAAGCGGGTCGCAAAGCTCAGAGGAAGAGACGCAGACCGTCTCCCAGCAGCTCAAG

CG island - number [1] starts at position 1b
CG island - number [2] starts at position 256b
CG island - number [3] starts at position 737b
CG island - number [4] starts at position 803b
CG island - number [5] starts at position 928b
CG island - number [6] starts at position 931b
CG island - number [7] starts at position 937b
CG island - number [8] starts at position 982b
CG island - number [9] starts at position 1062b
CG island - number [10] starts at position 1065b
CG island - number [11] starts at position 1151b
CG island - number [12] starts at position 1229b
CG island - number [13] starts at position 1335b
CG island - number [14] starts at position 1343b

Dinucleotide analysis
Dinucleotide: CG

Method used
● C+G content
○ CpG content

Edit sequence

□ DELETE PREVIOUS RESULTS
```

Figure 1. GCLUSTER program. The figure shows a screenshot of GCLUSTER program analyzing TorqueTenoVirus genome. The chart at the top shows either dinucleotide or nucleotide frequencies. The graph on the right shows the number of tests performed (x-axis) and the number of CpG islands and CpG clusters found for each test (y-axis).

Other types of dinucleotide islands may also be of interest for gene predictions. There are several applications and methods developed for locating CpG islands (CGIs) in DNA sequences, including EMBOSS CpGPlot [12], CpGProD (CpG Island Promoter Detection) [13], CpGIS [14], CpGIE [15], CpGcluster [16] or CpG MI [17]. Most of those methods rely on CGIs predetermined thresholds (ie. length, CpG Obs/Exp ratio, G+C content) [18]. All definitions of CGIs rely on ad hoc thresholds. We use several thresholds to elucidate CGIs. The evolutionary dynamics provided fault safe mechanisms in mammalian genes. DNA transcription starts near multiple alternative start sites after a CpG island region [19]. The same fault safe mechanism is observed for CGIs and CGCs positions. CG content varies continuously and CGIs decay or renew due to point mutations and selection pressure [20, 21, 22] both in the eukaryotic genomes and in smaller sequences like virus genomes [23].

Material and Method

We downloaded the assembled human genome (human build 37), several genes and viral (i.e. Human Immunodeficiency Virus, Torque Teno Virus genomes) sequences from NCBI database. In this study we examined the CGIs and CGCs that lie within gene regions. We used sliding window techniques and dynamic physical distances between two neighboring CpGs to detect GC clusters. Figure 1 shows the GCLUSTER program. It uses depth steps (repeated tests with different parameters) to elucidate the positions of CpG islands and CpG clusters.

The program requires two initial parameters. The first parameter is the sliding window length. The second parameter is the CG content used as a threshold (expressed in percentage).

GCLUSTER can analyze DNA sequences in two ways. The first type of test is the normal analysis of CGIs which involves setting a single CG threshold. The second type of test is the in-depth analysis involving several CG thresholds. For depth analysis, CG content thresholds are automatically incremented by the algorithm and the results are plotted on a graph. CpG sites are not a defining factor for GCLUSTER, the analysis can be performed on all dinucleotide combinations. The user can choose in depth analysis on dinucleotide frequencies or C+G percentage. GCLUSTER was tested on a computer equipped with a 2.8GHz processor, 500MB RAM and 80GB HDD.

Figure 2 shows extensive tests conducted for ten genes from Homo sapiens - GRCh37 primary reference assembly and viral genomes like Torque Teno Virus and Human Immunodeficiency Virus (HIV-1) (all downloaded from the NCBI FTP servers).

For all tests we used a sliding window of 100b in length. In depth analysis was performed for step thresholds from 1% up to 40% for HIV-1 and 10% up to 40% for GRCh37 primary reference assembly genes. The Y-axis maximum value for each diagram is recalculated according to the maximum number of CpG islands found.

GCLUSTER program has been written in Visual Basic. It runs on all Windows operating systems and does not require installation. The package size is 1.77Mb and the memory requirements are between 1.2Mb and 1.5Mb (Windows 7 and Windows XP). GCLUSTER can analyze DNA sequences up to 500kb. On average, GCLUSTER scan speed is 1.5Kb/s (i.e. CRHR2 gene of 30Kb is completely scanned in 20 seconds). Another feature worth mentioning is the GCLUSTER interface that makes a dynamic correlation between the diagram and the sequence. By moving the mouse over the diagram, GCLUSTER selects the appropriate plain text sequence. The aim of the project is to act like a platform for other future applications intended for different types of studies on nucleic acids.

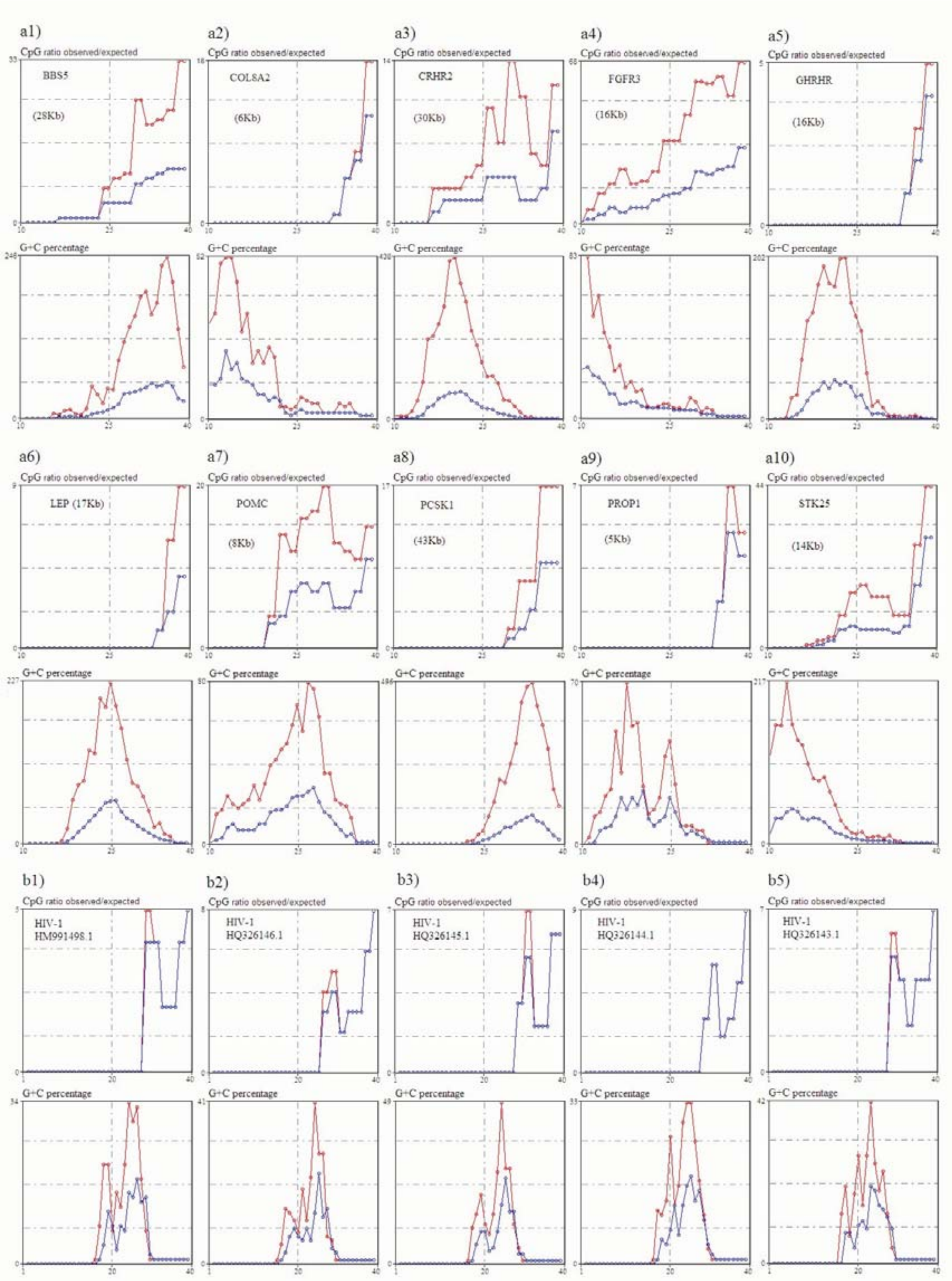


Figure 2. CpG Islands vs. CpG Clusters. The figure shows ten genes from Homo sapiens - GRCh37 primary reference assembly: BBS5, COL8A2, CRHR2, FGFR3, GHRHR, LEP, POMC, PCSK1, PROP1, STK25 (a1 - a10) and five different HIV-1 genomes (b1 - b5). a1) - a10) Each graph shows the number of tests performed (x-axis) and the number of CpG islands (red line) and CpG clusters (blue line) found for each test (y-axis). The step threshold used is 10% to 40%. b1) HIV-1 isolate SC24-40 envelope glycoprotein gene (HM991498.1), b2) HIV-1 isolate GX84-59 envelope glycoprotein gene (HQ326146.1), b3) HIV-1 isolate GX79-7 envelope glycoprotein gene (HQ326145.1), b4) HIV-1 isolate GX75-20 envelope glycoprotein gene (HQ326144.1), b5) HIV-1 isolate GX45-57 envelope glycoprotein gene (HQ326143.1). The step threshold used is 1% to 40%.

Conclusions

Hidden information of genetic architectures of complex diseases as diabetes, metabolic syndrome and obesity can be very hard to find without bioinformatics tools. Here we propose this software, whose value will be evaluated by processing genetic sequences associated with diabetes, which can lead to a better understanding of disease pathogenesis.

In this study, we showed a new method to study dinucleotide structures. We focused on clarifying the true size of CpG clusters and CpG islands on a given DNA sequence. The test included a systematic comparison of genes from *Homo sapiens* and several viral genomes. We also presented our program called GCLUSTER which incorporates 'in depth analysis' implementation.

Acknowledgements. This paper is supported by the Sectoral Operational Programme Human Resources Development, financed from the European Social Fund and by the Romanian Government under the contract *number POSDRU/89/1.5/S/ 64109 and* by the Romanian National Authority for Scientific Research as a part of the PNCDI 2, contract number 42161/2008.

REFERENCES

1. **Li S, Zhao JH, Luan J, Langenberg C, et al**. Genetic predisposition to obesity leads to incrised risk of type 2 diabetes. *Diabetologia* 54: 776-782. 2011.

2. **Ramos E, Chen G, Shriner D, et al.** Replication of genome-wide association studies (GWAS) loci for fasting plasma glucose in African-Americans, *Diabetologia* 54: 783-788. 2011.

3. **Jabbari K, Bernardi G**. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 333: 143-149. 2004.

4. **Ballestar E, Paz MF, Valle L, et al.** Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. *EMBO J* 22: 6335-6345. 2003.

5. **Strathdee G, Simand A, Brown R.** Control of gene expression by CpG island methylation in normal cells. *Biochem Soc Trans* 32: 913-915. 2004.

6. **Cross SH, Bird AP.** CpG islands and genes. *Curr Opin Genet Dev* 5: 309-314. 1995.

7. **Ioshikhes IP, Zhang MQ.** Large-scale human promoter mapping using CpG islands. *Nat Genet* 26: 61-63. 2000.

8. **Ponger L, Duret L, Mouchiroud D.** Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* 11: 1854-1860. 2001.

9. **Saxonov S, Berg P, Brutlag DL**. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA* 103: 1412-1417. 2006.

10. **Teodoridis JM, Strathdee G, Plumb JA, Brown R**. CpG-island methylation and epigenetic control of resistance to chemotherapy, *Biochem Soc Trans* 32. 916-917. 2004.

11. **Stancheva I, El-Maarri O, Walter J, Niveleau A, Meehan RR.** DNA methylation at promoter regions regulates the timing of gene activation in Xenopus laevis embryos. *Dev Biol* 243: 155-165. 2002.

12. **Rice P, Longden I, Bleasby A.** EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16: 276-277. 2000.

13. **Ponger L, Mouchiroud D.** CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18: 631-633. 2002.

14. **Takai D, Jones PA**. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA* 99: 3740-3745. 2002.

15. **Wang Y, Leung FC.** An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, 20: 1170-1177. 2004.

16. Hackenberg M, Previti C, Luque-Escamilla PL et al. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 7: 446. 2006.

17. Su J, Zhang Y, Lv J, et al. CpG MI: a novel approach for identifying functional CpG islands in mammalian genomes, *Nucleic Acids Res* 38: e6. 2010.

18. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* 196: 261-282. 1987.

19. Kawaji H, Frith MC, Katayama S, et al. Dynamic usage of transcription start sites within core promoters. *Genome Biol* 7: R118. 2006.

20. Antequera F. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci*, 60:1647-1658. 2003.

21. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, *Proc Natl Acad Sci USA* 103: 1412-1417. 2006.

22. Zhao H, Li QZ, Zeng CQ et al. Neighboring-Nucleotide E_ects on the Mutation Patterns of the Rice Genome, *Genomics Proteomics Bioinformatics*. 3: 158-168. 2005.

23. Greenbaum BD, Levine AJ, Bhanot G, et al. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog* 4: e1000079. 2008.

Correspondence Data:
Paul Gagniuc
Department of Genetics, Bucharest, Portocalelor Street, no 1-3, zip code 060101.
e-mail: paulgagniuc@yahoo.com