DNA PATTERNS AND EVOLUTIONARY SIGNATURES OBTAINED THROUGH KAPPA INDEX OF COINCIDENCE

PAUL GAGNIUC [1] , PAUL DAN CRISTEA [2], RODICA TUDUCE[3],
CONSTANTIN IONESCU-TÎRGOVIŞTE[4], LUCIAN GAVRILĂ[5]

Key words: Data mining, Kappa, DNA signatures, DNA patterns.

First used in cryptography, Kappa Index of Coincidence (Kappa IC) is an unexplored approach in molecular genetics. We used Kappa Index of Coincidence to determine new patterns and evolutionary signatures in DNA sequences. For analyzing genetic data through Kappa Index of Coincidence, we have created an open source project named DNAKAPPA. Extensive tests were conducted for eighteen genes from Homo sapiens and twelve Human Immunodeficiency Virus strains. We obtained three types of patterns, namely: KappaIC/*C*+*G*%, KappaIC/TM and KappaIC/CpG Obs/Exp.

1. **INTRODUCTION**

The examination of DNA sequences is a research area of great importance. The field of bioinformatics has rapidly developed into an essential asset for modern biology and powerful bioinformatics methods have been developed. We present a new method to analyze and interpret biological data through Kappa Index of Coincidence. The Index of Coincidence (IC) is a statistical measure first described by William F. Friedman [1–4]. Index of Coincidence principle is based on letter frequency distributions.

We used Index of Coincidence to discover correlations between DNA sequences (*e.g.* genes, viral genomes or promoter sequences). The order of nucleotide molecules influences the overall stability of a DNA sequence. For instance, two DNA sequences with identical *C*+*G* content can have different

[1] Human Genome and Molecular Diagnosis Laboratory, Institute of Genetics, University of Bucharest, Romania; E-mail: paulgagniuc@yahoo.com
[2] Biomedical Engineering Center, "Politehnica" University of Bucharest, Spl. Independenței 313, 060042 Bucharest, Romania, phone: +40 21 316 9569; E-mail: pcristea@dsp.pub.ro
[3] Biomedical Engineering Center, "Politehnica" University of Bucharest, Spl. Independenței 313, 060042 Bucharest, Romania, phone: +40 21 316 9569; E-mail: trodica@dsp.pub.ro
[4] "N. Paulescu" National Institute of Diabetes, Nutrition and Metabolic Diseases, 5-7 Ion Movila Street, Bucharest, Romania; E-mail: cit@paulescu.ro
[5] Human Genome and Molecular Diagnosis Laboratory, Institute of Genetics, University of Bucharest, Romania; E-mail: lucian.gavrila69@yahoo.com

melting temperatures (T_M). By extracting Kappa Index of Coincidence (Kappa IC) and T_M from a sliding window we can measure the localized values along the sequence. Kappa Index of Coincidence plotted on a graph against other types of signals (*e.g.* T_M or *C*+*G*%) form a recognizable pattern. These patterns (Figure 1) may have important implications for molecular genetics (*e.g.* selection pressure determination or gene prediction).

The evolutionary dynamics provided fault safe mechanisms in mammalian genes which can be highlighted by Kappa IC. For instance, DNA transcription starts near multiple alternative start sites, usually after a *CpG* island [5]. *CpG* islands are regions which show a *CpG* (Cytosine-phosphate-Guanine) ratio greater than 60% and are found near approximately 40% of promoters from mammalian genes [6, 7].

The same fault safe mechanism is observed for *CpG* island positions. *CG* content vary continuously and *CpG* islands decay or renew due to point mutations and selection pressure [8–10], both in the eukaryotic genomes and in smaller sequences like virus genomes [11].

2. **MATERIAL AND METHODS**

We downloaded the assembled human genome (human build 37) and several viral sequences from NCBI database. We used sliding window techniques for reading four types of signals: Kappa IC, T_M , *C*+*G*% and $CpG_{Obs/Exp}$ ratio. The sliding window (fixed length of 30b) starts at the beginning of a DNA sequence. All nucleotides within the new sliding window are processed in order to generate each of the four signal types. Once processing is complete, the sliding window is moved down by an offset of one nucleotide. The processing of all nucleotides from the sliding window is repeated and the window continues to move down on the DNA sequence until it reaches the end.

We begin by reading each signal in order to plot the data. First, we determine the T_M (defined as the dissociation temperature of the primer/template duplex) value for each DNA sequence in the sliding window. We used Marmur-Schildkraut-Doty formula (shown below) for determining the melting temperature of nucleic acids, also used for calculation of T_M on PCR primers and hybridization probes in Polymerase Chain Reaction (PCR) processes [12–14]

$$T_M = 81.5 + 16.6 \times (\log_{10}[\mathrm{Na}^+]) + 0.41 \times (\mathrm{mol}\%C + G) - \frac{675}{N}.$$

In above equation, the sodium ion concentration $[\mathrm{Na}^+]$ is 0.05 M and *N* represents the length of a DNA sequence. The *CG* percentage is also calculated for

each sliding window: $CG_{content} = (C+G/A+T+C+G)\times 100$ and is compared with the Kappa Index of Coincidence. We use a color scheme that make *Kappa*/*C*+*G*% spots more visible. The red and blue colors represent threshold values where *C*+*G*% > *Kappa* is plotted in red color and *C*+*G*% < *Kappa* is plotted in blue color (Fig. 1, section a)).

The third type of signal extracted from each sliding window is $CpG_{Obs/Exp}$. The *CpG* count represents the number of *CG* dinucleotides in the sequence. The ratio of observed and expected [15] *CpG* dinucleotides is calculated according to

$$CpG_{Obs/Exp} = \frac{CpG \times N}{C \times G},$$

where *N* represents the length of a DNA sequence. The last type of signal extracted from the sliding window is Kappa IC. The formula for Kappa IC is shown below, where sequences *A* and *B* have the same length *N*. Only if an A_i nucleotide from sequence *A* matches the B_i correspondent from sequence B, then $\sum$ is incremented by 1.

$$KappaIC = \frac{\sum_{i=1}^{N}[A_i = B_i]}{N/C}.$$

With small changes, the same method for measuring the Index of Coincidence was applied for only one sequence, in which the sequence was actually compared with itself, as shown below in the algorithm implementation.

```
Function IC(ByVal s1 As String) As Variant
        max = Len(s1) - 1
        For u = 1 To max
                s2 = Mid(s1, u + 1)
                For i = 1 To Len(s2)
                        If Mid(s1, i, 1) = Mid(s2, i, 1) Then
                                count = count + 1
                        End If
                Next i
                total = total + (count / Len(s2) * 100)
                count = 0
        Next u
        IC = Round((total / max), 2)
End Function
```

From what we observed, an input sequence of any size, must contain at least two types of nucleotides in order to obtain the Kappa Index of Coincidence below

100% (ie. sequence “AAAAAAAA” will generate a Kappa IC value of 100%, while a sequence of the same size “AAATAAAA” will generate a Kappa IC value of 81.87%.). After determining each signal separately, (*C*+*G*)%, $CpG_{Obs/Exp}$ and T_M values are compared with Kappa IC on a graph.

3. **RESULTS**

There are many signal types that can be obtained through Kappa Index of Coincidence which have distinctive features compared with other signals like *C*+*G*% (Figure 2, section d). DNA sequences that differ only by a single nucleotide molecule show completely different results.

Figure 2, section d2, shows two sequences – *(CG)n(TC)*42*(T)(CG)n* and *(GC)n(G)(TC)*43*(CG)n* which generate a Kappa IC signal pattern radically different from other signals generated by (*G*+*C)*%. Furthermore, a deletion of thymine (T) and other three insertions (a guanine and a TpC dinucleotide) in the first sequence shows a phase shift in Kappa IC signal (Fig. 2, section d2)). Another proof for the high sensitivity of Kappa IC are the sequences *(CG)n(G)*43*(CG)n* and *(CG)n(G)*44*(CG)n* which differ by one nucleotide (Fig. 2, section d5)). In this case, the middle of the slope passes from a noisy signal to a clean one. Moreover, sequences *(GC)n(TC)*43*(GC)n* and *(GC)n(TG)*43*(GC)n* differ by two types of dinucleotide structures, namely TpC and TpG. Interestingly, even with these small differences the first sequence shows a clean signal whereas the second sequence shows a noisy signal (Fig. 2, section d3)).

We used our DNAKAPPA software to plot Kappa Index of Coincidence values against *G*+*C* percentage, the melting temperature (T_M) and $CpG_{Obs/Exp}$. Extensive tests were conducted for eighteen genes from *Homo sapiens* – GRCh37 primary reference assembly: ALMS1, BBS5, PROP1, SLC2A2, BBS7, COL8A2, PPP1R3A, POMC, CRHR2, FGFR3, PCSK1, LMNA, GHRHR, GPC1, LEP, HSD11B1, GPR35, H6PD (Figure 1) and viral genomes like *Human Immunodeficiency Virus* (Fig. 3 Section a)) (all downloaded from the NCBI FTP servers).

Figure 1, subsection a, shows a color scheme. If (*C*+*G)*% > *Kappa IC* then *C*+*G*% values are plotted in red color and if *C*+*G*% < *Kappa* then *Kappa IC* values are plotted in blue color. The right side of all *Kappa IC/(C*+*G)*% patterns contain GC-rich sequences while their left side contain AT-rich sequences. Figure 1, subsection b, shows Kappa IC on *y*-axis and T_M ratio on *x*-axis.

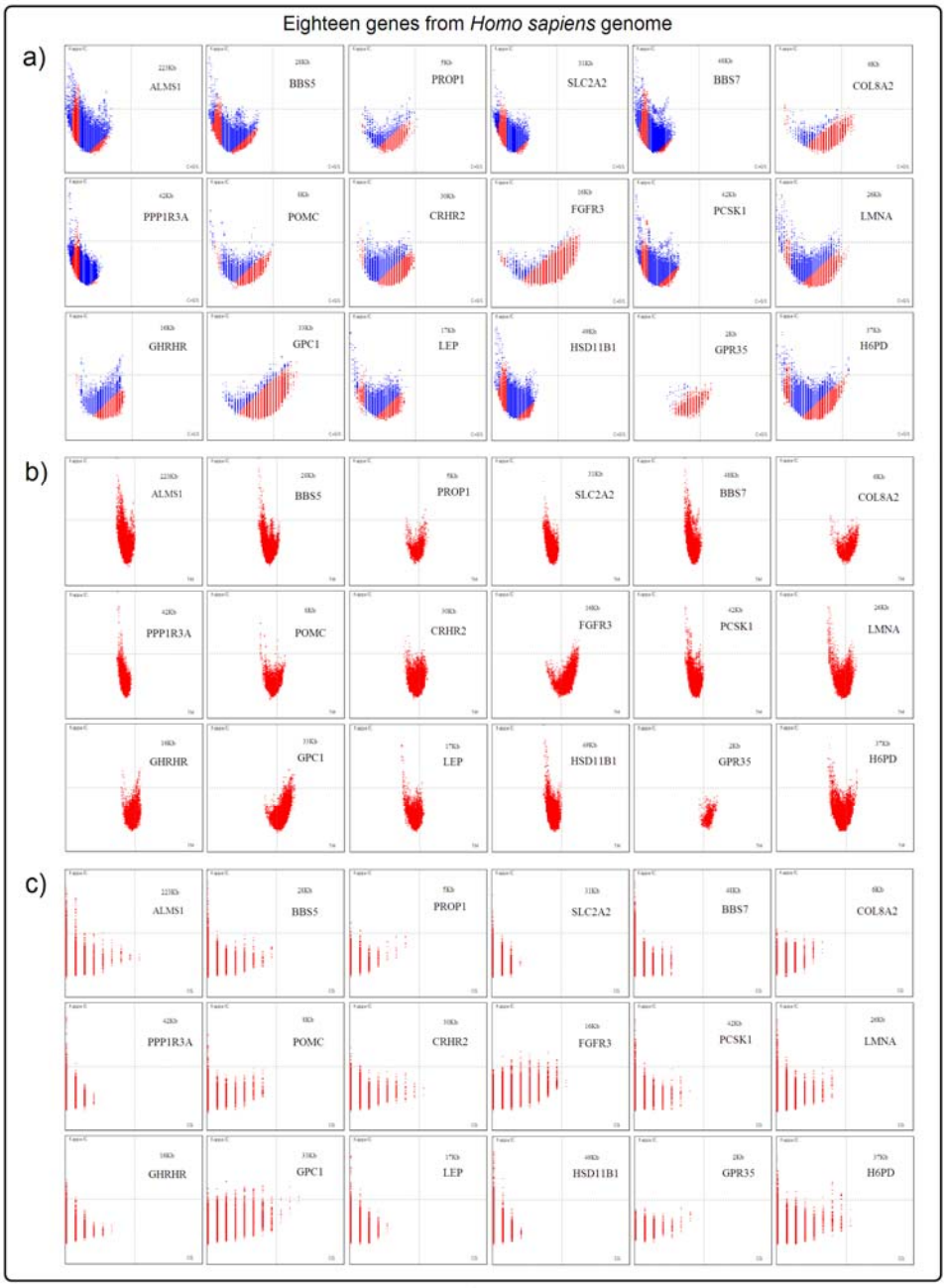

Fig. 1 – Three types of Kappa IC patterns. Eighteen genes from *Homo sapiens* (assembly GRCh37) were analyzed, Section: a) Kappa IC/(C+G)% patterns;
b) Kappa IC/TM patterns;
c) Kappa IC/CpG(Exp/Obs) patterns.

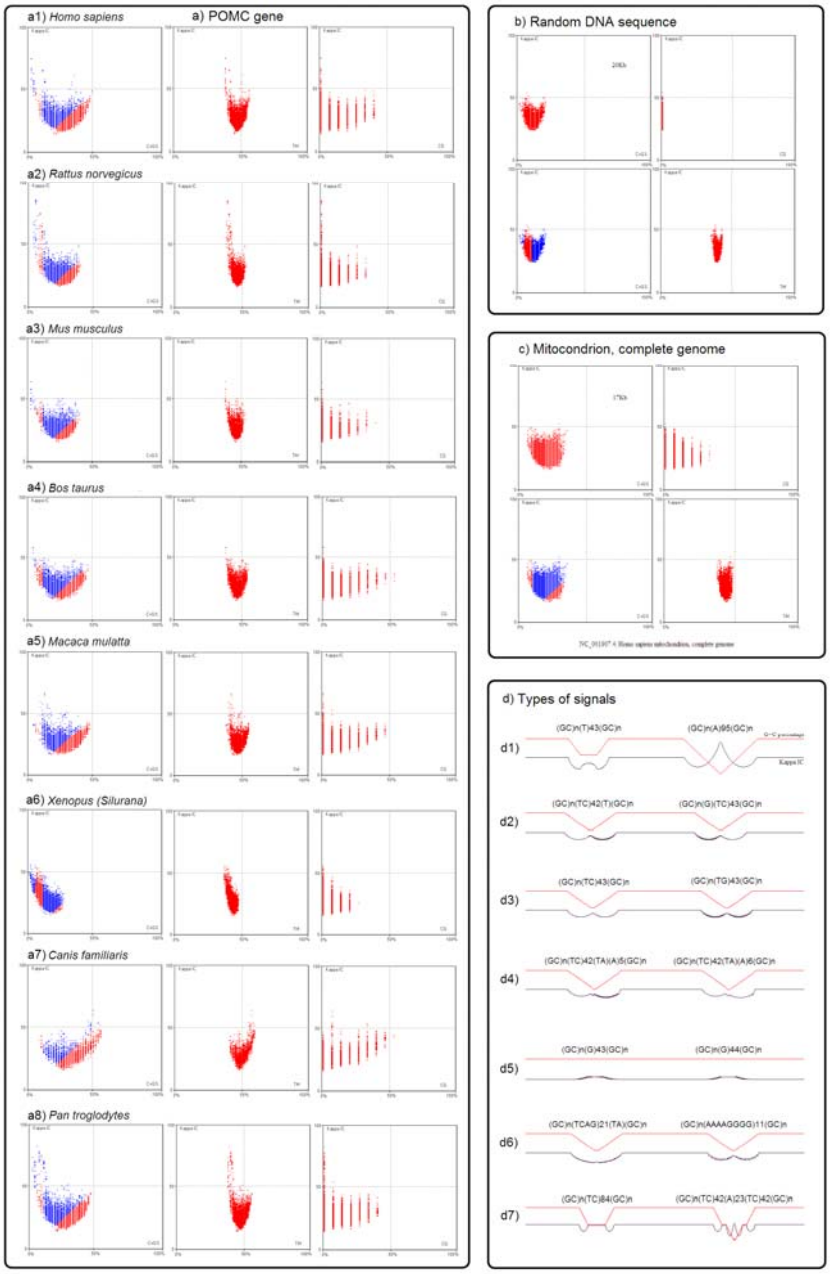

Fig. 2 – Comparative analysis, Section: a) Kappa IC patterns of POMC gene from eight different organisms; b) analysis of artificially generated DNA sequences; c) analysis of Homo sapiens mitochondrial genome; d) eight Kappa IC signal types.

Figure 1, subsection c, shows Kappa IC on *y*-axis and $CpG_{Obs/Exp}$ ratio on *x*-axis. In this case, as CpG dinucleotide structures are more frequent, these patterns show more lines toward the right side of the chart. Vertical positioning of these lines is determined by the CpG distribution inside the sequence. A more scattered distribution of CpG dinucleotides pushes these lines towards the bottom of the chart while a cluster formation of these structures positions these lines to the top of the chart.

Figure 2 shows a further analysis of POMC gene from eight different organisms [16] which may suggest the importance of Kappa IC in determining the selection pressure for a DNA sequence. POMC-derived peptides (Pro-opiomelanocortin) are associated with body weight regulation in the central nervous system [17, 18]. Diabetes has multiple causes and is commonly associated with obesity [19]. POMC neurons regulate glucose homeostasis and for this reason it is believed that POMC gene is often involved in diabetes [20]. In some species, these POMC patterns show a more prominent part in their left side, indicating a high presence of short AT-rich repetitions. These observations may be linked to other studies regarding the incidence of diabetes in different species [21–24]. Therefore, we suggest a possible correlation between short AT-rich repetitions and the predisposition for diabetes in different species.

Artificially generated sequences (with an equal probability of occurrence for A, T, C and G), regardless of the DNA sequence length, have a similar diagram type as shown in Fig. 2, section b. Figure 2, section c, shows a distinctive signature of *Homo sapiens* mitochondrial genome. Pattern similarities between mitochondrial genome and randomly generated sequences tend to confirm once more the oxidative stress and the lack of rules in the distribution of point mutations inside the mitochondrial genome.

4. **DISCUSSION**

We developed a program called DNAKAPPA (Fig. 3, section b). The properties of DNAKAPPA visualization make it a novel method of analyzing DNA sequences. The visualization process enables the identification of distribution differences along the sequence. DNAKAPPA is an open source program written in Visual Basic and it is freely available on the web at http://dnakappa.novusordo.ro.

It runs on all Windows operating systems and does not require installation. The package size is 1.77 Mb and the memory requirements are between 1.2 Mb and 1.5 Mb (Windows 7 and Windows XP). DNAKAPPA can analyze DNA sequences up to 500 kb. The program requires two initial parameters. The first parameter is the sliding window length. The second parameter is the sliding window step.

In Fig. 3, section b, the chart at the top of the DNAKAPPA window shows dinucleotide and nucleotide frequencies, Kappa Index of Coincidence values, T_M values and motif sequences found on 5'-3' and 3'-5' strands. Fig. 3, subsection b1) shows Kappa IC on *y*-axis and *C*+*G* % on *x*-axis. Fig. 3, subsection b2) shows the blue/red color scheme for Kappa IC (*y*-axis) and *C*+*G* % (*x*-axis). Fig. 3, subsection b3) shows Kappa IC on *y*-axis and $CpG_{Obs/Exp}$ ratio on *x*-axis. Fig. 3, subsection b4) shows Kappa IC on *y*-axis and T_M on *x*-axis.

DNAKAPPA was tested on a computer equipped with a 2.8 GHz processor, 500 MB RAM and 80 GB HDD. On average, DNAKAPPA scan speed is 1.5 Kb/s. Another feature worth mentioning is the DNAKAPPA interface that makes a dynamic correlation between the diagram and the sequence.

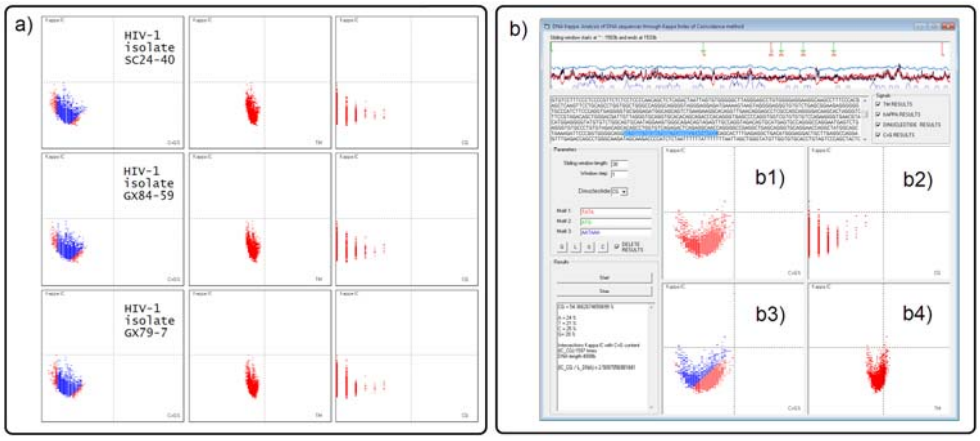

Fig. 3 – DNAKAPPA program. Section a) shows the analysis three envelope glycoprotein genes of HIV-1 isolate SC24-40, GX84-59 and GX79-7. Section b) shows a screenshot of DNAKAPPA program analyzing POMC gene from Homo sapiens genome.

By moving the mouse over the diagram, DNAKAPPA 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.

Nevertheless, in future tests we wish to point out possible correlations between Kappa IC patterns obtained from gene sequences and protein structures.

The prediction of gene structure in DNA sequences have been the focus of the scientific community over the past few years. Because DNA sequences have a probabilistic and non-deterministic structure, computational linguistic methods are effective for describing genomic structures. Many algorithms from electrical engineering and cryptography can be adapted for the field of molecular genetics and bioinformatics [25–28].

5. **CONCLUSIONS**

In this study we examined new patterns using Kappa Index of Coincidence and other parameters already used in molecular genetics: $G+C$ percentage, the nucleic acids melting temperature (T_M) and $CpG_{Obs/Exp}$ ratio. We showed possible applications for Kappa IC and we believe that these patterns may have important implications for molecular genetics in the near future.

ACKNOWLEDGEMENTS

This work was supported by the Romanian Ministry of Education and Research and represents a part of the Research Project PNII Partnerships 42-161/2008. Also was partially 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.

Received on 22 July 2011

REFERENCES

1. Friedman, W.F., *The index of coincidence and its applications in cryptology*, Department of Ciphers, Publ 22. Geneva, Illinois, USA, Riverbank Laboratories.
2. Mountjoy, Marjorie, *The Bar Statistics*. NSA Technical Journal, **VII**, *2*, *4*, 1963.
3. Friedman, W.F. and Callimahos, L.D., *Military Cryptanalytics*, Part I, **2**. Reprinted by Aegean Park Press, 1985.
4. Kahn, David, The Codebreakers, *The Story of SecretWriting*, New York, Macmillan, 1996.
5. Kawaji, H., Frith, M.C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., *Dynamic usage of transcription start sites within core promoters*, Genome Biol., **7**, R118, 2006.
6. Fatemi, M., Pao, M.M., Jeong, S., Gal-Yam, E.N., Egger, G., Weisenberger, D.J., Jones, P.A., *Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level*, Nucleic Acids Res., **33**, *20*, e176, 2005.
7. 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**, *5*, pp.1412-1417, 2006.
8. Antequera, F.*, Structure, function and evolution of CpG island promoters*, Cell Mol. Life Sci., **60**, *8*, pp.1647-58, 2003.
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**, *5*, pp. 1412-1417, 2006.
10. Hui Zhao, H., Li, QZ., Zeng, CQ., Yang, HM., Yu, J., *Neighboring-Nucleotide Effects on the Mutation Patterns of the Rice Genome*, Geno. Prot. Bioinfo., **3**, *3*, pp. 158-168, 2005.
11. Greenbaum, BD., Levine, AJ., Bhanot, G., Rabadan, R., *Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses*, PLoS Pathog., **4**, *6*, 2008.

12. Marmur, J., Doty, P., *Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature*, J. Mol. Biol., **5**, pp. 109-118, 1962.
13. Wetmur, JG., *DNA probes: applications of the principles of nucleic acid hybridization*, Crit Rev. Biochem. Mol. Biol., **26**, *3-4*, pp. 227-59, 1991.
14. von Ahsen, N., Wittwer, CT., Schütz, E., *Oligonucleotide Melting Temperatures under PCR Conditions: NearestNeighbor Corrections for Mg21, Deoxynucleotide Triphosphate, and Dimethyl Sulfoxide Concentrations with Comparison to Alternative Empirical Formulas*, Clinical Chemistry, **47**, *11*, pp. 1956-1961, 2001.
15. Gardiner-Garden, M., Frommer, M., *CpG islands in vertebrate genomes*, J. Mol. Biol. **196**, *2*, pp. 261-282, 1987.
16. Raffin-Sanson, M.L., de Keyzer, Y., Bertagna, X., *Proopiomelanocortin, a polypeptide precursor with multiple functions: from physiology to pathological conditions*, European Journal of Endocrinology, **149**, *2*, pp. 79-90, 2003.
17. Huszar, D., Lynch, C.A., Fairchild-Huntress, V., Dunmore, J.H., Fang, Q., Berkemeier, L.R., Gu, W., Kesterson, R.A., Boston, B.A., Cone, R.D., Smith, F.J., Campfield, L.A., Burn, P., Lee, F. *Targeted disruption of the melanocortin-4 receptor results in obesity in mice.* Cell, **88**, pp. 131-141, 1997.
18. Barsh, G.S., Farooqi, I.S., O'Rahilly, S. *Genetics of body-weight regulation.* Nature, **404**, pp. 644-651, 2000.
19. Smyth, S., Heron, A., *Diabetes and obesity: the twin epidemics*. **12**, *1*, pp. 75-80, 2006.
20. Parton, L.E., Ye, C.P., Coppari, R., Enriori, P.J., Choi, B., Zhang, C.Y., Xu, C., Vianna, C.R., Balthasar, N,, Lee, C.E., Elmquist, J.K., Cowley, M.A., Lowell, B.B., *Glucose sensing by POMC neurons regulates glucose homeostasis and is impaired in obesity*, **449**(**7159**), pp. 228-232, 2007.
21. Baral, R., Rand, J. S., Catt, M. & Farrow, H. A., *Prevalence of feline diabetes mellitus in a feline private practice*, J. Vet. Intern. Med., **17**, p. 433, 2003.
22. Panciera, D. L., Thomas, C. B., Eicker, S. W. & Atkins, C. E., *Epizootiological patterns of diabetes mellitus in cats: 333 cases (1980–1986)*, J. Am. Vet. Med. Assoc., **197**, pp. 1504-1508, 1990.
23. Guptill, L., Glickman, L., Glickman, N., *Time trends and risk factors for diabetes mellitus in dogs: analysis of veterinary medical data base records (1970–1999)*, Vet. J., **165**, pp. 240-247, 2003.
24. Wild, S., Roglic, G., Green, A., Sicree, R., King, H., *Global prevalence of diabetes: estimates for 2000 and projections for 2030*, Diabetes Care, **27**, *5*, pp. 1047-1053, 2004.
25. Teodor Leuca, Mihaela Nova, *Optimization of eddy-current heating process using genetic algorithm*, Rev. Roum. Sci. Techn. – Électrotechn. et Énerg., **54**, *4*, pp. 355-363, 2009.
26. Florea Ioan Hănțilă, Florin Constantinescu, Alexandru Gabriel, Gheorghe, Miruna Nițescu, Mihai Maricar, *A new algorithm for frequency domain analysis of nonlinear circuits*, Rev. Roum. Sci. Techn. – Électrotechn. et Énerg., **54**, *1*, pp. 57-66, 2009.
27. Felix Albu, Constantin Paleologu, Yuriy Zakharov, *An efficient algorithm for active noise control*, Rev. Roum. Sci. Techn. – Électrotechn. et Énerg., **55**, *4*, pp. 416-425, 2010.
28. Rui j.p. de Figueired, *A neural-network-based approach to speech signal predictio*, Rev. Roum. Sci. Techn. – Électrotechn. et Énerg., **55**, *1*, pp. 42-48, 2010.