International Journal of Computers, Communications & Control
Vol. I (2006), No. 3, pp. 25-32

Pathogen Variability. A Genomic Signal Approach

Paul Dan Cristea

Abstract: The conversion of genomic symbolic sequences into digital signals has been applied for the analysis pathogen variability. Results are given on the variability of Human Immunodeficiency Virus, type 1, subtype F, isolated in Romania, and of the type A avian influenza virus H5N1, for which sequences have been downloaded from GenBank [1]. Nucleotide sequence analysis is corroborated with techniques based on the genomic signal approach to detect pathogen resistance to antiretroviral treatment. In the case of protease (PR) inhibitors, it is found that the treatment induces single nucleotide polimorphisms (SNPs) in specific sites. For moderate resistance, the changes affect the PR enzyme only at the level of the protein, whereas for multiple drug resistance, the RNA gene secondary structure also changes.
Keywords: Genomic signals, Pathogen variability, HIV, Influenza, Orthomyxoviridae, Drug resistance

1 Introduction

As shown in a series of previous papers [2-4], the conversion of nucleotide and amino acid sequences into digital signals offers the possibility to apply signal processing methods for the analysis of genomic data. The genomic signal conversion used in our work is a one-to-one mapping of symbolic genomic sequences into complex signals, as described in [2]. The idea is to conserve all the information in the initial symbolic sequence, while bringing in foreground some features significant for the subsequent processing and analysis. This direct method has proven its potential in revealing large scale features of DNA sequences, maintained at the scale of whole genomes or chromosomes, including both coding and non-coding regions. One of the most conspicuous results is that the unwrapped phase of DNA complex genomic signals varies almost linearly along all investigated chromosomes, for both prokaryotes and eukaryotes. The slope is specific for various taxa and chromosomes. This regularity of the genomic signals reveals a corresponding large scale regularity in the distribution of pairs of successive nucleotides, which is similar to Chargaff's first order rules for the frequencies of occurrence of the nucleotides [5].

We applied the same genomic signal approach for studying the variability of several pathogens, including the Human Immunodeficiency Virus, type 1 (HIV-1), subtype F, isolated from Romanian patients at the National Institute of Infectious Diseases "Prof. Dr. Matei Bals", Bucharest [3], and the avian influenza virus type A, based on genomic sequences downloaded from GenBank [1].We have used mainly the phase analysis of the complex genomic signals attached to the nucleotide sequences describing viral genes, as well as the analysis of the corresponding secondary RNA structure and of the phylogenetic neighbor-joining trees for some of these genes.

The focus of the study is primarily on the enzyme changes involved in generating pathogen resistance to multiple drug treatment. A novel methodology for describing sets of related genomic signals, based on a common reference and on individual differences has been developed. Variability signals with respect to average, median and maximum flat references, and digital derivatives of genomic signals are applied to this purpose. Applying this method, it has been found that the mutations in the genes of the analyzed viruses occur only in some specific, well defined locations, while the largest part of their genome remains unchanged. The mutations conferring drug resistance are a subset of all mutations occurring in the studied viruses.

On the other hand, for the case of HIV protease, it has been shown that the changes in response to the antiretroviral drug treatment occur not only at the level of the final enzyme product, preventing the action

Selected from ICCCC 2006 (invited paper)

of the drug on the active protease catalytic site, but also at the level of protease gene RNA secondary structure. These type of changes have been found only for multiple drug resistant viruses.

2 Symbolic Sequence Conversion

For convenience we repeat here the mapping used in our work for the representation of the nucleotides [2]

$$a = 1 + j, c = -1 - j, g = -1 + j, t = 1 - j \tag{1}$$

Apart of the mapping of the four nucleotides (a,c,g,t), the complete genomic signal representation of nucleotide sequences also comprises the mapping of all the other IUPAC symbols for nucleotide classes: $s = \{c,g\}$ - strongly bonded, $w = \{a,t\}$ - weakly bonded, $r = \{a,g\}$ - purines, $y = \{c,t\}$ - pyrimidines, $m = \{a,c\}$ - amine, $k = \{g,t\}$ - ketone, $b = \{c,g,t\} = \neg a$, $d = \{a,g,t\} = \neg c$, $h = \{a,c,t\} = \neg g$, $v = \{a,c,g\} = \neg t$, and $n = \{a,c,g,t\}$ [2]. These symbols occur in the nucleotide sequences generated by genotyping because of the multiplicities determined either by the variability within the virus population or by noise. But this is not the case of the consensus sequences downloaded from GenBank [1], which are curated to contain only the (a,c,g,t) nucleotide symbols. The mapping in equation (1) has the advantage of conserving all the information in the initial symbolic sequence, as it uses a bijective mapping, while being as little biased as possible.

3 Representation by Reference and Variation

To study the variability of the genomic signals in a given set, for example, the signals for multiple resistant viruses, it is convenient to use a description comprising two types of components: (1) the reference - a certain signal considered to best describe the common variation of all components in the considered cluster; (2) the difference of each signal in the cluster with respect to the common reference. In such an approach, it is important to introduce in the common reference as much as possible of the variation shared by all the signals, and keep for the individual differences of each signal only the variations belonging actually to the that signal, without external variation.

The reference can be chosen as one of the following possibilities:

- average (mean) of the signals, or another linear combination of the signals;
- median - the signal in the central position, or the average of the pair of signals placed centrally;
- maximum flat signal - a modified median that keeps better local variations on the signals where they occur avoiding spurious transfers on other signals.

When the reference equals the average, the dispersion of the cluster of signals is minimum, i.e., the sum of the squares of the individual differences between each signal and the reference is minimized. But the average, as any other linear combination, has the important disadvantage that a localized variation of only one of the signals is transmitted to the reference, so that all the other signals will have an apparent variation of opposite sign in that point.

The median reference performs better, being a nonlinear function of the signals in the cluster, so that it decouples the common reference from the local variations of each of the individual signals. The median reference minimizes the sum of the absolute values of the differences between each signal and the reference. A variation localized on only one of the signals is no longer transmitted to the reference, so that it does not affect the variation with respect to the reference of the other signals. The exception occurs when the signal on which the localized variation occurs is just the median.

The maximum flat (MaxFlat) reference is equal to the median wherever the median has no variations which are not shared by other signals. Elsewhere, the MaxFlat reference assumes the minimal variation

that corresponds to its trend, if possible remaining constant. Consequently, the variation signals show better the changes that occur in each individual signal, with less "crosstalk". The digital derivatives of the variation signals show only the actual changes, caused by the variability in each of the signals and, for genomic signals, correspond directly to the SNPs.

4 HIV-1 Subtype F Variability

A phase analysis has been performed on a segment of about 1302 base pairs, approximately aligning with the standard sequence of HIV-1 (NC001802) in GenBank [1] over the interval 1799..2430 bp. This segment, which is currently used for the standard identification and assessment of HIV-1 strains, comprises the protease (PR) gene and almost two thirds of the reverse transcriptase (RT) gene. The PR and RT segments are contiguous and have been analyzed both together, as one entity, and independently, as two distinct encoding regions. The PR gene has the length 297 bp and is located in the first interval (1..297 bp) of the sequenced DNA segment, respectively along the 1799..2095 bp region of the NC001802 sequence. The RT encoding segment that has been analyzed has a length of 1005 bp and is located in the second interval (298..1302 bp) of the analyzed DNA segment, respectively along the 2096..3100 bp region of the NC001802 sequence. The entire RT gene has 1680 bp located in the interval 2096..3775 of the sequence.

Figures 1 and 2 show the cumulated and unwrapped phase of genomic signals for the protease (PR) genes from nine instances of HIV type 1, F clade [1, 6]. Three cases come from treatment naïve patients (S - sensitive), three from patients that developed resistance to one of the drugs (R), and three with multiple resistance to ther antiretroviral treatment (M). The cumulated phase is proportional to the unbalance in the number of nucleotides (statistics of first order) along the nucleic acid strand given by: $3(n_G - n_C) + (n_A - n_T)$, up to a $\pi/4$ factor, whereas the unwrapped phase is proportional to the difference between the number of direct and inverse nucleotide transitions (statistics of second order) along the nucleic acid strand $(n_+ - n_-)$, with a $\pi/2$ factor [2]. Figures 3 and 4 give the same informtion for the segment comprising 1005 bp of reverse transcriptase (RT) genes, out of the total of 1680 bp in this gene, for the same isolates in Figs. 1 and 2. As expected, the cumulated phase varies less than the unwrapped phase for these instances, as all mutations are of the SNP type and affect more the nucleotide pair distribution than the nucleotide distribution itself. Even for the unwrapped phase, the variation of the signal along the strand is quite similar for most of the sequences, but the local changes cumulate along the strands. Because of the mutations are local, the general shape of the phase signals are similar. It is also to be noticed that all the genomic material in these sequences is encoding and uses the same reading frame.

The vertical strips in these figures mark the positions of the mutations (SNPs) that induce resistance to protease inhibitors (Indinavir, Ritonavir, Saquinavir, Nelfinavir, and Amprenavir) [1]. The mutations that lead to multiple drug resistance are concentrated in several sites. In most of the remaining genome, the viruses have the same longitudinal structure. The sequences display mutations in several other locations. The effect of the mutations can easier be seen on the unwrapped phase, which is more sensitive to SNPs.

The successive mutations of the SNP type do not induce the divergence that could be expected, so that the signals do not actually diverge from one another. On the contrary, the signals tend to cluster, as the variations tend to compensate each other, so that the overall span of the signals does not increase directly with the number of mutations and the number of signals. This is another proof of the fact that, from the structural point of view, a genomic sequence satisfies more restrictions than a "plain text", which must just correspond to a certain semantics and to certain grammar rules, and resembles more to a “poem”, which additionally obeys rules of symmetry, giving its “rhythm” and “rhyme”. The recurrence of such patterned structures is reflected in simple mathematical rules satisfied by the corresponding genomic signals.

The representation can be improved by using the reference-difference description, choosing the maximum flat (MaxFlat) reference, as shown in Fig. 5 for the unwrapped phase in Fig. 4. In this case, the largest possible part of the common behavior of the signals is introduced in the reference signal, whereas each individual variation signal maintains only the changes occurring in that particular signal, or to the class it belongs to. The reference signal is no longer necessarily equal in each interval with one of the signals, even when the number of signals is odd. The digital derivatives of the difference signals, shown in Fig. 6 show only the actual changes caused by the variability in each of the signals. In the case of HIV, these changes correspond directly to the SNPs. For multiple resistant strains, the pulses correspond to the sites known from literature to confer resistance to various drugs.

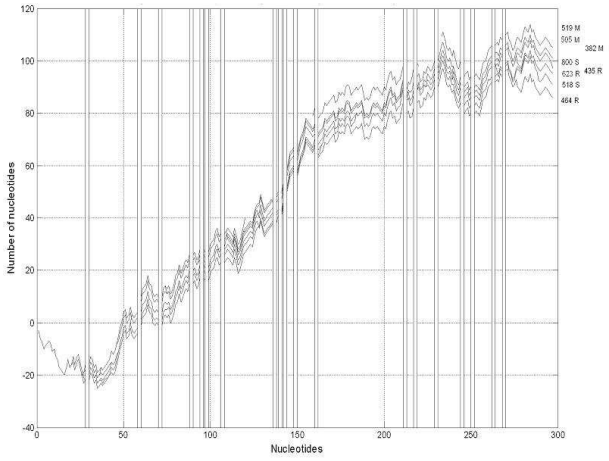

Figure 1: Cumulated phase expressed by $3(n_G - n_C) + (n_A - n_T)$ [2] for the protease (PR) gene of nine isolates of HIV-1, subtype F, showing sensitivity (S), resistance (R) and multiple resistance (M) to drugs.

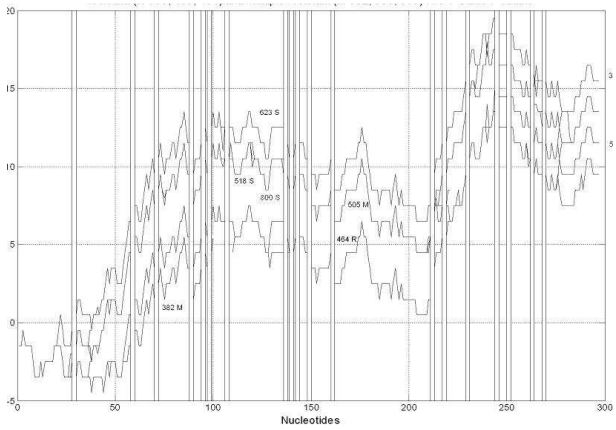

Figure 2: Unwrapped phase expressed by $n_+ - n_-$ [2] for the protease gene of the isolates of HIV-1 in Fig. 1.

HIV-1 makes many of its proteins in one long chain, and protease (PR) has the essential role of cutting this 'polyprotein' into the proper pieces, with the proper timing. Consequently, PR has been chosen as an important target for the current drug anti-HIV therapy. PR is a small enzyme, comprising two identical peptide chains, each of 99 amino acids long, which are encoded by the same gene of 297 nucleotides.

The two chains form a tunnel that holds the polyprotein, which is cut at an active site located in the center of the tunnel. Drugs bind to PR, blocking its action. Studying the estimated secondary structure

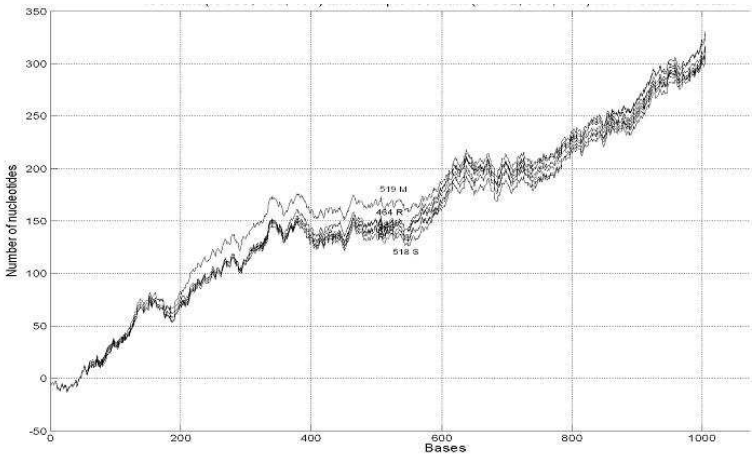

Figure 3: Cumulated phase of RT genomic signals for the isolates shown in Figs. 1 and 2.

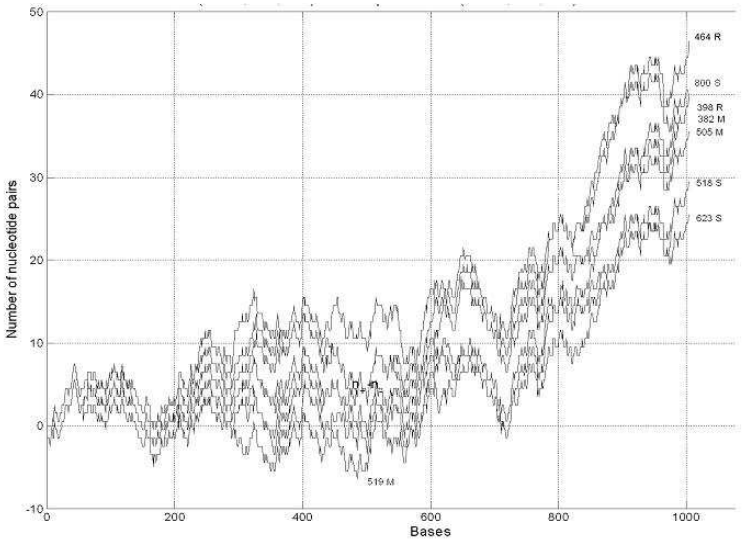

Figure 4: Unwrapped phase for the RT gene in the isolates shown in Figs. 1 and 2.

of the PR RNA for the nine virions previously analyzed, it can be shown [3] that the structures are quite similar for drug sensitive and drug simple resistant viruses. This result is consistent with the generally accepted model stating that the genomic changes of HIV, which induce resistance to drugs, operates at the level of the protein (the final protease enzime), preventing the blocking of its catalytic site. On the other hand, it is found the remarkable fact that, for drug multiple resistant strains, there is a significant change in the RNA secondary structure. Large loops and bulges are replaced with similar, but smaller, less vulnerable, closed-loop structures. These results indicate that there is a certain action of the drug at the level of the protease RNA, effect that becomes evident when mutations conferring multiple drug resistance occur.

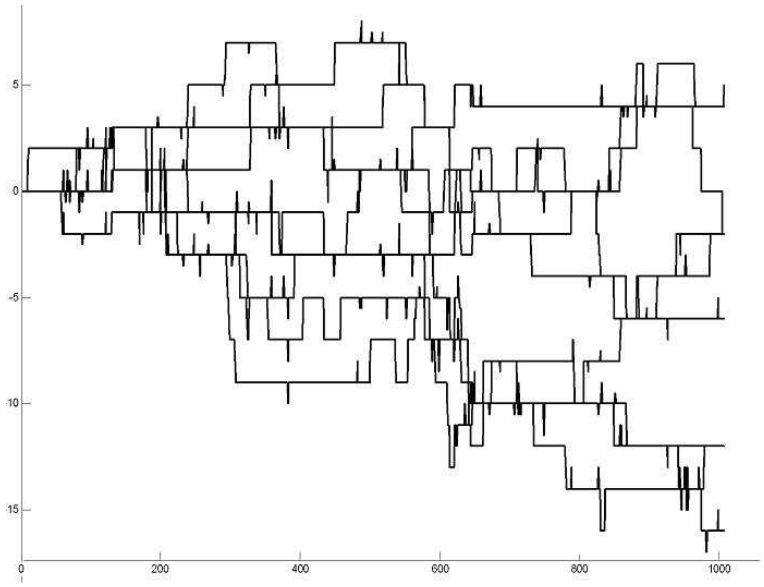

Figure 5: The unwrapped phase in Fig. 4 shown with respect to the *MaxFlat* reference.

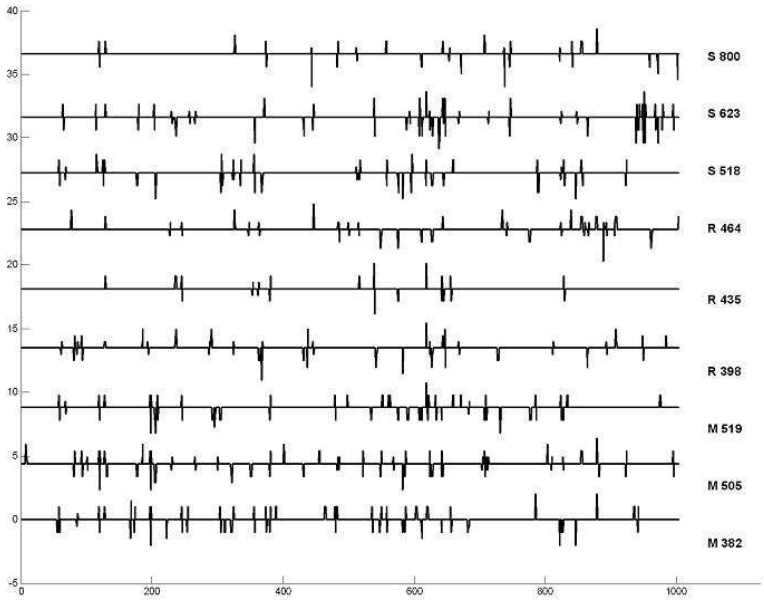

Figure 6: Digital derivatives of the variation signals in Fig. 5.

5 Variability of Hemagglutinin gene of influenza H5N1 virus

The influenza virus envelope embeds two specific antigenic glycoproteins that project out of the virion surface, the *Hemagglutinin* (HA) and the *Neuraminidase* (NA). Many different combinations of HA and NA proteins are possible, but only the H1N1 (Spanish endemic), H1N2 (Asian epidemic), and H3N2 (Hong Kong epidemic) subtypes have circulated worldwide among humans. HA protein selectively binds to the sialic acid of the host cell surface receptors, thus recognizing the cells that the virus can invade [4, 6]. Figure 7 gives the cumulated phase of the HA gene for H5N1 viruses isolated from two humans (AF046080, AF046097) and one chicken (AF046088), in Hong Kong, in 1997 [6, 7]. The genes for viruses isolated close in time are similar, even when crossing the inter-species barrier, whereas a large variation can be seen for genes isolated at larger time intervals. Only several SNPs are found in Fig. 8 which gives the difference cumulated phases with respect to the MaxFlat reference. The same result has been obtained for all the genes in the eight segments of the H5N1 virus [4, 6].

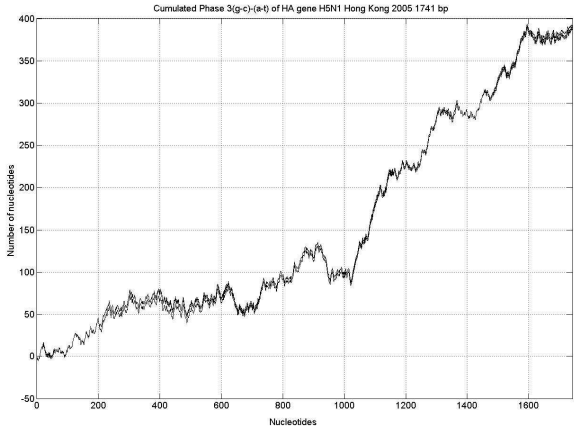

Figure 7: Cumulated phase of the HA gene, H5N1 virus (accessions AF046080, 88, 97 [1, 6]).

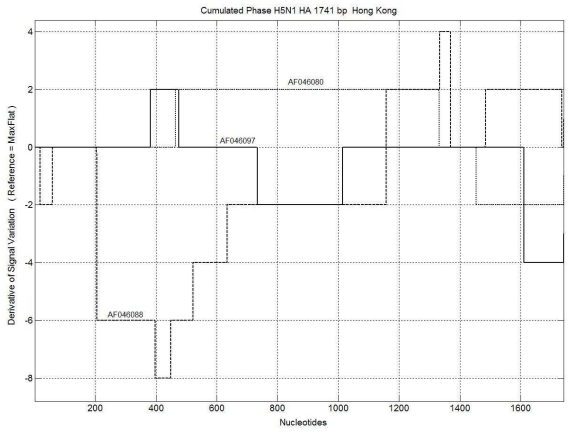

Figure 8: Differences of HA gene cumulated phases in Fig.7 with respect to the MaxFlat reference.

6 Further Work

Further work will be focused on:

- the dynamics of Influenza Type A viruses that have crossed till now the species barrier from birds to humans, and which hold the potential to become highly contagious and highly lethal in humans, including the H5N1 subtype,
- extending the study from the nucleotide to the amino acid level, which could be more significant from the phenotypic point of view,
- using genomic signals for helping clustering viruses in classes.

Acknowledgments

The sequences of HIV presented in this paper have been genotyped by Dr. Dan Otelea from the National Institute of Infectious Diseases “Prof. Dr. Matei Bals”, Bucharest, Romania. Results referring to the study of HIV variability have been previously jointly published [3].

References

[1] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, GenBank, http://www.ncbi.nlm.nih.gov/genoms

[2] P. D. Cristea, "Representation and Analysis of DNA sequences", in Genomic Signal Processing and Statistics, Editors E.G. Dougherty, I. Shmulevici, Jie Chen, Z. J. Wang, Book Series on Signal Processing and Communications, Hidawi, 2005, pp.15-65.

[3] P. D. Cristea, D. Otelea, Rodica Tuduce, "Study of HIV Variability Based on Genomic Signal Analysis of Protease and Reverse Transcriptase Genes", EMBC’05, Sept. 2005, Shanghai, China.

[4] P. D. Cristea, "Genomic Signal Analysis of Pathogen Variability", SPIE, BO24, paper 5699-52, San Jose, Jan, 2005, 12 pg.

[5] E. Chargaff, "Structure and function of nucleic acids as cell constituents", Fed. Proc., 10, pp. 654-659, 1951.

[6] D.L. Suarez,.et.al., Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong, J. Virology, vol. 72 (8), pp. 6678-6688, 1998. (AF046080-99).

[7] E. Ghedin, N. Sengamalay, M. Shumway et. al., "Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution", Nature, vol. 437, Oct. 2005, pp.1162-1166.

[8] M.S. Hirsch et al., "Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection", in Proc. Recommendations of an International AIDS Society - USA Panel, JAMA, vol. 283, no. 18, May 10, 2000, pp.2417-2426.

Paul Dan Cristea
University POLITEHNICA of Bucharest
Biomedical Engineering Center
Address: Spl. Independentei 313, sect. 6
060042 Bucharest, Romania
E-mail: pcristea@dsp.pub.ro