ISSN 0975 - 6299 Vol 2 / Issue 4 / Oct - Dec 2011

International Journal of Pharma and Bio Sciences

RESEARCH ARTICLE

BIOINFORMATICS

GENOME SIGNAL ANALYSIS: AN APPROACH FOR PREDICTING HIVI VARIABILITY BASED ON PHASE ANALYSIS OF PROTEASE GENE.

AKANKSHA KULSHRESHTHA*, V.M.KARTHIKA AND DR.ARCHANA TIWARI,

School of Biotechnology, Rajiv Gandhi Proudyogiki Vishwayidyalaya, Air Port bypass road,Gandhi nagar BHOPAL 462033 INDIA.

AKANKSHA KULSHRESHTHA
School of Biotechnology, Rajiv Gandhi Proudyogiki Vishwayidyalaya, Air Port bypass road, Gandhi nagar BHOPAL 462033 INDIA

***Corresponding author**

ABSTRACT

The sequencing of Human (*Homo sapiens*) including other eukaryotic "model organisms" and some prokaryotes have already been done. For handling this large amount of data, for further processing and analysis work, the conversion of genomic sequences into digital sequences i.e. Genomic signals are widely used. Genomic signal processing methods are powerful tool for the analysis of local and global features as well as for the analysis of genomic information. The review paper presents the different levels of HIV classification, genomic signal analysis of HIV I variability using phase analysis. It is significant to understand the evolution of HIV I deciphering its interaction with the immune system and developing effective control strategies. . It has been found that variability in HIVI can lead to multiple drug resistance due to altered RNA secondary structure in these cases

KEYWORDS

Variability, HIV, Genomic Signals, RNA secondary structure.

INTRODUCTION

As biology is a data rich science and we are having a large number of genomic data, genome of many organisms including eukaryotes, man (Homo sapiens [1-3] and several "model organisms" [4-8], have been completely sequenced. The genomic data i.e. structural and functional features of genome for various organisms are being accumulated, and analyzed all over the world, from the small university to the large laboratories. This genomic, proteomic and structural data is stored, managed, and analyzed on a large variety of computing systems, the database like N.C.B.I. and public access to genomic database offers the opportunity of data mining and exploring in depth this unique information depository and to convert this raw data into knowledge in ways that are useful to humankind.[9-11] The volume of genomic data is expanding at a huge rate, while its fundamental properties and relationships are not fully revealed. The genomic signal is one of the tool for revealing large scale features of DNA. Signal processing methods can be applied on genomic data after the conversion of basic form AGCT into digital signals.[10]-[15]. These basic symbols AGCT can be converted into digital signals by replacing the symbols with some values of certain physical or biochemical properties of the corresponding nitrogenous bases or amino acids, and genome signal processing methods can be applied on digital signals for the analysis of genomic data [12]. The basic theme is to conserve all the information present in genomic sequences in the form of symbolic sequences by using a one-to-one mapping [13].The most remarkable result obtained is that the unwrapped phase(distribution of pair of successive nucleotides) 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. The regularity of the genomic signals reveals that there is 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 [16].

The review paper presents the genomic signal approach for studying the variability of HIV type I. Analysis of secondary RNA structure for protease, specifically to characterize the variability of the F subtype HIV strain isolated in Romania.

Worldwide the AIDS epidemic started only several decades ago, before that HIV was unknown. An estimated 38.6(33.4-46.0) million people live with HIV I worldwide, while about 25 million have died already [18].Today there is no region of the world untouched by this pandemic,fig1 .Romania has a high total number of HIV/AIDS reported infection(12,559 in mid 2002 including 2699 AIDS related death)[19].A tragic feature of HIV epidemic in Romania is the large number of cases in children(9936 cases), out of which the vast majority (>70%) acquired HIV through blood transfusion or …infection. The first description of HIV I sub type f, typical for Romania and its drug susceptibility can be found in Apetrei *et al* [20,21].

Worldwide distribution of hiv-1 infections.

Figure I.

Worldwide distribution of hiv-1 infections, modes of transmission, and hiv-subtypes[1] HSex=heterosexual. MSM=Men who have sex with men. IDU=injection drug users. Based on Joint UNAIDS and WHO AIDS epidemic update December, 2005.

The rest of the review paper is organized as follows: section 2 presents the basic information about hiv1, its subtypes and structural biology of HIV1. Section 3 presents the genomic signal representation of the nucleotides, on the basis, the symbolic sequence of nucleotides in DNA molecule are converted into complex digital genomic signals. section4 presents phase analysis of hiv- 1 subtype f variability and some of the results of phase analysis of HIV I variability. section 5 presents the estimated secondary structure of RNA for nine virions, it has been shown that variability in protease enzyme can lead to change in secondary structure and confers that mutation leads to multiple drug resistance in some species.

2. HIV- ITS CLASSIFICATION AND EVOLUTION

Human immune deficiency virus a retrovirus of lentivirus family has double stranded rna genome.HIV can be divided into two major types HIV type 1 (HIV-1) and HIV type 2 (HIV-2). HIV-1 can be further divided into `Group M (major) the "outlier" group O and two new groups, N and P [22]. The HIV-1 group M viruses predominate and are responsible for the AIDS pandemic. Group M can be further subdivided into subtypes based on genetic sequence data.

Classification of HIV

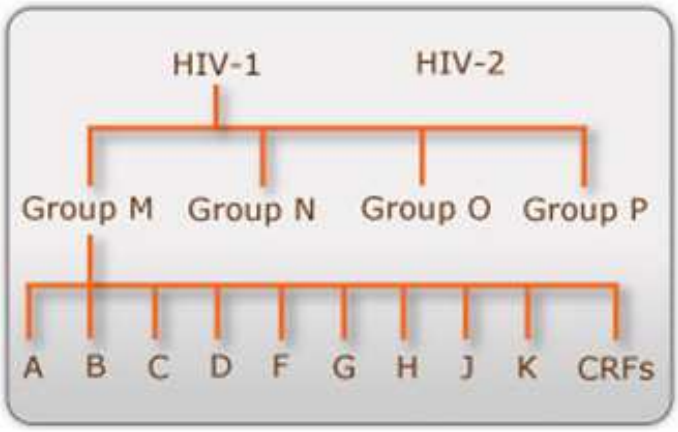


Fig-II

The different levels of HIV classification[22]

Group M

With 'M' for "major", this is the most common type of HIV, with more than 90% of HIV/AIDS cases deriving from infection with HIV-1 group M. The M group is subdivided further into clades, called subtypes that are also given a letter. There are also "circulating recombinant forms" or CRFs derived from recombination between viruses of different subtypes which are each given a number. CRF12_BF, for example, is a recombination between subtypes B and F.

- Subtype A is common in West Africa.[23]
- Subtype B is the dominant form in Europe, the Americas, Japan, Thailand, and Australia.[24]
- Subtype C is the dominant form in Southern Africa, India, and Nepal.[24]
- Subtype D is generally only seen in Eastern and central Africa.[24]
- Subtype E has never been identified as a non recombinant, only recombined with subtype A as CRF01_AE.[24]
- Subtype F has been found in central Africa, South America and Eastern Europe.[25]
- Subtype G (and the CRF02_AG) have been found in Africa and central Europe.[24]
- Subtype H is limited to central Africa.[25]
- (Subtype I) was originally used to describe a strain that is now accounted for as CRF04_cpx, with the cpx for a "complex" recombination of several subtypes
- Subtype J is primarily found in North, Central and West Africa, and the Caribbean.[26]
- Subtype K is limited to the Democratic Republic of Congo and Cameroon.[25]

Group N

The 'N' stands for "non-M, non-O". This group has only been seen in Cameroon and was discovered in 1998[27]

Group O

The O ("Outlier") group is not usually seen outside of West-central Africa. It is reportedly most common in Cameroon, where a 1997 survey found that about 2% of HIV-positive samples were from Group O.[28] The group caused some concern because it could not be detected by early versions of the HIV-1 test

kits. More advanced HIV tests have now been developed to detect both Group O and Group N.[29]

Group P

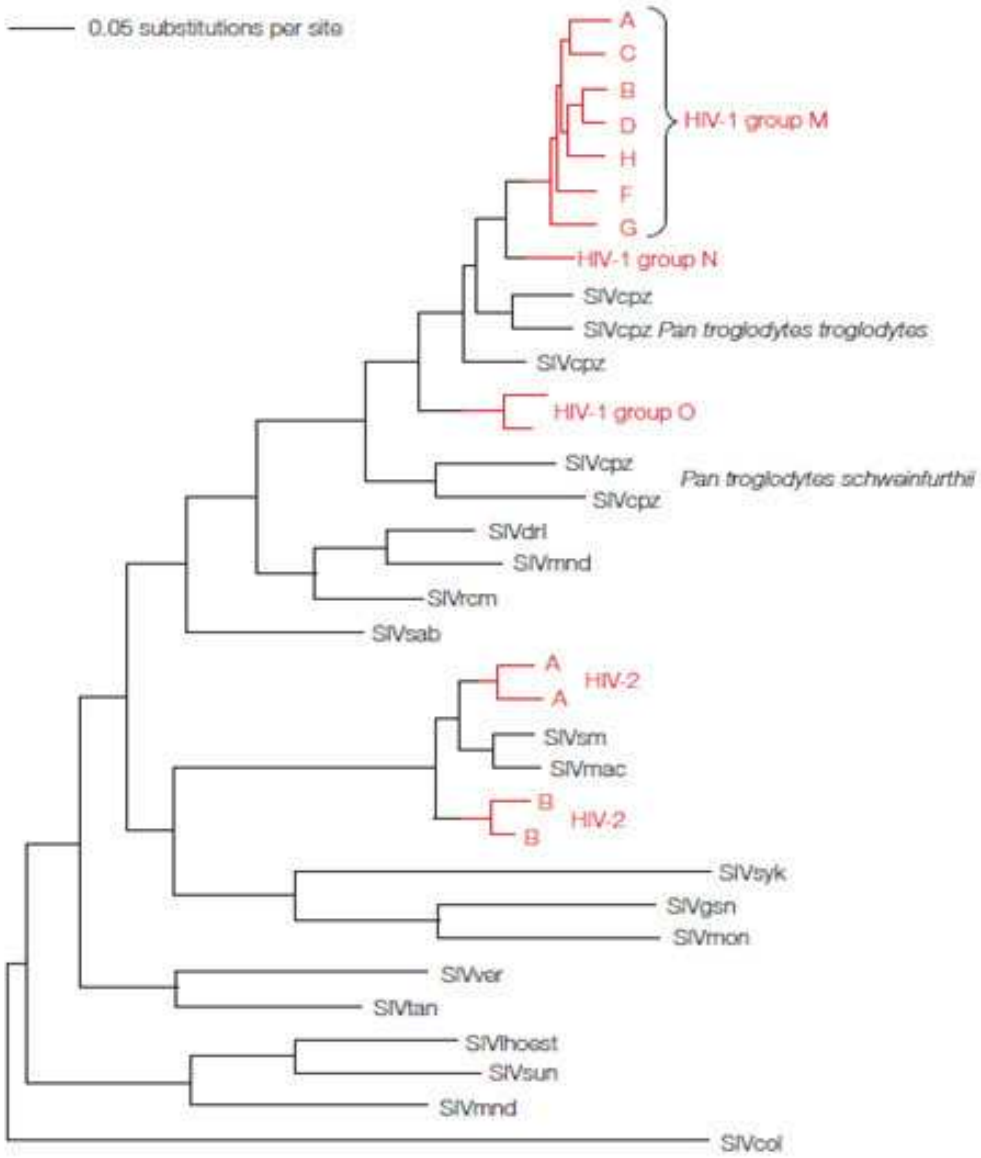
In 2009, a newly-analyzed HIV sequence was isolated from Cameroonian woman residing in France who was diagnosed with HIV-1 infection in 2004 and reported to have greater similarity to a simian immunodeficiency virus recently discovered in wild gorillas (SIVgor) than to SIVs from chimpanzees (SIVcpz).The scientists reporting this sequence placed it in a proposed Group P "pending the identification of further human cases"[30][31][32]

HIV-2

HIV-2 is mainly present in AFRICA. Many test kits are available for detecting HIV-1[33] as well as HIV-2.The first case found in United State in 1987. [34] There are 8 known groups of HIV-2 are present upto 2010, of these only 2 groups A and B are epidemic. Group A spread mainly in West Africa, but also to Angola, Mozambique, Brazil, India, and very limitedly to Europe or the US. Group B is mainly confined to West Africa.[35][36]

The origins of HIV

To find the interaction of HIV with immune system and to develop effective control strategies it is most important to understand the evolution of the human immunodeficiency virus (HIV). The evolutionary history of hiv1 and hiv2 has been reconstructed and further concluded that the two human viruses are related to different SIVs and therefore have different evolutionary origins.HIV-1 which is found in some sub-species of chimpanzee (*Pan troglodytes troglodytes* and *Pan troglodytes schweinfurthii*) that inhabit parts of equatorial Western and Central Africa, and most closely related to SIVcpz. HIV-2 is found at high prevalence in sooty mangabey monkeys (*Cercocebus atys*) and most closely related to SIVsm5,[37].



Phylogenetic tree showing history of primate lentivirus

Figure III
Evolutionary history of the primate lentiviruses.[37]

Both the human immunodeficiency virus type 1 (HIV-1) and HIV-2 lineages(red branches)fall within the simian immunodeficiency viruses (SIVs) come under lenti virus family The tree was reconstructed using a MAXIMUM LIKELIHOOD METHOD on an alignment of 34 published nucleotide sequences of the viral polymerase (*pol*) gene. Other abbreviations for viruses and their primate hosts are as follows: SIVcol, black and white colobus; SIVdrl, drill; SIVgsn, greater

spot-nosed monkey; SIVlhoest, L'Hoest monkey; SIVmac, macaque; SIVmnd, mandrill; SIVmon, Campbell's mona monkey; SIVrcm, red-capped monkey; SIVsab, Sabaeus monkey; SIVsun, sun-tailed monkey; SIVsyk, Sykes' monkey;SIVtan, tantalus monkey; SIVver, vervet monkey. For clarity, only some subtypes of HIV-1 and HIV-2 are shown. All gene sequences were taken from GenBank.[37]

3. GENOMIC SIGNAL REPRESENTATION

By using mapping the symbolic sequences have been converted to complex digital signals. According to the three main dichotomies in their biochemical properties the four nucleotides can be arranged in classes:

(1) *strength of link* – bases A and T are linked by two hydrogen bonds (W - weak bond), while C and G are linked by three hydrogen bonds (S - strong bond);

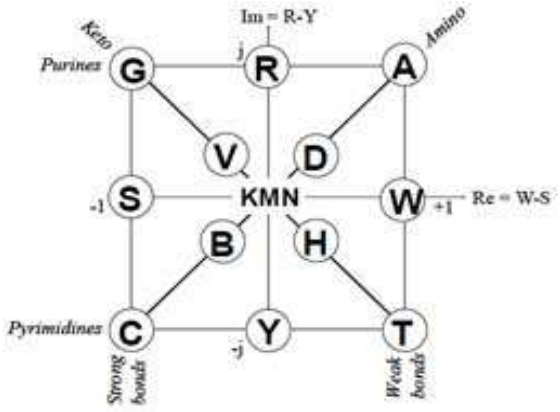
(2) *radical content* – A and C contain the amino (NH3) group (M class) in the major grove, while T and G the keto (C=O) group (K class).

(3) *molecular structure* – A and G are purines (R), while C and T are pyrimidines (Y);

On this basis A vector *tetrahedral representation* is done of the nucleotides [1-10].By giving less importance to amino-keto separation, the representation can be brought in a plane allowing to use the complex representation of nucleotides shown in Fig. 2 and expressed in equation 1.[38]

a=1+j; g=-1+j; t=1-j; c=-1-j; w = 1 y = - j s = -1 r = j, k=m=n=0

Representation of nucleotides in a tetrahedron



d=1/3(1+j)
h=1/3(1-j)
b=1/3(-1-j)
v=1/3(-1+j)

Fig. IV

Complex representation of nucleotides and Nucleotide classes(IUPAC SYMBOLS)[38]

4. PHASE ANALYSIS OF HIV- 1 SUBTYPE -F VARIABILITY

DNA sequences of Human immunodeficiency virus type 1 (HIV-1) from isolates of 121 Romanian patients have been genotyped by the laboratory of the National Institute of Infectious Diseases Prof. Dr. Matei Bals", Bucharest, Romania. On the basis of pol gene structure, the strains have been classified as subtype F. The sequenced segment has been of about 1302 base pairs, approximately aligning with the standard sequence of HIV-1 (NC001802) over the interval 1799..2430 bp.(37) The sequence which has been 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. Here protease and reverse transcriptase is analyzed as PR and RT which are two significant enzymes of HIV-1 and could be further used as a drug target for anti HIV therapy. HIV-1 protease is one of the important and best-studied enzymes.[38]

Phase Analysis

The property of the complex number is the phase of periodic magnitude, the complex number remains same even the multiple of 2π has been added or subtracted to or from its phase.[10] The phase of a complex number have been restricted to the domain (-π,π].In a sequence the sum of phases of complex number from the first element in the sequence, up to the current element is represented by cumulated phase.[10] Along a sequence of nucleotides at certain location the cumulated phase has the value :

$$\Theta_C=\pi/4[3(n_G-n_T) + (n_A-n_T)],$$

Where

$n_{A=}$ numbers of adenine, nucleotides in the sequence

$n_{C=}$ numbers of cytosine, nucleotides in the sequence

$n_{G=}$ numbers of guanine, nucleotides in the sequence

$n_{T=}$ numbers of thymine, nucleotides in the sequence

Along the DNA strand at a certain location the slope *sc* of the cumulated phase is linked to the frequencies of occurrence of the nucleotides around that location by the equation :

$$S_C = \pi/4[3(f_G-f_C)+(f_A-f_T)],$$

Where f_A, f_C, f_G, and f_T are the nucleotide occurrence frequencies. For the complex representations of DNA nucleotide sequences the current value of cumulated phase is giving an indication on the relative frequencies of the purines (A, G) vs. pyrimidines (C, T), and It always drifts between some positive and negative values, in the segment under consideration.[7] Due to the the[Delete] bias introduced by the conventional restriction of the phase to the domain (−π,π] which favors π over –π .By introducing a uniform complex noise, i.e., by adding uniformly distributed small random complex numbers to each of the nucleotide representations in the sequence. This unwanted effect, which would bias the phase analysis, could be avoided. For the noisy complex sequence, the phases close to −π are equally probable. With the phases close to + π and there is no spurious drift of the cumulated phase towards positive values.

Unwrapped phase

In a sequence of complex numbers, the corrected phase of the elements is represented by the unwrapped phase, in which the absolute value of the difference between the phase of each element in the sequence and the phase of its preceding element is kept smaller than *π* by

adding or subtracting an appropriate multiple of 2π to or from the phase of the current element. The value of the unwrapped phase gives an indication on the relative frequencies of the transitions between the nucleotides. For the complex representation given in equation (1.4).There is an increase of the unwrapped phase, corresponding to a rotation in the trigonometric sense by $\pi/2$ when positive *transitions* A → G, G → C, C → T, T → A is there and when the *negative transitions* A → T, T → C, C → G, G → A is there it determines a decrease, corresponding to a clockwise rotation by $-\pi/2$, while all other transitions are neutral. There are two types of neutral transition first type is exactly neutral transitions A ↔ A, C ↔ C,G ↔ G, T ↔ T, for which the difference of phase is zero in each instance, so that the unwrapped phase does not change, and the "second type on average" neutral transitions A → C, C → A, G → T, T → G, for which the difference of phase is $\pm\pi$. Because of the bias introduced by the conventional restriction of the phase to the domain $(-\pi, \pi]$, which favors π over $-\pi$, the standard unwrapped phase function and the corresponding functions implemented in most commercial software mathematics libraries, which apply the basic convention for the phase mentioned above, attach $+\pi$ to all the "on average" neutral transitions.[10]

Two solutions have been used to avoid this unwanted effect,

1. By introducing a uniform complex noise i.e.by addition of uniformly distributed small random complex numbers have been added to each nucleotide complex representation, so that phases and differences of phase close to $-\pi$ are equally probable with the phases close to π.
2. Primarily for medium or small sequences, for example, when studying virus genomes, but also for large and very large sequences, a custom unwrapped phase function has been used that attaches zero phase change for all neutral transitions.

Using artificial sequences the accuracy of both procedures has been thoroughly verified.

For the complex representation (1.4), taking the precautions mentioned above, at a certain location along a sequence of nucleotides the unwrapped phase has the value:

$$\Theta u = \pi/2(n_+ - n_-),$$

Where n_+ represents the number of positive transitions

n_- represents the numbers of the negative transitions,

The slope su of the variation of the unwrapped phase along a DNA strand is given by the relation:

$$S = \pi/2(f_+ - f_-)$$

Where f+ represents the frequencies of the positive transition and f− are the frequencies of the and[Delete] negative transitions.

In previous work it has been shown that there are two basic mutation types by which the cumulated phase and the unwrapped phases are influenced. [39]. The unwrapped phase is sensitive to the punctual quasi-random mutations of the SNP type, which alter the nucleotide pair distribution but remains unaffected by crossover and similar types of mutations, including the reversal of exchanged segments accompanied by strand switching. 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 π/2 factor [38]. On the contrary, the cumulated phase is changed significantly for the crossover-type mutations but less sensitive to SNPs. The cumulated phase is proportional to the differences in the number of nucleotides (statistics of first order) along the nucleic acid strand: 3(nG- nc) + (nA - nT), with a π/4 factor.[11]

As expected, the unwrapped phase varies more than that of cumulated phase for these instances, as all mutations are of the SNP type and affect more the nucleotide pair distribution than the nucleotide distribution itself.

The symbolic sequences of the PR gene for 30 patients have been converted into complex genomic signals using (1). The nucleotide imbalance N (cumulated phase $\theta c = \pi N / 4$) and the nucleotide pair imbalance P (unwrapped phase $\theta u = \pi P / 2$) have been computed for the genomic signals using (2) and (3), respectively. The signals have been classified into three groups taking into account the clinical behavior of the patients with respect to the response to current antiretroviral treatment:

- Sensitive (S) - the patient responds to the antiretro-viral treatment (wild type pathogen),
- Resistant (R) - patient resistant to one anti-HIV drug.
- Multiresistant (M) - patient resistant to several anti-HIV drugs.[17

5. PROTEASE RNA SECONDARY STRUCTURE

PR has the essential role of cutting of 'polyprotein' into the proper pieces, with the proper timing, actually HIV-1 is making many of its proteins in one long chain,and needs PR for making them functional. 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.

Phase Analysis

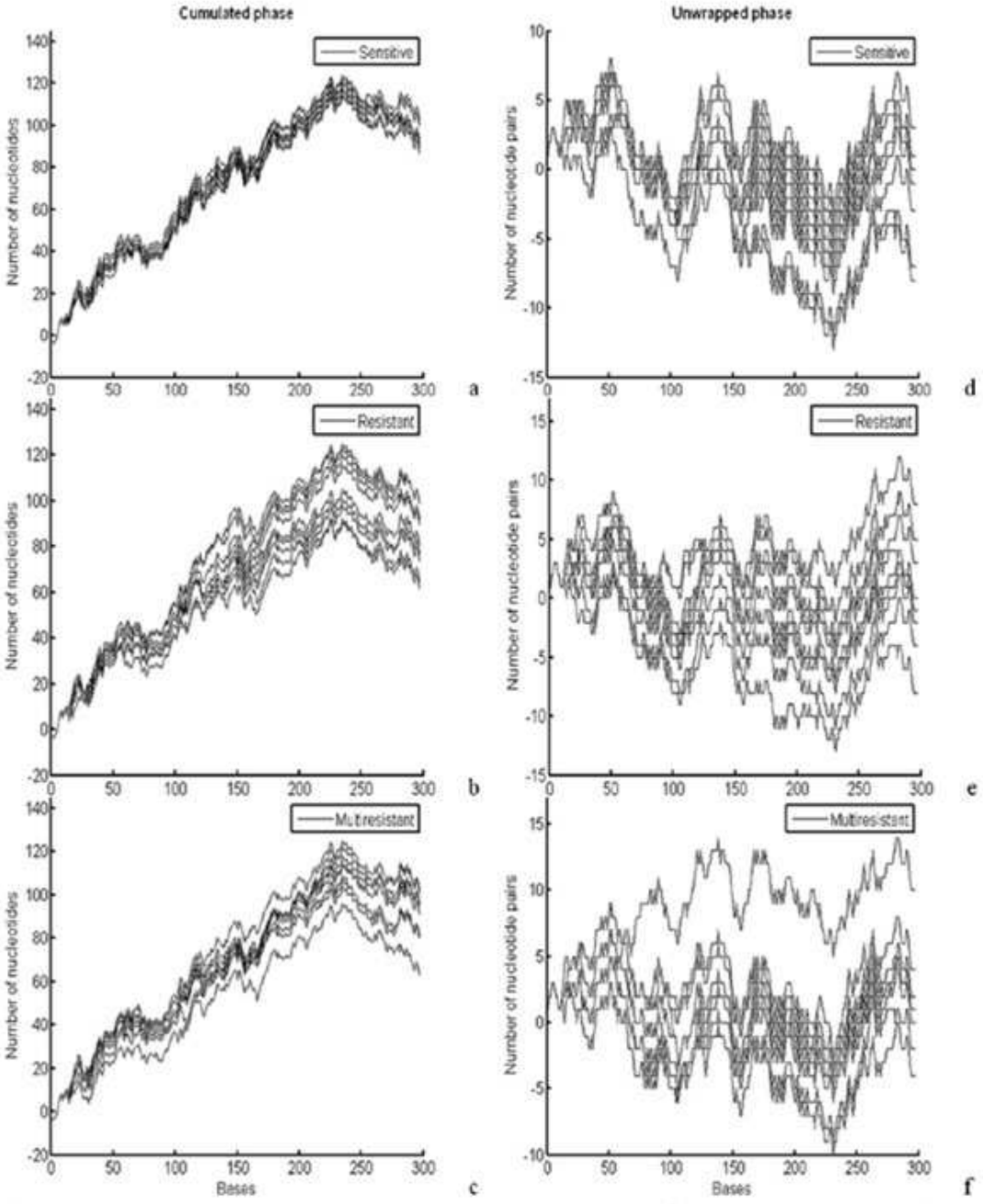
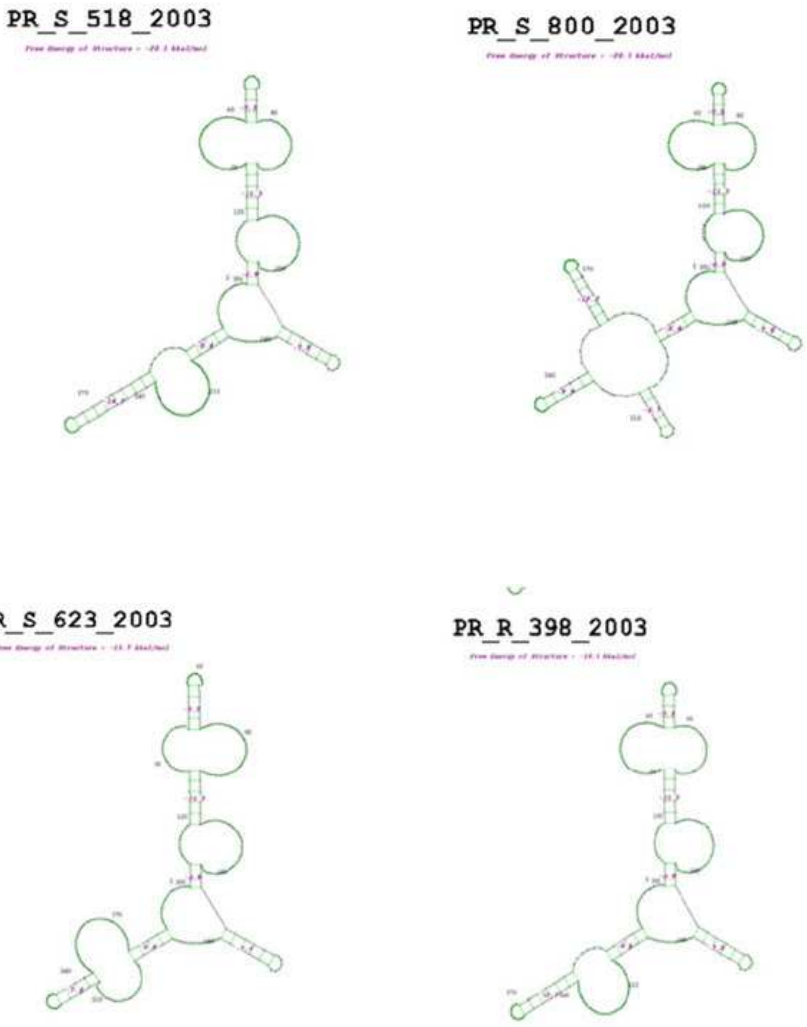


Figure V

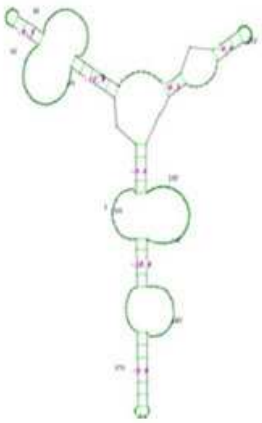
Nucleotide imbalance (cumulated phase) and nucleotide pair imbalance P (unwrapped phase) of PR gene genomic signals for sensitive, resistant and multiresistant patients.[40]

Figure 9 shows the estimated secondary structure of RNA for the nine virions previously discussed. It can be noticed that the structure is quite similar for drug sensitive and drug simple resistant viruses but it is significant that for multiple drug resistant strains, there is a marked change in the RNA secondary structure. Here in case of multidrug resistant pathogen large loops and bulges are replaced with similar, but smaller, less vulnerable, closed-loop structures. These results show 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. [17]

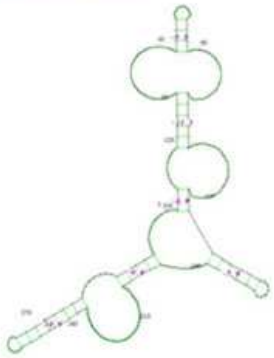
RNA Secondary Structures



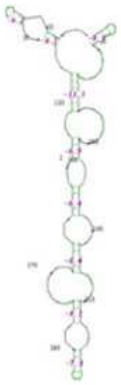
PR_R_435_2003



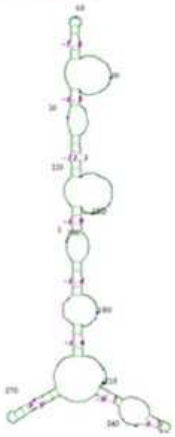
PR_R_464_2003



PR_M_382_2003



PR_M_519_2003



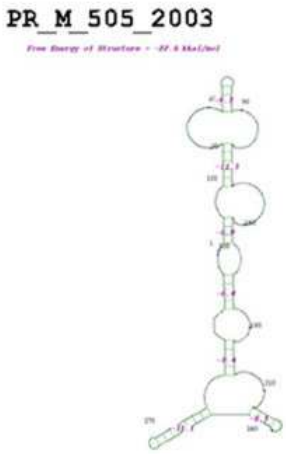


Figure-VI
Estimated secondary structure of the RNA segments corresponding to the PR gene for the analyzed isolates.[17]

CONCLUSIONS

Understanding the evolution of the human immunodeficiency virus (HIV) is crucial for reconstructing its origin, deciphering its interaction with the immune system and developing Effective control strategies.
Symbolic genomic sequence can be converted to digital sequence or genomic signals with the help of complex quadrantal representation of nucleotides; the main advantage of genomic signal over symbolic sequence is that they can be analyzed by using signal processing methods. Tools for genomic signal analysis including the use of phase analysis, aggregated phase, unwrapped phase, nucleotide path analysis, independent component analysis, cluster analysis, phylogenetic analysis etc. the cumulated phase and unwrapped phase of genomic signal reflects the statical distribution of bases and base pairs. While independent component analysis is used to reveal SNP and other variability's induced changes in a set of related nucleotide sequence and for investigating the reliability of ICA estimates cluster analysis is performed. The review paper presents the results of phase analysis of complex genomic signals for protease gene and reverse transcriptase gene of nine isolates of HIV I subtype f, showing different drug resistance. [13]
For the case of protease, it has been shown that the changes in response to antiretroviral drug treatment occur not only at the level of the final enzyme product, preventing the blocking of the protease catalytic site under the effect of the drug, but also at the level of the protease gene RNA secondary structure. These types of changes have been found only for multiple drug resistant types of viruses and consist in the replacement of vulnerable large loops and bulges by similar, but smaller and less vulnerable closed-loop structures.
Genomic signal analysis is a promising technique for the analysis of variability with the help of phase, unwrapped phase, Independent component analysis etc. The kind of variability in any pathogen is responsible for the resistance or multi drug resistance; it shows against the drugs, so new drug can be made according to the kind of variability and can work better on pathogens. Genome signal analysis can be made for

appropriate clustering of viruses, bacteria's and other organisms in the classes. Extending the studies from nucleotide to amino acid level, and sequence to structure and then function level, could be more significant for in depth work and analysis in the future.

REFERENCES

1. International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome", Nature, 409, pp. 860-911, February 15, 2001.
2. 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
3. J. C. Venter, et al., "Draft Analysis of the Human Genome by Celera Genomics", *Science*, 291, pp. 1304-1351, February 16, 2001.
4. C. elegans Sequencing Consortium, "Genome sequence of the nematode C. elegans: a platform for investigating biology",*Science*, 282 (5396), pp. 2012-2018, 1998.
5. Genome Sequencing Center, Washington Univ. Medical School, http://www.genome.wustl.edu/ projects/chicken/, 1 March 2004.
6. A. Theologis, J. R. Ecker, *et al.*,"Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana", Nature, 408 (6814), pp. 816-820, 2000.
7. Rat Genome Sequencing Consortium, http://www.ncbi.nlm.nih.gov/genoms, August 30, 2003
8. RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, "Functional annotation of a fulllength mouse cDNA collection", ***Nature,*** 409, pp. 685-689, February 8, 2001.
9. Cristea P., Genetic signals: An emerging concept, Proceedings of IWSSIP 2001 pp. 17-22, 2001
10. P. D. Cristea, "Representation and analysis of DNA sequences", in: Genomic Signal Processing and Statistics, Editors: E.G. Dougherty, I. Shmulevich, Jie Chen, Z. Jane Wang, EURASIP Book Series on Signal Processing and Communications, Hidawi Publishing Corporation, 2005, Ch. 1, pp. 15-65
11. D. Anastassiou, "Frequency-domain analysis of biomolecular sequences", *Bioinformatics*, vol.16, no. 12, pp. 1073-1081, 2000.
12. P. D. Cristea, "Genomic Signals of Re-Oriented ORFs", EURASIP – Journal on Applied Signal processing, Special Issue on Genomic Signal Processing, vol. 2004, no.1, pp. 132-137, January 1, 2004
13. P. D. Cristea, "Large Scale Features in DNA Genomic Signals", Elsevier, Signal Processing, Special Issue on Genomic Signal Processing, 83, pp. 871-888, 2003.
14. P. D. Cristea, "Invariants of DNA Genomic Signals", in 2004 Proc. SPIE - AU104 Biomedical Applications of Micro- and Nanoengineering II, Biocomputation & Biomodelling Conference.
15. P. D. Cristea, "Conversion of Nitrogenous Base Sequences into Genomic Signals", Journal of Cellular and Molecular Medicine, 6, 2, pp. 279-303, April – June 2002.
16. E. Chargaff, "Structure and function of nucleic acids as cell constituents", *Fed. Proc*., 10, pp. 654-659,1951.
17. Paul Dan Cristea, Genomic Signal Analysis of HIV-1clade F Gene Variability, EUROCON 2005, Serbia & Montenegro, Belgrade, November 22-24, 2005

18. UNAIDS. 2006 report on the global AIDS epidemic: a UNAIDS 10th anniversary special edition. [(accessed July 20, 2006)]. http://www.unaids.org/en/HIV_data/2006GlobalReport/default.asp

19. T. Novotny, D. Haazen and O. Adeyi, "HIV / AIDS in Southeastern Europe: Case Studies from Bulgaria, Croatia and Romania",
ECSHD / ECC05, Washington, D.C., February 11, 2003, http://hivinsite.ucsf.edu

20. C. Apetrei et al., "Human Immunodeficiency Virus Type 1 Subtype F Reverse Transcriptase Sequence and Drug Susceptibility", J Virol, 72, pp. 3534-3538, 1998.

21. C. Apetrei *et al.*, "HIV Type 1 Diversity in Northeastern Romania in 2000-2001 Based on Phylogenetic Analysis of pol Sequences from Patients Failing Antiretroviral Therapy", *AIDS* Research and Human Retroviruses, Vol. 19, No. 12, pp. 1155-1161, December 2003.

22. WWW.AVERT.ORG/HIV TYPES.HTM

23. Bobkov AF, Kazennova EV, Selimova LM, *et al.* (October 2004). "Temporal trends in the HIV-1 epidemic in Russia: predominance of subtype A". *J. Med. Virol.* **74** (2): 191–6

24. Goudsmit, Jaap. Viral Sex; The Nature of AIDS. Oxford University Press. New York, New York, 1997. Pg. 51-58. Retrieved May 25, 2008.

25. Hemelaar J, Gouws E, Ghys PD, Osmanov S. (March 2006). "Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004.". *AIDS* **20** (16): W13–23.

26. Peeters M, Gueye A, Mboup S, Bibollet-Ruche F, Ekaza E, Mulanga C, Ouedrago R, Gandji R, Mpele P, Dibanga G, Koumare B, Saidou M, Esu-Williams E, Lombart JP, Badombena W, Luo N, Vanden Haesevelde M, Delaporte E (March 1997). "Geographical distribution of HIV-1 group O viruses in Africa". *AIDS* **11** (4): 493–8

27. Julie Yamaguchi, Ruthie Coffey, Ana Vallari, Charlotte Ngansop, Dora Mbanya, Nicaise Ndembi, Lazare Kaptué, Lutz G. Gürtler, Pierre Bodelle, Gerald Schochetman, Sushil G. Devare, -Catherine A. Brennan (January 2006). "Identification of HIV Type 1 Group N Infections in a Husband and Wife in Cameroon: Viral Genome Sequences Provide Evidence for Horizontal Transmission". AIDS Research and Human Retroviruses 22 (1): 83–92.

28. Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F (August 2009). "A new human immunodeficiency virus derived from gorillas".*Nature Medicine* **15** (8): 871–2.

29. http://www.abbottmolecular.com/PDF/E0608633_RealTimeHIV_rev.pd

30. Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F (August 2009). "A new human immunodeficiency virus derived from gorillas".Nature Medicine **15** (8): 871–2. doi:10.1038/nm.2016. PMID 19648927

31. "New HIV strain discovered". Associated Press (CBC News). 2009-08-03. Retrieved 2009-08-03.

32. Donald G. McNeil, Jr. (September 16, 2010). "Precursor to H.I.V. Was in Monkeys for Millennia". New York Times. Retrieved 2010-09-17. "But P appears to have crossed over from a gorilla; it was discovered only last year, and in only one woman, who was from Cameroon, where lowland gorillas are hunted for meat."

33. CBER - Donor Screening Assays for Infectious Agents and HIV Diagnostic Assays

34. http://www.hivworkshop.com/hiv-2.htm

35. Santiago, M. L.; Range, F.; Keele, B. F.; Li, Y.; Bailes, E.; Bibollet-Ruche, F.; Fruteau, C.; Noe, R. et al. (2005). "Simian Immunodeficiency Virus Infection in Free-Ranging Sooty Mangabeys (Cercocebus atys atys) from the Tai Forest, Cote d'Ivoire: Implications for the Origin of Epidemic Human Immunodeficiency Virus Type 2". Journal of Virology **79** (19): 12515–27.

36. Marx PA, Alcabes PG, Drucker E (2001). "Serial human passage of simian immunodeficiency virus by unsterile injections and the emergence of epidemic human immunodeficiency virus in Africa". Philos Trans R Soc Lond B Biol Sci **356**

37. Andrew Rambaut*,David Posada, Keith A. Crandall and Edward C.Holmes" THE CAUSES AND CONSEQUENCES OF HIV EVOLUTION" JANUARY 2004 VOLUME 5 www.nature.com/reviews/genetics

38. P. D. Cristea, D. Otelea, Rodica Tuduce, "Genomic signal analysis of HIV variability", in Proc. SPIE - BIOS 2005 – Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues II Conf. , vol. 5699, January 22-27, paper 52

39. C. Apetrei et al., "Human Immunodeficiency Virus Type 1 Subtype F Reverse Transcriptase Sequence and Drug Susceptibility", J Virol, 72, pp. 3534-3538, 1998.

40. P. D. Cristea, Rodica Tuduce, Valeri Miladenov, Georgi Tsenov, and Simona Petrakievai "Prediction of Nucleotide Sequences by using Genomic Signals" 9th WSEAS International Conference on NEURAL NETWORKS (NN'08), Sofia, Bulgaria, May 2-4, 2008