Available online at www.pharmacie-globale.info ISSN 0976-8157

Review Article

PHARMACIE GLOBALE

INTERNATIONAL JOURNAL OF COMPREHENSIVE PHARMACY

GENOME VARIABILITY OF HIV 2 USING VARIOUS ANALYTICAL TECHNIQUES

V M Karthika*, Archana Tiwari and Akanksha Kulshreshtha

School of Biotechnology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, University of Technology, Madhya Pradesh, India.

Received: 19 September 2011; Revised: 14 October 2011; Accepted: 28 November 2011; Available online: 5 December 2011

ABSTRACT

Human Immunodeficiency Virus 2 (HIV 2) is pandemic in West African countries. High genetic variability is an important characteristic of HIV due to its fast replication cycle, high error rate and recombinogenic properties of reverse transcriptase. Infectivity in HIV 2 is found in *vif*, *vpx*, *vpr* and *nef* genes. Signal processing tools which includes symbolic to digital mapping of nucleotides and amino acids on the basis of genetic code, reflecting better their structure and degeneracy, is used for the conversion of genomic data into digital sequences. It is investigated by phase analysis, nucleotide path analysis, independent component analysis, cluster analysis and phylogenetic tree analysis. HIV variability can be described by these digital derivative signals and variability signals with respect to average, median and maximum flat references. Various mathematical tools like Fourier transform, Digital filtering technique are used as signal processing methods. Based on genomic signal methods and on statistical techniques, this review defines the study of infective gene of HIV 2 variability. Since the past studies have been focused upon molecular and genetic level, now the researchers emphasize on the signal analysis of HIV 2 gene variability. So, the aim of this review is to correlate the present and future aspects of variability of HIV-2 at genomic level.

Keywords: genetic variation, genome analysis, sequence analysis, viruses, computational molecular biology.

INTRODUCTION

Among the RNA Viruses, Human Immunodeficiency Virus 2 (HIV 2) is a member of Lentivirus genus of the *Retroviridae* family, and is closely related to the prototype AIDS virus, Human Immunodeficiency Virus 1 (HIV 1).[1] HIV-2 is endemic in West Africa, with the highest prevalence in Guinea-Bissau[2-4], where it infects ~1% of the population. HIV-2 closely resembles the strain of simian immunodeficiency virus (SIV) found in sooty mangabey monkeys.[5] Viruses of the HIV-2 subgroup were discovered by Kanki et al. (1986) by a serological study on healthy people from Senegal[6], West Africa, in the mid – 1980s which prompted numerous studies to determine the geographic distribution and biological significance.[7] During an eight - year period of follow-up in Senegal, HIV 2 infection rates remained relatively constant.[8,9] HIV-2 is associated with AIDS but is less pathogenic than HIV-1.[2-4,10-13] Phylogenetic analyses have revealed five sequence subtypes (A-E), Gao e*t. al* (1994) have also identified HIV-2, derived from different regions of their genome.[14] Subtype-specific differences may exist in HIV-2 biology and that certain variants may be more virulent.[14] With respect to the biological properties and molecular variability of HIV 2, spectrum of HIV 2 is being investigated.[15] The major obstacle to the development of a broadly efficacious human immunodeficiency virus (HIV) vaccine is Genetic Variability.[16,17] This variation is more pronounced in the genes that encode the outer envelope regions compared with the polymerase (*pol*) and group specific antigen (*gag*) genes, which are more genetically constrained from variability.[1] The *env* gene of HIV and related primate *lentiviruses* is a major site for viral variation.[18-20] During the course of infection, Variability in V1, V2, V3 and V4 has been observed.[21-23] This *env* gene encodes glycoprotein which helps in determining the viral cell host range, replication rate, and induction of cytopathic effects.[19-21] A close and controlled examination of the kinetics of viral variation and replication reveals the relationship between genetic variation and pathogenicity, of HIV 2 strain (a human virus), which infects and produces disease in macaques and baboons.[24-28] HIV-1 and HIV-2 have existed in their present populations for similar lengths of time; it is proved by the *env* gene variance. Discrepancy between HIV-2 studies is due to differences in the populations examined and that HIV-2 infection has a lower morbidity rate. Though the genetic structure of HIV 1 and HIV 2 are similar, the only exception is the extra Open Reading Frame (ORF) in HIV 2 designated as X, which translates into a protein of 16 kDa in HIV 2 whose function is still unknown.[29]

The extraordinary information in the form of genomes present in the database[30,31] offers an unprecedented chance to data mine and explore in depth to convert data into knowledge. Studies involving feature extraction at the scale of chromosomes, multi resolution analysis, comparative genomic analysis or quantitative variability analysis[32-34] is not possible by representation of genomic signals as symbolic sequences. So, the conversion of deoxyribonucleic acid (DNA) sequences into digital signals[34] is mandatory so that Signal processing methods

***Corresponding Author:**
V M Karthika,
School of Biotechnology, Rajiv Gandhi Proudyogiki Vishwavidyalaya.
University of Technology of Madhya Pradesh,
Airport Bypass Road, Gandhi Nagar, Bhopal-462 033, India..
Contact no: +91-9407522674; Email: vmkarthika.manoharan@gmail.com

can be applied for genome data analysis[35-44] by the conversion of nucleotide and amino acid sequences into digital signals[35,44-46] and vector representation, conserving the information present in initial symbolic sequences; and thus revealing the features of chromosomes and DNA present at a distance of 10^6 – 10^8 bp.[39] The average unwrapped phase of complex genomic signals varies almost linearly along all investigated chromosomes for prokaryotes and eukaryotes, the magnitude and slope being specific for various taxa and chromosomes; whereas the cumulated phase, the variation is piecewise in prokaryotes and drifts towards zero in eukaryotes. Helicoidal coiling of nucleotide complex representations proves rule similar to Chargaff's Rule.[47] Signal processing techniques had a role of digital filtering technique in gene identification which arising via the protein coding region which exhibits period 3 behavior, that is not present in other parts of DNA molecule.[48,49] The Long - range correlation corresponding to a $1/f$ type of power spectrum, within the pairs prevails.[50] For the study of proteins, Fourier transforms is used and the role of Karhunen-Loeve like transforms[51] in the interpretation of DNA microarray[52-54] data for gene expression. New tools for genomic signal analysis[35,46] are presented, including the use of phase, cumulated phase, unwrapped phase, sequence path, stem representation of sequence components relative frequencies, as well as the transition analysis at the nucleotide, codon and amino acid levels.

CONVERTING THE NUCLEOTIDES INTO NUMBERS

The Chargaff's Rule for the distribution of nucleotides[47] depends upon the first order statistics, but the statistical regularity for the succession of nucleotides depends on second order statistics.[36,38,55] The genetic code (GC) is used by most known organism only with the exception in mitochondria and in certain microbes, and is majorly applicable to nuclear genetic material.[34]

Cartesian Representation

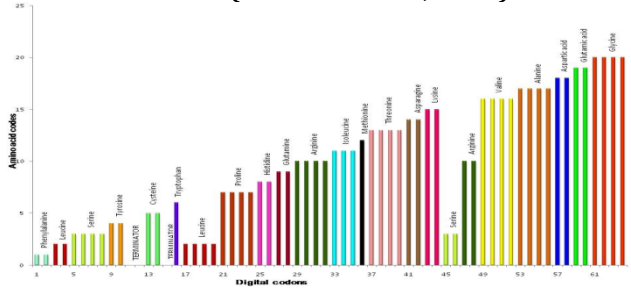
For one dimensional mapping the digits 0,1,2,3 are attached to four nucleotides A, C, G and T. Thus there are 4! = 24 choices to attach them. The three base codons along the DNA are interpreted as 3 – digit numbers written in base 4.[56] Therefore, it implies the biological activity to interpret each codon as a distinct sample of genomic signal distributed along the strand. (Table **1**)

Table 1. Mapping of nucleotides to digits in base four

Purines	Pyrimidines
Adenine = A = 0	Thymine = T = 2
Guanine = G = 1	Cytosine = C = 3

This forms the basis of Real representation of the nucleotides. As for the amino acids, the codons from 0 - 63 are assigned to numerical values.[57] There are: Two single codons for Met and Trp, Nine double codon, one triple codon, five quadruple, Three sextuple and Three codons for Terminator represented by Figure **1**.

Figure 1. Optical correspondence of numerical codes to amino acid codes (Source: Cristea, 2002).



The classic table does not include in their structure, its characteristic symmetries and degeneration. An exhaustive search for all the 24 possible correspondences of the nucleotides to the digits 0-3 has shown that there does not exist a more monotonic mapping[39] shown in Table **2**.

Table 2. Optical correspondnce of numerical codes to amino acid codes (Source : Cristea, 2002)

Digital codon	Amino acid code	Long name
11, 12, 15	0	Terminator
1, 2	1	Phenylalanine
3, 4, 17, 18, 19, 20	2	Leucine
5, 6, 7, 8, 45, 46	3	Serine
8, 10	4	Tyrosine
13, 14	5	Cysteine
16	6	Tryptophan
21, 22, 23, 24	7	Proline
25, 26	8	Histidine
26, 27	9	Glutamine
29, 30, 31, 32, 47, 48	10	Arginine
33, 34, 35	11	Isoleucine
36	12	Methionine
37, 38, 39, 40	13	Thereonine
41, 42	14	Aspargine
43, 44	15	Lysine
49, 50, 51, 52	16	Valine
53, 54, 55, 56	17	Alanine
57, 58	18	Aspartic acid
59, 60	19	Glutamic acid
61, 62, 63, 64	20	Glycine

Tetrahedral Representation

On arranging the nucleotides with respect to the dichotomies in biochemical properties[58,59] i.e. according to:

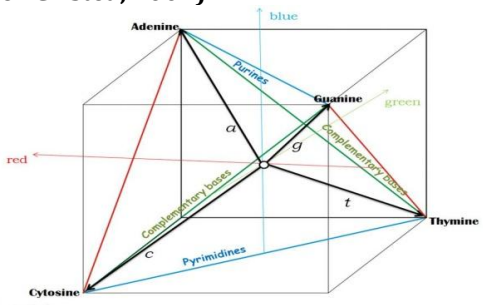
(i) Molecular Structure: **A** and **G** are Purines [**R**]
C and **T** are Pyrimidines [**Y**]

(ii) Strength of Link: **A** and **T** have two Hydrogen bonds, i.e. they are Weak [**W**]
G and **C** have three Hydrogen bonds, i.e. they are Strong [**S**]

(iii) Radical content: **A** and **C** contain Amino Group [**M**]
G and **T** contain Keto group [**K**]

A vector tetrahedral representation[34] can be made with four equal length vectors symmetrically placed with respect to each other as shown in Figure **2**.

Figure 2. Tetrahedral representation of nucleotides (Source : Cristea, 2002)



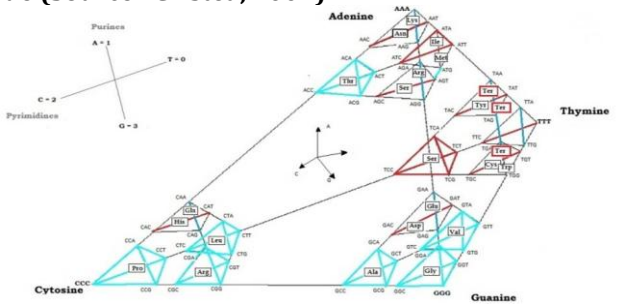
In the above diagram,

- The distance between any two nucleotides is the same which is represented by Gray Coding.[56]
- The 3 D representation expresses the differences x = W – S, y = M – K, z = R – Y and the axes; x, y and z shown in red, green and blue lines respectively.[45] The six edges of the tetrahedron correspond to the pair of nucleotides.[58]

The zero order tetrahedron of the overall Genetic Code is composed by the first base in the codon which selects one of the four, first order 16-codon tetrahedron. Similarly, the First order tetrahedron comprises the second base chosen

from the second order 4-codon tetrahedron and the third base identifies one of the vertices.[44] On the edges along the purines' and pyrimidines' directions, since most of the interchangeable bases are distributed, Degeneracy prevails to the second order tetrahedrons. From the frequency of amino acids in proteins, it implicates that Genetic code has some of the characteristics of Huffman (entropy) coding[57] depicted in Figure **3**.

Figure 3. Tetrahedral representation of the genetic code (Source : Cristea, 2002)



Vectorial Representation

Features of DNA sequence were revealed by resulting digital signals. With truthful and minimum bias, a mere bijective numeric representation of symbolic genomic sequences have been done so as information is conserved and significant features are brought forth for analysis.[59] To make the mathematical description simpler, {±1} integer is chosen as co-ordinates of the vertices of the cube, (even where the bases are present). The base vectors take the simpler form.[57]

$$\vec{a} = \vec{\imath} + \vec{\jmath} + \vec{k} \quad \text{.....(1)}$$

$$\vec{c} = -\vec{\imath} + \vec{\jmath} - \vec{k} \quad \text{.....(2)}$$

$$\vec{g} = -\vec{\imath} - \vec{\jmath} - \vec{k} \quad \text{.....(3)}$$

$$\vec{t} = \vec{\imath} - \vec{\jmath} - \vec{k} \quad \text{.....(4)}$$

$$\vec{h} = \frac{\vec{t}+\vec{a}+\vec{c}}{3} = -\frac{\vec{g}}{3} \quad \text{.....(5)}$$

$$\vec{d} = \frac{\vec{g}+\vec{t}+\vec{a}}{3} = -\frac{\vec{c}}{3} \quad \text{.....(6)}$$

$$\vec{u} = \frac{\vec{a}+\vec{c}+\vec{g}}{3} = -\frac{\vec{t}}{3} \quad \text{.....(7)}$$

$$\vec{b} = \frac{\vec{c}+\vec{t}+\vec{g}}{3} = -\frac{\vec{a}}{3} \quad \text{.....(8)}$$

Apart from the above vectors, IUPAC symbols have to be used when a noise is generated or variability occurs when experimental entries are enrolled for a particular set of genome sequence. IUPAC conventions include symbols for classes (S, W, R, Y, M, K) as well as for classes comprising three nucleotides B = {C, G, T} = ~A; D = {A, G, T} = ~C; H = {A, C, T} = ~G; V = {A, C, G} = ~ T; or unspecified nucleotide N.[57]

$$\vec{w} = \frac{\vec{a}+\vec{t}}{2} = \vec{\imath} \quad \text{.....(9)}$$

$$\vec{k} = \frac{\vec{g}+\vec{t}}{2} = -\vec{\jmath} \quad \text{.....(10)}$$

$$\vec{r} = \frac{\vec{a}+\vec{g}}{2} = \vec{k} \quad \text{.....(11)}$$

$$\vec{y} = \frac{\vec{c}+\vec{t}}{2} = -\vec{k} \quad \text{.....(12)}$$

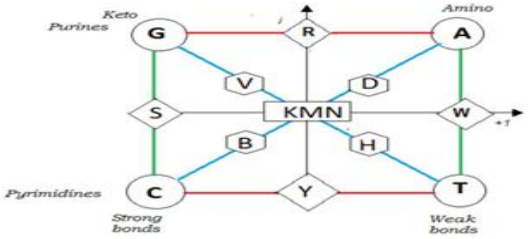
$$\vec{s} = \frac{\vec{c}+\vec{g}}{2} = -\vec{\imath} \quad \text{.....(13)}$$

$$\vec{m} = \frac{\vec{a}+\vec{c}}{2} = \vec{\jmath} \quad \text{.....(14)}$$

Complex Representation

The complex mapping conserves the meaning of distances between the codons as resulting from the genetic code. The projection of the vectors on x*0*z plane brings down the tetrahedron to planar form[35,37] as brought about in Figure **4**.

Figure 4. The tetrahedron in 2- dimensional form (Source : Cristea, 2002)



The less important amino – keto separation can be eliminated and can be expressed in equations as:

$$a = 1 + j \quad \text{.....(15)}$$

$$g = -1 + j \quad \text{.....(16)}$$

$$t = 1 - j \quad \text{.....(17)}$$

$$c = -1 - j \quad \text{.....(18)}$$

Moreover, the IUPAC symbols can be vectorally represented as:

$$w = 1 \quad \text{.....(19)}$$

$$y = -j \quad \text{.....(20)}$$

$$s = -1 \quad \text{.....(21)}$$

$$r = j \quad \text{.....(22)}$$

$$k = m = n = 0 \quad \text{.....(23)}$$

$$d = \frac{1}{3}(1 + j) \quad \text{.....(24)}$$

$$h = \frac{1}{3}(1 - j) \quad \text{.....(25)}$$

$$b = \frac{1}{3}(-1 - j) \quad \text{.....(26)}$$

$$v = \frac{1}{3}(-1 + j) \quad \text{.....(27)}$$

Representation by Reference

The variability signals for multiple resistant viruses can be described by 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.

A common reference can be introduced for the variations shared by all the signals and the individual differences of each signal can be kept along the variations belonging to each signal, without external variation.[59]

Thus, by these various ways described, the symbolic representation of the nucleotides has been converted to digital form for further analyzing the genomic sequence and testing the variability. It can be interpreted from the following categorization that converting the nucleotides into numbers, involves geometrical and statistical knowledge for pursuing the work of genome data into signals.

NUCLEOTIDE SIGNAL ANALYSIS

Phase Analysis

The phase analysis of complex genomic signals, which have been analyzed focusing on the extraction of Large scale features of DNA sequences and up to scale of whole chromosomes is important for understanding functions of chromosomes like replication, transcription and crossing over. The phase of a complex number has a periodic magnitude, which does not change on addition or subtraction of any multiple of 2π. So, the phase of complex number is restricted to the domain (-π, π] so that it covers all possible orientations once. The phase of the nucleotide representations can have the values {-3π/4, -π/4, π/4, and 3π/4} radians. The absolute values of the nucleotide complex representations are same, while the phases are

given by:

$a = 1+ j = \sqrt{2} \angle \pi/4$ .....(28)

$g = -1+ j = \sqrt{2} \angle 3\pi/4$ .....(29)

$c = -1 - j = \sqrt{2} \angle -3\pi/4$ .....(30)

$t = 1 - j = \sqrt{2} \angle - \pi/4$ .....(31)

The cumulated phase is the sum of the phases of the complex numbers in a sequence from the first element in the sequence, up to the current element. The cumulated phase at a certain location along a sequence of nucleotides has the value:

$$\Theta_c = \pi/4 [3(n_G - n_C) + (n_A - n_T)] \quad .....(32)$$

where n_A, n_C, n_G, and n_T are the numbers of adenine, cytosine, guanine, and thymine nucleotides in the sequence, from the first to the current location. The slope is given by:

$$s_c = \pi/4 [3(f_G-f_C) + (f_A-f_T)] \quad .....(33)$$

where f_A, f_C, f_G, and f_T are the nucleotide occurrence frequencies.

The aggregated phase is the sum of the phases of all the complex base representations starting from the beginning of a segment. Its value is never zero and it helps to indicate the frequency of Purines and Pyrimidine present in a segment.

The unwrapped phase is the corrected phase of the elements in a sequence of complex numbers, 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 *positive transitions* A→G, G→C, C→T, T→A determine an increase of the unwrapped phase, corresponding to a rotation in the trigonometric sense by π/2, the *negative transitions* A→T, T→C, C→G, G→A determine a decrease, corresponding to a clockwise rotation by −π/2, while all other transitions are *neutral*; so that

$$\Theta = \pi/2 (n+ - n-) \quad .....(34)$$

Where n+ and n- are the numbers of the positive and negative transitions, respectively.[60] Due to the restriction of the phase from (-π, π), a phase jump has to be introduced, which is eliminated by the unwrapped phase. This unwrapped phase allows global phase trends along a sequence. A clockwise or counter wise helix is formed spinning the axis. In all chromosomes, helicoidal wrapping of complex representation of base can be maintained over distances of tens of millions of bases. This contradicts with Ohno's assertion and also maintains a statistical regularity of succession of bases, not distribution of nucleotides.[61-64] The artificial drift of the unwrapped phase towards positive values is eliminated by adding small random complex numbers to each nucleotide complex representation, to make phase and difference of phase close to − π, equal probable to phase close to π. For virus genomes, unwrapped phase function which attaches zero phase change to all neutral transitions, are used for study.

Nucleotide Path Analysis

Based on the construction of the cumulated sum of the vectors representing the nucleotides in a sequence, Nucleotide path analysis is done. The 3-D representation in this analysis is done by

$x = n_W - n_S$ .....(35)

$y = n_M - n_K$ .....(36)

$z = n_R - n_Y$ .....(37)

On Projecting the 3-D representation along 0y plane, the x-coordinate becomes the real component and the z-coordinate becomes the imaginary component, resulting into 2-D representation. A third dimension, corresponding to the advancement along the DNA strand, can be added to the previous 2D nucleotide path. This 3D diagram shows the gradual accumulation of the differences of nucleotide species along the DNA sequence and allows an easy comparison of various isolates of the same virus.

Independent Component Analysis

A convenient way to reveal single nucleotide polymorphisms (SNPs) and other variability induced changes in a set of related nucleotide sequences, and to establish possible links between such events, is the Independent Component Analysis (ICA). The set of nucleotide sequences could correspond to isolates of the same virus in different patients or in different phases of change under the combined selection pressure of the immune response and of treatment. ICA allows separating statistically independent variations, thus showing the links between simultaneously occurring changes, sometimes at rather distant locations along the DNA strands.

ICA assumes that a set of observable signals $x_1, x_2, \ldots, x_n$ are linear mixtures of some not directly accessible, but statistically independent, source signals $s_1, s_2, \ldots, s_n$. Equations of the form:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \Delta + a_{jn}s_n = \sum_{k=1}^{n} a_{jk}s_k \quad .....(38)$$

are assumed for all j = 1,..., n, with constant or almost constant entries of the mixing matrix entries a_{jk}; *j,k = 1,..., n*. The problem is the "blind source separation", i.e., finding (all or some of) the independent components s_k, when knowing only the mixed signals $x_1, x_2, \ldots, x_n$, but not the mixing matrix entries. The solution is based on a hypothesis which infers a reciprocal of the Central Limit Theorem, i.e., the hypothesis that independent (non-mixed) signals should be non-gaussian.[65] Various measures of non-gaussianity or "contrast functions" can be used, such as kurtosis, negentropy, mutual information, Kullback-Leibler divergence, or maximum likelihood estimation. The results in this paper have been obtained by using the Fast ICA algorithm, a free (GPL) MATLAB package available at[66], which implements a fast fixed-point algorithm for ICA and projection pursuit.

Cluster analysis

For investigating the reliability of the ICA estimates, a clustering analysis by using *icasso* has been done, an interactive visualization method and software package available. Its basic principle is to repeatedly run ICA algorithm, and visually estimate the clustering of the estimated independent components in the signal space. Basically, the clusters should be compact, containing components that are close to each other and well separated from the rest.

Phylogenetic analysis

The analyzed DNA sequences has to be visually inspected, aligned, gap stripped and validated using the multiple sequence editor BioEdit . A pairwise distance matrix can be generated with the DNA Dist program of the Phylip package. The tree topology has been inferred by the neighbor-joining method with the Fitch-Margoliash and least-square methods.[59]

The above analysis performed with the signals obtained, reveals various characteristics of the sequences and helps in understanding the infected genes where variability is obtained and its origin along with its prevalence.

TOOLS USED FOR SIGNAL PROCESSING

Digital Filtering Technique

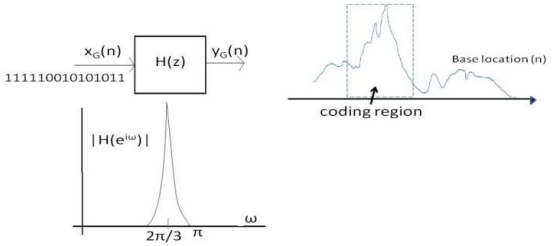
It is used for gene identification. The exons exhibit a

period -3 behavior that is not found in other parts of the DNA molecule, in prokaryotes and most eukaryotes. This period -3 property is due to codon bias.[67] Techniques which exploit this property for gene prediction, proceeds by computing the discrete Fourier Transform (DFT) and exhibits a peak at the frequency of $2\pi/3$ due to the periodicity. The information of gene location is given as a function of base location, by the output signal of antinotch filter with a sharp frequency of $2\pi/3$. Here there is a compromise between sharpness of notch filter and base domain resolution achieved, but this the four bases is calculated as binary sequence, at a particular point and the strength of peak is observed at any particular point where it is immensely pronounced.[68] Background spectrum should be dominated by periodicity, for which the window length is maintained large. From the background information like *1/f* noise, period-3 behavior can be effectively isolated, if the filter in the sliding window having a simple impulse response

$$\omega(n) = e^{j\omega on}\ 0 \le n \le N-1;\ 0 \text{ otherwise} \qquad .....(39)$$

is designed such that it has a pass band centre at $\omega_o = 2\pi/3$ and minimum stop band attenuation of about 13dB. Computational complexity can be reduced for the better design and implement of the filter. The filter can be designed such that narrow band pass digital filter H (z) is taken centered at $\omega_o = 2\pi/3$ and $x_G(n)$ is taken as input and $y_G(n)$ as output, n being the base location. $y_G(n)$ is expected to be large in coding regions and $x_G(n)$ to have period -3 component i.e. it has large energy in filter pass band exemplified below in Figure **5**.

Figure 1. A digital filter H(z) with indicator sequence xG(n) as its input. (Source: Yoon, 2004)



The further designing and development of notch filters has been better described by Vaidhyanathan and Yoon.[68] Base **G** dominates at certain codon position in coding regions, attributing this feature to period-3 property.

Long Range Correlations in DNA

Base pairs far away from each other in a DNA sequence were found correlated. The genes and introns present far away from each other in long strand of DNA were found correlated. But on the contrary, no correlation was found in intronless genes and complementary DNA, resulting into concept called DNA walk. Auto correlation was determined for the indicator sequence thus: If $x_A(n)$ be indicator for base A, then

$$r_A(k) = \sum x_A(n)\, x_A(n-k) \qquad .....(40)$$

Here the sum extends to all 'n' and the product is nonzero. Fourier transforms of the above equation gives the power spectrum for base A:

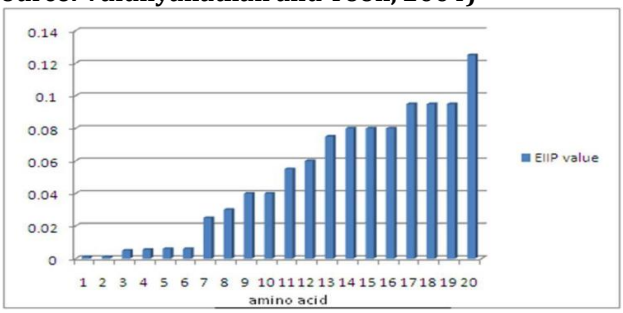
$$S_A(e^{j\omega}) = S_A(e^{j2\pi f}) = \sum r_A(k) e^{-j2\pi kf} \qquad .....(41)$$

Here $S_A(e^{j\omega}) = |X_A(e^{j\omega})|^2$. Explanation of *1/f* β behavior or the power law of the power spectrum where β value varies from base to base is done by Voss (1992). Power spectra has the property that it has peaks at frequency $f = 1/3$ due to codon structure. The *1/f* behavior implies an unbounded component at $f = 0$. At zero frequency, an impulse in Fourier transform represents a constant component in the time domain. This shows a slow decay term in auto correlation sequence, giving rise to long range correlation.[69]

Fourier Transform

A mathematical way to identify a common function among the given set of proteins, is based on Fourier techniques; known as Resonant Recognition model (RRM).[70] A non negative number ranging from 0.0 to 0.1263 called Electron ion Interaction Potential (EIIP: average – ion electron potential) is associated with each amino acid molecule as depicted in the plot of the electron – ion interaction potential as shown below in Figure **6**.

Figure 2. Plot of Electron ion Interaction Potential. (Source: Vaidhyanathan and Yoon, 2004)



Then for any given protein, the value of nth amino acid in protein is given by equating the numerical sequence $x(n)$ of the protein to EIIP thus,

$$X(e^{j\omega}) = \Sigma x(n) e^{-j\omega n} \qquad .....(42)$$

is calculated to get fourier transform, where N is length of amino acid sequence determining the protein.[69]

Karhunen – Loeve Transforms

Gene expression can be best analyzed with the data recorded on the microarray. Applying the signal processing methods, this includes normalization[71], data clustering, denoising, and data interpretation by linear transformations. A matrix is constructed from expression levels of N genes at M different occasions; and using Eigen arrays, Eigen genes are formed. The Eigen vectors corresponding to dominant singular values σ_0 and σ_1 closely approximate sines and cosines.[53] They can capture gene expression as a function of time. Wave pattern of the ordered genes is expressed in the micro array data.[70,71]

Such electrical based techniques, when integrated with genes and signals are obtained in digital form, are interpreted for variability, graphs obtained are studied for test of variance and statistical analysis is done.

CONCLUSION

The review paper explicates about the sequences of infected genes and their representation in complex quadrantal form by forming tetrahedron of the nucleotides. The tetrahedron is representing symmetry and degeneracy. Real representations are also made using geometrical make up. Large scale features i.e. at scales of 10^6 -10^8 base pairs have been detected for analyzing signals. Symbolic sequence having one – base resolution can be constructed if the information of DNA strand is conserved by both cumulated and unwrapped phase. Interesting behavior of the cumulated phase has been found for most sequenced prokaryote genomes. The circular chromosomes are divided in two almost equal domains in which the slope of the variation of the cumulated phase along the strand has opposite signs and which can be put in correspondence with the *replichores* of the chromosome, in a way similar to the approach based on skew diagrams. Genome signals corresponding to virus displaying similar characteristically property is curbed

and the obtained signals are decomposed to common reference. Variability signals are applied with description to average, mean and maximum flat references. Gene prediction done using Hidden Markov Models (HMM) is further processed for their signals using genomics techniques.

REFERENCES

1. Phyllis J Kanki, Seema Thakore Meloni,. Biology and Variation in HIV-2 and HIV-1. Text Serial Journal. 2009; 2: 1-24.
2. Nauclér A, Andreasson P A, Costa C M, Thorstensson R, Biberfeld G;. HIV-2-associated AIDS and HIV-2 seroprevalence in Bissau, Guinea-Bissau. *J Acquir Immune Defic Syndr*. 1989; 2(1):88-93.
3. Poulsen A G et al. Prevalence of and Mortality from Human Immunodeficiency Virus type 2 in Bissau. West Africa. 1989; 333(8642):827-831.
4. Brun Vezinet F, De Cock K M, Soro B; HIV-1 and HIV-2 infections and AIDS in West Africa. AIDS. 1991; 5(1):21-28.
5. Jenny Buckland; Partners in crime. Nat Rev Immunol 2003; 3(6):442.
6. Kanki P J, Barin F et al. New human T-lymphotropic retrovirus related to simian T-lymphotropic virus type III (STLV-IIIAGM). 1986; 232(4747):238- 243.
7. Donnelly C, Leisenring W, Kanki P, Awerbuch T, Sandberg S and Fellow B; Comparison of transmission rates of HIV-1 and HIV-2 in a cohort of prostitutes in senegal. Bulletin of Mathematical Biology. 1993; 55(4):731-743.
8. P J Kanki et al. Slower heterosexual spread of HIV-2 than HIV-1. The Lancet. 1994; 343(8903):943-946.
9. M Grez et al. Genetic analysis of human immunodeficiency virus type 1 and 2 (HIV-1 and HIV-2) mixed infections in India reveals a recent spread of HIV-1 and HIV-2 from a single ancestor for each of these viruses. The Journal of Virology. 1994; 68(4):2161-2168.
10. Wilkins A, Ricard D, Todd J, Whittle H, Dias F and Paulo Da Silva A; The epidemiology of HIV infection in a rural area of Guinea-Bissau. AIDS. 1993; 7(8):1119-1122.
11. R Marlink et al. Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science. 1994; 265(5178):1587-1590.
12. Pepin J, Morgan G et al. HIV-2-induced immunosuppression among asymptomatic West African prostitutes: evidence that HIV-2 is pathogenic, but less so than HIV-1. AIDS. 1991; 5(10):1165-1172.
13. F Gao et al. Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. The Journal of Virology. 1994; 68(11):7433-7447.
14. Beatrice H Hahn, David L Robertson and Paul M Sharp; Intersubtype Recombination in HIV-1 and HIV- 2. Recombination in HIV 1 and HIV 2. 1995; 3:22 -29.
15. T F Schulz et al. Biological and molecular variability of human immunodeficiency virus type 2 isolates from The Gambia. The Journal of Virology. 1990; 64(10):5177-5182.
16. C Cheng-Mayer; HIV-1 variation: consequences for disease progression and vaccine strategies. Cancer Research Institute, University of California, School of Medicine, San Francisco. 1993; 1(9):353-355.

ACKNOWLEDGEMENT

Corresponding author thanks faculty at School of Biotechnology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India for constant accompaniment, during the tenure of paper writing.

17. Koff W C, Schultz A M; Prospects for an AIDS vaccine. Semin Immunol. 1990; 2(5):351-359.
18. Almond N, Jenkins A, Taffs L F, Heath A B and Kitchln P; The genetic evolution of the envelope gene of simian immunodeficiency virus in cynomolgus macaques infected with a complex virus pool. Virology. 1992; 191(2):996-1002.
19. Campbell B J and Hirsch V M; Extensive envelope heterogeneity of simian immunodeficiency virus in tissues from infected macaques. The Journal of Virology. 1994; 68(5):3129-3137.
20. B H Hahn et al. Genomic diversity of the acquired immune deficiency syndrome virus HTLV-III: different viruses exhibit greatest divergence in their envelope genes. Proceedings of the National Academy of Sciences. 1985; 82(14):4813-4817.
21. M H Bayon-Auboyer et al. Evolution of the human immunodeficiency virus type 2 envelope gene in preimmunized and persistently infected rhesus macaques. The Journal of Virology. 1994; 68(5):3415-3420.
22. Z Chen et al. Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. The Journal of Virology. 1996; 70(6): 3617-3627.
23. T Tolle et al. Variability of the env gene in cynomolgus macaques persistently infected with human immunodeficiency virus type 2 strain ben. The Journal of Virology. 1994; 68(4):2765-2771.
24. S W Barnett et al. An AIDS-like condition induced in baboons by HIV-2. Science. 1994; 266(5185):642 -646.
25. G Franchini et al. Persistent infection of rhesus macaques with a molecular clone of human immunodeficiency virus type 2: evidence of minimal genetic drift and low pathogenetic effects. The Journal of Virology. 1990; 64(9):4462-4467.
26. Otten R A, Brown B G et al. Differential replication and pathogenic effects of HIV-1 and HIV-2 in Macaca nemestrina. AIDS. 1994; 8(3):297-306.
27. Putkonen P et al. Experimental infection of cynomolgus monkeys (Macaca fascicularis) with HIV-2. J Acquir Immune Defic Syndr. 1989; 2(4):366 -373.
28. Stahl-Hennig C et al. Experimental infection of macaques with HIV-2ben, a novel HIV-2 isolate. AIDS; 1990; 4(7):611-617.
29. Franchini G et al. The human immunodeficiency virus type 2 (HIV-2) contains a novel gene encoding a 16 kD protein associated with mature virions. AIDS research and human retroviruse. 1988; 4(4):243-250.
30. The Genome Data Base, http://gdbwww.gdb.org/, Genome Browser, http://genome.ucsc.edu, European Informatics Institute, http://www.ebl.ac.uk, Ensembl., http://www.ensembl.org.
31. National center for Biotechnology Information, NLM,

NIH, ftp://ncbi. nlm. nih. gov/ genoms/ H.sapiens/, Genbank, http:// www. ncbi.nlm.nih.gov/ Genbank/ index.html.

32. Jean-Michel Claverie; Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences. Human Molecular Genetics. 1997; 6(10):1735-1744.
33. W Ford Doolittle, Phylogenetic Classification and the Universal Tree. Science. 1999; 284(5423):2124-2128.
34. R Durbin; Biological sequence analysis: Probabilistic models of proteins and nucleic acids (Cambridge university press) 1998.
35. Cristea P D; Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine. 2002; 6(2):279-303.
36. Cristea P D; Genomic Signals of Reoriented ORFs. EURASIP Journal on Advances in Signal Processing. 2004; 1:132-137.
37. Cristea P D; Multiresolution phase analysis of genomic signals. in First International Symposium on Control, Communications and Signal Processing, (presented at the First International Symposium on Control, Communications and Signal Processing, Hammamet, Tunisia). 2004; 743-746.
38. Cristea P D; Genomic signals of chromosomes and of concatenated reoriented coding regions. in Proceedings of SPIE (presented at the Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues II, San Jose, CA, USA). 2004; 29-41.
39. Cristea P D; Large scale features in DNA genomic signals. Signal Processing. 2003; 83(4):871.
40. Cristea P D; Large-scale and global features of complex genomic signals. (SPIE) 2003.
41. Cristea P D; Whole chromosome features of genomic signals. Presented at the 2002, 6th Seminar on NEUREL 2002, Belgrade, Yugoslavia. 2002; 1-4.
42. Cristea P D; Genetic signal analysis. Presented at the ISSPA 2001. Sixth International Symposium on Signal Processing and its Applications, Kuala Lumpur, Malaysia: IEEE. 2001; 703-706.
43. Cristea P D; Genomic signals for whole chromosomes. Presented at the Manipulation and Analysis of Biomolecules, Cells, and Tissues, San Jose, CA, USA. 2003; 194-205.
44. Cristea P D; Genetic signal representation and analysis. Presented at the Functional Monitoring and Drug-Tissue Interaction, San Jose, CA, USA: Manfred D Kessler. 2002; 77-84.
45. Dimitris Anastassiou; "Frequency-domain analysis of biomolecular sequences," Bioinformatics. 2000; 16(12):1073-1081.
46. Cristea P; Real and complex genomic signals. Presented at the 14th International Conference on Digital Signal Processing (DSP2002), Santorini, and Greece: IEEE. 2002; 543-546.
47. Chargaff E; Structure and function of nucleic acids as cell constituents. 1951; 10(3):654-659.
48. Vaidyanathan P P and Yoon B J; Gene and exon prediction using allpass-based filters. In Workshop on Genomic Signal Processing and Statistics (USA). 2002; 3-9.
49. Vaidyanathan P P and Yoon B J; Digital filters for gene prediction applications. Presented at the Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA. 2002; 306-31.
50. Peng C K et al. Long-range correlations in nucleotide sequences. Nature. 1992; 356(6365):168-170.
51. Wornell G W; A Karhunen-Loeve-like expansion for 1/f processes via wavelets. IEEE Transactions on Information Theory. 1990; 36(4)859-861.
52. Alter O, Brown P O and Botstein D; Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences. 2000; 97(18):10101-10106.
53. Alter O, Brown P O and Botstein D; Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proceedings of the National Academy of Sciences. 2003; 100(6):3351-3356.
54. Patrick O, Brown and David Botstein; Exploring the new world of the genome with DNA microarrays. Nat Genet. 1999; 21:33-37.
55. Cristea P D; Invariants of DNA genomic signals (SPIE). 2005.
56. Matthew X H, Sergei V P and Ricci P; Genetic code, hamming distance and stochastic matrices. Bulletin of Mathematical Biology. 2004; 66(5):1405-1421.
57. Cristea P D; Representation and Analysis of DNA sequences. Genomic Signal Processing and Statistics, Editors EG Dougherty, I Shmulevici, Jie Chen, ZJ Wang, Book Series on Signal Processing and Communications, Hidawi. 2005; 15-65.
58. Cristea P D; Pathogen Variability. A Genomic Signal Approach. 2006; 1(3):25-32.
59. Cristea P D, Dan Otelea M D and Tuduce R A; Genomic signal analysis of HIV variability. in Proceedings of SPIE (presented at the Imaging, Manipulation, and Analysis of Biomolecules and Cells: Fundamentals and Applications III, San Jose, CA, USA: SPIE). 2005; 362-372.
60. Jan O Andersson, W Ford Doolittle and Camilla L Nesbø; Are There Bugs in Our Genome? Science. 2001; 292(5523):1848 -1850.
61. Henry Gee; A journey into the genome: what's there? nature news. 2001.
62. The Genome Data Base, http://gdbwww.gdb.org/, Genome Browser, http://genome.ucsc.edu, European Informatics Institute, http://www.ebl.ac.uk, Ensembl, http://www.ensembl.org.
63. National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, http://www.ncbi.nlm. nih.gov/, ftp://ftp.ncbi.nlm.nih.gov/genoms/, GenBank, http: //www.ncbi.nlm.nih.gov/Genbank/index.html.
64. Hyvärinen A, Karhunen J and Oja E; Independent component analysis. Adaptive and learning systems for signal processing. Communications and control. (Canada, USA: Wiley-Interscience). 2001; 26.
65. H Herzel et al. Interpreting correlations in biosequences. Physica A: Statistical Mechanics and its Applications. 1998; 249(1-4)449-459.
66. Helsinki University of Technology, Laboratory of Computer and Information Science, Neural Networks Research Center, "The FastICA algorithm for independent component analysis and projection pursuit", http://www.cis.hut.fi/projects/ica/fastica/.

67. Vaidyanathan P P and Byung-Jun Yoon. The role of signal-processing concepts in genomics and proteomics. Journal of the Franklin Institute. 2004; 341(1-2):111-135.
68. Maria de Sousa Vieira; Statistics of DNA sequences: A low-frequency analysis. Physical Review E 60, 5, Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1999; 5932-5937.
69. Richard F Voss; Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical Review Letters. 1992; 68(25):3805-3808.
70. Cosic I; Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications. IEEE Transactions on Biomedical Engineering. 1994; 41(12):1101-1114.
71. Yue Wang et al. Iterative normalization of cDNA microarray data. IEEE Transactions on Information Technology in Biomedicine. 2002; 6(1):29-37.