Graphical Method to Determine Sequence Variation on NA Protein of H5N1 Virus using Discrete Wavelet Transform

Shiwani Saini
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, India
shiwani_saini76@yahoo.com

Lillie Dewan
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, India
l_dewanin@yahoo.com

Received: June 16, 2019. Revised: September 27, 2021. Accepted: October 19, 2021. Published: November 23, 2021.

***Abstract*—Influenza A virus belongs to the Orthomyxoviridae family and its genome is divided in eight distinct linear segments of negative-sense single stranded ribonucleic acid (RNA). Of all the eight influenza protein sequences, mutations in hemagglutinin and neuraminidase proteins show significant variations in their sequences. The threat of Influenza pandemic is ever rising due to its constant antigenic drift. Thus there is a need to characterize the genomic information in these viruses and signal processing methods offer the advantage of faster analysis in comparison to conventional techniques. Genomic information is converted into digital form by representation of the nucleotide bases in the form of mathematical sequences. In this paper, sequence variations of H5N1 virus have been studied using wavelet transforms as a signal analysis technique. Nucleotide sequences of neuraminidase protein of influenza virus occurring in different regions, in different hosts and over different years have been downloaded from National Centre for Biotechnology Information (NCBI) database. Sequences are aligned and converted into mathematical sequences then transformed using wavelet transforms. Graphical representations of the transformed sequences have been used to localise the regions of mutations along the sequence length.**

Keywords*— Influenza A virus, wavelet transforms, genomic sequences, signal processing, neuraminidase protein

I. Introduction

A typical feature of influenza A viruses is that its genome is divided in eight distinct linear segments of negative-sense single stranded RNA [1] including: HA (hemagglutinin), NA (neuraminidase), NP (nucleoprotein), M (two matrix proteins, M1 and M2), NS (two distinct non-structural proteins, NS1 and NEP), PA (RNA polymerase), PB1 (RNA polymerase and PB1-F2 protein), and PB2 (RNA polymerase) [2]. Of all influenza proteins, mutations in hemagglutinin (HA) and neuraminidase (NA) are significant [3]. Since the genome is segmented, it favours the exchange of entire genes between different viral strains which increases the probability of mutations between strains cohabitating the same cell.

In general, an influenza virus infects only a single species; however, whole viruses may occasionally be transmitted from one species to another, and genetic reassortment between viruses from two different hosts can produce a new virus capable of infecting a third host [4] and thus causing a pandemic. With the growing volume of deoxyribonucleic acid (DNA) sequence database of the human and model organisms since the completion of Human Genome Program (HGP), signal processing methods offer the advantage of faster analysis.

This genomic sequence information can be converted into digital form by representation of the nucleotide bases in the form of mathematical sequences [5]. Using signal processing principles to analyze genomic sequences requires defining an adequate representation of the nucleotide bases by numerical values, converting the nucleotide sequences into an equivalent of time series [6] wherein nucleotide bases are represented on the x-axis and mathematical values assigned to the sequence are represented on y-axis.

There are several mathematical transforms such as Fourier Transforms, Short Time Fourier Transforms, Wavelet Transforms, Hilbert Transforms, etc. which can be applied to the digital data to obtain information from that signal that is not readily available in the raw signal [7]. Furthermore graphical representation methods for the transformed DNA sequences help in the better interpretation of the biological properties in the graphical domain such as sequence visualization and analysis. Wavelet Transforms of genomic signals can be used for graphical representation of the data and offer the advantage of analysing the regions of interest within the DNA sequence without the use of pattern matching methods. Several methods have been used for determining the variations of sequences that include molecular characterisation [8, 9], Z curve method [10], phase analysis of genomic sequences [11], graphical sliding window techniques [12].

In the present work, the variations undergone by viral strains of NA protein of H5N1 virus on account of mutations have been determined. The paper gives an introduction about wavelet transforms. The nucleotide sequences are converted into integer number representations and then transformed using Haar wavelets. Plots of wavelet coefficients of different sequences show significant variations in sequences occurring in different regions and in different hosts.

II. Wavelet Transforms

Wavelet is a waveform of finite duration and zero average value. Wavelet transform (WT) is obtained using a wavelet function ψ(t), in which the original signal is convolved with the scaled and shifted version of the mother wavelet. Wavelet transforms are capable of transforming the signal simultaneously in both time and frequency domain and hence offer the advantage of time frequency localisation of any event. There are two types of wavelet transforms: Continuous wavelet transforms (CWT) and Discrete wavelet transforms (DWT)

Continuous wavelet transforms generate a large amount of data as the transform is calculated at all possible scales and positions. In discrete wavelet analysis, scales and positions are chosen based on powers of two called the dyadic scales. After discretization the wavelet function is defined as (2):

$$\psi_{\mathrm{m,n}}(t) = \frac{1}{\sqrt{a_0^{m}}}\psi^*\left(\frac{t - nb_0a_0^{m}}{a_0^{m}}\right) \qquad (2)$$

where a0 and b0 are constants. The scaling term is represented as a power of a0 and the translation term is a factor of a0^{m}. The most common choice for the parameters a0 and b0 are 2 and 1 (dyadic grid scaling). The dyadic grid wavelet is expressed as (3):

$$\psi_{\mathrm{m,n}}(t) = \frac{1}{\sqrt{2^{m}}}\psi\left(\frac{t - n2^{m}}{2^{m}}\right) = 2^{-m/2}\psi(2^{-m}t - n) \qquad (3)$$

An efficient way to implement this scheme using filters was developed by Mallat [13] in 1988. For many signals, the low-frequency content is the most important part. It is what gives the signal its identity. The high-frequency content, on the other hand, imparts flavour or nuance. The most basic filtering process is represented by (Fig. 1).

The original signal passes through a pair of high pass and low pass filters, is down sampled to get the decomposed signal through each filter which is half the length of the original signal. The signal S can be expressed as S = cD + cA.

After the analysis of the signal, the original signal can be synthesised using inverse discrete wavelet transform. The signal is reconstructed as shown in Fig. 2. Reconstruction involves the up sampling of the decomposed signal followed by filtering through two complementary filters. The complete decomposition and reconstruction process of the signal with the low- and high-pass decomposition filters (L and H) and their associated reconstruction filters (L' and H') form a system of quadrature mirror filters (Fig. 3).

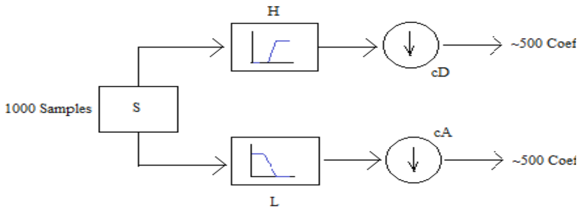

Fig.1 Signal decomposition using DWT

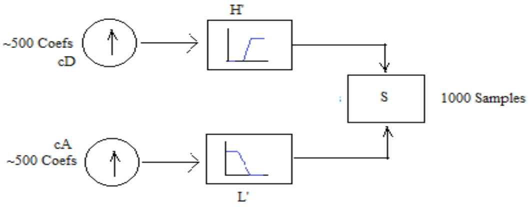

Fig.2 Signal reconstruction using DWT

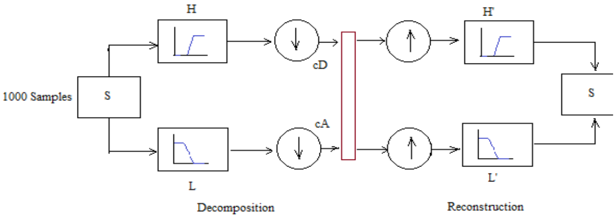

Fig.3 Signal decomposition and reconstruction

III. Genomic Signals and Representation

The main nucleic genetic material of cells is represented by DNA molecules that have double helix structure comprising of two antiparallel intertwined complementary strands, each a helicoidally coiled heteropolymer. Four kinds of nitrogenous bases are found in DNA that constitute the genomic sequences: thymine (T) and cytosine (C)—which are pyrimidines, adenine (A) and guanine (G)—which are purines. A pyrimidine in one chain always faces a purine in the other along the two strands of DNA double helix, and only the base pairs T−A and C−G exist. As a consequence, the two strands of a DNA helix are complementary, store the same information, and contain exactly the same number of A and T bases and the same number of C and G bases.To express the genomic sequences mathematically, there are several methods such as Voss representation [14], purine (A, G = -1) and pyrimidines (C, T = +1) representation [15], mapping of the nucleotides onto a complex tetrahedral plane [16], complex number representation [17], electron ion interaction potential (EIIP) [18] and integer number representation (Bioinformatics toolbox in Matlab).

.

IV . Method

Different nucleotide sequences of NA protein of H5N1 virus taken from the different hosts occurring in different years and in different regions (Vietnam, India and Egypt) were downloaded from NCBI database [19]. The sequences were first aligned using nucleotide BLAST. The aligned sequences were represented in mathematical form using integer number representation (A=1, C=2, G=3, T=4) and then transformed by discrete wavelet transform using Haar wavelet. The decomposition level was chosen to be 4. Graphical comparison of the transformed sequences at different levels were used to investigate the variations in the sequence composition of different strains of Influenza virus.

Multi resolution analysis up to level 4 decomposes the sequences into approximations and details at various levels (1-4). Whereas wavelet coefficients at coarse scales correspond to low frequency components in the signal (approximations) and capture gross and global features of the sequence, wavelet coefficients at fine scales (details) correspond to high frequency information and contain local varaitions. The local variation that is point to point variation of the signal can be determined by plotting the difference of the detail coefficients at level 4 of the transformed reference sequence and subject sequence. The gross variation trend in the sequence can be determined by comparing the approximation plots at level4.

V. Results

The approximation and detail wavelet coefficients plots of NA protein at level 4 , occurring in the same region but in different years were compared to determine variation in nucleotide sequences. The approximation coefficients plots of NA protein of H5N1 occurring in Vietnam are shown in Fig.4. The graph shows global variations in the sequences of the protein originating in different years. The differences appear due to the different years of characterisation (2003-2012). These variations correspond to the mutations undergone by the sequences. The approximation plots only represent the global trend in a particular sequence. However the regions where the sequences have undergone mutations can be localized by comparing the detail coefficients plots for all the sequences occurring in a particular region with respect to a reference sequence, as the detail coefficients represent the high frequency information. As shown in Fig. 5, 6, different sequences of NA protein sampled from different hosts and occurring in different years in Vietnam were compared taking one sequence as a reference sequence (Accession Number AB741568.1). Since the sequences are aligned, the difference in the detail plots of the reference sequence and subject sequence show rectangular peaks that give information about the local variations undergone by the sequences. If the reference and subject sequences are similar, the difference plots of detail coefficients do not show any peaks. Thus by visual analysis of the differences in the detail coefficients plots, the regions where the sequences have undergone mutations can be identified. Similarly, the coefficients plots of sequences sampled from India over different years show the regions where the sequences have undergone mutations (Fig.7-9). The reference sequence was chosen as CY089477.1. Fig.10,11 give the plots of sequences sampled from Egypt. The reference sequence was chosen as CY126266.1.

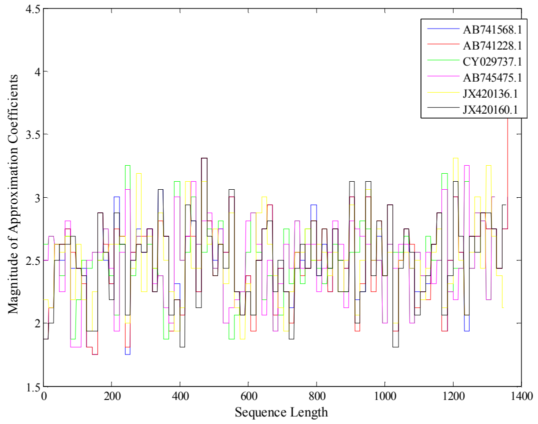

Fig. 4 Approximation coefficients plot of NA protein (Vietnam)

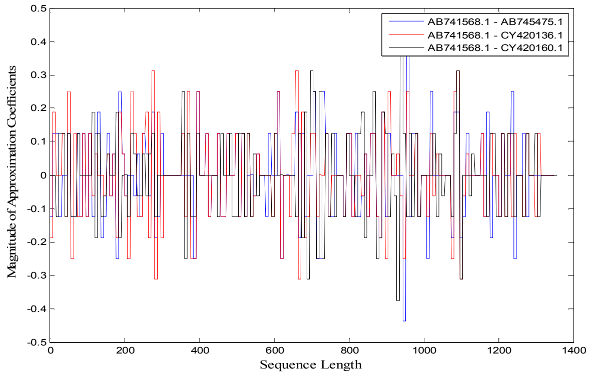

Fig.5 Detail coefficients plot of NA protein (Reference sequence – Subject sequence,Vietnam)

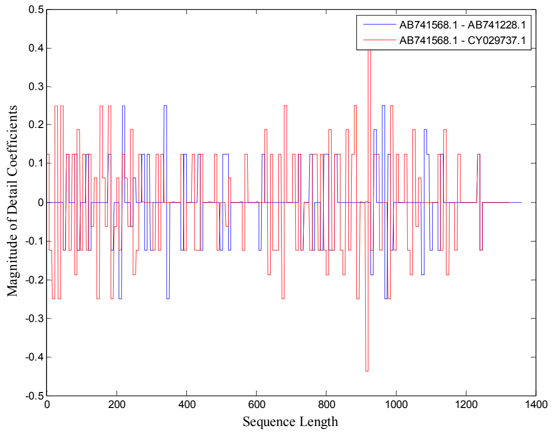

Fig.6. Detail coefficients plot of NA protein (Reference Sequence – Subject Sequence, Vietnam)

.

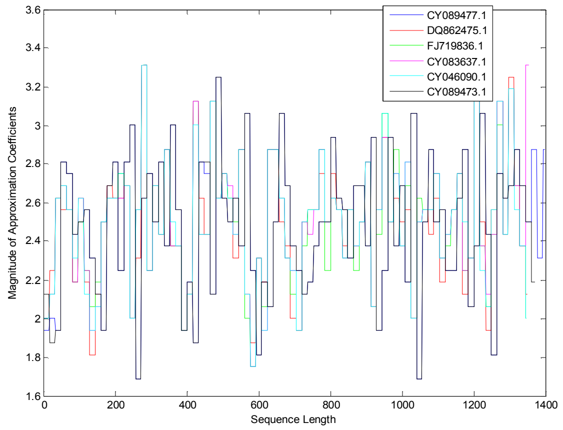

Fig. 7 Approximation coefficients plot of NA protein (India)

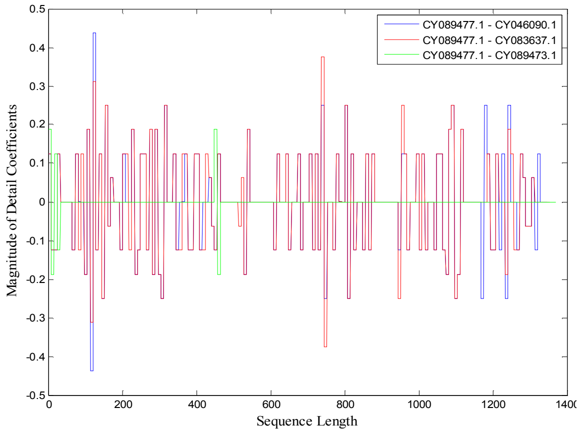

Fig.8. Detail coefficients plot of NA protein (Reference Sequence – Subject Sequence, India)

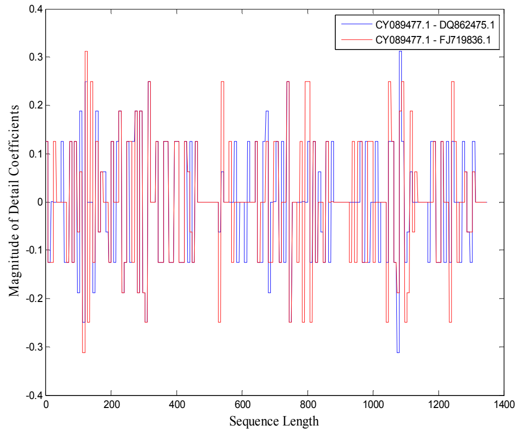

Fig.9 Detail coefficients plot of NA protein (Reference Sequence– Subject Sequence, India)

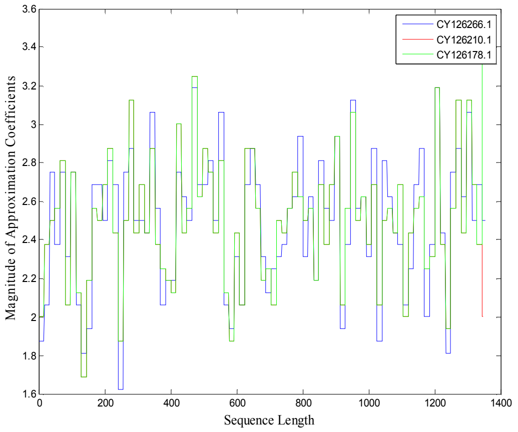

Fig.10 Approximation coefficients plot of NA protein (Egypt)

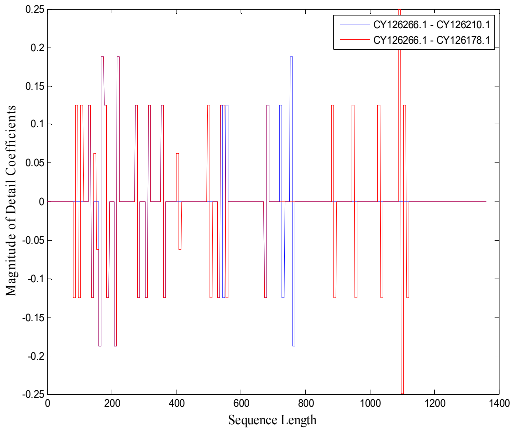

Fig.11 Detail coefficients plot of NA protein (Reference Sequence – Subject Sequence, Egypt)

VI. Conclusions

Wavelet transforms offer the advantage of reducing the computational complexity as compared to mathematical calculations and faster identification of sequence changes by visual representation of the plot of the wavelet coefficients. The positions where the sequences undergo changes can be interpreted by graphically representing the difference of the detail plots of a reference sequence and subject sequence. The mutations are evident in the form of rectangular peaks. Identical sequences do not show any peaks in the difference plots. These graphical plots help in visual identification of relatively stable regions in different protein sequences of H5N1 that can be used as the target regions in diagnosis for determining drug resistance and also in vaccine manufacturing.

Thus the combined use of the digital signal processing methods and bioinformatics provides a simple tool for analyzing the interactions in the viral sequences accumulating in unprecedented large numbers from throughout the world during the epidemics and can be used for vaccine design and new diagnosis development.

References

[1] E. Fodor and G.G. Brownlee. in (Potter, C.W., ed.) “Influenza”, Elsevier, Amsterdam, pp. 1-29, 2002.

[2] Claas E C, Osterhaus A D, and van Beek R , “Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus,” Lancet, 351, pp.472–477, 1998.

[3] Scholtissek, C., H. Burger, O. Kistner, and K. F. Shortridge, “The nucleoprotein as a possible major factor in determining host specificity of influenza H3N2 viruses”, Virology, 147, pp. 287–294, 1985.

[4] Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, Xu X, “Triple-reassortant swine influenza A (HI) in humans in the United States, 2005-2009”, N Engl J Med., 360, pp. 2616-25, 2009.

[5] P. Cristea, “Conversion of nitrogenous base sequences into genomic signals”, Journal of Cellular and Molecular Medicine, 6, 2, pp. 279-303 April - June, 2002.

[6] Juan V. Lorenzo-Ginori, Aníbal Rodríguez-Fuentes, Ricardo Grau Ábalo and Robersy Sánchez Rodríguez,” Digital signal processing in the analysis of genomic sequences”, Current Bioinformatics, pp.28-40, 4, 2009.

[7] Robi Polikar (1999), The Wavelet Tutorial [online]. Available: http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html

[8] David L. Suarez et al., “Comparisons of highly virulent h5n1 influenza a viruses isolated from humans and chickens from Hong Kong”, Journal Of Virology, pp. 6678–6688, Aug. 1998.

[9] Elodie Ghedin et al., “Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution”, Vol 437(20 October 2005) doi:10.1038/nature04239.

[10] Yan-ling Yang1, “A geometrical analysis of the avian influenza virus at different areas”, Proc. International Conference on Engineering Computation, pp. 229-231, 2009 [International Conference on Engineering Computation, May 2-3, Hong Kong, 2009].

[11] Paul Dan Cristea, Rodica Tuduce,, “Genomic signal analysis of avian influenza virus variability”, Proc. Second International Symposium on Communications, Control and Signal Processing, 2006, paper 1369, ISBN 2-00848-17-8 [International Symposium on Communications, Control and Signal Processing, Marrakech, Morocco, 13-15 March 2006,].

[12] Ghosh et al, “Computational analysis and determination of a highly conserved surface exposed segment inH5N1 avian flu and H1N1 swine flu neuraminidase”, BMC Structural Biology, 10: 6,2010.

[13] S. Mallat, A Wavelet Tour of Signal Processing, 2nd ed, Academic Press, New York, 2000.

[14] Richard F. Voss, “Evolution of long-range fractal correlations and 1/f noise in dna base sequences,” Physical Review Letters, vol. 68, pp. 3805-3808, June 1992.

[15] Swarna Bai Arniker and Hon Keung Kwan, “Graphical representation of dna sequences”, Proc. IEEE International Conference on Electro/Information Technology, EIT 2009, pp.311-314 [IEEE International Conference on Electro/Information Technology, EIT , Windsor, ON, Canada, June 7-9, 2009].

[16] P. D. Cristea, “Conversion of nucleotides sequences into genomic signals,” J. Cell. Mol. Med., vol. 6, no. 2, pp. 279–303, 2002.

[17] J.A. Berger, S.K. Mitra, M. Carli, A. Neri, “New approaches to genome sequence analysis based on digital signal processing”, in: Workshop on Genomic Signal Processing and Statistics (GENSIPS), IEEE, Raleigh, North Carolina, USA, 11–13 October 2002, pp. 1–4, CP2-08.

[18] Achuthsankar S Nair and Sivarama Pillai Sreenadhan’ “A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)”, Bioinformation 1(6), pp. 197-202,2006.

[19] The National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine website. [Online]. Available : http://www.ncbi.nlm.nih.gov/genoms/,ftp://ftp.ncbi.nlm.nih.gov/genoms/,GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html.