Graphical Method to Determine Base Change Locations in Genomic Sequences of Influenza A Virus Using Wavelets

SHIWANI SAINI
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, Haryana
INDIA
shiwani_saini76@yahoo.com

LILLIE DEWAN
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, Haryana
INDIA
l_dewanin@yahoo.com

Abstract: - Single base substitutions occur in genomic sequences when a single base gets replaced by another base. These base substitutions are also called point mutations. Mutation detection is important as these are linked to genetic disorders. Several direct and indirect methods are available to detect sequence variation in specific regions of DNA (Deoxyribonucleic acid) sequences. However indirect methods cannot perfectly identify the mutation location in the sequence. Also these methods require mutation confirmation by visual analysis. Thus for automated detection of base substitutions, signal processing methods offer the advantage of simpler, faster and more accurate localisation of point mutations. In this paper, a novel graphical method using wavelet transforms to identify single base changes in H5N1 Influenza A virus has been proposed. The paper discusses the graphical plots of wavelet transformed Hemagglutinin (HA) and Neuraminidase (NA) nucleotide sequences to identify the locations of base changes.

Key-Words: -Base substitutions, genomic sequences, wavelet transforms, multiresolution decomposition, Influenza A virus, signal processing

1 Introduction

Most of the genetic disorders are linked to nucleotide substitutions thus their accurate identification is particularly important [1]. DNA sequencing is a direct method for detecting mutations as well as their locations in the DNA sequence. There are several indirect methods also for mutation detection such as denaturing gradient gel electrophoresis (DGGE), denaturing high performance liquid chromatography (DHPLC), temperature gradient gel electrophoresis (TGCE), single stranded DNA conformation analysis (SSCP), chemical or enzyme cleavage mismatch (CECM) to name a few. These indirect methods are not capable of localizing the mutation location and require confirmation by DNA sequencing. Fluorescence based sequencing instruments are capable of automated detection of single point mutations by direct comparison of sequence chromatograms of a reference and suspected mutant. Several software based methods such as Trace Difference,SeqDoc [3]ABI SeqEd, PE/ABI Factura[4] Mutation Surveyor and Mutation Explorer [5] determine mutation points by trace subtraction of a reference and a given sequence. These trace subtraction methods have the disadvantage of context effects and variations in the intensity of the sequences which further requires normalizing the traces before their comparison.

Although sequencing method is sufficient for the discovery of single-nucleotide variations, but simpler, faster, and more automated methods are needed for analyzing the large volumes of sequenced genomic data available data post the Human Genome Project. Methods utilizing signal processing techniques have the advantage of faster processing of already sequenced DNA sequences in comparison to conventional laboratory methods. Signal processing methods are able to reveal large scale features of DNA sequences at the scale of the whole genome or chromosomes [6]. Genomic signal processing approach has been used to study multiresistance mutations in HIV virus [7], H5N1 virus [8] and Mycobacterium Tuberculosis [9] to analyze and track the development of drug resistance.

A wavelet transform based graphical approach to identify base substitutions and hence mutation points in DNA sequences of Influenza A virus has been described in this paper. This method provides automated detection of sequence changes by visual analysis of the peaks in the plots of two compared DNA sequences: a reference and a subject. These graphs are the plots of the difference of wavelet coefficients of the transformed DNA sequences of Influenza A virus. Section 2 gives a brief introduction about the Influenza A virus whose analysis has been performed. An insight of wavelet transforms is given in section 3 followed by the method of sequence analysis using wavelet transforms (WT) in section 4 and observations in section 5.

2 Influenza A Virus and Representation

Influenza A virus is a member of the orthomyxoviridae family. The genome of influenza A virus is divided into eight distinct linear segments of negative-sense single stranded RNA [Fodor and Brownlee[10]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).

Influenza A and B viruses contain surface glycoproteins: the hemagglutinin (HA) and the neuraminidase (NA). Both proteins recognize the same host cell molecule, sialic acid. The HA segment of Influenza virus binds to sialic acid-containing receptors and initiates virus infection. NA segment helps to spread of the infection to neighboring cells by releasing progeny virus. The influenza virus undergoes a larger number of mutations in HA and NA proteins as compared to the other protein segments. These mutations in HA and NA proteins help in easy replication of virus in the host by allowing them to evade the host's immune response [11]. Thus studying the mutations in the HA and NA proteins is of utmost importance as they can help to track the efficacy of specific drugs in containing the infection during viral epidemics.

DNA is the main nucleic genetic material of the cells with a double helix structure and two antiparallel intertwined complimentary strands. There are four kinds of nitrogenous bases found in DNA that constitute the genomic sequences: thymine (T) and cytosine (C) - called pyrimidines, adenine (A) and guanine (G) - called purines. Base A always pairs with base T while base C always pairs with base G. Hence, the two strands of a DNA helix are complementary and contain exactly the same number of A-T bases and the same number of C-G bases.

To apply a signal analysis method, the genomic sequences have to be expressed mathematically first. There are several methods of mathematical representation such as Voss representation [12] purine – pyrimidine representation[13], mapping of the nucleotides onto a complex tetrahedral plane[14] complex number representation [15],electron ion interaction potential (EIIP)[16] and integer number representation.

3 Wavelet Transform

A brief insight of WT is being reproduced here from wavelet documentation [17]. A waveform of finite duration and zero average value is called a wavelet. WT is calculated using a mother wavelet function ψ(t), by convolving the original signal with the scaled and shifted version of the mother wavelet using Eq. (1). Mathematical transforms such Fourier Transforms (FT) and Short Time Fourier Transform (STFT) are also used in signal processing and analysis. Whereas FT only give information about the various frequency components in a particular signal, STFT does provide the time-frequency localization of the signal but in a fixed window frame. Wavelet transforms in comparison to FT and STFT, offer the advantage of time frequency localisation of a signal by using windows of varying sizes and hence are capable of multi resolution of signals. There are two types of wavelet transforms: continuous wavelet transforms (CWT) and discrete wavelet transforms (DWT).

$$Cab = \int_t f(t)\frac{1}{\sqrt{a}}\frac{\psi^*(t-b)}{a}dt \qquad (1)$$

Continuous wavelet transforms generate a large amount of data as the transform is calculated at all possible scales and positions and require larger computation time. 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 given in Eq(2), where a_0 and b_0 are constants.

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

The scaling term is represented as a power of a0 and the translation term is a factor of a_0^m.Values of the parameters a_0 and b_0 are chosen as 2 and 1 respectively and is called as dyadic grid scaling. The dyadic grid wavelet is expressed in Eq (3).

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

This scheme, implemented using filters was developed by Mallat[18]. The basic filtering process is represented in Fig. 1. The original signal is filtered through a pair of high pass and low pass filters and then down sampled to get the decomposed signal through each filter which is half the length of the original signal. The signal S can bewritten as s=cD+cA.

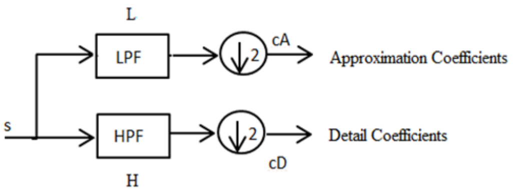

Fig. 1Signaldecomposition

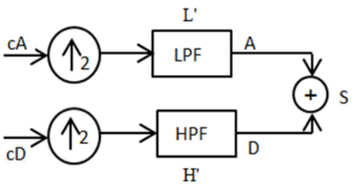

Fig. 2 Signalreconstruction

After the analysis of the signal, the original signal can be synthesized using inverse discrete wavelet transform. The signal is reconstructed as shown in Fig. 2 by up sampling of the decomposed signal followed by filtering through two complementary filters and is expressed as A + D = S. The low-pass and high-pass decomposition filters (L and H) and reconstruction filters (L' and H') together form a set of quadrature mirror filters as shown in Fig. 3.

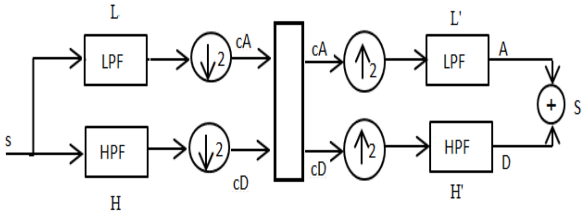

Fig. 3Signaldecomposition and reconstruction

The resolution of the signal is a measure of the amount of detail information in the signal, can be changed by the filtering operations, and the scale can be changed by up (2 sampling and down sampling operations.The decomposed signal can be broken down into lower resolution components by decomposing the successive approximations iteratively. Signal decomposition at different frequency bands is successive high-pass and low-pass filtering and forms the basis of multi resolution decomposition (Fig. 4). The signal can be analyzed at different frequency bands and resolutions by decomposing the signal into a coarse approximations and details. Similar relationships also hold for the reconstructed signal (Fig. 5).

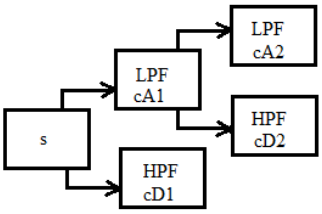

Fig. 4Multilevel signal decomposition

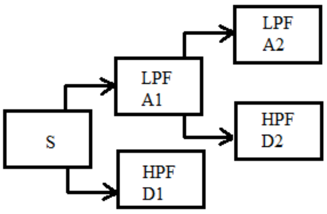

Fig. 5Multilevel signal reconstruction

The decomposed signal can be written as s = cA2+cD2+cD1. Similarly the signal can be reconstructed from the successive approximations and details as A2+D2+D1 = S.There are two functions associated with the low pass and high pass filters and are called scaling functions and wavelet functions respectively.

4 Method

Different nucleotide sequences of HA and NA proteins of H5N1 virus taken from the different hosts occurring in different regions (India) were downloaded from NCBI (National Center for Biotechnology Information)[19] database for comparison. The sequences were first aligned using nucleotide BLAST (Basic Local Alignment Search Tool) [20] and converted into mathematical form using integer number representation (A=1, C=2, G=3, T=4).These sequences were then transformed using discrete Haar wavelet transform.

Several comparisons for the synthesized signal were made at different levels of multi resolution decomposition using WT. The multi resolution analysis decomposes the signal of length 2^n into approximations and details at various levels (1-n). Whereas wavelet coefficients at higher scales correspond to low frequency components in the signal (approximations) and determine the global features of the sequence, wavelet coefficients at smaller scales (details) determine high frequency information and therefore local variations. The overall trend of the sequence can be visualized by the plot of approximation coefficients of the entire sequence. The sequences were analyzed at various levels of decomposition and best results were obtained at 4th level of decomposition of the signal, hence the decomposition level was chosen to be 4.

To locate the positions of base changes in a sequence, a pair of sequences were compared - one sequence taken as a reference and other as a subject. The reference sequence was chosen randomly. Both the sequences were first decomposed and then reconstructed upto level 4 using discrete Haar wavelet transform. To determine point mutations, the reference and subject sequences were compared by plotting the difference of wavelet detail coefficients of level 1 and level 2 respectively for both the sequences. The locations on the plots along the sequence length showed peaks at the places where base changes had occurred.

For determining point mutations in NA proteins of Influenza sequences, a reference sequence (accession number CY090110.1) was compared with two subject sequences (accession numbers CY090126.1 and CY090118.1). The difference of the detail coefficients at levels 1 and 2 for the complete sequence length (CY090126.1 and CY090118.1) were plotted (shown in Figs. 6 and 7 respectively). To locate point mutations in HA sequences, a subject HA protein sequence with accession number GQ917229.1 was compared with a reference sequence GQ917227.1. Point mutations for the complete sequence length are shown in Fig. 8. Another subject HA sequence (accession number JQ319658.1) was compared with a reference sequence (accession number JQ319657.1). The point mutations for this sequence are shown in Fig. 9. These plots appear flat at the places where the two sequences are perfectly identical as can be seen in Fig. 6 along sequence length between 200-400 bases, 500-600 bases and 1000-1100 bases. In Fig. 7, the two compared sequences are identical in regions around 1-100 bases, 450-550 bases, 600-700 bases. In Fig. 8 identical sequence regions are 0-300 bases, 400-600 bases, 900-1100 bases, 1500-1600 bases. Fig. 9 shows similar sequence regions around 0-700 bases, 800-1300 bases and 1350-1500 bases. The plots show peaks at places where the sequence bases have undergone changes.

The peaks of the difference coefficients exhibit both positive and negative magnitudes, so the exact location of the base changes can be identified by seeking coincident peaks of detail coefficients of both the decomposition levels 1 and 2 in the plot. The position along the sequence where the peaks of the level 1 and level 2 difference coefficients plot are either coincident or significantly overlapping are the locations of mutations. The magnified views for different subject NA sequences CY090126.1 and CY090118.1 (compared to a reference sequence) are plotted in Figs. 6a-6f and 7a-7g respectively. The magnified views of the two subject sequences of HA GQ917229.1 and JQ319657.1 compared to reference sequences are shown in Figs. 8a-8f and 9a-9c respectively.

The algorithm for the implementation of the above method is listed below.

1. Align the downloaded viral sequences to be compared using BLAST.
2. Convert the nucleotide sequences into mathematical sequences.
3. Using Haar wavelet transform, decompose and reconstruct the mathematical sequences (considering one sequence as reference and one sequence as subject) into details at levels 1 and 2.
4. Compare the detail coefficients for reference and subject sequence at levels 1 and 2.
5. Plot the difference in detail coefficients at both the levels.
6. Identify the positions of base changes that the locations of the peaks for level 1 detail

coefficients which are either coincident or have major overlap with the peaks for level 2 details.

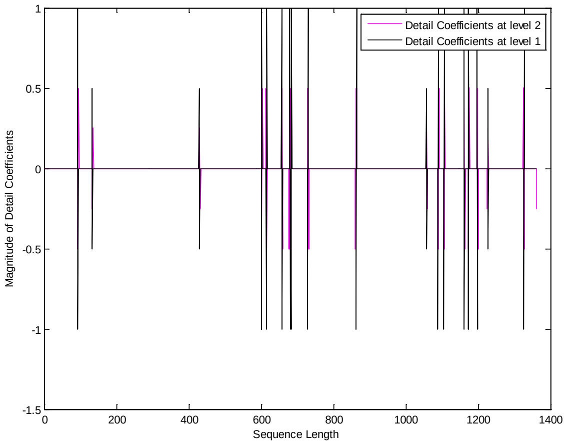

Fig.6 Sequence comparison CY090110.1 and CY090126.1

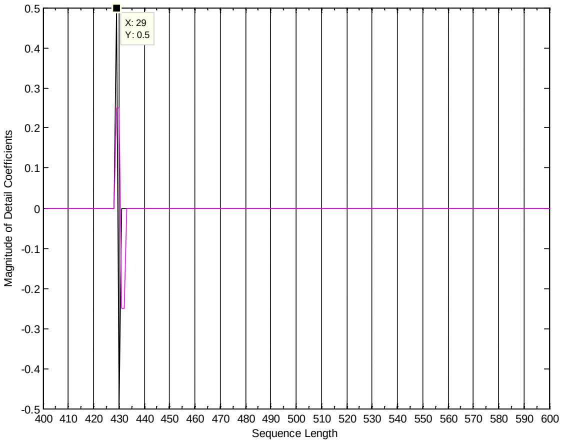

Fig.6b Sequence comparison CY090110.1 and CY090126.1 (400-600 bases)

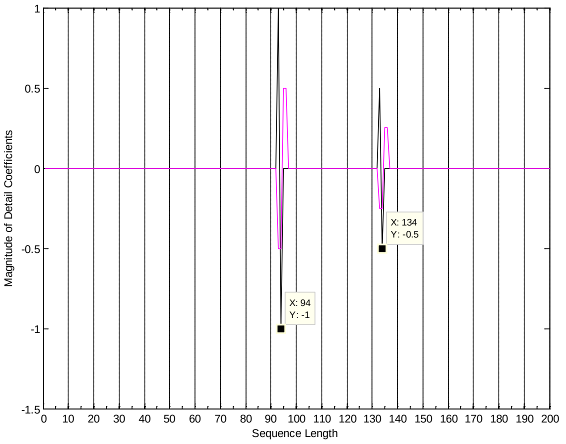

Fig.6a Sequence comparison CY090110.1 and CY090126.1 (0-200 bases)

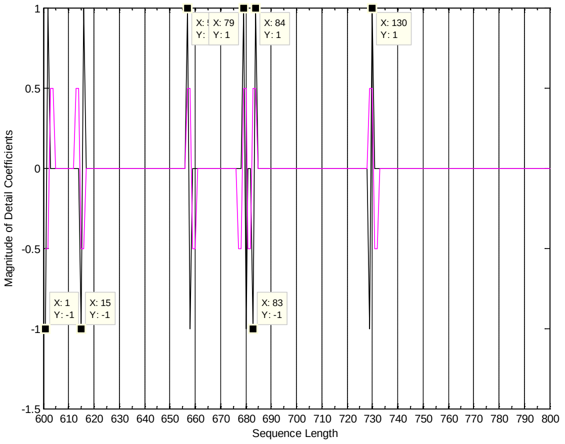

Fig.6c Sequence comparison CY090110.1 and CY090126.1 (600-800 bases)

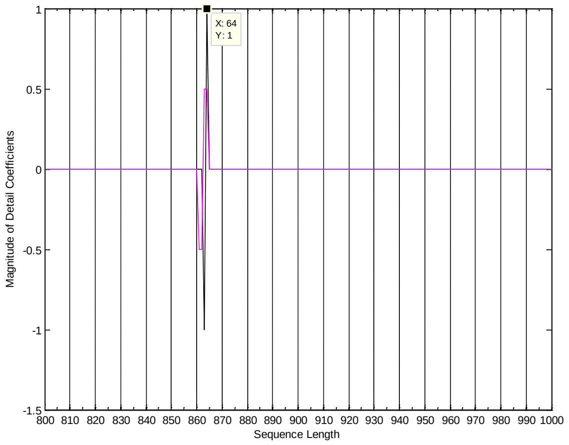

Fig.6d Sequence comparison CY090110.1 and CY090126.1 (800-1000 bases)

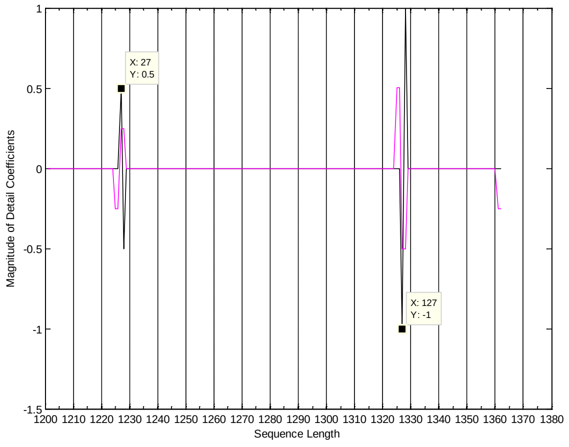

Fig.6e Sequence comparison CY090110.1 and CY090126.1 (1000-1200 bases)

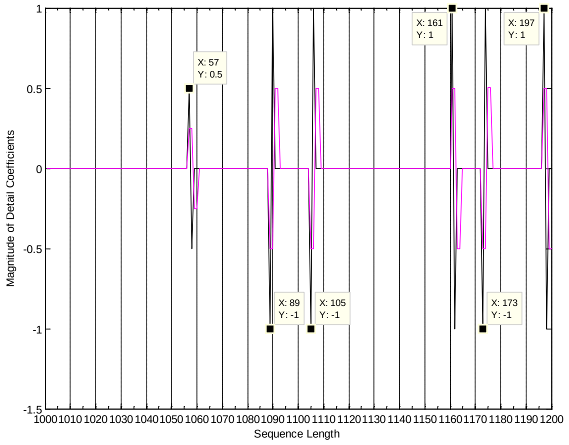

Fig.6f Sequence comparison CY090110.1 and CY090126.1 (1200-1380 bases)

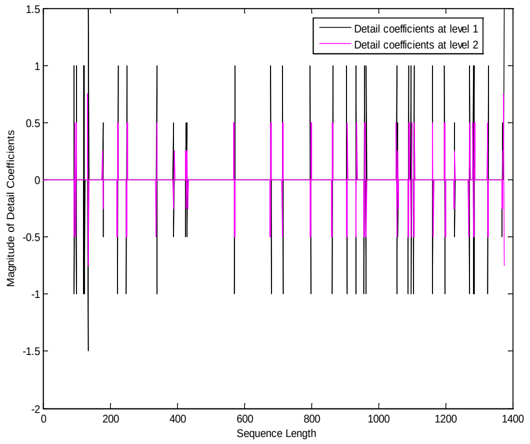

Fig.7 Sequence comparison CY090110.1 and CY090118.1

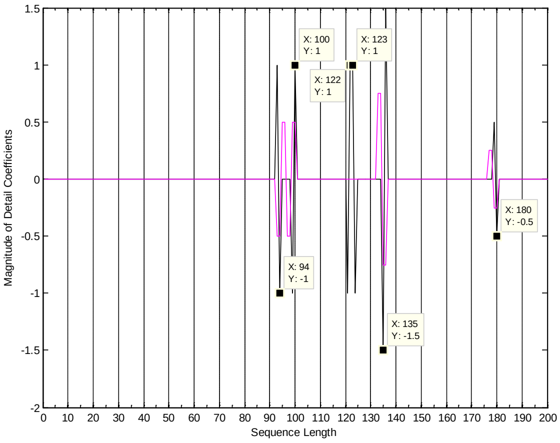

Fig.7aSequence comparison CY090110.1 and CY090118.1 (0-200 bases)

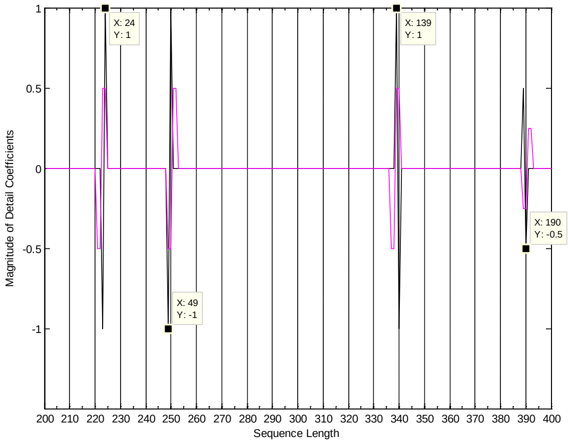

Fig.7b Sequence comparison CY090110.1 and CY090118.1 (200-400 bases)

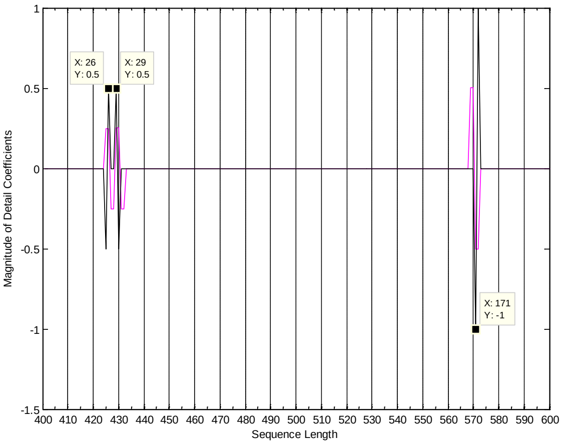

Fig.7c. Sequence comparison CY090110.1 and CY090118.1 (400-600 bases)

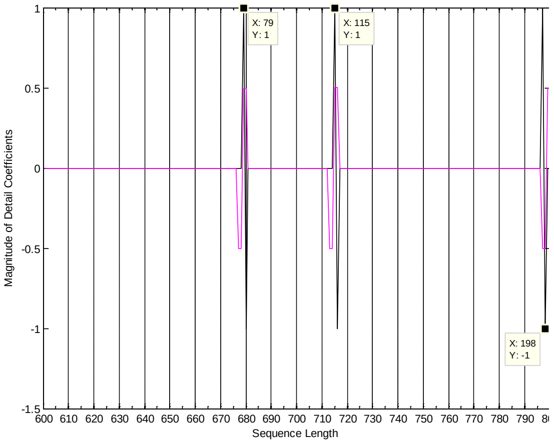

Fig.7d Sequence comparison CY090110.1 and CY090118.1

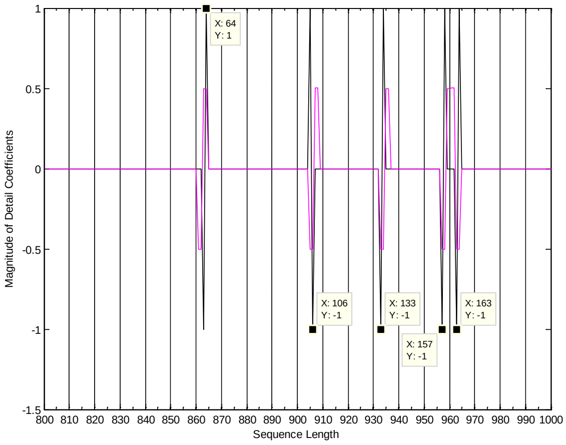

Fig.7eSequence comparison CY090110.1 and CY090118.1 (800-1000 bases)

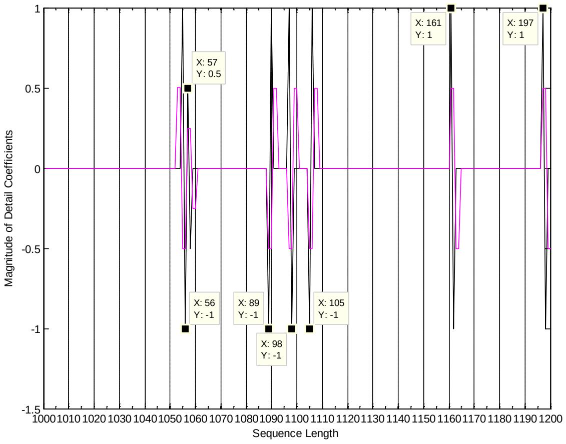

Fig.7f Sequence comparison CY090110.1 and CY090118.1 (1000-1200 bases)

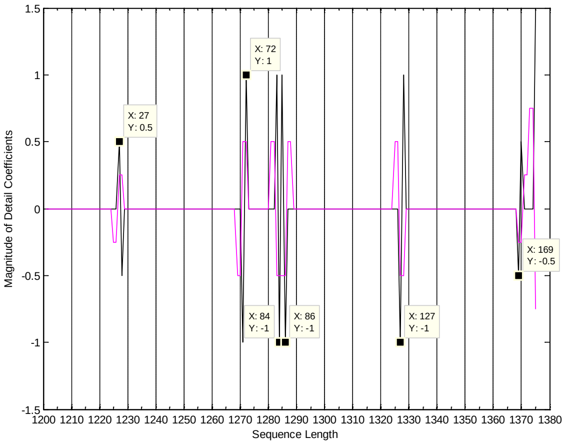

Fig.7g. Sequence comparison CY090110.1 and CY090118.1 (1200-1380 bases)

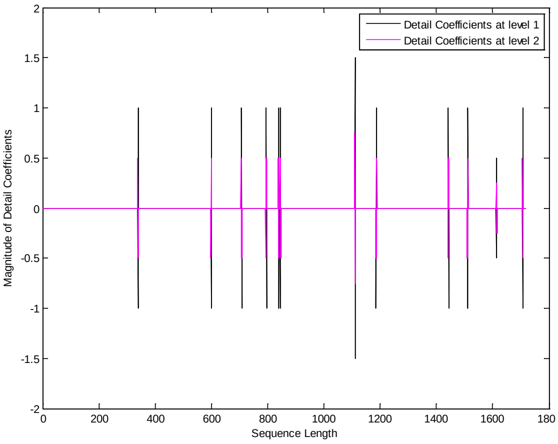

Fig.8 Sequence comparison HA917227.1 and HA917229.1

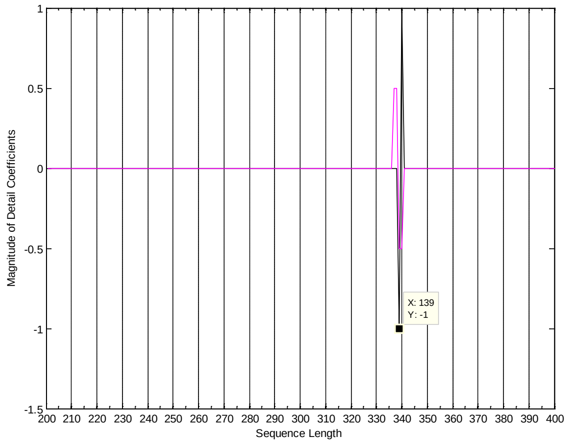

Fig.8a Sequence comparison HA917227.1 and HA917229.1 (200-400 bases)

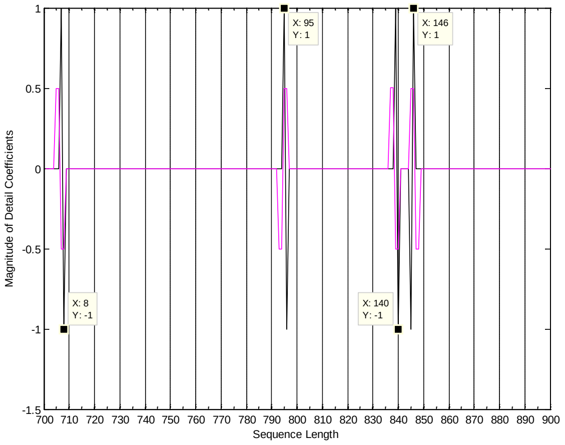

Fig.8cSequence comparison HA917227.1 and HA917229.1 (700-900 bases)

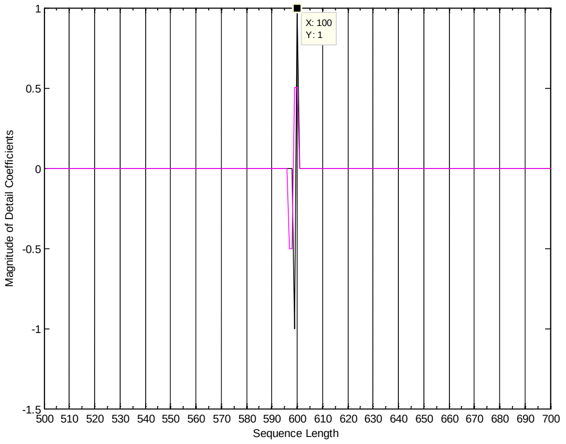

Fig.8b Sequence comparison HA917227.1 and HA917229.1 (500-700 bases)

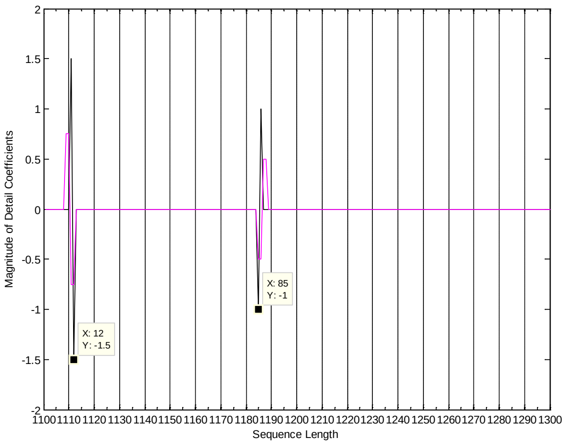

Fig.8d Sequence comparison HA917227.1 and HA917229.1 (1100-1300 bases)

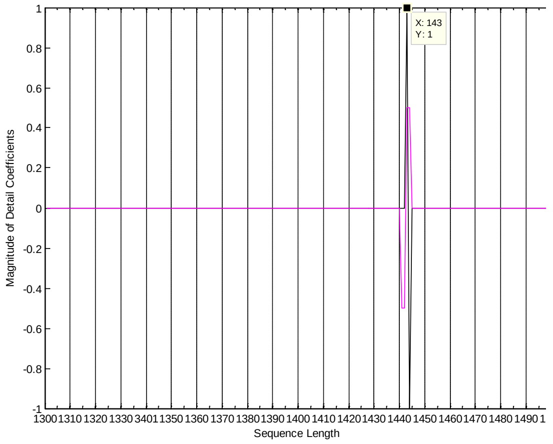

Fig.8e Sequence comparison HA917227.1 and HA917229.1 (1300-1500 bases)

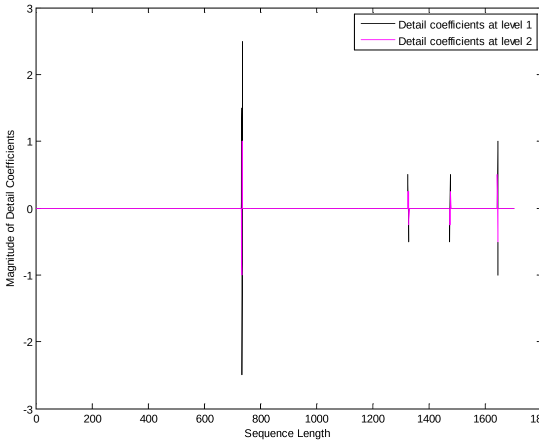

Fig.9Sequence comparison HA319657.1 and 319658.1

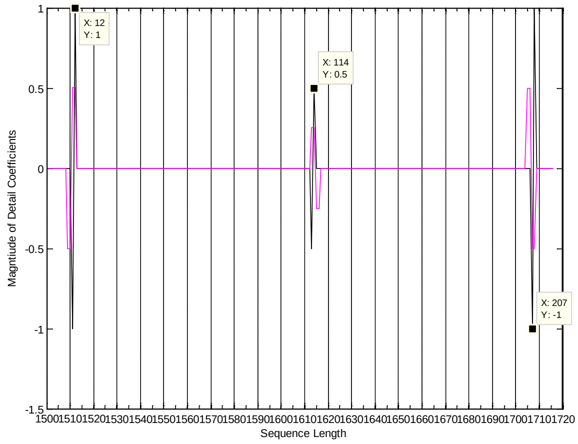

Fig.8fSequence comparison HA917227.1 and HA917229.1 (1500-1720 bases)

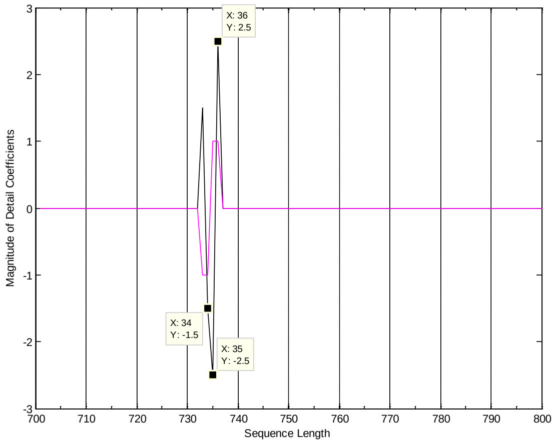

Fig.9a Sequence comparison HA319657.1 and 319658.1 (700-800 bases)

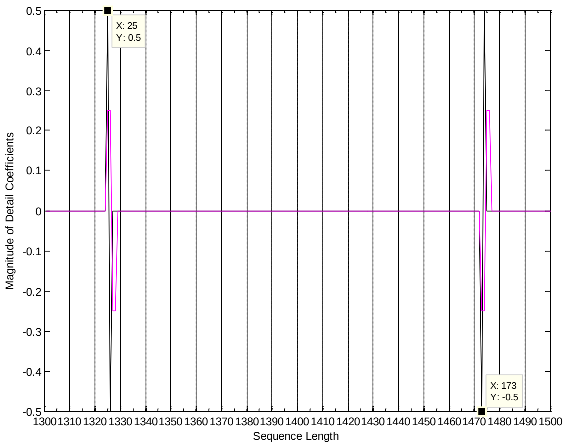

Fig.9b Sequence comparison HA319657.1 and 319658.1 (1300-1500 bases)

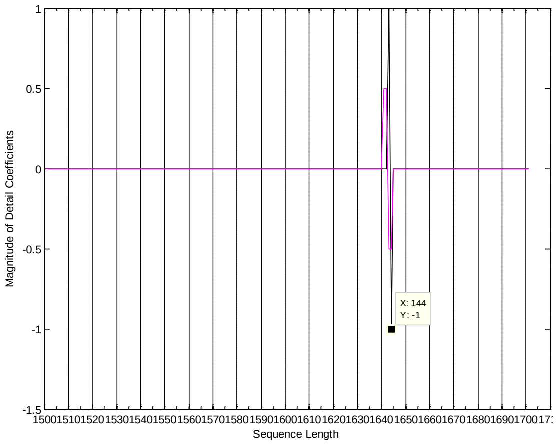

Fig.9cSequence comparison HA319657.1 and 319658.1 (1500-1710 bases)

5 Observations

Based on the proposed method using wavelet transforms, following observations were made.

1. The difference plots for wavelet transformed NA sequences with accession numbers CY090110.1 (reference) and CY090126.1 (subject) show 18 peaks hence 18 mutations. Thus it can be seen that even though these viruses were sequenced in the same year (2008), they have undergone 18 base changes. The exact locations of base changes were localized by the coincident or overlapping peaks of details coefficients at levels 1 and 2 (Figs. 6a-6f). The number of mutations and their locations identified by WT method were confirmed bycomparison with the base change locations of sequence alignment tool, BLAST and were found to be perfectly similar.

2. The NA protein sequences CY090110.1 (reference sequence) and CY090118.1 (subject sequences) when compared perfectly localized 32 out of 34 base changes on comparison with BLAST. The exact location in the sequence where base changes occurred is shown in figures (7a-7g). Of all the base changes detected, the method however, did not perfectly locate only 2 base changes in the difference plots at positions 122 and 123 along sequence. Though the detail coefficients at level 1 were existing and the peaks for detail coefficients at level 2 were missing, the exact locations could not be identified. But the existence of peaks at level 1 meant that mutations had occurred at these locations.

3. The wavelet detail coefficients plots for HA protein sequences GQ917227.1 (reference) and GQ917229.1 (subject) identified the location of all the 12 base changes perfectly as can be seen in Figs. 8a- 8f.

4. The wavelet detail coefficients plots of JQ319657.1 (reference) and JQ319658.1 (subject) also perfectly identified all the 6 base changes and their locations (shown in Figs. 9a-9c).

From the above observations, it can be summarized that the advantage of using wavelet transform method is that only by visual analysis of the plot of the detail coefficients; the point mutations are evident as compared to BLAST where one has to manually count and locate the point changes. Of all the base change locations identified using wavelet method, no false positives or false negatives were observed.

The methods of sequence trace comparisons of DNA chromatograms, such as SeqDoc also uses the difference technique to highlight the point changes between a reference and subject sequence. But chromatograms of the subject sequences need to be normalized before subtraction since they are traces of intensity. Moreover, these intensity traces are subject to background noise. Wavelet approach, on the contrary is advantageous as their plots do not require any normalization since the text based DNA sequences are converted into discrete mathematical sequences and then transformed and reconstructed using WT. These sequences are also not affected by signal noise. WT method can analyze any varying length of sequences of the order of 10^8 bases or more.

6 Conclusions

In this paper, a wavelet transform based method has been applied to identify base change locations in Influenza A virus. Unlike the sequence trace comparison method, this method does not require any normalization of sequences and is immune to any background noise. Thus wavelet transforms offer the advantage of reducing the computational complexity and faster identification of sequence changes by visual representation of the plot of the wavelet coefficients.

These graphical plots help in determining the relatively stable regions in different protein sequences of H5N1 only by visual analysis. These stable regions can be used as the target regions for determining the effect of various drugs on different strains of the viruses to identify drug resistance and also in vaccine manufacturing. Thus wavelet based signal analysis methods and bioinformatics together provide a simple tool for analyzing the changes undergone by the viral sequences. The availability of signal processing methods help in providing accurate and faster results for huge amount of already sequenced genomic data collected from throughout the world during the epidemics, these signal data which can be used as a basis for drug design and new diagnosis development.

References:

[1] Collins, F. S., Guyer , M.S. and Chakravarti, A., Variations on a Theme: Cataloging Human DNA Sequence Variation, *Science*, 278 (5343)(1997),pp. 1580–1581.

[2] James K. Bonfield, Cristina Rada and Rodger*Staden*, Automated detection of point mutations using fluorescent sequence trace subtraction,*Nucleic Acids Research*,26(14),1998, pp.3403-3409.

[3] Crowe, M. L., SeqDoc: Rapid SNP and Mutation Detection by Direct Comparison of DNA Sequence Chromatograms, *BMC Bioinformatics*, 6:133(2005).

[4] www.appliedbiosystems.com

[5] Sheng, C and Ni, S. Mutation Detection from DNA Sequence Traces with Mutation Surveyor and Mutation Explorer Software, *Application Note:* *www.softgenetics.com*

[6] Cristea, P.D. Genomic Signal Processing and Analysis: Applications in the Study of Pathogen Variability, *Proceedings CODEC (2006) – International Conference on Computers And Devices for Communication (Institute of Radio Physics and Electronics, University of Kolkata, India).*

[7] Cristea, P. D., Otelea,D. and Tuduce,R., Study of HIV Variability Based on Genomic Signal Analysis of Protease and Reverse Transcriptase Genes, *in Proceedings EMBC'05*(Shanghai, China, Sept 2005).

[8] Cristea, P. D.,Pathogen Variability*: A Genomic Signal Approach, International Journal of Computers, Communication and Control,* 1(3)(2006),pp.25–32.

[9] Cristea, P.D., Tuduce, R.,Banica,D. and *Rodewald, K.*, Genomic Signals for the Study of Multiresistance Mutations in M Tuberculosis,*in Proceedings – ISSCS 2007, International Symposium on Signals, Circuits and Systems,*1:1-4,Digital Object Identifier: 10.1109/ISSCS.2007.4292708.

[10] Fodor, E. and Brownlee, G.G. (2002). Perspectives in medical virology, Influenza virus replication, edPotter,C.W.*Elsevier, Amsterdam, The Netherlands),* 7,(2002),pp. 1:29.

[11] Taubenberger, J. K. and Kash, J. C., Influenza Virus Evolution, Host Adaptation, and Pandemic Formation, *Cell Host and Microbe*, 7(6),2010, pp. 440–451.

[12] Voss, R.F. (1992). Evolution of Long-range Fractal Correlations and 1/f noise in DNA base sequences, *Physical Review Letters*, 68,1992, pp.3805–3808.

[13] Arniker, S.B. and Kwan H. K.,Graphical Representation of DNA sequences, *Proceedings of IEEE International Conference on Electro/Information Technology*, EIT,2009, pp. 311–314.

[14] Cristea, P. D.,Conversion of nucleotides sequences into genomic signals, *J. Cell.Mol. Med.*, 6(2),2002,pp.279–303.

[15] Berger, J.A., Mitra, S. K., Carli, M. and Neri, *A.*, New approaches to genome sequence analysis based on digital signal processing, *Workshop on Genomic Signal Processing and Statistics GENSIPS*,2002.

[16] Nair, A. S. and Sreenadhan, S. P. (2006). A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), *Bioinformation*1(6),2006, pp.197–202.

[17] Wavelet Documentation, *www.mathworks.com/help/wavelet/*.

[18] Mallat, S., A Wavelet Tour of Signal *Processing, Second edition, Academic Press, New York*,2000.

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

[20]www.blast.ncbi.nlm.nih.gov/