Original Article

A Nonlinear Pattern Recognition of Pandemic H1N1 Using a State Space Based Methods

Mai S. Mabrouk

Biomedical Engineering Department, Misr University for Science and Technology (MUST), Egypt

Abstract

Genomic Signal Processing is a relatively new field in bioinformatics, in which signal processing algorithms and methods are used to study functional structures in the DNA. An appropriate mapping of the DNA sequence into one or more numerical sequences enables the use of many digital signal processing tools in the analysis of different genomic sequences. Also, a novel *Influenza A* (H1N1) virus of swine origin emerged in the spring of 2009 and spread very rapidly among people. The severity of the disease and the number of deaths caused by a pandemic virus varies greatly and can change over time. Throughout this work, Pandemic H1N1 genomic sequences were characterized according to nonlinear dynamical features such as moment invariants and largest Lyapunov exponents and then compared to those features that extracted from classical H1N1 genomic sequences. The proposed methods were applied to a number of sequences encoded into a time series using a coding measure scheme employing Electron-Ion Interaction Pseudo-potential (EIIP). The aim of this work is to extract genomic features that can distinguish the new swine flu from the classical H1N1 existed before using sequences from segment 8 of the influenza genome that consists of 8 RNA segments which encodes two important proteins for immune system attack (NS1 and NS2). According to the obtained results it is evident that variability is present based on a significance test in both groups; pandemic and classical H1N1 sequences.

*** Corresponding author:**
Mai S. Mabrouk, Ph.D.,
Department of Biomedical Engineering, Misr University for Science and Technology (MUST), Egypt
E-mail:
msm_eng@k-space.org
Received: 27 Oct 2010
Accepted: 28 Feb 2011

Avicenna J Med Biotech 2011; 3(1): 25-29

Keywords: DNA, Genome, H1N1 Subtype, Pandemics, Sequence

Introduction

The variations of pandemic *H1N1 influenza* virus are caused as a result of different mutations occurring during viral replication [(1)]. The polymerase of this RNA virus lacks proof reading activity [(2)]; this gives rise to considerable viral variability culminating in 3 different types A, B and C, in addition to many subtypes based on variations in the hemagglutinin (HA) and the neuraminidase (NA) surface proteins [(3)]. The influenza genome consists of 8 RNA segments and encodes for 10 polypeptides; the internal structural proteins, nucleocapsid protein (NP), the two matrix protein (M) are used for the classification of the influenza virus into A, B and C. The surface proteins neuraminidase (NA) and hemagglutinin (HA) have been studied extensively and the antigenic variations in the these surface glycoproteins are used to subtype *Influenza A*. Additionally, three of the influenza polypeptides are associated with RNA polymerase activity (PA, PB1, PB2), and the

RNA binding non-structural protein (NS) that contribute to viral pathogenicity and play a central role in the prevention of interferon mediated antiviral response. The *Influenza A* Virus (IAV) undergoes major and minor genetic variations, the yearly antigenic drift resulting in as minor as a single amino acid mismatch. Major variations known as antigenic shifts are the cause of serious outbreaks and pandemics as the 1918, 1957, and 1968 worldwide outbreaks [4]. Changes in the genetic and antigenic composition result in challenges in the development of influenza vaccines and antiviral medications [5].

In the last two decades, there has been an increasing interest in applying techniques from the domains of nonlinear analysis and chaos theory in different fields of research. In this work, the chaos theory was applied to both pandemic H1N1 and classical H1N1 genomic sequences in order to discriminate between them according to their non linear dynamical features as moment invariants, and Largest Lyapunov Exponents (LLE).

Material and Methods

The conversion of the DNA sequences into digital signals offers the possibility of applying signal processing methods to the analysis of genomic data [6,7]. The genomic signal processing applications in bioinformatics provides an efficient tool used to extract features of DNA sequences maintained over the whole genomes [8]**.** In this work, the EIIP sequence indicators were used, the energy of delocalized electrons in amino acids and nucleotides has been calculated as the Electron-Ion Interaction Pseudopotential (EIIP). The EIIP values of amino acids were used to substitute for the corresponding amino acids in protein sequences, whose power spectrum is taken to extract the information contents [9]. To study the dynamics of the proposed system, the state space trajectory was first reconstructed. Phase space reconstruction is the fundamental for analyzing nonlinear signals, by which a time series can be embedded to n-dimensional space.

Briefly the basic steps of the reconstruction of the phase space were demonstrated. First, different sequences of the pandemic H1N1 and classical H1N1 which existed before were encoded into a time series signal using EIIP sequence indicators. A good choice for a delay time was yielded by using the first minimum of the auto mutual information function. The first minimum of the auto mutual information could be found at four. The minimal embedding dimension for the pandemic H1N1 and classical H1N1 time series signals were calculated using Cao's method with a delay time of four, a maximal dimension of eight, three nearest neighbors and reference point depending on the length of each signal. There was a kink produced by Cao's method at 3. This kink represents the time delay reconstruction of pandemic and classical H1N1 time series signals with embedding dimension of 3 and delay of 4. Finally, the phase space trajectory was obtained for both time series signals of the two types of H1N1 genomic sequences (pandemic and classical). The step following obtaining the phase trajectory is the step of feature extraction [10].

Feature extraction

TSTOOL software package is used to estimate the extracted nonlinear dynamical features; it is a software package for signal processing with emphasis on nonlinear time-series analysis [11].

Moment invariants

Features obtained by moment invariants are simple calculated features that do not change under translation, scaling or rotation [12]. These invariants are constructed using the generalized fundamental theorem of moment invariants (GFTMI), which was formulated as in [13]. The *n*-dimensional moments of order p of a function of intensity $\rho(x_1, ..., x_n) = \rho(x)$ are defined in terms of Rieman integral as:

$$1)\ m_{p_1 \cdots p_n} = \int \cdots \int x_1^{p_1} \cdots x_n^{p_n} \rho(x) dx_1 \cdots dx_n$$

Where $p_i + ... p_n = p,\ 0 < p < \infty$. It is assumed that $\rho(x)$ is piecewise continuous and therefore bounded function, and it can

have nonzero values only in a finite part of the R^n; then the moments of all orders exist. The central moments:

$$2)\ \mu_{P_1 \ldots P_n} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \left(x_1 - \bar{x}_1\right)^{P_1} \ldots\ldots \left(x_n - \bar{x}_n\right)^{P_n} \rho(x) dx_1 \ldots . dx_n$$

Where

$$3)\ \bar{x}_1 = \frac{m_{1\ldots0}}{m_{0\ldots0}}, \ldots\ldots\ldots, \bar{x}_n = \frac{m_{0\ldots1}}{m_{0\ldots0}}$$

The seven features of moment invariants:

$$4)\ \phi_1 = \frac{1}{\mu^4} \begin{vmatrix} \mu_{2\ldots0} & \ldots & \mu_{1\ldots1} \\ \ldots & \ldots & \ldots \\ \mu_{1\ldots1} & \ldots & \mu_{0\ldots2} \end{vmatrix}$$

$$5)\ \phi_2 = \frac{1}{\mu^4}(\mu_{20}\mu_{02} - \mu^2{}_{11})$$

$$6)\ \phi_3 = \frac{1}{\mu^{10}}[(\mu_{30}\mu_{03} - \mu_{21}\mu_{12})^2 - 4(\mu_{30}\mu_{12} - \mu_{21}{}^2)(\mu_{21}\mu_{03} - \mu^2{}_{12})]$$

$$7)\ \phi_4 = \frac{1}{\mu^6}(\mu_{40}\mu_{04} - 4\mu_{31}\mu_{13} - 3\mu^2{}_{22})$$

$$8)\ \phi_5 = \frac{1}{\mu^9}(\mu_{40}\mu_{22}\mu_{04} + 2\mu_{31}\mu_{22}\mu_{13} - \mu_{40}\mu^2{}_{13} - \mu^2{}_{31}\mu_{04} - \mu^3{}_{22})$$

$$9)\ \phi_7 = \frac{1}{\mu^5}(\mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011} - \mu_{200}\mu^2{}_{011} - \mu^2{}_{110}\mu_{002} - \mu^2{}_{101}\mu_{020})$$

$$10)\ \phi_8 = (\mu^2{}_{20}\mu_{04}) - 4\mu_{20}\mu_{11}\mu_{13} + 2\mu_{20}\mu_{02}\mu_{22} + 4\mu^2{}_{11}\mu_{22} - 4\mu_{11}\mu_{02}\mu_{31} + \mu^2{}_{02}\mu_{40}$$

Largest lyapunov exponent (LLE)

In this work, a set of genomic sequences from segment 8 of the influenza genome of both pandemic and classical H1N1 was downloaded from the NCBI. The length of these sequences was chosen to be 800-1000 *bp*. These sequences are first encoded using EIIP sequence indicators. Then, the phase space trajectory was reconstructed for each time series of both of them. The TSTOOL larglyap algorithm was used to estimate the Largest Lyapunov Exponent (LLE). This algorithm is similar to Wolf's algorithm and provides an efficient estimation of the Largest Lyapunov Exponent through the calculation of the rate of increase of the prediction error versus the pre-diction time [14].

Results

Results of moment invariants

Features based on moment invariants were computed after the construction of phase space of both pandemic and classic H1N1 EIIP encoded sequences. The seven features are arranged as *($\varphi 1$, φ_2, φ_3, φ_4, φ_5, φ_7,* and *$\varphi 8$)*. A significance test (t-test) was performed on the proposed features to assess the use of such parameters for discriminating between them. The result of the t-test is presented and the p value is calculated for all seven features; they are all less than 0.05 as shown in table 1. Figure 1 shows the result of comparing the average features extracted based on moment invariants for pandemic and classical H1N1. There is a significant difference between the two types of H1N1 as shown in the figure. Also, small vertical bars represent a standard deviation across features.

Results of largest lyapunov exponent (LLE)

The LLE estimates of a set of pandemic and classical H1N1 genomic sequences were calculated using TSTOOL largelyap algorithm as shown in table 2. It is an algorithm very similar to the Wolf algorithm; it computes the average exponential growth of the distance of neighboring orbits via the prediction error. The increase of the prediction error versus the prediction time allows an estima-

Table 1. P-value of t-test on a set of pandemic and classical H1N1 EIIP encoded sequences for feature extracted using moment invariants

Moment invariants feature	p-value
φ1	1.8235e-004
Φ2	2.2912e-005
Φ3	1.3674e-010
Φ4	0.0012
Φ5	0.0288
Φ7	2.2912e-005
Φ8	6.7141e-005

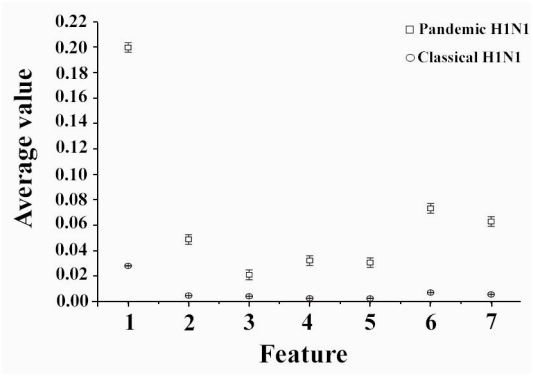

Figure 1. Features extracted based on moment invariants for pandemic and classical H1N1, the small vertical bars represent standard deviations across features

tion of the Largest Lyapunov Exponent. A significance t-test was applied to assess the use of LLE estimates in the discrimination between pandemic and classical H1N1.

Significance test

The accuracy of a test was evaluated to discriminate between pandemic H1N1 and classical H1N1 by moment invariants and Largest Lyapunov Exponent dynamical system features). These features were divided into three feature vectors as follows:

V1= {Φ1, Φ2, Φ3, Φ4, Φ5, Φ7, Φ8}
V2= {LLE}
V3= {Φ1, Φ2, Φ3, Φ4, Φ5, Φ7, Φ8, LEE}

Table 2. Largest Lyapunov Exponent estimates of pandemic and classical H1N1 encoded sequences

LLE (Pandemic H1N1)	LLE (Classical H1N1)
2.6932	0.3428
2.7103	0.3601
2.7113	0.3628
2.7142	0.3667
2.7153	0.3795
2.7280	0.3854
2.7392	0.4429
2.7475	0.4491
2.7506	0.4533
2.7567	0.5569
2.7577	0.6274
2.7722	0.7417
2.8972	0.7509
2.9075	0.8078
2.9178	0.8891
2.9472	0.9501
2.9508	0.9677

The feature vectors were fed into the classification process using K-means clustering classifier. Results of applying the significance test are shown in table 3.

Table 3. Accuracy of the proposed nonlinear pattern recognition method using K- means classifier

	V1	V2	V3
Pandemic H1N1	100%	100%	100%
Classical H1N1	70.8%	100%	100%

Discussion

The proposed techniques were implemented and applied to a number EIIP encoded sequences of pandemic and classical H1N1 from segment 8 of the influenza genome to identify their genomic signatures as continuous detection of these signatures is important in the analysis of the adaptation process from nonhumans to humans.

As to chaotic features extracted based on moment invariants, the seven features are arranged as ($\varphi 1$, φ_2, φ_3, φ_4, φ_5, φ_7, and $\varphi 8$). Considering the p-values: if $p< 0.05$ there is a significant difference, if $p>0.05$ there is no significant difference. The results show that these features generally support the hypothesis that they have a potential to discriminate between pandemic and classical H1N1 as they all <0.05.

As to chaotic features based on LLE estimates, the p-value of the t-test was calculated as 2.1546e-019 which is < 0.05. To validate this result, a random DNA sequence of length 1000 *bp* was generated, the Largest Lyapunov Exponent (LLE) of this random sequence was estimated at 1.4046 and compared to the average LLE estimates of pandemic H1N1 (2.8218) and the average LLE estimates of classical H1N1 (0.4697). The results confirm that pandemic H1N1 genomic sequences can be statistically differentiated from classical H1N1 genomic sequences by LLE dynamical features.

Conclusion

The analysis of different genomic mutations of the pandemic H1N1 genomic se-

quences is very important to study the possibility of virus adaptation from non-humans to humans.

A study of nonlinear dynamics of pandemic and classical H1N1 genomic sequences of segment 8 of the influenza genome was presented to discriminate between them by their moment invariants and Largest Lyapunov Exponent (LLE) estimates. The results of this work were supported by statistical analysis indicating that the discrimination between these two types of H1N1 provides a clear outline for the potential of using such nonlinear dynamical features with high accuracy. The study shows that using these nonlinear dynamical features will open the door to extract more patterns to be used in monitoring and extracting all H1N1 genomic signatures.

Acknowledgement

The author really expresses a deepest thanks to Dr. Yasser Kadah at the Department of Systems and Biomedical Engineering, Cairo University for his valuable discussions and continuous support; he is actively engaged in bioinformatics, including sequence analysis data mining, genomic signal analysis and software development. Also, the author presents many thanks to Dr. Mahmoud El-Hefnawi at National Research Center (NRC) for his support and help.

References

1. Cox NJ, Subbarao K. Influenza. Lancet 1999;354: 1277-1282.
2. Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, et al. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. Am J Epidemiol 2008; 167(7):775-785.
3. Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, Xu X, et al. Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. N Engl J Med 2009;360:2616-2625.
4. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 2009;325(5937):197-201.
5. Chen GW, Chang SC, Mok CK, Lo YL, Kung YN, Huang JH, et al. Genomic signatures of human versus avian influenza A viruses. Emerg Infect Dis 2006;12(9):1353-1360.
6. Cristea PD. Large scale features in DNA genomic signals. Signal Processing 2003;83(4):871-888.
7. Cristea PD. Genomic signals of re-oriented ORFs. EURASIP JASP 2004;2004(1):132-137.
8. Cristea PD. Conversion of nitrogenous base sequences into genomic signals. J Cell Mol Med 2002; 6(2):279-303.
9. Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation 2006;1(6):197-202.
10. Mabrouk MS, Solouma NH, Youssef AM, Kadah YM. Eukaryotic gene prediction by an investigation of nonlinear dynamical modeling techniques on EIIP coded sequences. Int J Biol Sci 2007;3(4): 225-230.
11. http://www.physik3.gwdg.de/tstool/.
12. Mamistvalov AG. n-dimensional moment invariants and conceptual mathematical theory of recognition n-dimensional solids. IEEE Trans Pattern Anal Mach Intell 1998;20(8):819-831.
13. Mamistvalov AG. On the construction of affine invariants of n-dimensional patterns. Bull Acad Science Georgian SSR 1974;76(1):61-64.
14. Owis MI, Abou-Zied AH, Youssef AM, Kadah YM. Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification. IEEE Trans Biomed Eng 2002;49(7): 733-736.