Early-warning signals for an outbreak of the influenza pandemic*

Ren Di(任 迪) and Gao Jie(高 洁)†

School of Science, Jiangnan University, Wuxi 214122, *China*

(Received 14 May 2011; revised manuscript received 25 July 2011)

Over the course of human history, influenza pandemics have been seen as major disasters, so studies on the influenza virus have become an important issue for many experts and scholars. Comprehensive research has been performed over the years on the biological properties, chemical characteristics, external environmental factors and other aspects of the virus, and some results have been achieved. Based on the chaos game representation walk model, this paper uses the time series analysis method to study the DNA sequences of the influenza virus from 1913 to 2010, and works out the early-warning signals indicator value for the outbreak of an influenza pandemic. The variances in the CGR walk sequences for the pandemic years (or + −1 to 2 years) are significantly higher than those for the adjacent years, while those in the non-pandemic years are usually smaller. In this way we can provide an influenza early-warning mechanism so that people can take precautions and be well prepared prior to a pandemic.

Keywords: influenza virus, early-warning signals, chaos game representation (CGR) walk model, DNA sequence

PACS: 87.10.Vg, 02.50.Fz **DOI:** 10.1088/1674-1056/20/12/128701

1. Introduction

Influenza (flu) is an acute respiratory infection caused by the influenza virus. It is a recurrent and contagious disease with a fast transmission spread and can cause high global morbidity and mortality.[1] In recent years, with the further spread of flu, the World Health Organization raised the swine flu alert level to its sixth grade.[2,3] This has led many experts and scholars from all over the world to pay more attention to the biological characteristics, chemical characteristics and external environmental factors of flu viruses, and some results have been achieved.

In 2007 Liu and Luo discussed apoptosis induced by the molecular mechanism of the bird flu virus.[4] In 2009 Rao and Liu discovered the "heart" – the working mechanism of the RNA polymerase PA subunit at the atomic level of the bird flu virus, and achieved a major breakthrough in the study of the virus, which provided a target for the research of the related drugs.[5] In 2007 Shaman found that the effect of absolute humidity on the survival, transmission and seasonality of the flu virus is more than that of the relative humidity, and increasing absolute humidity may contribute to an inhibition of the indoor flu pandemic.[6] Integrating the data for the years 1890–2004, the researchers concluded six climate characteristics for the flu pandemic as follows: the Pacific decade oscillation and its boundary; the previous year, or more than two years ago, as moderate La Nina years; in the 20th century, the age 50–70 years, while there is a strong sandstorm in China; before and after the year or the years there is a cold summer in northeast China (in the 20th century, while there were 50–70 years of severe cold damage); the year for the moderate or more El Nino; the year for the sunspot valley m or the peak sunspot years of M, $m-1$, $m+1$, $M+1$, and so on.[7−9]

In 2003 Scheffer *et al.* found that before the critical transition, complex dynamic systems such as ecosystems, climate, financial markets and other complex power systems would show the presence of a universal early-warning signal. The results show that the so-called critical point is the critical value for change from one state to another state in a system, and the critical point is associated with a well-known critical slowing down theory of dynamical systems. In the bifurcation of the stochastic forcing system close to a controlled parameter threshold value, the critical slowing down often leads to increasing variance of

*Project supported by the Fundamental Research Funds for the Central Universities (Grant No. JUSRP21117) and the Program for Innovative Research Team of Jiangnan University (Grant No. 2008CX002).

†Corresponding author. E-mail: ezhun6669@sina.com

© 2011 Chinese Physical Society and IOP Publishing Ltd

volatility.[10,11] This is a revelation for studying the flu virus.

In 2007 Gao *et al.* proposed a method, based on the chaos game representation (CGR) coordinate, for transforming a DNA sequence into a time series (CGR-walk sequence), and used long memory ARFIMA (p, d, q) models to fit and verify the solution to the problem of missing data, simultaneously.[12] In 2009 this model was applied to the simulation of the real CGR-walk sequence data of genomic and protein sequences.[13,14]

Based on the CGR walk model, this paper uses the time series analysis method to study the DNA sequences of the flu virus from 1913 to 2010, and works out the early-warning signals indicator value for the outbreak of a pandemic. That is, the variances of the CGR walk sequences for the pandemic years (or $+-1$ to 2 years) are significantly higher than those for the adjacent years, while those for the non-pandemic years are usually smaller. In this way we can construct a flu early-warning mechanism so that people can take precautions and get well prepared before the outbreak of a flu pandemic.

2. CGR walk model

CGR was proposed as a scale-independent representation for genomic sequences by Jeffrey in 1990.[15] The CGR coordinates are calculated iteratively by moving a pointer to half the distance between the previous position and the current binary representation (Eq. (1)). The binary CGR vertices are assigned to the four nucleotides as $A = (0,\ 0)$, $C = (0,\ 1)$, $G = (1,\ 1)$, $T = (1,\ 0)$ and $(0.5,\ 0.5)$ as an arbitrary starting position.

For a sequence, $S = S_1, S_2, \ldots, S_N, S_i \in \{A, T, G, C\}$, through the following iterative process we can gain the CGR walk sequence

$$\begin{aligned} &CGR_i = CGR_{i-1} - 0.5 \cdot (CGR_{i-1} - g_i), \\ &i = 1, \ldots, N, CGR_0 = (0.5, 0.5), \end{aligned} \tag{1}$$

where $g_i \in \{(0,0), (1,0), (1,1), (0,1)\}$, g_i is corresponding with S_i.

Here, by setting up base coordinates $CGR_n(x_n,\ y_n)$, let

$$t_n = y_n / x_n, \tag{2}$$

then we can obtain a CGR-walk sequence $\{t_n : n = 1, 2\ldots, N\}$.[11]

Let

$$\begin{aligned} \mathrm{Var}(t_n) &= \frac{(t_1 - m)^2 + (t_2 - m)^2 + \cdots + (t_n - m)^2}{n - 1} \\ &\left(m = \frac{t_1 + t_2 + \cdots + t_n}{n}\right) \end{aligned} \tag{3}$$

be the variance of the CGR-walk sequence $\{t_n : n = 1, 2\ldots, N\}$.

The so-called critical point is the critical value of a proper parameter to signify a system changing from one state to another. Work in different scientific fields has shown that before the critical transition, complex dynamic systems such as ecosystems, financial markets and climate would imply the presence of a universal early-warning signal. The results show that whether the system is approaching the critical point is associated with the well-known critical slowing down theory of dynamical systems. In the bifurcation of the stochastic forcing system close to a controlled parameter threshold value, the critical slowing down often leads to an increased variance of volatility.[10,11]

Based on the above conclusion, we regard a flu outbreak as a critical transition and regard an outbreak point as a critical point. So our purpose is to verify that the variances of the pandemic years are significantly higher than those of the adjacent years based on a certain model.

3. Pandemic early-warning indicator value

We chose the flu virus DNA sequences from 1913 to 2010, which have a relatively high homology (data from the NCBI website at: www.ncbi.nlm.nih.gov/). These virus sequences are translated into CGR walk sequences. For each CGR walk sequence $\{t_n : n = 1, 2\ldots, N\}$, we calculate its variance $\mathrm{var}(t_n)$.

For scientificity and universality, and in view of the collection and integrity of the data, we selected swine flu virus DNA sequences, as the homology is relatively high after 1970. We also selected the avian flu DNA sequences for which the homology is relatively high, and the swine flu DNA sequences before 1970 for which the homology was relatively high. In the same 10 years, we compare the entire sequence with the gene in the same paragraph. Coupled with the prevailing scientific and medical standards, as well as the environmental effects that can have some impact on the flu, we can scientifically compare the date from the outbreak points with their vicinity. Because the data

from 1919 are incomplete, we can only collect part of the gene sequences from that period, but because the variance of all the sequences is larger than the part, and the year is only the pandemic year, this does not affect the analysis. The analysis shows that the variances in the pandemic year (or + −1 to 2 years) are significantly higher than those of the adjacent years. The 1919 Spanish flu is the most striking one in history, and infected about 500 million people globally. The number of deaths was at least 40 to 50 million, which is almost five times more than those in World War I. Its variance (partial, if the whole is considered it is certainly higher than this) is 2.0252×10^6, much higher than the adjacent years. The 1919 observation figure data is plotted in Fig. 1. The 1913 variance is 3404.9, 1915 is 9566.8, 1916 is 182.6732, 1917 is 17.08, 1918 is 870.6748, 1922 is 455.8428 and 1923 is 3228.9.

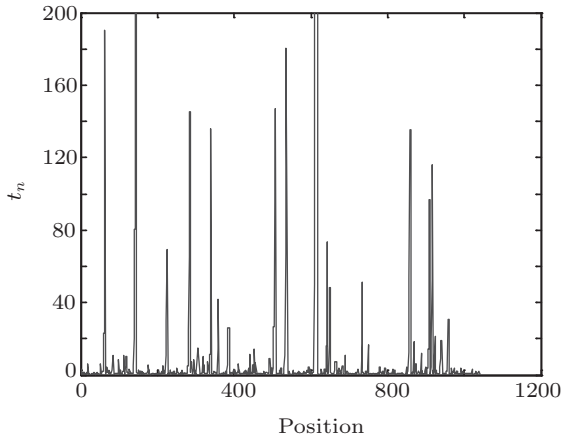


Fig. 1. The 1919 observation figure data.

Between 1957 and 1959 the outbreak of the so-called Asian flu caused the deaths of around 2 million people worldwide, which is considered a very serious flu pandemic. The variance of 1953 is 2.1715×10^4, 1954 is 2.1698×10^4, 1955 is 4.3238×10^3, 1956 is 2.1733×10^4, 1957 is 2.1913×10^4, 1959 is 3.2586×10^4, 1960 is 2.1707×10^4 and 1961 is 1.5387×10^3. The variance in 1959 is also significantly higher than that of the adjacent years. The observation figure of the 1959 data is plotted in Fig. 2.

The Hong Kong flu outbreak in 1968 is considered the "lightest" one in the last century. This virus first appeared in Hong Kong, but in the next two years it spread to other parts of the world and the global mortality rate reached about one million. The variances of 1968 and 1969 are significantly higher than those of the adjacent years, and are about 2.2×10^4. The variances of the years around 1968 and 1969 are at most about 1.2×10^4, and some are way below this. The observation figure of the 1969 data is plotted in Fig. 3.

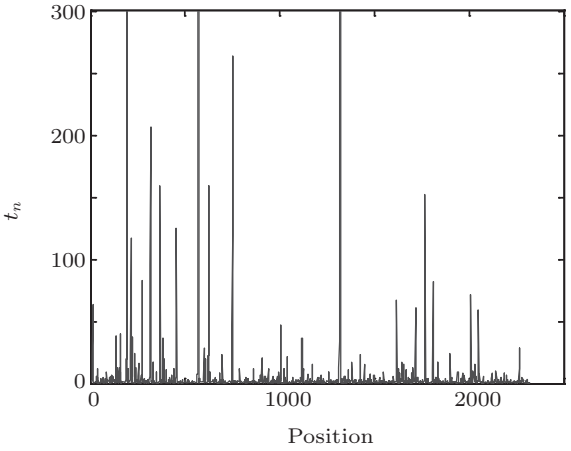


Fig. 2. The observation figure of the 1959 data.

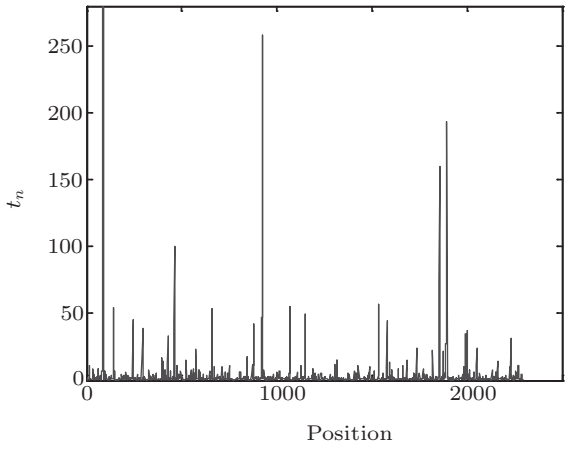


Fig. 3. The observation figure of the 1969 data.

The Russian flu outbreak occurred between 1977 and 1980. The incidence of it was relatively small, i.e. less harmful than some of the other outbreaks. The variances of these years are higher than those of the adjacent years, but not as significant as those in the serious outbreak years mentioned above. In the period 1986 to 1993, many people were infected with the swine flu virus in different parts of the world, so the variances of these years also show the characteristics of a large variance relative to the adjacent years. There was bird flu in Hong Kong in 1997, but there was neither an outbreak nor a variance in the abnormal value. The main reason for this is the relevant environmental conditions. The La Nina event occurred in 1995–1996, the weak cold damage occurred in 1995, 1996 is sunspot valley year m and 1997 is year $m+1$. In 1997–1998 the strongest El Nino events of the 20th century occurred. In addition to Bradley in 1997, in the warm phase of Ramallah there were no serious injuries due to chilling, and the abnormal duration of

the days, during which the high tides of the winter and summer periods continued to shock between the north and the south, the other features of the flu outbreak in 1997 are available.[16] But the warnings of the 1997 avian flu outbreak in Hong Kong cannot be ignored. In 2003 the flu is most widely spread, while its variance also appears abnormal, that is, its variance is higher than in the adjacent years. For the recent 2009 flu pandemic, its variance, which is 3088.4, is also significantly higher than those of the adjacent years, at 1703.2, 719.4759, 1016.0, 977.2245, 621.8030, and so on. The observation figure of the 2009 data is plotted in Fig. 4.

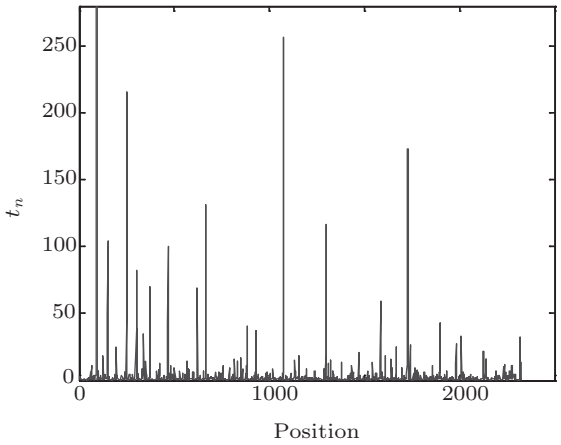


Fig. 4. The observation figure of the 2009 data.

4. Conclusions

Flu outbreaks can prove disastrous to human beings, so we need to establish a warning system prior to these outbreaks. This so-called early-warning mechanism can sensitively and accurately predict risk, so that we can provide timely warnings and use the information to strengthen preventative work, while enhancing targeted prevention. The role of the warning system is to feedback in advance, and to enable timely preparation in the first place. Over the years, environmental disasters such as drought and water logging, snow problems, earthquakes and tsunamis are occurring more and more, and cause human society significant damage. So early-warning systems are particularly important. We can analyse them in the following way. In previous studies on early warning systems, we find that they only give an explanation of a phenomenon from a macro perspective. For example, they observe how the surrounding environment changes when a major disasters happens, which can be a bit vague. In this article, we have used calculations to gain the values of the early-warning signals indicator for the outbreak of a flu pandemic. That is, the variances of the pandemic years (or $+ -1$ to 2 years) are significantly higher than those of the adjacent years. We take the quantitative and qualitative study from the micro to reveal the large intrinsic information contained in a flu outbreak, and use the specific form of mathematical expression to explain the nature of the change. Unfortunately, the data for some individual years are missing or not easy to collect, so there is some impact of this on the study.

On this basis, if the DNA sequence of the flu virus is predicted for the next few years, the future situation can be analysed so that we can take early prevention and control. In addition, we also find that the variance change has important links with the strength of the solar activity cycle, the relative number of sunspots, and the size and extent of the outbreak. How the relationship between these factors can be expressed in a mathematical form needs further study from researchers.

References

[1] Morens D, Folkers G and Fauci A 2004 *Nature* **430** 242
[2] Liu J and Gao J 2011 *Acta Phys. Sin.* **60** 783 (in Chinese)
[3] Lou H R 2009 *Modern Hospital* **9** 6 (in Chinese)
[4] Liu H J and Luo Y L 2007 *Studies of Avian Influenza virus-Induced Apoptosis and HA Protein Expression* (Taiwan: National Pingtung University of Science and Technology) pp. 1, 2 (in Chinese)
[5] Rao Z H and Liu Y F 2009 *Nature* **458** 909
[6] Jeffrey Shaman and Melvin Kohn 2009 *PNAS* **106** 3243
[7] Wei F S 2000 *Science and Technology Daily* **6** 19 (in Chinese)
[8] Yang D H and Yang X N 2006 *Northwestern Seismological Journal* **28** 95 (in Chinese)
[9] Yang D H and Yang X N 2007 *The Desert and the Oasis Weather* **1** 1 (in Chinese)
[10] Scheffer M, Bascompte J, Brock W A, Brovkin V, Carpenter S R, Dakos V, Held H, van Nes E H, Rietkerk M and Sugihara G 2003 *Nature* **461** 53
[11] Carpenter S R and Brock W A 2006 *Ecol. Lett.* **9** 308
[12] Gao J, Xu Z Y and Zhang L T 2007 *Comput. Theor. Nanoscience* **4** 1237
[13] Gao J and Xu Z Y 2009 *Chin. Phys.* B **18** 370
[14] Gao J, Jiang L L and Xu Z Y 2009 *Chin. Phys.* B **18** 4571
[15] Jeffrey H J 1990 *Nucleic Acids Res.* **18** 2163
[16] Yang X N 2009 Web of Science http://www.sciencenet.cn/blog/user_content.aspx?id=228052