EARLY-WARNING MODEL OF INFLUENZA A VIRUS PANDEMIC BASED ON PRINCIPAL COMPONENT ANALYSIS

GAO, J.[1,2,*] – XU, H. X.[1] – DING, T.[1] – WANG, K.[1]

[1]*School of science, Jiangnan University, Wuxi 214122, China*
(phone/ fax: +86-510-85910532)

[2]*Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
(phone/ fax: +86-21-64365937)

**Corresponding author*
e-mail: gaojie@jiangnan.edu.cn; phone:+86-510-85912033

(Received 28th Oct 2016; accepted 28th Feb 2017)

Abstract. During the courses of the human history, several major influenza pandemics caused great disasters to human beings. In this paper, we choose all PA protein sequences of influenza A virus from 1933 to 2013, and these PA protein sequences are translated into chaos game representation walk sequences. For each CGR walk sequence we calculate 25 index values related to sequence structure characteristics, and principal component analysis method is used to study these values. Then we construct a principal component model to early-warn influenza A virus pandemic, compute the comprehensive index values CIV, and sort these values in descending order. Through computing the comprehensive index values based on the model and sorting, the comprehensive index value in 2009 is 30.274, ranked first; the comprehensive index value in 1969 is 8.383, ranked second; and the comprehensive index value in 1959 is 5.684, ranked third. And there were well-known influenza A virus pandemics in 2009, 1969, 1959. It is found that those influenza A virus pandemic years are almost at the top of the list, so we can draw a conclusion using the model that there maybe an influenza A virus pandemic when the CIV in the year is significantly bigger than that in the nearby years and the CIV in the year is more than 3.

Keywords: *PA protein sequences, chaos game representation (CGR) walk model, PCA, eigenvalue, comprehensive index value (CIV)*

Introduction

Influenza A is an acute respiratory infection. It is a disease with a fast transmission and may cause high morbidity and mortality in the world (Kobasa et al., 2004; Morens et al., 2004). In recent years, with the further spread of the flu, the World Health Organization has raised the flu alert level to its sixth grade (Sokolov et al., 2012). This has led many experts and scholars from all over the world to study the influenza from different aspects, and search for the methods to forecast influenza A virus pandemic. For example, Gao et al. (2013) found that a T160A mutation was identified at the 150-loop in the HA gene by means of real-time reverse-transcriptase–polymerase-chain-reaction assays, viral culturing and sequence analysis for clinical, epidemiologic, and virologic data from those patients. Pu et al. (2015) conclude that the prevalence and variation of H9N2 influenza

DOI: http://dx.doi.org/10.15666/aeer/1503_891899
© 2017, ALÖKI Kft., Budapest, Hungary

virus in farmed poultry could provide an important early-warning of the emergence of novel reassortants with pandemic potential.

In 1990, chaos game representation (CGR) for DNA sequences has been proposed by Jeffrey (1990). In 2004, Yu et al proposed a new CGR algorithm of protein sequences based on the detailed hydrophobic-hydrophilic (HP) model (Yu et al., 2003, 2010). In 2009, Gao and Xu proposed a chaos game representation (CGR) walk model based on the new CGR coordinates for the protein sequences from complete genomes. In 2009, Scheffer et al. found that before the critical transition, some complex system such as ecological system, financial market and climate will show the general characteristics, such as leading to the increasing of variance and autocorrelation. In 2011, Ren and Gao used the variance of the influenza virus data as the early-warning signal to analyze and predict the pandemic year of influenza.

We choose all protein sequences of influenza A virus from 1933 to 2013, and the result of PA protein sequences is the most significant. These PA protein sequences are translated into CGR walk sequences. For each CGR walk sequence we calculate 25 index values related to sequence structure characteristics, and principal component analysis (PCA) method is used to study these values. Then we construct a principal component model to early-warn influenza A virus pandemic, compute the comprehensive index values (CIV), and sort these values in descending order. We will establish a warning system to forecast outbreak of influenza A pandemic.

Material and Methods

Dataset

There are 10 proteins PB2, PB1, PA, HA, NP, NA, M1, M2, NS1, NS2 in influenza A virus. Polymerase composed by protein PA and PB1 and PB2 often decides the difficulty of the virus infected host. So we analyze protein PA and PB1 and PB2, and find that the characteristic information of PA protein is the most obvious. And before 1933, there are less than 5 PA protein sequences of influenza A virus only in 1902, 1918, 1927, 1931. So we select all the influenza A virus PA protein sequences from 1933 to 2013. (data from the NCBI website: http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database).

CGR walk model

In 1990 Jeffrey proposed CGR for DNA sequence. The CGR has been extended to protein sequence (Fisher et al., 1994; Basu et al., 1997).

For a given protein sequence $s = s_1 s_2 \cdots s_n$ with length n, where s_i is one of the 20 kinds of amino acids for $i = 1, \cdots, n$, we define

DOI: http://dx.doi.org/10.15666/aeer/1503_891899
© 2017, ALÖKI Kft., Budapest, Hungary

$$c_i = \begin{cases} A0, & if \quad s_i \quad is \quad non-polar, \\ A1, & if \quad s_i \quad is \quad negative \quad polar, \\ A2, & if \quad s_i \quad is \quad uncharged \quad polar, \\ A3, & if \quad s_i \quad is \quad positive \quad polar, \end{cases}$$

and then obtain a sequence $X(s) = c_1 c_2 \cdots c_n$, where c_i is a letter of an alphabet $\{A0, A1, A2, A3\}$. Next, we define a CGR for a sequence $X(s)$, similar to that of DNA sequence, in a square $[0,1]\times[0,1]$, where the four vertices correspond to the four letter $A0$, $A1$, $A2$ and $A3$, $A0=(0,0)$, $A1=(0,1)$, $A2=(1,1)$, $A3=(1,0)$. The first point of the plot is placed half way between the center of the square and the vertex corresponding to the first letter of the sequence $X(s)$; the i-th point of the plot is then placed half way between the (i-1)th point and the vertex corresponding to the i-th letter

$$CGR_i = CGR_{i-1} - 0.5\cdot(CGR_{i-1} - c_i),\quad i = 1,...,n,\quad CGR_0 = (0.5, 0.5) \qquad \text{(Eq.1)}$$

For a given protein sequence, we construct a CGR-walk model , where y_k is the y-coordinate of *CGR$_k$*, x_k is the x-coordinate of *CGR$_k$*.

PCA method

PCA is a commonly used and effective multivariate statistic analysis method (Johnson and Wichern, 2001). Multiple original variables can be reduced into a few comprehensive indexes. Due to too many indexes and a certain correlation between each other, there is duplicate information in the observation data. It is difficult to study on the distribution of the sample in high dimension space. Therefore people want to substitute a few comprehensive indexes for plenty of original variables, and these comprehensive indexes contain enough information to reflect the original variables and they are independent each other.

Here we want to make comprehensive analysis and evaluation for the multiple indexes of protein data of influenza A virus each year using PCA method. The few comprehensive indexes can provide most of the information in the original indexes, therefore we can simplify the analysis of virus protein data. Finally we can make quantitative and comparison evaluation for each year’s data.

Results

Indexes

Based on the CGR-walk model for the protein sequence of influenza A virus, we can translated PA protein sequences into numerical sequences. Thus we can get the specific numerical data. Based on the principal of scientific and feasibility, we choose 25 indexes reflecting the sequence structure characteristic information of protein data of influenza A virus: average(x_1), variance(x_2), standard deviation(x_3), coefficient variation(x_4), lag *k* autocorrelation coefficient(x_k).

For each numerical sequence we calculate 25 index values.

Data analysis based on PCA

We compute the eigenvalue of correlation matrix, contribution and cumulative contribution of every principal component (*Table 1*).

According to *Table 1*, we can see that the contribution of the first principal component is 41.17%, the contribution of the second principal component is 19.51%, and the contribution of the third principal component is 10.11%. This also shows that the three principal components have gathered about 70% data information of 25 indexes.

The selected principal component numbers should meet the general requirement that the cumulative contribution had better be greater than 85%. We extract the first six principal components because their cumulative contribution has been up to 87.62% (*Table 1*). Therefore the first six principal components represent about 87.62% data information of all the original indexes.

Table 1. *Eigenvalue, proportion and cumulative contribution of every principal component*

item	eigenvalues	proportion	cumulative
z1	9.88178292	0.4117	0.4117
z2	4.68336852	0.1951	0.6069
z3	2.42705361	0.1011	0.7080
z4	1.62607717	0.0678	0.7758
z5	1.42185141	0.0592	0.8350
z6	0.98931913	0.0412	0.8762
z7	0.77035157	0.0321	0.9083
z8	0.48546446	0.0202	0.9286
z9	0.46061874	0.0192	0.9477
z10	0.29273349	0.0122	0.9599
z11	0.24461422	0.0102	0.9701
z12	0.13972687	0.0058	0.9760
z13	0.10908996	0.0045	0.9805
z14	0.09633166	0.0040	0.9845
z15	0.07787623	0.0032	0.9878
z16	0.06715905	0.0028	0.9906

z17	0.05626723	0.0023	0.9929
z18	0.04961278	0.0021	0.9950
z19	0.03837153	0.0016	0.9966
z20	0.02963200	0.0012	0.9978
z21	0.02454026	0.0010	0.9988
z22	0.01251669	0.0005	0.9993
z23	0.00920517	0.0004	0.9997
z24	0.00643533	0.0003	1.0000
z25	0.00000000	0.0000	1.0000

Early-warning model of influenza A virus pandemic

Here we use $x_i\left(i=1,2,...,25\right)$ as the 25 indexes, $z_j\left(j=1,2,...,6\right)$ as the first six principal components, y as the CIV.

After normalizing the eigenvalues as the weights, we can construct a principal component model to make comprehensive analysis.

$$y=0.470z_1+0.223z_2+0.115z_3+0.077z_4+0.068z_5+0.047z_6 \qquad \text{(Eq.2)}$$

where the first six principal components are

$$z_1=0.155x_1+0.082x_2+0.168x_3-0.273x_5-0.209x_6+...+0.110x_{25}$$

$$z_2=-0.346x_1-0.147x_2-0.200x_3+0.158x_5+0.181x_6+...-0.378x_{25}$$

$$z_3=0.214x_1+0.129x_2+0.159x_3+0.120x_5+0.119x_6+...+0.251x_{25}$$

$$z_4=0.007x_1+0.0777x_2+0.058x_3+0.170x_5+0.378x_6+...+0.018x_{25}$$

$$z_5=0.118x_1+0.326x_2+0.188x_3+0.087x_5+0.040x_6+...-0.037x_{25}$$

$$z_6=0.045x_1+0.506x_2+0.070x_3+0.131x_5+0.268x_6+...-0.004x_{25}$$

Then a principal component model is obtained, i.e.

$$\begin{aligned}y=&0.036x_1+0.073x_2+0.051x_3-0.054x_5+0.0002x_6+0.119x_7+0.095x_8+0.156x_9+\\&0.161x_{10}+0.166x_{11}+0.037x_{12}+0.099x_{13}+0.155x_{14}+0.051x_{15}+0.155x_{16}+0.066x_{17}+\\&0.091x_{18}+0.129x_{19}+0.137x_{20}+0.095x_{21}+0.153x_{22}+0.121x_{23}+0.090x_{24}-0.005x_{25}\end{aligned} \qquad \text{(Eq.3)}$$

Through computing the CIVs based on the model and sorting, the CIV in 2009 is 30.274, ranked first; the CIV in 1969 is 8.383, ranked second; and the CIV in 1959 is 5.684, ranked third (*Table 2*).

Table 2. *Principal component value, CIV and ranking(R) of every year*

year	z1	z2	z3	z4	z5	z6	CIV	R
2009	35.45261	-64.3871	55.6356	32.82124	135.5943	208.8902	30.27	1
1969	10.75625	-20.3654	16.9396	9.437207	37.34962	56.40566	8.383	2
1959	7.795853	-15.1847	12.27423	6.59448	25.2774	37.55668	5.684	3
2008	6.92281	-13.5436	11.01679	5.961404	22.97595	34.28074	5.135	4
2010	6.539319	-12.886	10.38519	5.55199	21.17855	31.42996	4.741	5
2011	6.425023	-12.678	10.22931	5.473158	20.82713	30.89184	4.661	6
1993	5.596687	-11.1373	8.946197	4.743167	17.84305	26.32073	3.993	7
1991	5.219861	-10.5587	8.360391	4.335047	16.01551	23.41813	3.635	8
1943	5.214887	-10.4623	8.328537	4.348227	16.22068	23.80802	3.586	9
1986	5.002644	-10.0965	7.957286	4.099997	15.19687	22.18209	3.409	10
2012	4.851384	-9.89545	7.8407	4.053103	14.88543	21.75071	3.324	11
1987	4.760841	-9.71039	7.653763	3.932538	14.4574	21.0958	3.232	12
1942	4.71931	-9.66417	7.616887	3.931356	14.35521	20.91896	3.203	13
1998	4.63469	-9.38142	7.38231	3.786482	13.95378	20.32245	3.133	14
1935	4.573606	-9.384	7.399063	3.798539	13.84638	20.15095	3.091	15
1984	4.525075	-9.20382	7.21974	3.692189	13.56738	19.71275	3.040	16
2006	4.443647	-9.08068	7.137881	3.653652	13.3598	19.42347	2.989	17
2004	4.408305	-9.03353	7.096972	3.627309	13.22039	19.21263	2.979	18
2007	4.398081	-8.9953	7.084407	3.633615	13.27443	19.30528	2.968	19
2000	4.394602	-9.01406	7.095766	3.638541	13.24885	19.25685	2.960	20
1990	4.384001	-8.95316	7.07441	3.640189	13.33133	19.4291	2.957	21
2013	4.338363	-9.0209	7.109714	3.631746	13.14415	19.12288	2.919	22
1980	4.312083	-8.85831	6.949018	3.547294	12.87069	18.68706	2.879	23
1940	4.278652	-8.90727	7.025017	3.587957	12.96613	18.84638	2.878	24
2001	4.260032	-8.76218	6.873552	3.503521	12.69635	18.42958	2.843	25
1981	4.232171	-8.70129	6.848415	3.503136	12.72917	18.49092	2.840	26
1989	4.222301	-8.67173	6.787246	3.450174	12.53508	18.1458	2.804	27
2002	4.196715	-8.62875	6.764026	3.442491	12.49758	18.11223	2.794	28
1982	4.187976	-8.60569	6.735987	3.41699	12.45142	18.04004	2.784	29
2003	4.111168	-8.4786	6.667246	3.404529	12.32938	17.88454	2.751	30
1966	4.110836	-8.4605	6.650805	3.400187	12.26497	17.7928	2.744	31
1946	4.065945	-8.39225	6.574759	3.341303	12.05336	17.45467	2.712	32
1988	4.063561	-8.48662	6.676848	3.400406	12.21633	17.70217	2.695	33
1999	4.037999	-8.32806	6.518398	3.313335	11.96569	17.30743	2.675	34
1985	3.997718	-8.27397	6.472493	3.283428	11.82137	17.09198	2.640	35
1934	3.982795	-8.23895	6.447109	3.272446	11.7926	17.05317	2.634	36

1949	3.969248	-8.29151	6.50847	3.310986	11.85032	17.15423	2.634	37
2005	3.963369	-8.2032	6.418088	3.252412	11.71281	16.91381	2.633	38
1947	3.961196	-8.25633	6.496668	3.300623	11.8421	17.13525	2.618	39
1983	3.942828	-8.25215	6.48484	3.29192	11.78705	17.0612	2.615	40
1954	3.94202	-8.16981	6.393387	3.244238	11.67373	16.88136	2.608	41
1994	3.929615	-8.23147	6.465072	3.286152	11.75578	17.00483	2.605	42
1977	3.925433	-8.05524	6.302699	3.179317	11.3804	16.49163	2.569	43
1996	3.881433	-8.02709	6.267735	3.166216	11.35729	16.4135	2.558	44
1963	3.879425	-8.07637	6.322384	3.203014	11.47917	16.57627	2.551	45
1992	3.877836	-8.02063	6.255938	3.182188	11.41601	16.48086	2.547	46
1995	3.86972	-8.03878	6.283142	3.18044	11.41713	16.48211	2.546	47
1979	3.865806	-8.04053	6.294495	3.18636	11.41509	16.47804	2.544	48
1951	3.845849	-7.95636	6.214164	3.135211	11.24216	16.23212	2.540	49
1997	3.844208	-8.05372	6.319254	3.202019	11.43375	16.51276	2.525	50
1971	3.838855	-7.91574	6.193847	3.135487	11.26491	16.27181	2.524	51
1978	3.823466	-7.96203	6.247162	3.170972	11.3238	16.35059	2.519	52
1975	3.772341	-7.83217	6.110679	3.082703	11.01468	15.87567	2.464	53
1976	3.755126	-7.79189	6.088791	3.069454	10.93845	15.79496	2.452	54
1974	3.713367	-7.78417	6.090982	3.081197	10.95042	15.77402	2.435	55
1973	3.701735	-7.75512	6.060881	3.063336	10.88844	15.68703	2.423	56
1970	3.677573	-7.70947	6.035923	3.05188	10.82081	15.59016	2.409	57
1957	3.659814	-7.68056	6.008265	3.03068	10.75799	15.48909	2.393	58
1972	3.620107	-7.61571	5.956285	3.009259	10.65492	15.33504	2.367	59
1950	3.56955	-7.4756	5.865539	2.907946	10.49566	15.14866	2.337	60
1965	3.555612	-7.49582	5.853291	2.95037	10.4289	15.00284	2.316	61
1968	3.473366	-7.29213	5.686112	2.856599	10.04583	14.44506	2.277	62
1961	3.455395	-7.31377	5.716101	2.870219	10.10658	14.50721	2.244	63
1958	3.448371	-7.29522	5.743725	2.908215	10.27138	14.86336	2.242	64
1960	3.443321	-7.28738	5.690758	2.85632	10.06019	14.43522	2.232	65
1933	3.44021	-7.26985	5.680508	2.85143	10.03667	14.39776	2.230	66
1948	3.432887	-7.25711	5.670356	2.846259	10.01652	14.36421	2.224	67
1964	3.431035	-7.24787	5.658146	2.846155	10.00246	14.33896	2.222	68
1945	3.426377	-7.24317	5.660667	2.83876	9.992328	14.32875	2.219	69
1962	3.41329	-7.21923	5.636199	2.831483	9.948574	14.26059	2.209	70
1936	3.412416	-7.21813	5.630371	2.825652	9.930939	14.25011	2.206	71
1956	3.405262	-7.20874	5.644031	2.822356	9.931835	14.25194	2.206	72
1967	3.349235	-7.06397	5.514002	2.75762	9.702004	13.90095	2.160	73
1953	3.293649	-6.9902	5.470127	2.739002	9.570284	13.68923	2.125	74
1952	*	*	*	*	*	*	*	
1955	*	*	*	*	*	*	*	

* represents the influenza A virus PA protein sequence data missing.

Discussion

From *Table 2*, we can see that the CIVs in 2009, 1969 and 1959 have the highest ranking. Also the CIVs in 1993, 1991 are very high.

During 1957 to 1959, the Asian flu pandemic caused two million people's death all over the world, and it is one of the most serious outbreak in the history. The CIV in 1959 is significant bigger than that in nearby several years. The CIV in 1954 is 2.608, the CIV in1956 is 2.206, the CIV in 1957 is 2.393, the CIV in 1958 is 2.242, 1959 is 5.684, the CIV in 1960 is 2.232, the CIV in 1961 is 2.244.

Hong Kong flu in 1969 is similar to the Asian flu in 1959. Maybe the affected people had accumulated related antibody in the Asian flu, so the Hong Kong flu had relatively fewer deaths than other epidemic. Estimated that there were 750,000 people died (in America 34,000 people died). And Hong Kong flu caused more than one million deaths. The CIV in 1969 is 8.383, which is higher than 2.160, 2.277, 2.409, 2.524 in the nearby years.

In 2009 there was well-known serious bird flu. The CIV in 2009 is 30.274 which is the highest in all years, and it is higher than 5.135, 2.524, 4.741, 4.661 in nearby years .

During 1986 to 1993, there are a lot of human infected the swine flu epidemic in many areas of the world. From *Table 2* we can see the CIV in these years are bigger than that in near years.

It is found that those influenza A virus pandemic years are almost at the top of the list, so we can verify those influenza A virus pandemic years based on the model, and draw a conclusion that there maybe an influenza A virus pandemic when the CIV in the year is significantly bigger than that in the nearby years and the CIV in the year is more than 3 (*Fig. 1*). Therefore we can regard the principal component model as an early-warning model to predict the pandemic year, and also verify its validity and objectivity. In this way we can take prevention and control.

In this paper, we might have not choose the best indexes that can completely reflect the protein data information of influenza A virus, so we should make further research and improvement in the future work. In addition, due to the lack of partial year data (e.g.1952, 1955), we cannot calculate the CIVs of these mentioned years, which it will affect our prediction and analysis of the pandemic years.

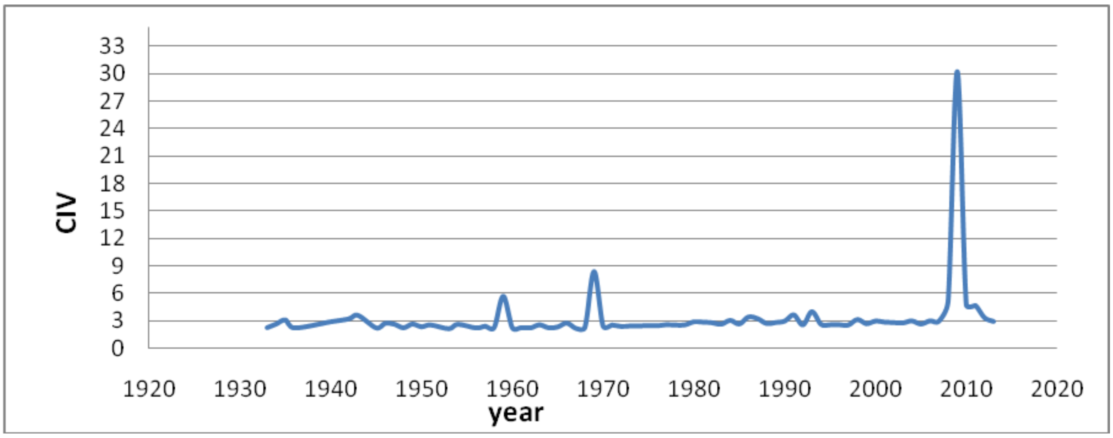

Figure 1. *The CIV data from 1933 to 2013*

Acknowledgments. The project was supported by the National Natural Science Foundations of China (Grant No. 11271163 and No. 11371174).

REFERENCES

[1] Basu, S., Pan, A., Dutta, C., Das, J. (1997): Chaos game representation of proteins.–J. Mol. Graph. Model 15: 279-289.

[2] Fisher, A., Tusnady, G.E., Simon, I. (1994): Chaos game representation of protein structures.–J. Mol. Graphics 12: 302-304.

[3] Gao, J., Xu, Z.Y. (2009): Chaos game representation walk model for the protein sequences.–Chin. Phys. B 18: 4571-4579.

[4] Gao, R.B., Cao, B., Hu, Y.W., Feng, Z.J., Shu, Y.L., et al. (2013): Human infection with a novel Avian-Origin influenza A(H7N9) Virus.–N Engl J Med 368: 1888 -1897.

[5] Jeffrey, H. J. (1990): Chaos game representation of gene structure.–Nucleic Acid Res 18: 2163-2170.

[6] Johnson, R. A., Wichern, D. W. (2001): Applied multivariate statistical analysis. –Tsinghua Press, Beijing, 388. [in Chinese]

[7] Kobasa, D., Takada, A., Shinya, K., et al. (2004): Enhanced virulence of influenza A viruses with the haemagglutinin of the 1918 pandemic virus. –Nature 431: 703-707.

[8] Morens, D., Folkers, G., Fauci, A. (2004): The challenge of emerging and re-emerging infectious disease.–Nature 463: 242-249.

[9] Pu, J., Wang, S., Yin, Y., Zhang, G., Carter, R.A., et al. (2015): Evolution of the H9N2 influenza genotypes that facilitated the genesis of the novel H7N9 virus.–PNAS 112: 548-553.

[10] Ren, D., Gao, J. (2011): Early-warning signals for an outbreak of the influenza pandemic.–Chin. Phys. B 12: 128701.

[11] Scheffer, M., Bascompte, J., Brock, W. A., et al. (2009): Early-warning signals for critical transitions.–Nature 461: 53-59.

[12] Sokolov, D. N., Zarubaev, V. V., Shtro, A. A., et al. (2012): Anti-viral activity of (−)- and (+)-usnic acids and their derivatives against influenza virus A(H1N1)2009.–Bioorganic & Medicinal Chemistry Letters 22: 7060-7064.

[13] Yu, Z.G., Anh, V.V., Lau, K.S. (2003): Multifractal and correlation analysis of protein sequences from complete genome.–Phys. Rev. E 68: 021913.

[14] Yu, Z.G., Anh, V.V., Lau, K.S. (2010): Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses.–J. Theor. Biol 226: 341-348.

DOI: http://dx.doi.org/10.15666/aeer/1503_891899
© 2017, ALÖKI Kft., Budapest, Hungary