广西科学 Guangxi Sciences 2010,17(3):247～254

Fitting Evolutionary Process of Polymerase Acidic Protein Family from Influenza A Virus with Analytical Solution of System of Differential Equations*

用微分方程组的解析解拟合甲型流感病毒聚合酶酸性蛋白家族的进化过程

YAN Shao-min[1],WU Guang[2**]

严少敏[1],吴　光[2]

(1. National Engineering Research Center for Non-food Biorefinery, Guangxi Academy of Sciences, Nanning, Guangxi, 530007, China; 2. Computational Mutation Project, DreamSciTech Consulting,Shenzhen, Guangdong, 518054, China)

(1.广西科学院国家非粮生物质能源工程技术研究中心,广西南宁　530007;2.深圳市追梦科技咨询有限公司,广东深圳　518054)

Abstract:We fitted the evolution of polymerase acidic protein family from influenza A virus using a mathematical model: (i) we used the amino-acid pair predictability to quantify 2433 polymerase acidic proteins isolated from 1918 to 2008 to represent their evolution, (ii) we determined if the uphill half-life is similar to the downhill one as a pre-request for fitting, (iii) we used the analytical solution of system of differential equations to fit this evolution, and (iv) we simulated the possible evolutionary process from 2009 to 2018 using the obtained fitted parameters. The results showed a good-of-fit for polymerase acidic protein family and its different subtypes, indicating that the study on protein evolution begins to move forward dynamically mathematical modeling from passively empirical data-collection.

Key words: amino-acid pair predictability,evolution, fitting, influenza A virus, polymerase acidic protein,differential equation

摘要:用数学模型对甲型流感病毒聚合酶酸性蛋白家族的进化进行拟合:(1)用氨基酸对的可预测性量化 1918 年至 2008 年分离的 2433 个聚合酶酸性蛋白以表示其演变,(2)确定上升半寿期和下降半衰期是否相似作为拟合的前提条件,(3)用微分方程组的解析解拟合进化,(4)用所获得的拟合参数模拟 2009 年至 2018 年可能的进化过程。结果呈现了对聚合酶酸性蛋白家族及其不同亚型的良好拟合,这标志着蛋白质进化的研究已经从经验性的实验分析向动力学的数学建模迈进。

关键词:氨基酸对的可预测性　进化　拟合　甲型流感病毒　聚合酶酸性蛋白　微分方程

中图法分类号:R373.1+3　　**文献标识码**:A　　**文章编号**:1005-9164(2010)03-0247-08

The unpredictable mutations in influenza A viruses frequently threaten the world with possible flu pandemics or epidemics, this is so because we do not have many ways to predict the mutation positions, would-be-mutated amino acids, and when a mutation would occur[1~9]. In broad sense, this is so because we do not know much about the evolution of influenza A virus in a predictable form, although we have already accumulated a large amount of samples of influenza A viruses at different time points in different geographical locations[10].

For better understanding of the evolution of influenza A virus, we not only need to have a considerable amount of data, but also we need to find out a mathematical way to describe this evolutionary

收稿日期:2009-11-05

作者简介:严少敏(1958-),女,博士,研究员,主要从事定量诊断病理学和计算变异学研究。

* This study was partly supported by National Science and Technology Platform Construction (2008FU115XB), Guangxi Science Foundation (0907016 and 0991080) and Guangxi Academy of Sciences (09YJ17SW07).

** 通讯作者。

process, then we would be in the position to have a basic concept where the influenza A virus would be likely to go, because scientific history clearly shows that a mathematical description is a landmark for the development of scientific discipline.

The evolution of influenza A virus is a process of mutations along the time course, thus the first task is to represent influenza A viruses along the time course. Actually the influenza A virus contains ten protein families, therefore it is necessary to represent each protein family over time. In coordinates, it should be that x -axis is the sampled time while y -axis marks the proteins in terms of meaningful values because it would be meaningless if we use their accession number for y -axis.

This means that we need a way to meaningfully represent a protein family along the time course, which can be done using the computational mutation approach developed by our group[5]. Thereafter, we have attempted to use the fast Fourier transform to determine the periodicity in the evolution of hemagglutinins from influenza A viruses[11,12].

Very recently, we have explored the possibility to use a system of differential equations to describe the evolution of hemagglutinin family from influenza A virus[13], because this mathematical description reasoned the underlined mechanism for evolution[14~16]. The basic assumption is that each mutation literally brings in a mutating amino acid and takes away a mutated amino acid. The difference between bring-in-amino-acid and take-away-amino-acid can be quantified using our approach in terms of randomness (entropy)[1~5, 11~13], which constructs a standard mass-balance relationship suited for the description of differential equation[13].

The polymerase acidic protein (PA) is one of ten proteins found in influenza A virus and plays an important role in all RNA synthesizing activities associated with influenza virus[17] as they replicate and transcribe their segmented negative-sense single-stranded RNA genome in the nucleus of the infected host cell. The PA subunit is also involved in the conversion of RNA polymerase from transcriptase to replicase[18] and contains the endonuclease active site. A recent study strongly implicates the viral RNA polymerase complex as a major determinant of the pathogenicity of the 1918 pandemic virus[19].

Therefore, it is very meaningful to use the analytical solution of system of differential equations, which describe the mutation process for a protein family, to fit the evolutionary process of PA proteins from influenza A viruses, in order to pave the way for timing mutation and understanding the underlined mechanism of protein evolution from influenza A virus.

1 Materials and methods

1.1 Data

5165 full-length PA proteins of influenza A viruses sampled from 1918 to 2008 were obtained from the influenza virus resources[10]. After excluded identical sequences, 2433 PA proteins were actually used in this study.

1.2 Presentation of PA evolution

We use the amino-acid pair predictability to convert each PA protein as a single number that really represents the instinct of protein[1~5, 11, 13]. For example, a swine H1N1 influenza virus was isolated in 1976, and strain A/swine/Tennessee/15/1976 (H1N1). Its PA protein (accession number ABQ45443) has 716 amino acids. The first and second amino acids can be counted as an adjacent amino-acid pair, the second and third as another adjacent amino-acid pair, the third and fourth, until the 715th and 716th, thus there are totally 715 amino-acid pairs. This PA protein has 39 aspartic acids (D) and 75 glutamic acids (E): if the permutation can predict the appearance of amino-acid pair DE, it must appear 4 times ($39/716 \times 75/715 \times 715 = 4.09$); actually it does appear four times, so the pair DE is predictable. By contrast, this protein has 60 leucines (L): if the permutation can predict the appearance of amino-acid pair LL, it must appear five times ($60/716 \times 59/715 \times 715 = 4.94$); however, it appears nine times in realty, so the pair LL is unpredictable. In this way, all amino-acid pairs in ABQ45443 PA protein can be classified as predictable and unpredictable, which are 28.53% and 71.47%.

Another swine influenza virus was isolated in 1979, and its PA protein (accession number ABR28643) has only one amino acid different from ABQ45443 one at position 323. However, its predictable and unpredictable portions are 29.51% and 70.49%. Thus, the amino-acid pair predictability distinguishes one PA protein from another in terms of numbers rather than the letters that represent amino acids in proteins.

In this manner, we can use 28.53% to represent ABQ45443 PA protein and 29.51% to represent ABR28643 PA protein in a coordinates, where x -axis is the time of isolated year and y-axis is the predictable portion. This method is applied to all 2433 PA proteins involved in this study.

1. 3 Analytical solution of system of differential equations

In our previous studies, we have shown that the possibly analytical solution for n differential equations is a sum of decaying exponential and sinusoidal functions $y(t) = \sum_{i=1}^{n} A_i e^{-k_i t}\cos(\alpha_i t + \varphi_i) + C$, where y is the predictable portion over time, A, α and k are parameters, t is time, φ is phase difference, and C is a constant[13].

1. 4 Fitting

The above analytical solution suggests that there was an input at zero time, then the exchange of entropy from generation to generation results in decaying but fluctuating. The symmetry in the fluctuation is an important feature, indicating the uphill half-life is similar to the downhill one. All of these are the basis for fitting[20], which was conducted using SigmaPlot.

1. 5 Statistics

The Mann-Whitney U-test was used to compare the difference between uphill and downhill half-life, and $P < 0.05$ was considered significant.

2 Results

Figure 1 shows the evolutionary process of PA proteins over 90 years in terms of all 2433 PA proteins and their subtype classifications, which can be read as follows. For example, the solid curve in the top panel of figure 1 represents the evolution of 2433 PA proteins from 1918 to 2008, and each point is the mean value of predictable portions of all PA proteins in a given year with its standard deviation (vertically grey line). The similar reading can be applied to other panels.

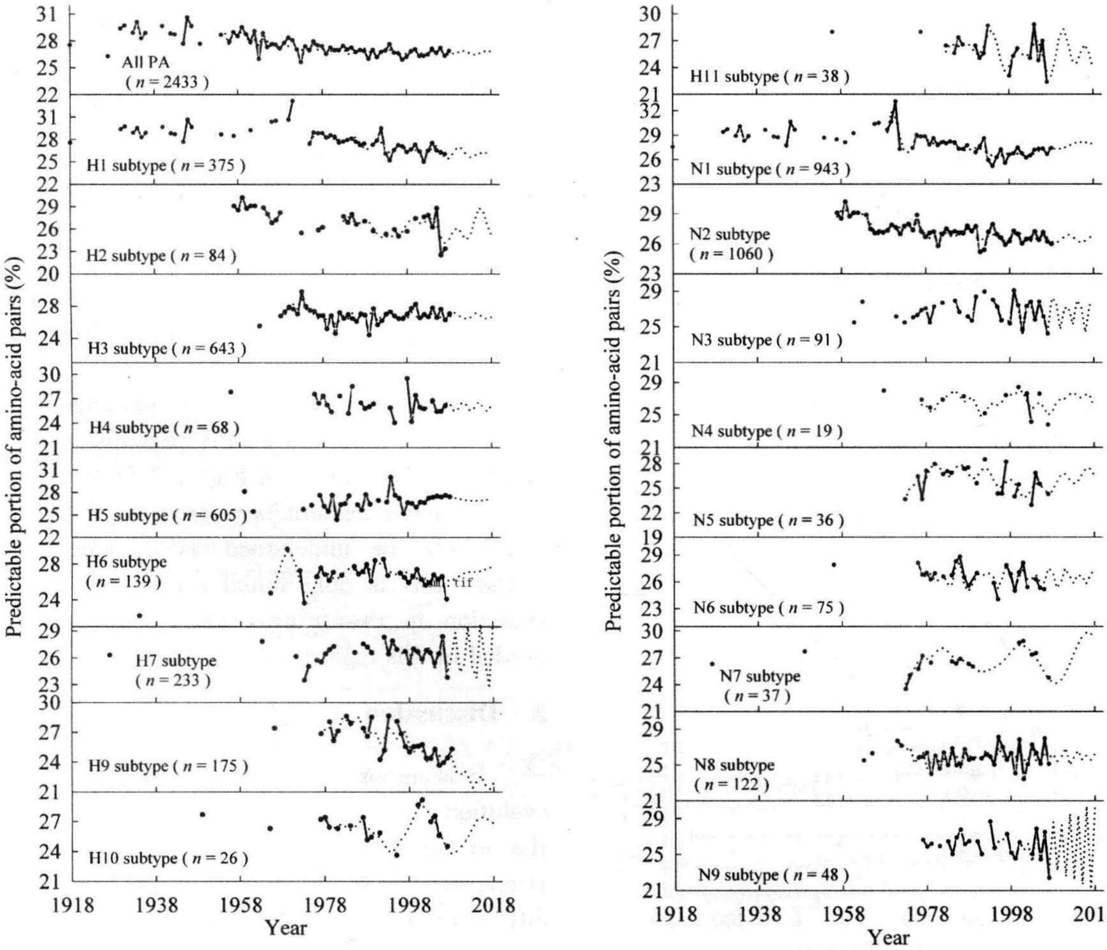

Fig. 1 Evolution of PA proteins with respect to all PA proteins and different substypes from influenza A viruses

The solid curves from 1918 to 2008 are the evolution of PA proteins in terms of randomness (entropy), the filled cycles with vertical solid lines are mean ±SD of predictable portion in all the PA proteins in given year, the dotted curves from 1918 to 2008 are fitted curves, and the dotted curves from 1918 to 2018 are the simulated curves representing the evolution of PA proteins in the future.

With decaying exponential, the half-life is $T_{1/2} = \frac{\ln(2)}{k} = \frac{0.696}{k}$, where $k = \frac{\ln(y_{\text{peak}}) - \ln(y_{\text{trough}})}{t_{\text{interval}}}$, which is the downhill half-life. Symmetrically, we can also compute the uphill half-life. All possibly stratified peaks and troughs are recorded to compute the half-life. Figure 2 displays no statistical difference between the uphill half-lives and downhill ones, which provides the theoretical basis for fitting. The calculated half-life can also serve as initial estimates for fitting.

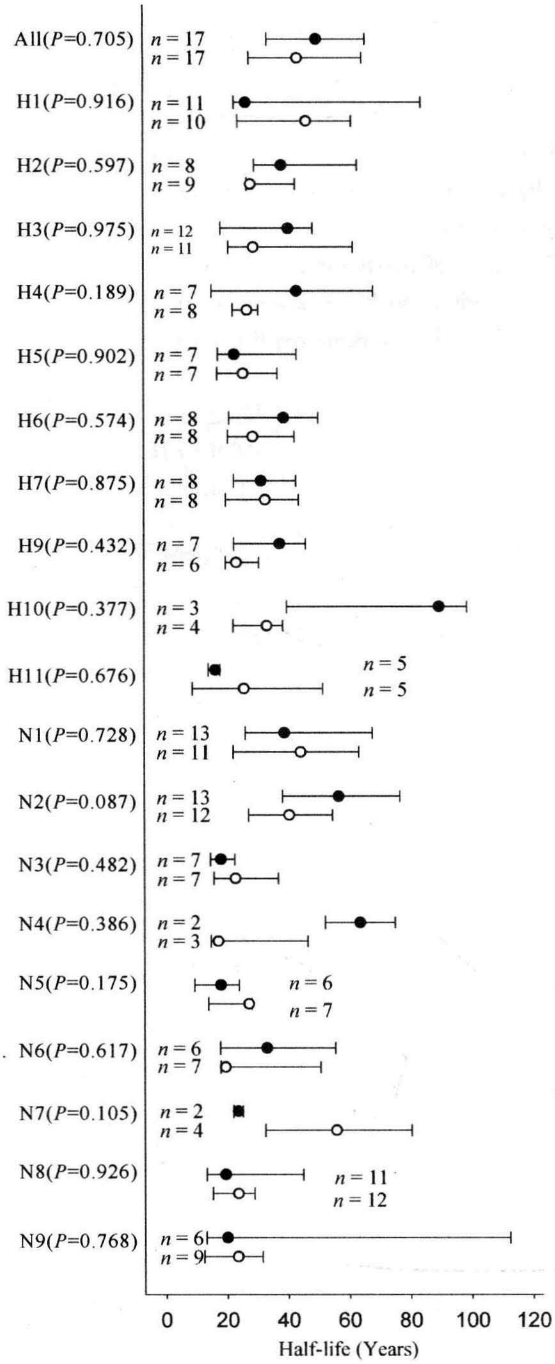

Fig. 2 Comparison of uphill half-life with downhill half-life

The data are presented as median with interquatile range. No statistical differences were found in the half-life between uphill and downhill (Mann-Whitney U -test).

The dotted curves in Figures 1 are fitted curves using the analytical solution, whose fitted parameters are listed in Table 1. As can be seen, the dotted curves generally are much approximate to the evolutionary trend presented by the solid curve, indicating that the analytical solution can present the evolutionary process of PA proteins from influenza A viruses. In general, the fitted results differ in different subtypes. For example, the fitted values quite match to their actual ones in H4, H5, H7, H10, N3, N7 and N9 subtypes.

After fitting, an important step is to determine if the conducted fitting is good, which can generally be done using many different methods[20, 21]. For example, the residuals reflect the difference between actual and fitted values. Figures 3, 4 and 5 show the residual versus the time, fitted value, and actual value, respectively. In Figure 3, we cannot find the residuals either increase or decrease over time. In other words we cannot find any trend of residual along the time course suggesting a good-of-fit. In Figure 4, we also cannot find the residuals either increase or decrease over fitted value suggesting a good-of-fit again. In Figure 5, we can see the trend that the residuals increase as actual values increase in some cases, which is understandable because some samples might be located far beyond the general trend found by fitting and far beyond the mean ± SD for the samples in the given year.

With the obtained fitted parameters, we can simulate the evolution of PA proteins in the future, and the dotted curves in Figure 1 from 2009 to 2018 are simulated evolutionary process of PA proteins, which can be understood as possible trends. Of course, this is only initial attempt to simulate the evolution in the future, and much more work is needed in this regard.

3 Discussion

It is important to understand the mechanism of evolution, and it is also equally important to describe the evolutionary process, thus in this study we attempted to use the analytical solution of system of differential equations to fit the evolutionary process of PA family from influenza A virus. For the past, the mathematical description of evolution can help us understand all factors which impact the evolution. For the future, the mathematical description can help us

Table 1 Fitted parameters in analytical solution of system of differential equations

Subtype	A_1	k_1	α_1	φ_1	A_2	k_2	α_2	φ_2	A_3	k_3	α_3	φ_3	C	R^2
All	−0.752	0.034	2.681	−108.381	−40.634	0.136	0.009	1.617	0.527	0.015	1.081	−0.006	26.689	0.728
H1	−18.995	0.113	0.020	1.700	−0.528	0.000	49.301	7.242	0.395	0.000	1.325	−3.682	25.960	0.711
H2	−1.269	0.000	0.416	−4.841	−2.472	0.000	1.126	2.361	2.239	0.000	1.093	2.753	26.096	0.576
H3	−0.331	0.000	1.226	−11.057	0.864	0.032	2.625	−22.016	3.612	0.149	0.364	−8.474	26.740	0.469
H4	0.624	0.000	−4.601	95.344	−45.051	0.861	3.144	−4.929	—	—	—	—	26.037	0.982
H5	−6.333	0.449	1.492	−0.681	−1.828	0.099	0.414	3.318	−3.391	0.304	2.521	5.752	27.103	0.996
H6	−0.713	0.000	0.215	10.173	1.356	0.043	0.909	0.637	4.567	0.144	0.571	3.459	26.896	0.727
H7	−8.666	0.000	1.987	4.750	−6.345	0.381	−18.853	−98.498	−9.492	0.019	1.965	8.034	26.288	0.878
H9	−0.804	0.016	0.571	−14.257	−6.979	0.000	0.039	3.015	−0.659	0.000	−10.833	−57.939	20.735	0.721
H10	−0.472	0.035	1.361	1.276	−1.669	0.012	0.340	7.360	−1.221	0.000	0.488	9.539	26.040	0.872
H11	2.903	0.115	0.278	10.288	1.213	0.000	0.693	4.463	0.668	−0.033	1.134	−0.979	25.643	0.606
N1	−7.732	0.224	0.849	0.979	−0.738	0.030	0.793	4.002	1.117	0.009	0.147	4.950	27.408	0.754
N2	−0.380	0.000	1.045	1.171	−25.950	0.048	0.000	1.676	1.236	0.051	0.476	3.893	26.278	0.734
N3	−3.884	0.394	−2.218	13.275	1.793	0.000	16.779	6.765	1.851	0.062	1.303	−1.565	26.606	0.968
N4	−1.058	0.000	0.438	5.569	0.395	0.000	1.148	−2.163	0.475	0.000	7.865	10.510	26.614	0.676
N5	2.828	0.030	−0.201	2.981	0.891	0.000	1.035	−2.046	2.296	0.088	0.340	4.266	25.862	0.657
N6	−0.851	0.000	−1.810	−2.658	−1.905	0.145	0.909	0.870	1.336	0.033	1.244	5.531	26.489	0.621
N7	4.117	0.000	−18.450	−41.978	3.121	0.000	0.430	4.696	0.821	0.089	1.072	−4.335	26.633	0.878
N8	−0.630	0.000	−10.178	138.003	−2.376	0.134	0.217	3.543	0.467	0.000	−1.672	−13.896	25.957	0.405
N9	−1.813	0.082	1.638	14.247	17.538	0.000	−3.128	−20.505	−2.715	0.240	−0.699	−7.177	25.772	0.923

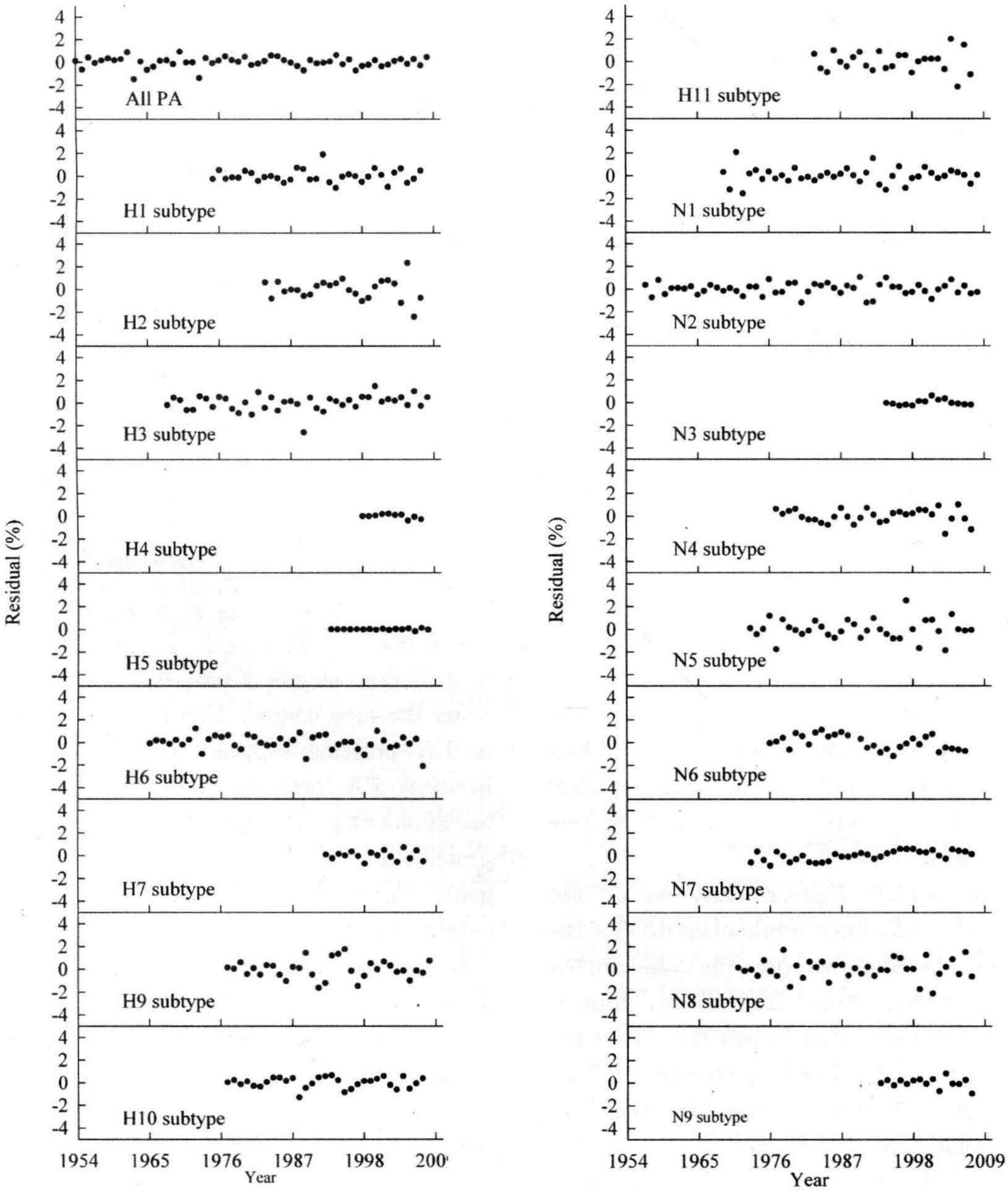

Fig. 3 Residual versus time in all and different subtypes of PA proteins from influenza A virus

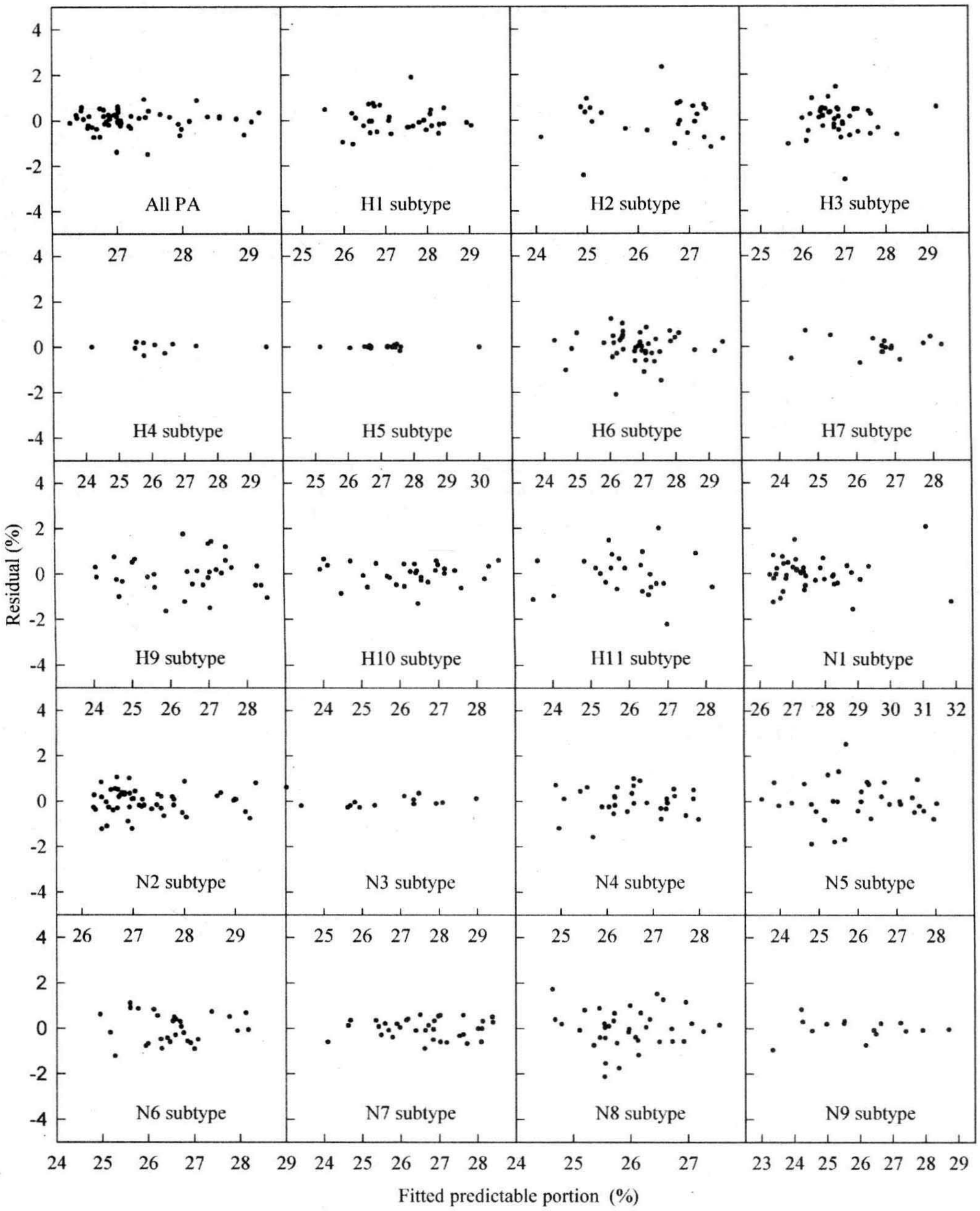

Fig. 4 Residual versus fitted value of predictable portion for all and different subtypes of PA proteins from influenza A virus

know where the biological evolution will go. For example, we would know whether and when a species will extinct or be prosperous at species level[14], while we would know when a protein family will mutate more at protein level.

For our mathematical description, we do not require uncountable underlined mechanism driving the protein evolution because we use the randomness (entropy) to cooperate them all[1~5, 11~13], thus it would be easy to correlate other factors that affect the evolution and can be modeled as perturbation[22]. This may warrant the systematical and mathematical description of evolution[14~16].

In Figures 1, the solid curves actually represent the measured randomness (entropy) in PA proteins along the time course. The fluctuation of the entropy in PA proteins suggests the exchange of entropy between PA proteins and their environments, so mutations play the role to balance the entropy between proteins and their environments, and drive evolution going on. On the other hand, the decaying of the entropy in PA proteins suggests the discharge of the entropy that was stored at the very beginning of formation of the first PA protein, and the exchange of entropy among generations of PA proteins. In both cases, we would expect to see a similar length for both downhill half-lives and uphill ones, otherwise we would expect to have seen irregular half-lives if random perturbations would be added to this evolutionary process of PA family.

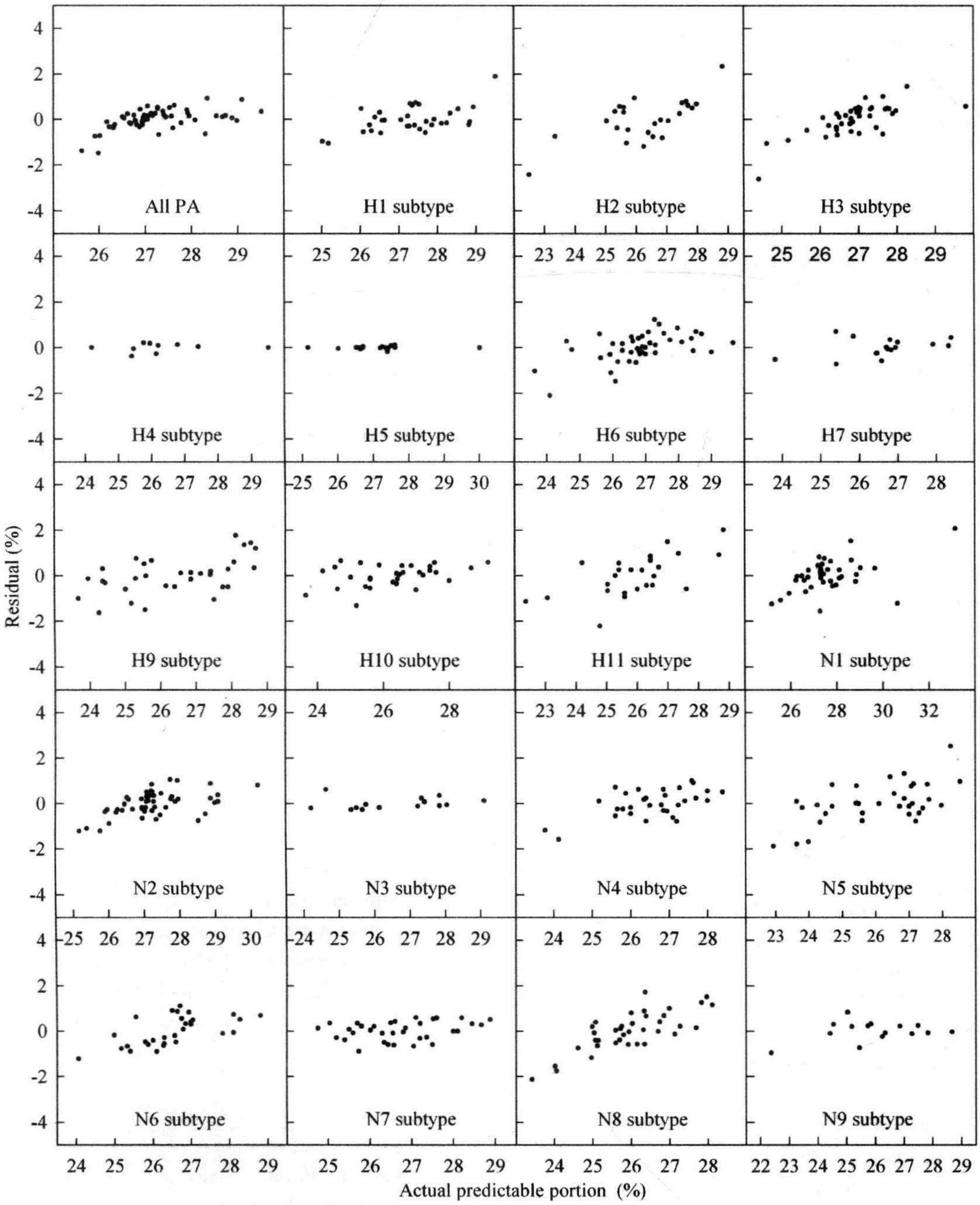

Fig. 5 Residual versus actual value of predictable portion for all and different subtypes of PA proteins from influenza A virus

An important application of this differential description of PA evolution is that we can use this analytical solution with fitted parameters to predict the PA evolution in the future, where the fluctuation would be a spike of mutations. Then it would be possible to connect the spike of mutations with possible flu outbreaks. This approach would have some advantage over our approach using fast Fourier transform to stratify the evolutionary process, and then to compare each stratified segments to time the mutation[1,5,11,12].

The results of this study demonstrate that the analytical solution can fit the PA evolution from influenza A virus, which is very important for predicting its general trend. Currently, the evolution of a protein or its gene has been conducted according to all sequencing data, which can provide information on the phylogenetic trees, evolutionary rate and selective pressure[23]. In this context, our study provides the information on dynamic process because neither phylogenetic tree nor evolutionary rate nor selective pressure can easily and clearly be represented along the time course, which consequently is hard to be modeled. Thus, our approach can be considered the first step to analyze the evolution of protein family from empiric description to mathematical modeling[24].

References:

[1] Wu G,Yan S. Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint[J]. Acta Pharmacol Sin,2006,27:513-526.

[2] Wu G,Yan S. Prediction of mutations in H1 neuraminidases from North America influenza A virus engineered by internal randomness[J]. Mol Divers,2007,11:131-140.

[3] Wu G,Yan S. Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus[J]. Amino Acids,2008,34:81-90.

[4] Wu G,Yan S. Prediction of mutations engineered by randomness in H5N1 hemagglutinins of influenza A virus [J]. Amino Acids,2008,35:365-373.

[5] Wu G,Yan S. Lecture notes on computational mutation [M]. New York:Nova Science Publishers,2008.

[6] Duvvuri V R,Duvvuri B,Cuff W R,et al. Role of positive selection pressure on the evolution of H5N1 hemagglutinin[J]. Genomics Proteomics Bioinformatics, 2009,7:47-56.

[7] Dunham E J,Dugan V G,Kaser E K,et al. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses[J]. J Virol, 2009,83:5485-5494.

[8] Hill A W,Guralnick R P,Wilson M J,et al. Evolution of drug resistance in multiple distinct lineages of H5N1 avian influenza[J]. Infect Genet Evol,2009,9:169-178.

[9] Xia Z,Jin G,Zhu J,et al. Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus[J]. Bioinformatics,2009,25: 2309-2317.

[10] The National Center for Biotechnology Information. Influenza virus resources [EB/OL]. [2009-09-20]. http://www. ncbi. nlm. nih. gov/genomes/FLU/ Database/multiple. cgi,2009.

[11] Wu G,Yan S. Timing of mutation in hemagglutinins from influenza A virus by means of unpredictable portion of amino-acid pair and fast Fourier transform [J]. Biochem Biophys Res Commun,2005,333:70-78.

[12] Wu G,Yan S. Timing of mutation in influenza A virus hemagglutinins by means of amino-acid distribution rank and fast Fourier transform [J]. Protein Pept Lett, 2006,13:143-148.

[13] Yan S,Wu G. Describing evolution of hemagglutinins from influenza A viruses using a differential equation [J]. Protein Pept Lett,2009,16:794-804.

[14] Ao P. Laws of Darwinian evolutionary theory [J]. Phys Life Rev,2005,2:117-156.

[15] Ao,P. Metabolic network modelling: Including stochastic effects [J]. Compt Chem Eng, 2005, 29: 2297-2303.

[16] Ao P,Lee L W,Lidstrom M E,et al. Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AM1 growth as validation[J]. Chinese J Biotechnol,2008,24:980-994.

[17] Engelhardt O G,Fodor E. Functional association between viral and cellular transcription during influenza virus infection[J]. Rev Med Virol,2006,16:329-345.

[18] Honda A,Ishihama A. The molecular anatomy of influenza virus RNA polymerase[J]. Biol Chem,1997, 378:483-488.

[19] Watanabe T,Watanabe S,Shinya K,et al. Viral RNA polymerase complex promotes optimal growth of 1918 virus in the lower respiratory tract of ferrets[J]. Proc Natl Acad Sci USA,2009,106:588-559.

[20] Wu G. Fit fluctuating blood drug concentration: a beginner's first note[J]. Pharmacol Res,1996,33:379-383.

[21] Wu G,Cossettini P,Furlanut M. Prediction of blood cyclosporine concentrations in haematological patients with multidrug resistance by one-, two- and three-compartment models using Bayesian and non-linear least squares methods[J]. Pharmacol Res,1996,34:47-57.

[22] Wu G,Yan S. Searching of main cause leading to severe influenza A virus mutations and consequently to influenza pandemics/epidemics [J]. Am J Infect Dis, 2005,1:116-123.

[23] Dugan V G,Chen R,Spiro D J,et al. The evolutionary genetics and emergence of avian influenza viruses in wild birds[J]. PLoS Pathog,2008,4:e1000076.

[24] Wu G,Yan S M. Creation and application of computational mutation[J]. J Guangxi Acad Sci 2010, 26:130-139.

（责任编辑：尹　闯）