流感病毒组成蛋白质序列的分析与预测

靳佩轩， 高 洁*
（江南大学 理学院，江苏 无锡 214122）

摘要：在NCBI数据库中获得1902—2013年关于流感病毒10种组成蛋白的所有氨基酸序列，在MATLAB中采用大数据编程分析，结合详细的HP模型，并基于CGR-WALK模型的方法将全部流感病毒蛋白质序列转化为数据形式，引入时间序列ARFIMA(p,d,q)模型来拟合所有序列，分析10种组成蛋白的序列在近80年的变化趋势，并对其未来10年的发展趋势进行预测. 通过分析可以发现，其对流感病毒变异趋势的预测有很好的效果，这为基于大数据分析流感病毒蛋白质序列，预测流感病毒的爆发提供一定的的研究参考价值.

关键词：流感病毒；蛋白质序列；详细HP模型；CGR-WALK模型；ARFIMA(p,d,q)模型
中图分类号：Q 51　　**文献标志码**：A　　**文章编号**：1673—1689(2016)04—0393—06

Sequence Analysis and Prediction of the Influenza Virus Protein

JIN Peixuan， *GAO Jie**
（School of Science，Jiangnan University，Wuxi 214122，China）

Abstract：Ten protein amino acid sequences of influenza virus were obtained from the National Center for Biotechnology Information （NCBI） from 1902 to 2013，which was analyzed using big data in MATLAB programming with the detailed HP model. Meanwhile，the protein sequences were converted to the data series based on the CGR - WALK model. The time series ARFIMA(p,d,q) was introduced to fit all the sequences. The analysis results indicated a good model with accurate prediction for the variation tendency in the next 10 years，which also provided a reference for the prediction of influenza virus using the big data analysis.

Keywords：influenza virus，protein sequence，the detail HP model，CGR-WALK model，ARFIMA(p,d,q) model

流感病毒为负向单链RNA病毒[1]，自身具有很强的变异性，历史上多次流感大流行都是由其新的亚型和以往出现过的亚型经过变异再次出现，人类由于缺乏对其的免疫力而导致流感病毒在人群中快速传播。

目前对流感病毒蛋白质序列的研究上，刘娟等用时间序列模型识别，预测流感病毒的DNA序列。张玲用时间序列模型识别，预测甲型H1N1流感病毒蛋白质序列在未来年份的变异情况取得很好的预测效果[2-5]。作者以时间序列分析研究为基础，分别选取数据库中现有的从1902—2013年间近100年的流感病毒的10种组成蛋白：Hemagglutinin，Matrix

收稿日期：2014-06-09
基金项目：国家自然科学基金项目(11271163)；中央高校基本科研业务费专项资金项目(JUSRP21117)。
* **通信作者**：高 洁(1972—)，女，江苏无锡人，工学博士，副教授，主要从事生物信息学研究。E-mail：ezhun6669@sina.com

Protein 1,Matrix Protein2,Neuraminidase,Nonstructural Protein1,Nonstructural Protein2,Nucleocapsid Protein,Polymerase PA,Polymerase PB1,Polymerase PB2 等作为研究对象。运用大数据处理方法将全部序列以 HP 模型为基础进行数据化转换，并利用 CGR-WALK 建模，再采用时间序列 ARFIMA(p,d,q) 模型分析流感病毒每种蛋白质的变异规律和未来的发展趋势，以蛋白质序列整体为研究对象，从宏观研究分析的角度为研究流感病毒在之后几年内的变异情况提供预测依据，并能够为相关流感病毒的预测研究提供重要的研究思路和方法。

1 材料与方法

1.1 材料

选取 NCBI 数据库中 1902—2013 年之间所有流感病毒的 10 种组成蛋白质序列，即 Hemagglutinin,Matrix Protein 1,Matrix Protein 2,Neuraminidase,Nonstructural Protein 1,Nonstructural Protein 2,Nucleocapsid Protein,Polymerase PA,Polymerase PB1,Polymerase PB2 蛋白质序列作为作者的研究对象进行分析。(NCBI:http:/ /www.ncbi.nlm.nih.gov /.)

1.2 方法

1.2.1 蛋白质序列基于详细的 HP 模型数据化构建 CGR-WALK 模型 Jeffrey 在 1990 年提出的一种将序列数据化的 CGR-WALK 方法[6],其是一种迭代映射技术,可以将蛋白质序列中的每一个位置上氨基酸投影到一个连续坐标空间上,由此将序列进行可视化图形表示,同时可以进行有效的进行独立的精确尺度的序列分析研究。

在详细的 HP 模型中将 20 种氨基酸分成 4 大类,即非极性氨基酸(non-polar),极性带负电荷的氨基酸 (negative polar), 极性不带电荷的氨基酸(uncharged polar),极性带正电荷的氨基酸(positive polar), 在此分别记作 NP,NEP,UP,PP。因此将 20 种氨基酸{A,I,L,M,F,P,W,V,N,C,Q,G,S,T,Y,D,E,R,H,K} 按照在详细的 HP 格点模型中依据氨基酸的生物特性的分类方法将氨基酸依次分类:

NP ={A,V,L,I,P,F,W,M},NEP ={D,E},UP ={G,S,T,C,Y,N,Q},PP={K,R,H}。

经过分类之后,则可将任意含有 n 个氨基酸的蛋白质序列进行数据化定义:蛋白质序列 $s=s_1s_2s_3\cdots s_n$,其中 s_i,i=1,2,…,n 为组成此蛋白质序列的氨基酸,

$$\alpha_1=\begin{cases}A_0,\text{若 } s_i\in \text{NP};\\ A_1,\text{若 } s_i\in \text{NEP};\\ A_2,\text{若 } s_i\in \text{UP};\\ A_3,\text{若 } s_i\in \text{PP}\end{cases}$$

由上方法即可将任意一条蛋白质序列转化为一条由 A_0,A_1,A_2,A_3 构成的四元序列,记作:$X(s)=\alpha_1\alpha_2\alpha_3\cdots\alpha_n$。

定义序列 $X(s)$的 CGR-WALK:

(1)在二维坐标平面上作[0,1]×[0,1]正方形,标记四个顶点为 $A_0(0,0)$,$A_1(0,1)$,$A_2(1,1)$,$A_3(1,0)$。

(2)以正方形中心(0.5,0.5)作为 CGR-WALK 的初始点。

(3)设置目标蛋白质序列的第一个数据作为当前迭代目标，并将初始点与当前起始目标坐标连线,并标记此线段中点。

(4)以此规律依次迭代,继续以蛋白质序列的下一个数据作为当前迭代目标,循环执行(3)过程,直到将整条蛋白质序列循环运算结束,最终得到在坐标平面上的一个可视化 CGR-WALK 模型视图。

在此给出 CGR 迭代函数公式: 含有 n 个氨基酸的蛋白质序列:$s=s_1s_2s_3\cdots s_n$,其中 s_i,i=1,2,…,n,并且有 $s_i\in${A,V,L,I,P,F,W,M,D,E,G,S,T,C,Y,N,Q,K,R,H}, 由详细的 HP 模型的分类得由 A_0,A_1,A_2,A_3 构成的序列:$X(s)=\alpha_1\alpha_2\alpha_3\cdots\alpha_n$。通过以下迭代过程得到此蛋白质序列的 CGR:令 $A_0(0,0)$,$A_1(0,1)$,$A_2(1,1)$,$A_3(1,0)$,即:

$\text{CGR}_i=\text{CGR}_{i-1}-0.5(\text{CGR}_{i-1}-g_i)$,$i$=1,2,…N,$\text{CGR}_i=(0.5,0.5)$。

其中 $g_i\in\{(0,0),(1,0),(1,1),(0,1)\}$,$g_i$ 与 s_i 相对应。

对于流感病毒的蛋白质序列研究,在此定义变量 $t_k=y_k/x_k$,其中 x_k 和 y_k 分别是 CGR_k 的 x 和 y 对应坐标值, 由此可以将甲型 H1N1 流感病毒蛋白质序列数据化为一条具有统计意义的数据序列 $\{t_k:k=1,2,\cdots,n\}$, 即视作一条时间序列, 由于其是经过 CGR-WALK 得到,则在此记为“CGR-WALK 序列”。

1.2.2 ARFIMA 模型 定义 1 $\{\varepsilon_t\}$为白噪声序列[7],记作 $\varepsilon_t\sim\text{WN}(\mu,\sigma_2)$。如果时间序列满足如下性质:

(1)任取 $t\in T$,有 $E\varepsilon_t=u$;

(2)任取 $t,s\in T$,有 $\gamma(t,s)=\begin{cases}\sigma^2,t=s;\\ 0,t\neq s;\end{cases}$

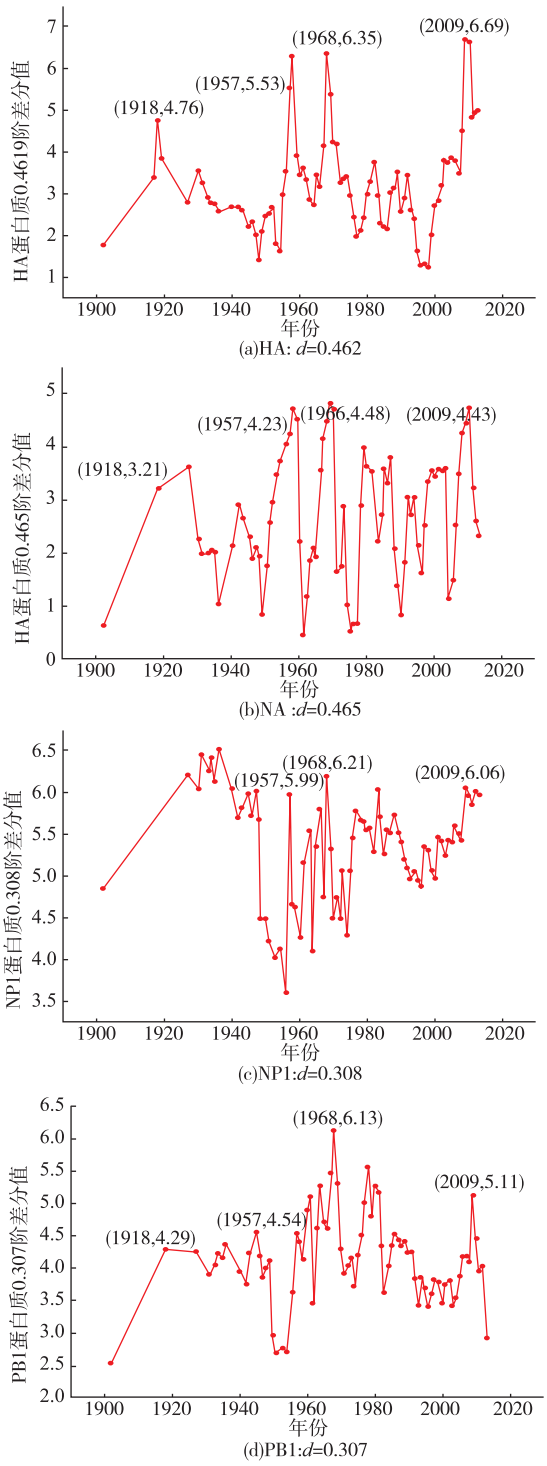

图 1　蛋白质 HA,NA,NP1,PB1 差分时序图

Fig. 1　Differenced model of HA, NA, NP1, PB1 sequence timing diagram

定义 2　如果随机序列$\{X_t\}$满足差分方程$(1-B)^d x_t=\varepsilon_t$，其中 $-0.5<d<0.5$，$\{\varepsilon_t\}$为白噪声序列，$E\varepsilon_t=0$，$E\varepsilon_t^2=\sigma\varepsilon^2<\infty$，称$\{X_t\}$服从$-0.5<d<0.5$ 的 ARFIMA(0, d, 0)模型[7]。

定义 3　如果随机过程$\{X_t\}$是平稳的，且满足差分方程 $\Phi(B)\nabla^d x_t=\Theta(B)\varepsilon_t$，其中$\{\varepsilon_t\}$为白噪声序列，$E\varepsilon_t=0$，$E\varepsilon_t^2=\sigma_\varepsilon^2<\infty$，$\Theta(B)=1-\Phi_1B-\cdots-\Phi_pB^p$，为 p 阶自回归系数多项式；$\Theta(B)=1-\theta_1B-\cdots-\theta_qB^q$，为 q 阶移动平均系数多项式，$-0.5<d<0.5$，则称$\{X_t\}$服从$-0.5<d<0.5$ 的 ARFIMA(p,d,q)模型[8]。

2　结果与分析

2.1　流感病毒蛋白质序列数据集构造

选取数据库中 1902—2013 年的所有关于流感病毒蛋白质序列，对所取的数据集中每一条序列从第一个位置开始，进行数据化处理，即将 $A_0\rightarrow0$；$A_1\rightarrow1$；$A_2\rightarrow2$；$A_3\rightarrow3$。则可将由初始的氨基酸序列转化为由 0,1,2,3 构成的四元序列，并对序列进行 CGR-WALK 转化，获得 CGR-WALK 后的$\{t_k\}$数据集。在此对每种蛋白质对应的 t 值序列分别求其变异系数，得到 10 组相应蛋白质的变异系数数据集。

2.2　甲型 H1N1 流感病毒蛋白质数据集特征分析

以流感病毒蛋白质 HA、NA、NP1、PB1（分别为 Hemagglutinin, Neuraminidase, Nonstructural Protein 1, Polymerase PB1 等 4 种蛋白简写）为例进行分析。计算得其差分值分别为：$d_1=0.462$，$d_2=0.465$，$d_3=0.308$，$d_4=0.307$。

对应其 $d_i(i=1,2\cdots,10)$阶差分分别得到 4 组对应的差分序列。然后分别作对应阶差分的时间序列图(如图 1)，可以看到在从所选择数据中 4 次爆发年份 1918,1957,1968,2009 年所对应差分值较高，此数据结构与实际情况相符。

在图 2 中分别做了关于 4 种蛋白对应阶差分值的自相关函数(ACF)和偏自相关函数(PACF)图像，可以发现他们的自相关函数曲线衰减迅速，偏相关函数曲线衰减缓慢，则其具备长记忆的特征。

对 4 种蛋白质对应的差分序列进行白噪声检验，(结果见表 1)，均有 $P<0.000\ 1<0.05$，则可知它们不是白噪声序列，则可利用 ARFIMA(p,d,q)模型对 4 组序列进行拟合。

由 Akaike 信息判别准则，选取 ARFIMA (1,-0.047,1)，ARFIMA (4,0.308,0)，ARFIMA (2, -0.146,0)，ARFIMA(2,0.307,3)对 4 条序列进行拟合，在表 2 中给出模型中相应的参数估计值。从表中可以看到，P 值都小于 0.1，则表明 ARFIMA(p,d,q)模型的对序列能够取得很好的拟合效果。

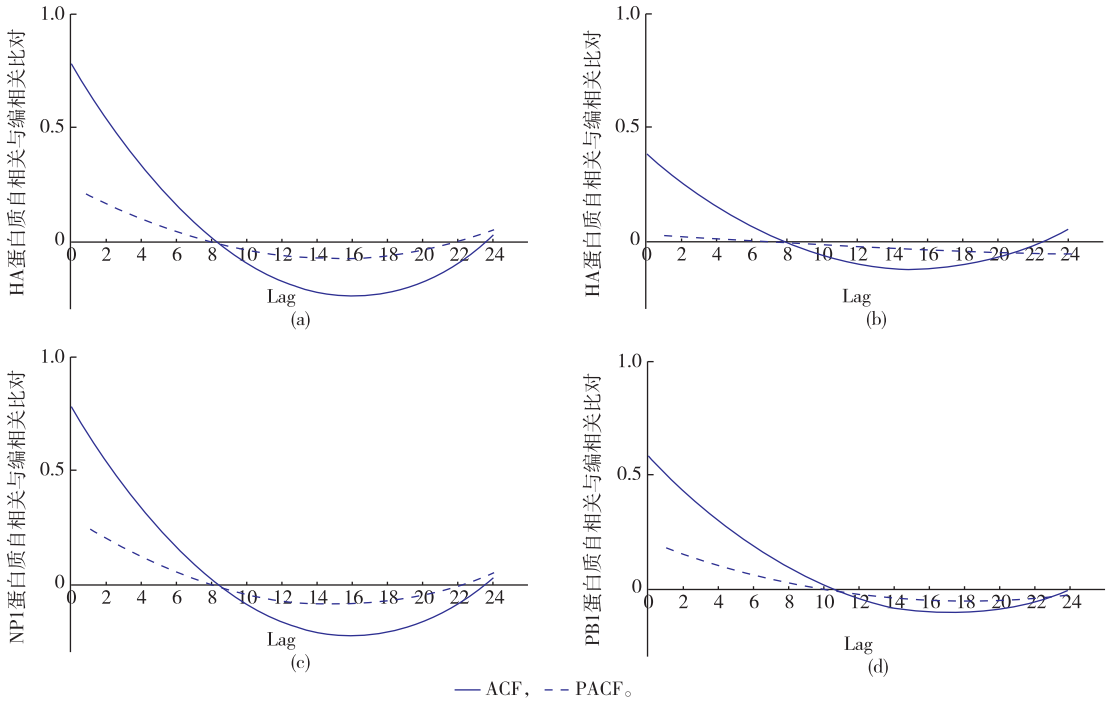

图 2　蛋白质 HA、NA、NP1、PB1 的 ACF 与 PACF 曲线图

Fig. 2　ACF and PACF of HA、NA、NP1、PB1

表 1　蛋白质 HA、NA、NP1、PB1 的白噪声检验

Table 1　White noise test of HA、NA、NP1、PB1

	HA				NA			
To Lag	6	12	18	24	6	12	18	24
Chi-Square	85.94	87.68	117.81	128.01	60.55	79.04	103.50	116.05
DF	6	12	18	24	6	12	18	24
Pr >ChiSq	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	NP1				PB1			
To Lag	6	12	18	24	6	12	18	24
Chi-Square	75.67	86.31	112.09	117.86	50.53	55.91	66.99	82.28
DF	6	12	18	24	6	12	18	24
Pr >ChiSq	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

最后对模型的合理性进行检验，选取 LB 检验统计量：

$$LB=n(n+2)\sum_{k=1}^{M}\frac{r_k^2}{n-k}\sim X^2(M-p-q-1)$$

其中 r_k 是滞后的样本自相关函数，n 为样本容量，M 是一个比 n 小的常数，且 n 为正整数。

由此可得表 3 中为各滞后阶数的相关统计量，其中 LB 统计量的 P 值都显著大于 0.1，则可知所拟合模型的残差序列为白噪声序列，则 ARFIMA(p,d,q)模型可以进行正确的数据拟合，即此模型在流感病毒蛋白质序列分析是合理的。

在表 4 中给出由 ARFIMA(p,d,q)模型进行流感病毒 4 种组成蛋白在未来 10 年中的趋势预测。并从对应图 3 中的预测图中综合观察，自 1902 年起到未来 10 年中关于 4 种流感病毒组成蛋白 HA、NA、NP1、PB1 的变化趋势。在 1918,1957,1968,2009 年 4 次流感爆发年份中，在图中都有大幅度的波动，表明该模型的建立符合实际情况。从 2014—2023 这未来 10 年中波动情况存在差异，蛋白质 HA 与NP1 较平缓，蛋白质 NA 与 PB1 有明显起伏，因

表 2　蛋白质 HA、NA、NP1、PB1 的参数的最小二乘估计

Table 2　Least-square estimation of HA、NA、NP1 and PB1

	HA			NA				
Parameter	MU	MA1,1	AR1,1	MU	AR1,1	AR1,2	AR1,3	AR1,4
Estimate	2.910 22	−0.680 85	0.553 05	2.600 78	0.814 15	−0.350 25	0.189 42	−0.269 5
Standard Error	0.268 11	0.093 08	0.107 99	0.152 82	0.111 58	0.144 16	0.144 84	0.114 49
t	10.85	−7.31	5.12	17.02	7.3	−2.43	1.31	−2.35
Approx Pr > |t|	<.000 1	<.000 1	<.000 1	<.000 1	<.000 1	0.017 5	0.019 49	0.021 2
Lag	0	1	1	0	1	2	3	4
	NP1			PB1					
Parameter	MU	AR1,1	AR1,2	MU	MA1,1	MA1,2	MA1,3	AR1,1	AR1,2
Estimate	5.379 68	0.411 29	0.285 81	4.021 11	0.549 9	−0.19753	−0.58236	1.203 46	−0.649
Standard Error	0.177 19	0.111 67	0.112 42	0.147 6	0.136 43	0.128 59	0.107 8	0.148 12	0.142 93
t	30.36	3.68	2.54	27.24	4.03	−1.54	−5.4	8.12	−4.54
Approx Pr > |t|	<.000 1	0.000 4	0.013 1	<.000 1	0.000 1	0.012 89	<.000 1	<.000 1	<.000 1
Lag	0	1	2	0	1	2	3	1	2

表 3　蛋白质 HA、NA、NP1、PB1 的参数的自相关检验

Table 3　Auto-correlation test of the HA、NA、NP1 and PB1

	HA				NA			
To Lag	6	12	18	24	6	12	18	24
Chi-Square	3.7	6.36	11.46	15.48	0.79	2.19	8.62	12.91
DF	4	10	16	22	2	8	14	20
Pr >ChiSq	0.448 5	0.7843	0.780 1	0.841	0.674 4	0.974 8	0.854 8	0.881 2
	NP1				PB1			
To Lag	6	12	18	24	6	12	18	24
Chi-Square	6.71	15.41	32.52	34.34	0.49	2.78	11.73	12.72
DF	4	10	16	22	1	7	13	19
Pr >ChiSq	0.151 9	0.117 7	0.008 5	0.045 3	0.484 8	0.905	0.549 8	0.852 7

表 4　蛋白质 HA、NA、NP1、PB1 在未来 10 年中的预测值

Table 4　Forest values of HA、NA、NP1 and PB1

HA		NA		NP1		PB2	
年份	预测值	年份	预测值	年份	预测值	年份	预测值
2014	4.148	2014	1.921	2014	5.806	2014	3.015
2015	3.595	2015	1.976	2015	5.727	2015	3.299
2016	3.289	2016	2.282	2016	5.645	2016	3.312
2017	3.120	2017	2.507	2017	5.588	2017	3.636
2018	3.026	2018	2.701	2018	5.541	2018	4.018
2019	2.975	2019	2.823	2019	5.506	2019	4.268
2020	2.946	2020	2.815	2020	5.478	2020	4.320
2021	2.930	2021	2.742	2021	5.456	2021	4.220
2022	2.921	2022	2.656	2022	5.439	2022	4.067
2023	2.916	2023	2.577	2023	5.426	2023	3.947

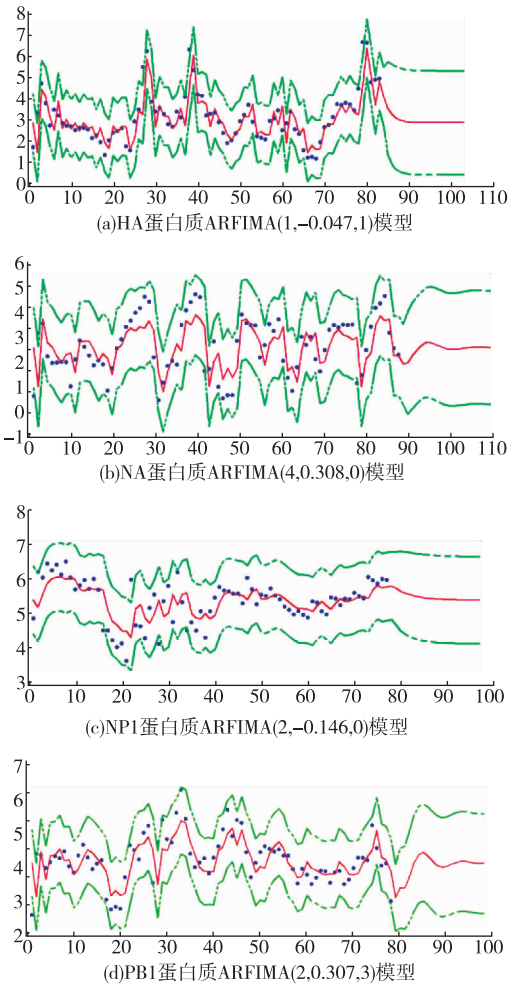

图 3　蛋白质 HA、NA、NP1、PB1 差分时序模型与预测图

Fig. 3　Forest model of HA、NA、NP1、PB1

此在未来十年里要对蛋白质 NA 与 PB1 的变异情况进行重点的研究和检测，为流感病毒预防和临床治疗提供指导与帮助。

3 结　语

利用大数据分析处理方法获得所要研究的对象，并利用详细的 HP 模型将全部蛋白质序列数据化，以 CGR-WALK 模型和分数阶差分模型对目标数据进行处理，最终求的相应的变异系数序列，分析后显示其具有明显的长记忆特征，并对相应序列的差分化序列运用 ARFIMA(p,d,q)模型对其进行预测，从理论上取得了较好的预测分析结果；又由预测结果可知，利用从 1902—2013 年流感病毒的 10 种组成蛋白的氨基酸序列数据得到的时间序列模型结果很好的与历史上 4 次较大流感爆发年份 1918,1957,1968,2009 相吻合，并接着进行了 2014—2023 年序列的变异情况进行预测，从理论和实际分析应用中表明时间序列分析结合大数据的思想对流感病毒蛋白质序列的分析预测有重要的应用和研究意义，由此表明最终获得很好的预测效果，在下一步的研究中要进一步完善数据收集和整理工作，以更加全面的数据作为研究分析工作的基础，并对模型在精确度上进行优化，提高该研究的实际应用价值和科研价值。

参考文献：

[1] Hilleman,MauriceR. Realities and enigmas of human viral influenza:pathogenesis,epidemiology and control [J]. **Vaccine**, 2002,20(25-26):3068-3087.

[2] 刘娟,高洁. 甲型流感病毒 DNA 序列的长记忆 ARFIMA 模型[J]. 物理学报,2011,60(4):702-707.
LIU Juan,GAO Jie. Long-memory ARFIMA model for DNA sequences of influenza A virus [J]. **Acta Physica Sinica**,2011,60(4):702-707.(in Chinese)

[3] 刘娟,高洁. 甲型 H1N1 流感病毒 DNA 序列碱基的预测[J]. 生物信息学,2011,9(3):259-262.
LIU Juan,GAO Jie. Forecasting bases for DNA sequences of influenza virus A/H1N1 [J]. **China Journal of Bioinformatics**, 2011,9(3):259-262.(in Chinese)

[4] 刘娟,高洁. 乙型、丙型流感病毒 DNA 序列的长记忆 ARFIMA 模型[J]. 生物信息学,2011,9(2):97-101.
LIU Juan,GAO Jie. Long-memory ARFIMA model for DNA sequences of influenza B,C virus [J]. **China Journal of Bioinformatics**,2011,9(2):97-101.(in Chinese)

[5] 刘娟. 基于时间序列理论方法的流感病毒 DNA 序列特征分析[D]. 无锡:江南大学,2011.

[6] Jeffrey H J. Chaos game representation of gene structure[J]. **Nucleic Acids Research**,1990,18(8):2163-2170.

[7] 王燕. 应用时间序列分析[M]. 北京:中国人民大学出版社,2008.

[8] GAO Jie,XU Zhenyuan. Chao game representation (CGR)-walk model for DNA sequences [J]. **Chinese Physics B**,2009,18(11):370-376.