研究论文

甲型流感病毒 HA 蛋白质序列的预测

张 玲， 高 洁*

（江南大学 理学院，江苏 无锡 214122）

摘要：基于 CGR-游走模型和分数阶差分，用 ARFIMA 模型预测甲型流感病毒 HA 蛋白质序列。对所选取 1943~2013 年同源性相对较高的 71 条蛋白质序列，用 ARFIMA(p,d,q)模型对前 20 个位置去拟合并且预测，除极个别外由预报区域图显示原始数据都在预报区域内，表明模型建立的比较合理，预报效果很好。这对流感病毒的研究和预测有着重要的意义。

关键词：甲流；HA 蛋白质序列；预测；CGR 游走模型；ARFIMA(p,d,q)

中图分类号：Q 516　**文献标志码**：A　**文章编号**：1673—1689(2013)08—0828—04

Prediction for Bases of Influenza Virus A /HA Protein Sequence

ZHANG Ling, *GAO Jie**

(School of Science, Jiangnan University, Wuxi 214122, China)

Abstract: Based on the chaos game representation walk model and the integer-order difference, the purpose of this paper is prediction for bases of influenza virus A /HA protein sequence. For the 71 selected protein sequences with high homology from 1943 to 2013, we use ARIMA(p,d,q) model to fit and predict its former 20 positions, almost all the raw data are in the forecast region except a few, showing that this model is more reasonable, and prediction of the effect is very good. It's important significance for the study and prediction of influenza virus.

Keywords: influenza virus, HA protein sequence, prediction, CGR walk model, ARFIMA(p,d,q)

流感是一种传染病，它具有一定的周期性和反复性，其在全球引起的发病率和死亡率非常高[1-2]。流感病毒基因组由 8 个独立的 RNA 片段组成，分别编码多个与病毒结构和病毒复制有关的蛋白质分子。其中，最受关注的蛋白分子是血凝素(hemagglutinin，HA)。血凝素蛋白与病毒易感的宿主范围和宿主对病毒感染产生的免疫反应有直接联系[3]。且蛋白质是生命的物质基础，没有蛋白质就没有生命[4]。因此，许多学者和专家研究甲型流感病毒并取得了一些成就。

他们从不同的角度研究流感病毒序列，且在寻找着最佳模型。李晓燕等对新型 H1N1 流感病毒 HA 基因特性分析[5]。基于聚类算法 ModuleFind，梅娟等通过最大化蛋白质网络的模块性来寻找具有较强蛋白质结构的划分[6]。刘娟、任迪用时间序列的方法及整数阶差分和分数阶差分预测 DNA 序列[7-8]。

收稿日期：2012-06-20

基金项目：国家自然科学基金项目(11002061)；中央高校基础研究专项项目(JUSRP21117)。

* **通信作者**：高 洁(1972—)，女，江苏无锡人，工学博士，副教授，硕士研究生导师，主要从事应用统计学、生物信息学方面的研究。

E-mail：ezhun6669@sina.com

基于 CGR-游走模型,作者用分数阶差分的 ARFIMA 模型预测 HA 蛋白质序列。对所选取的 1943~2013 年同源性相对较高的 71 条蛋白质序列,我们用 ARFIMA(p,d,q)模型对前 20 个位置去拟合并且预测除极个别外由预报区域图显示原始数据都在预报区域内,表明模型建立的比较合理,预报效果很好,这对流感病毒的研究和预测有着重要的意义。

1 方 法

1.1 基于详细 HP 模型的蛋白质序列的 CGR-游走模型

CGR 是一种迭代映射技术[9],2004 年,喻祖国等人提出了基于详细 HP 模型的蛋白质序列的 CGR 方法[10]。在详细的 HP 模型中,将 20 种氨基酸分为四大类:非极性、负极性、无电荷极性和正极性,分别用 np、nep、up、pp 来表示,np=$\{A,I,L,M,F,P,W,V\}$,nep $=\{D,E\}$,up $=\{N,C,Q,G,S,T,Y\}$,$pp=\{R,H,K\}$。对于一个给定长为 n 的蛋白质序列 $S=s_1s_2Ks_n$,其中 s_i,i=1,2,…,n 是 20 种氨基酸中的一种,定义如下:

$$C_i=\begin{cases}A_0,s_i\in \text{np}\\A_1,s_i\in \text{nep}\\A_2,s_i\in \text{up}\\A_3,s_i\in \text{pp}\end{cases}$$

则得到一条序列 $X(s)=c_1c_2\cdots c_n$,其中 $c_i\in\{A_0,A_1,A_2,A_3\}$。

再定义序列 $X(s)$的 CGR,类似于 DNA 序列的 CGR 定义,通过迭代过程得到该序列 CGR $\text{CGR}_i=\text{CGR}_{i-1}-0.5\cdot(\text{CGR}_{i-1}-c_i)$,$i$=1,…,$n$,$\text{CGR}_0=(0.5,0.5)$。

对于一个蛋白质序列,定义 $t_k=y_k/x_k$,其中 y_k 是 CGR_k 的 y 坐标值,x_k 是 CGR_k 的 x 坐标值,则得到一个数据序列$\{t_k:k=1,2,\cdots,n\}$,把它作为一个时间序列,并称它为“CGR-游走序列”。

1.2 ARFIMA 模型

定义 1 $\{\varepsilon_t\}$称为白噪声(white noise)序列[11],简记为 $X_t\sim\text{WN}(\mu,\sigma^2)$。如果时间序列满足如下性质:

(1)任取 $t\in T$,有 $\text{EX}t=u$;

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

定义 2 如果随机序列$\{X_t\}$满足差分方程$(1-B)^dx_t=\varepsilon_t$,其中$-0.5<d<0.5$,$\{\varepsilon_t\}$为白噪声序列,$E\varepsilon_t=0$,$E\varepsilon_t^2=\sigma^2_\varepsilon<\infty$,称$\{X_t\}$服从$-0.5<d<0.5$ 的 ARFIMA$(0,d,0)$模型,也$\{X_t\}$为分数差分噪声[11]。

定义 3 如果随机过程$\{Xt\}$是平稳的,且满足差分方程 $\Phi(B)\nabla^dx_t=\Theta(B)\varepsilon_t$,其中$\{\varepsilon_t\}$为白噪声序列,$E\varepsilon_t=0$,$E\varepsilon_t^2=\sigma^2_\varepsilon<\infty$ $\Theta(B)=1-\phi_1B-K-\phi_pB^p$ 为 p 阶自回归系数多项式;$\Theta(B)=1-\theta_1B-K-\theta_qB^q$,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[4]。

2 数据分析

选取同源性相对较高的 71 条流感病毒 HA 蛋白质序列(1943~2013 年),数据来自 NCBI 网站,其网址:http:/ /www.ncbi.nlm.nih.gov /.

对于这 71 条蛋白质序列,选每一条序列的第三个位置为例来如下分析,首先把 HA 蛋白质序列转换成数值,定义:

$$f=\begin{cases}A_0\to 0\\A_1\to 1\\A_2\to 2\\A_3\to 3\end{cases}$$

得到相对应的 HA 蛋白质数值序列为:[0 0 0 1 0 0 3 0 0 0 3 0 2 2 0 3 2 0 3 0 0 0 0 2 3 3 0 0 2 2 2 0 3 3 0 0 0 0 0 3 2 2 2 0 3 0 0 0 0 3 0 0 2 0 0 0 0 0 3 0 0 3 2 0 2 0 3 0 3 3 3 2],再 CGR 混沌游走,经过计算得到分数阶差分 d=0.151,再取对数及 0.151 阶差分,可以得到一组新的数值序列:[0 0 0 0 −2.833 −2.833 −2.833 2.026 6 2.026 6 2.026 6 2.026 6 4.852 5 4.852 5 0.233 6 0.084 16 0.084 16 0.084 16 1.322 5 0.361 1 0.361 1 1.569 1.569 1.569 1.569 2 0.067 0 1.109 5 1.940 23 1.940 23 1.940 23 0.168 8 0.060 7 0.026 64 0.026 64 1.194 6 2.060 9 2.060 9 2.060 9 2.060 9 2.060 9 2.060 9 5.700 8 0.412 71 0.157 0.070 55 0.070 55 1.211 4 1.211 4 1.211 4 1.211 4 1.211 4 4.337 4.337 4.337 0.121 16 0.121 16 0.121 16 0.121 16 0.121 16 0.121 16 4.174 7 4.174 7 4.174 7 6.356 4 0.446 03 0.446 03 0.106 61 0.106 61 1.461 4 1.461 4 2.839 6 3.754 3 4.542 2],这 71 条序列的各个位置都要经过这样的预处理。

然后把这一组新的 HA 蛋白质序列转化成如图 1 所示的时序图 (CGR 游走和取对数 0.151 阶差

分),图中显示蛋白质序列在零上下波动,呈现出基本的平稳性。

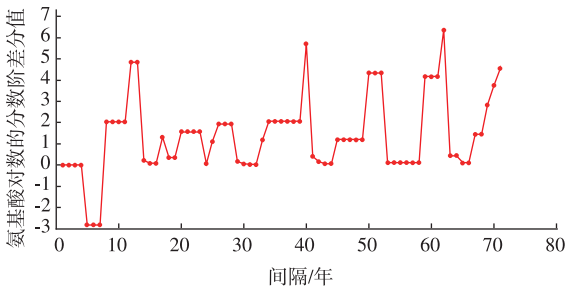

图 1　第三个位置对数差分后的时序图

Fig. 1　Time series figure of the third position sequence

图 2(ACF)和(PACF)为第三位置序列取对数再 0.151 阶差分后的自相关函数图和偏自相关函数图。序列的自相关图衰减迅速,而序列的偏自相关图衰减缓慢,这意味着原序列具有长记忆特征。

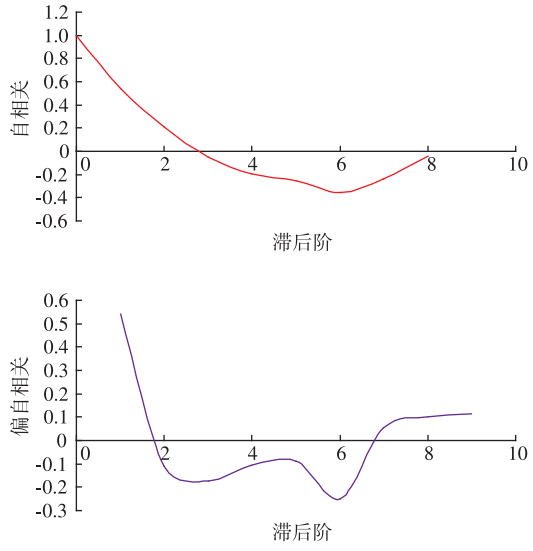

图 2　对数 0.151 阶差分序列的自相关函数图和偏自相关函数图

Fig. 2　Sample ACF and PACF of the 0.151 order differenced log data of protein sequences

我们以纯粹的随机测试的分数差分序列,结果见表 1。P<0.000 1<0.05,所以经取对数分数阶差分后此序列不是白噪声序列。因此,我们可以用 ARFIMA(p,d,q)模型拟合该序列。

根据 Akaike 信息判别准则[12],可选择 ARFIMA(4,0.151,0)模型来拟合此每条序列的第三个位置序列。表 2 给出了模型的参数估计。

表 1　白噪声检验

Table 1　Autocorrelation check for white noise

滞后阶数	χ^2 统计量	自由度	p 值
6	43.26	6	0.000 1
12	65.24	12	0.000 1
18	67.69	18	0.000 1
24	71.66	24	0.000 1

表 2　参数最小二乘估计

Table 2　Conditional least squares estimation

参数	估计值	标准误差	T 统计量值	P 值	滞后阶数
MU	1.396 32	0.281 28	4.96	<.0001	0
AR1,1	0.584 65	0.123 51	4.73	<.0001	1
AR1,2	−0.005 09	0.142 44	−0.04	0.071 6	2
AR1,3	−0.112 93	0.142 55	−0.79	0.031 1	3
AR1,4	−0.116 62	0.125 25	−0.93	0.055 2	4

从表中可以看出参数的 P 值都小于 0.1,这表明 ARFIMA(4,0.151,0)模型能有效地拟合该序列。为检验该模型的合理性,选择了一个合适的检验统计量 LB 检验统计量[13]:

$$LB=n(n+2)\sum_{k=1}^{M}\frac{r^2k}{n-k}\overset{appr}{\sim}\chi^2(M-p-q-1)$$

其中,r_k 是滞后的样本自相关函数,n 是样本容量,M 是一个取定的比 n 小的正整数。

表 3 显示了对于各滞后阶数,LB 统计量的 P 值均显著大于 0.1,意味着拟合模型的残差序列应为白噪声(纯随机),因而可认为 ARFIMA(4,0.151,0)模型能合理有效地拟合该序列。

表 3　残差的自相关检验

Table 3　Autocorrelation check of residuals

滞后阶数	χ^2 统计量	自由度	P 值
6	4.78	2	0.1915
12	11.21	8	0.1899
18	13.77	14	0.4666
24	18.58	20	0.5496

表 4 则利用 ARFIMA(4,0.151,0)模型对 HA 序列进行短期预测,即预测后十年 2014~2023 年第三个位置的预报值。(注:这些预报数据是混沌游走后取对数差分得到的,所以我们可以推出相应的数值。)

表 4 2014–2023 年的第三个位置的预报值

Table 4 Forecast value of the third position for 2014–2023 year

年份	预报值	标准误差	95% 的置信度	界限
2014	3.053 0	1.535 9	0.042 6	6.063 4
2015	1.914 3	1.779 2	−1.572 9	5.401 4
2016	1.060 5	1.852 8	−2.571 0	4.691 9
2017	0.443 4	1.847 0	−2.996 3	4.183 0
2018	0.706 1	1.868 9	−2.956 8	4.369 1
2019	0.974 2	1.890 8	−2.731 7	4.680 0
2020	1.277 2	1.906 1	−2.458 6	5.013 0
2021	1.494 6	1.910 5	−2.249 9	5.239 1
2022	1.582 5	1.910 6	−2.162 1	5.327 2
2023	1.567 4	1.911 3	−2.178 6	5.313 4

从图 3 可以看出,除极个别外由预报区域图显示原始数据都在预报区域内,表明该模型建立的比较合理,预报效果很好。

根据上面对蛋白质序列的第三个位置的选择模型、参数估计、LB 统计量检验,我们可以预测 2014–2023 年第三个位置的 HA 蛋白质序列的预报值为:3 2 1 0 1 1 1 1 2 2,进而可以预知 2014–2023 年第三个位置 HA 蛋白质是非极性、负极性、无电荷极性还是正极性。

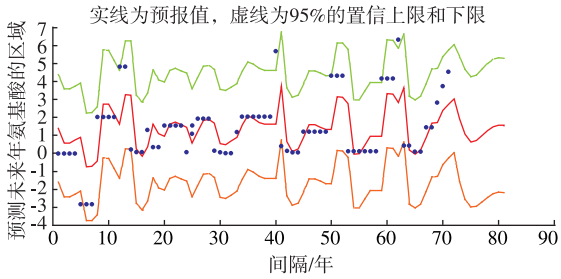

图 3 预报区域图

Fig. 3 Forecast figure

3 结 语

基于 CGR 混沌游走模型和分数阶差分模型,作者采取是一种纵向的预测方法,即用 ARFIMA(p,d,q) 模型预测未来年甲型流感病毒 HA 蛋白质序列。以 1943~2013 年这 71 条蛋白质序列的第三个位置为例,我们得到用 ARFIMA(p,d,q)模型对其前 20 个位置去拟合并且预测,发现除极个别外原始数据都在预报区域内除极个别外,表明模型建立的比较合理,预报效果很好。

我们可以用此方法分析和研究预测未来年的流感病毒蛋白质序列,这样节省大量的精力和财力。当然这种方法有一定的缺点,如选定的位置不足够多等,会影响预测值的精确性,还有待改进。

参考文献:

[1] Morens D,Folkers G,Fauci A. The challenge of emerging and re-emerging infectious diseases[J]. **Nature**,2004,430:242–249.
[2] Muzaffar S B,Ydenberg R C,Jones I L. Avian influenza:an ecological and evolutionary perspective for waterbirdd scientists[J]. **Waterbirds**,2006,29:243–257.
[3] Kobasa D,Takada A,Shinya K,et al. Enhanced virulence of influenza A viruses with the hae-magglutinin of the 1918 pandemic virus[J]. **Nature**,2004,431(7017):703–707.
[4] GAO Jie,XU Zhen-yuan. Chao game representation (CGR)-walk model for DNA sequences [J]. **Chinese Physics B**,2009,18(11):370–376.
[5] 李晓燕,孔梅,陈锦英,等. 新型 H1N1 流感病毒 HA 基因特性分析[J]. 中国卫生检验杂志,2010,20(12):3121–3124.
LI Xiao-yan,KONG Mei,CHEN Jin-ying. Analysis on the HA gene characteristics of novel influenza A (1H1N1) virus[J]. **Chinese Journal of Health Laboratory Technology**,2010,20(12):3121–3124.(in Chinese)
[6] 梅娟,何正,王正祥,等. 基于网络模块性的蛋白质序列聚类[J]. 食品与生物技术学报,2010,29(1):123–127.
MEI Juan,HE Zheng,WANG Zheng-xiang. Clustering protein sequences through modularity optimization [J]. **Journal of Food Science and Biotechnology**,2010,29(1):123–127.(in Chinese)
[7] 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)
[8] REN Di,GAO jie. Early-warning signals for an outbreak of the influenza pandemic[J]. **Chin Phys B**,2011,20(12):128701–4.
[9] Jeffrey H J. Chaos game representation of gene structure[J]. **Nucleic Acid Res**,1990,18:2163–2170.
[10] YU Zu-guo,Anh V V,Lau Ka-sing. Fractal analysis of measure representation of large proteins based on the detailed HP model [J]. **Physicas A**,2004,337:171–184.
[11] 王燕. 应用时间序列分析[M]. 北京:中国人民大学出版社,2008.
[12] Akaike H. A new look at statistical model identification[J]. **IEEE transaction on Automatics Control**,1974,19:9–14.
[13] Ljung G M,Box G E P. On a measure of lack of fit in time series models[J]. **Biomertrica**,1978,30A:9–14.