甲型流感病毒 DNA 序列的长记忆 ARFIMA 模型*

刘 娟 高 洁†

(江南大学理学院,无锡 214122)

(2010 年 4 月 16 日收到;2010 年 8 月 4 日收到修改稿)

流感病毒分为三类:甲型(A 型),乙型(B 型),丙型(C 型). 在这三种类型中甲型(A 型)流感病毒是最致命的流感病毒,对人类引起了严重疾病. 本文对甲型流感病毒 DNA 序列建立了一种新的时间序列模型,即 CGR(Chaos Game Representation)弧度序列. 利用 CGR 坐标将甲流病毒 DNA 序列转换成 CGR 弧度序列,且引入长记忆 ARFIMA 模型去拟合此类序列,发现随机找来的 10 条 H1N1 序列,10 条 H3N2 序列都具有长相关性且拟合很好,并且还发现这两种序列可以尝试用不同的 ARFIMA 模型去识别,其中 H1N1 可用 ARFIMA(0,d,5)模型去识别, H3N2 可用 ARFIMA(1,d,1)模型去识别.

关键词: 甲型流感, 时间序列模型, CGR, ARFIMA(p,d,q)模型

PACS: 87.10.Vg, 02.50.Fz

1. 引 言

流感是一种反复出现的传染病,在全球引起了高发病率和高死亡率[1]. 流感病毒分为三类:甲型(A 型),乙型(B 型),丙型(C 型). 在这三种类型中甲型(A 型)流感(以下简称甲流)病毒是最致命的流感病毒,给人类带来了严重的疾病. 甲流病毒根据其表面的血凝素(hemagglutinin, HA)和神经氨酸酶(neuraminidase, NA)基因的不同又可分成 16 个 HA 亚型(H1-H16)和 9 个 NA 亚型(N1—N9),不同的 HA 和 NA 形成了甲流病毒的许多亚型, 如 H1N1, H3N2, H5N1 等等[2—4]. 笔者参看了许多文献几乎没有看到用时间序列模型来挖掘甲型流感病毒的特性的,因而本文采用时间序列模型来分析甲型流感病毒.

1992 年,Peng 等[5]提出了 DNA 一维游走模型. 同年 Voss 等[6]提出了不同的观点,他们发现 DNA 序列的谱密度显示的 $1/f^{\beta}$ 噪声无处不在,意味着当 $0<\beta<1$ 存在长相关性,认为不仅在非编码区序列中在编码序列中也存在长相关性. 另一方面 Buldyrev 等[7,8]设计了一个广义-Lévy 游走模型去生成一个模型序列,使用所有可用的 DNA 序列发现主要在非编码序列中呈现长相关性. 基于该模型,Tai 等[9]提出了一个二维修正-Lévy 游走模型. 为区分 C 和 T, A 和 G, Lou 等建立了二维和三维游走模型[10],Yu 等则建立了图谱[11,12],来研究 DNA 序列的相关性. 2006 年, Lopes 和 Nunes[13]引入长记忆 ARFIMA(0,d,0)模型去拟合 DNA 序列的一维游走序列. 2009 年, Gao 等[14]基于 CGR (chaos game representation)坐标提出了一种 DNA 序列转换成一个时间序列(CGR-游走序列)的方法,并引入长记忆 ARFIMA(p,d,q)模型来分析.

本文对甲流病毒 DNA 序列提供了一种新的时间序列模型,即 CGR 弧度序列. 利用 CGR 坐标将甲流病毒 DNA 序列转换成 CGR 弧度序列,且引入长记忆 ARFIMA 模型去拟合此类序列,发现随机找来的 10 条 H1N1 序列,10 条 H3N2 序列都具有长相关性且拟合很好,并且还发现这两种序列可以尝试用不同的 ARFIMA 模型去识别, 其中 H1N1 可用 ARFIMA(0,d,5)模型去识别, H3N2 可用 ARFIMA(1,d,1)模型去识别.

2. ARFIMA 模型

如果随机过程 $\{x_t\}$ 是平稳的,且满足方程

* 江南大学创新团队发展计划(批准号:2008CX002)中央高校基本科研业务经费专项资金(批准号:JUSRP21117)资助的课题.

† 通讯联系人. E-mail:ezhun6669@sina.com

$\Phi(B)\ \nabla^d x_t = \Theta(B)\varepsilon_t$, 其中, $-0.5 < d < 0.5$, $\{\varepsilon_t\}$ 为白噪声序列, $E\varepsilon_t = 0, E\varepsilon_t^2 = \sigma_\varepsilon^2 < \infty$, $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^P$, 为 p 阶自回归系数多项式; $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$, 为 q 阶移动平均系数多项式.

则称 $\{x_t\}$ 服从 $-0.5 < d < 0.5$ 的 ARFIMA (p,d,q) 模型. 如果 $\Theta(B) \neq 0$, $|B| \leqslant 1$, 则随机序列满足

$$\nabla^d y_t = \varepsilon_t,$$
$$\Phi(B)x_t = \Theta(B)y_t,$$

因此, $\{x_t\}$ 可看作是分数差分噪声导出的 ARMA (p,q) 过程. 当 $2d - 1 = -1$ 时, $d = 0$ 即为短记忆过程; 所以当 $2d - 1 > -1$ 时, $d \in (0,0.5)$ 具有长记忆的特征.

3. 基于 CGR 的时间序列模型

1990 年 Jeffrey 提出了一种 DNA 序列可视化的方法即 CGR 方法[15]. CGR 是一种迭代映射技术, 它把序列中的每个单元, 如蛋白质序列中氨基酸, DNA 中的核苷酸, 映射到一个连续的坐标空间中去.

正方形的四个顶点对应四种核苷酸. 在这里, 用 DNA 序列代替随机数, 每一个碱基的坐标都可以来确定下一个碱基的位置. 我们取 A(0,0), T(1,0), G(1,1), C(0,1), 并且取点(0.5,0.5)为起始点.

下面给出 DNA 迭代函数, 也可以认为是 CGR 算法的公式化形式[15,16]. 对于一个序列 $S = s_1 s_2 \cdots s_N, s_i \in \{A,T,G,C\}$,

$$\mathrm{CGR}_i = \mathrm{CGR}_{i-1} - 0.5(\mathrm{CGR}_{i-1} - g_i),$$
$$i = 1,\cdots,N, \mathrm{CGR}_0 = (0.5,0.5)$$

其中 $g_i = \{(0,0),(1,0),(1,1),(0,1)\}$, g_i 和 s_i 相对应.

对于一个 DNA 序列, 定义

$$\mathrm{R}_n = \mathrm{arccot}(y_n/x_n),$$

其中 y_n 是 CGR_n 的 y 坐标值, x_n 是 CGR_n 的 x 坐标值. 则得到一个数据序列 $\{\mathrm{R}_n : n = 1,2,\cdots,N\}$, 我们把它作为一个时间序列, 并称它为"CGR 弧度序列".

以甲流病毒 H1N1 序列 CY056890 为例, 数据来自 NCBI 网站, 其网址: http://www.ncbi.nlm.nih.gov/.

它的 CGR 弧度序列"游走图"如下表 1.

表 1 CY056890 序列所选部分前 8 个 ACATGGTA 游走结果

				x	y	y/x	Arccot(y/x) (radians)
1	A	0	0	0.25	0.25	1	0.785397
2	C	0	1	0.125	0.625	5	0.197396
3	A	0	0	0.0625	0.3125	5	0.197396
4	T	1	0	0.53125	0.15625	0.294118	1.284745
5	G	1	1	0.765625	0.578125	0.755102	0.924038
6	G	1	1	0.882813	0.789063	0.893805	0.841414
7	T	1	0	0.941406	0.394531	0.419087	1.173945
8	A	0	0	0.470703	0.197266	0.419087	1.173945

4. 甲流 H1N1 型病毒 CY056890 的数据分析

图 1(a)是 CY056890 序列 CGR 弧度序列图(位置 380—2170), 样本容量 1791. 这些数据变动较大, 呈现非平稳特征. 考虑对此过程作 d 阶差分. 先对原序列作对数变换然后再做一阶差分结果如图 1(b)所示, 可见除少数地方呈现异方差外, 基本呈现平稳性.

图 2(a)(ACF)和图 2(b)(PACF)为样本取对数再一阶差分后的自相关函数图形和偏自相关函数图形. 可见 ACF 衰减迅速, 而 PACF 衰减缓慢, 这意味着原序列具有长记忆特征.

图 3 给出了方差图[17]是一个估计长记忆参数 d 的有用工具. 对于一个长记忆时间序列 $\{\mathrm{R}_n\}$, 它的均值 $\bar{R}_k$ 的方差满足

$$\mathrm{Var}(\bar{R}_k) \sim k^{2d-1},$$
$$\frac{\log[\mathrm{Var}(\bar{R}_k)]}{\log(k)} \sim 2d - 1,$$

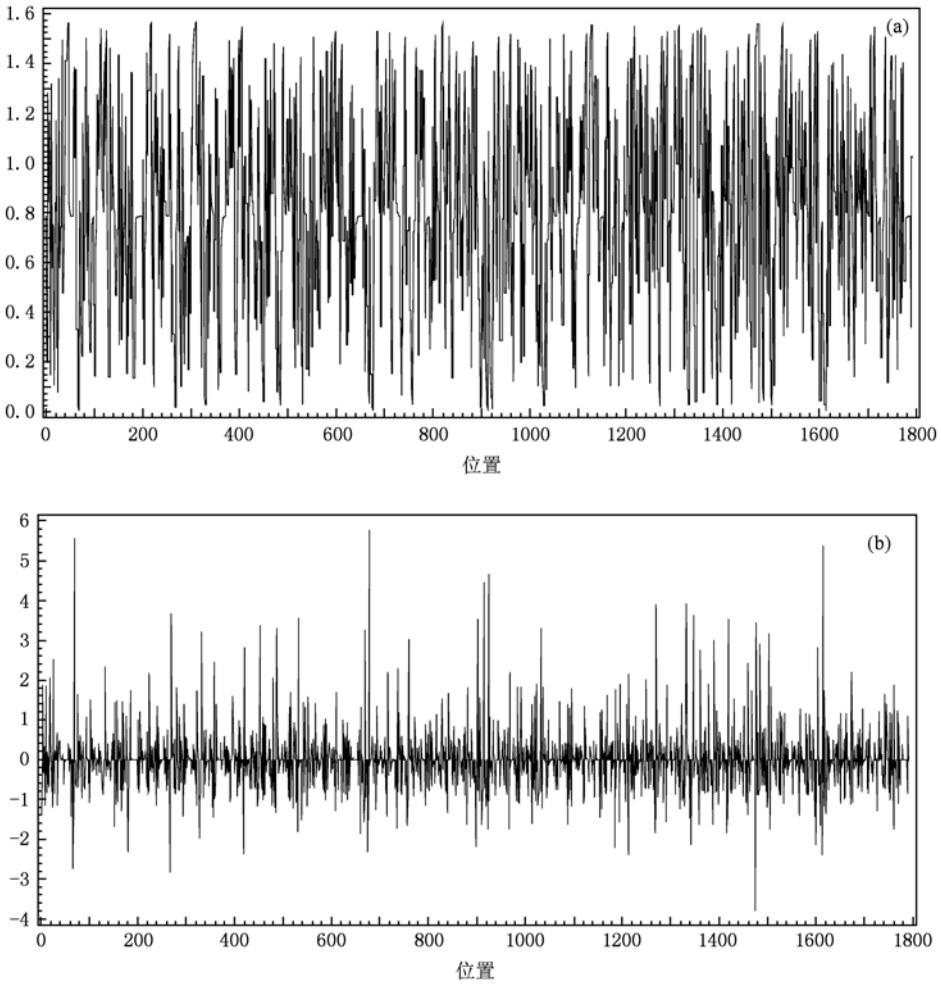

图 1 (a)甲流 H1N1 型病毒 CY056890 的弧度序列图;(b)取对数再一阶差分图

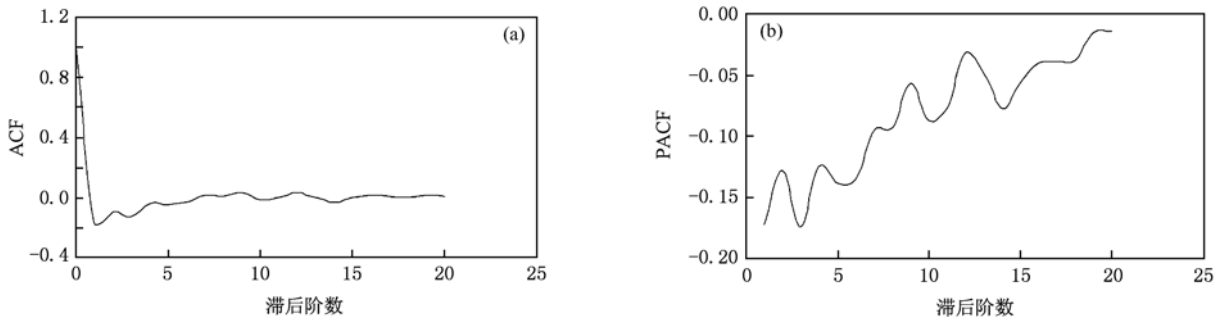

图 2 (a)取对数再一阶差分的样本自相关图;(b)取对数再一阶差分的样本偏自相关图

作 $\log[\mathrm{Var}(\bar{R}_k)]$ 关于 $\log(k)$ 的散点图,对散点图线性拟合,可估计得到线性方程的斜率为 -0.6877,令该斜率为 $2d-1=-0.6877$,即可得 d 的估计值 0.156.

根据上述理由我们可选择 CGR 弧度序列显示长记忆特征.目的是利用上述特点为序列建立一个合适的模型.因此,可以考虑长记忆 ARFIMA(p,d,q)模型($d\in(0,0.5)$),p,q 定阶时为考虑实用性,仅考虑 p,q 均小于等于 5 的 ARFIMA(p,d,q)模型.由 Akaike 信息判别准则[18,19],可选 ARFIMA(0,0.156,5)模型来拟合.

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[20,21]

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

其中 r_k 是滞后 k 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 2 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），因而可以认为 ARFIMA(0, 0.156,5)模型能很合理地拟合 CY056890 序列的 CGR-游走序列.

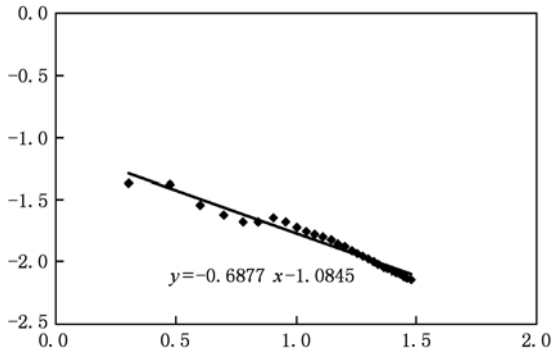

图 3 CY056890DNA 序列的 CGR 弧度序列方差图

表 2 残差的自相关检验

滞后阶数	χ^2 统计量	自由度	p 值
6	1.59	1	0.2071
12	6.80	7	0.4499
18	8.23	13	0.8286
24	12.61	19	0.8578
30	16.08	25	0.9123
36	20.07	31	0.9343
42	22.57	37	0.9702
48	28.46	43	0.9570

表 3 给出了被选择的 ARFIMA(0,0.156,5)模型的参数估计，5 个参数的 T 检验统计量的 p 值均显著小于 0.005. 这意味着 ARFIMA(0,0.156,5)模型能有效地拟合这个 CGR 弧度序列.

表 3 条件最小二乘估计

参数	估计值	标准误差	t 统计量值	p 值
MA1,1	0.37456	0.02358	15.88	<0.0001
MA1,2	0.23100	0.02512	9.20	<0.0001
MA1,3	0.21480	0.02519	8.53	<0.0001
MA1,4	0.08660	0.02513	3.45	0.0006
MA1,5	0.08686	0.02360	3.68	0.0002

5. 其余 9 条 H1N1 序列和 10 条 H3N2 序列数据分析

表 4 和表 5 分别给出了随机选的 9 条 H1N1 序列和 10 条 H3N2 序列的数据信息、被选择的 ARFIMA(p,d,q)模型及参数估计. 从计算结果可得 d 均位于(0,0.5)；对于各滞后阶数，LB 统计量的 p 值除极个别外其余均显著大于 0.1；且每个被选择的模型中各参数的 T 检验统计量的 p 值均显著小于 0.01. 所有这些结果都显示 ARFIMA(p,d,q)模型能很合理很有效地拟合这些不同的 CGR 弧度序列且还发现所选 H1N1 序列均为 ARFIMA(0,d,5)模型，所选 H3N2 序列均为 ARFIMA(1,d,1)模型. 所以我们可尝试用 ARFIMA(0,d,5)模型，ARFIMA(1,d,1)模型分别去识别 H1N1 序列，H3N2 序列.

表 4 9 条 H1N1 序列的数据信息、被选择的 ARFIMA 模型和参数估计

序列号	样本容量	位置	被选择模型	参数估计
GQ365655	1658	283—1940	ARFIMA(0,0.18,5)	$\theta_1=0.35808,\theta_2=0.22925,\theta_3=0.23525,\theta_4=0.08623,\theta_5=0.08505$
GQ329074	1601	1—1601	ARFIMA(0,0.13,5)	$\theta_1=0.36664,\theta_2=0.26738,\theta_3=0.16470,\theta_4=0.10600,\theta_5=0.08634$
CY057284	1536	626—2161	ARFIMA(0,0.443,5)	$\theta_1=0.32674,\theta_2=0.27829,\theta_3=0.15550,\theta_4=0.14634,\theta_5=0.08650$
GQ232082	1815	216—2030	ARFIMA(0,0.111,5)	$\theta_1=0.35973,\theta_2=0.22188,\theta_3=0.22919,\theta_4=0.08235,\theta_5=0.10038$
GQ402179	2025	136—2160	ARFIMA(0,0.379,5)	$\theta_1=0.36273,\theta_2=0.21926,\theta_3=0.23053,\theta_4=0.08217,\theta_5=0.09832$
HM014330	974	353—1326	ARFIMA(0,0.373,5)	$\theta_1=0.32700,\theta_2=0.24028,\theta_3=0.16637,\theta_4=0.11517,\theta_5=0.13136$
GQ265538	1085	1—1085	ARFIMA(0,0.290,5)	$\theta_1=0.35185,\theta_2=0.24899,\theta_3=0.14188,\theta_4=0.10144,\theta_5=0.14762$
HM006717	910	491—1400	ARFIMA(0,0.223,5)	$\theta_1=0.29892,\theta_2=0.25869,\theta_3=0.12455,\theta_4=0.14679,\theta_5=0.15932$
CY055628	1790	1—1790	ARFIMA(0,0.203,5)	$\theta_1=0.34917,\theta_2=0.20703,\theta_3=0.24961,\theta_4=0.05821,\theta_5=0.10855$

表 5 10 条 H3N2 序列的数据信息、被选择的 ARFIMA 模型和参数估计

序列号	样本容量	位置	被选择模型	参数估计
CY013107	1259	141—1399	ARFIMA(1,0.412,1)	$\theta_1 = 0.99009$, $\phi_1 = 0.56516$
EU097810	1254	147—1400	ARFIMA(1,0.321,1)	$\theta_1 = 0.99642$, $\phi_1 = 0.60669$
EU103940	1410	1—1410	ARFIMA(1,0.466,1)	$\theta_1 = 0.99378$, $\phi_1 = 0.72564$
EU103951	1295	74—1368	ARFIMA(1,0.235,1)	$\theta_1 = 0.99826$, $\phi_1 = 0.74827$
EU097802	1005	105—1109	ARFIMA(1,0.307,1)	$\theta_1 = 0.99209$, $\phi_1 = 0.63956$
EU103956	976	75—1050	ARFIMA(1,0.150,1)	$\theta_1 = 0.99787$, $\phi_1 = 0.76254$
CY013115	1113	218—1330	ARFIMA(1,0.366,1)	$\theta_1 = 0.98762$, $\phi_1 = 0.55766$
CY031566	1042	79—1120	ARFMA(1,0.316,1)	$\theta_1 = 0.99670$, $\phi_1 = 0.55635$
CY031568	1020	424—1443	ARFIMA(1,0.324,1)	$\theta_1 = 0.99624$, $\phi_1 = 0.62372$
CY031560	1049	492—1540	ARFIMA(1,0.235,1)	$\theta_1 = 0.99632$, $\phi_1 = 0.60412$

6. 结 论

本文基于 CGR 坐标提出了一种将甲流病毒 DNA 序列转换成时间序列(CGR 弧度序列)的方法,并引入长记忆模型 ARFIMA 模型来分析,首先分析了甲流 H1N1 型病毒 CY056890 序列,从图 1 到图 3 可知弧度序列显示长记忆特征,并选择了 ARFIMA(0,0.156,5)模型去拟合它,从表 2 到表 3 发现拟合合理有效.

然后又分析了随机找来的 19 条序列的 CGR 弧度序列,从表 4 和表 5 可知所有 ARFIMA(p,d,q)模型都有效合理. 并且从表 4 中还发现所选 H1N1 序列均为 ARFIMA(0,d,5)模型,表 5 所选 H3N2 序列均为 ARFIMA(1,d,1)模型.

由此可见,DNA 序列的 CGR 弧度序列能由长记忆 ARFIMA(p,d,q)模型有效合理地拟合,并且还可尝试用 ARFIMA(0,d,5)模型,ARFIMA(1,d,1)模型分别去识别 H1N1 序列、H3N2 序列. 作为具有完善算法的经典时间序列模型,不仅可以帮助我们得到甲流病毒 DNA 序列清晰的结构,而且还可帮助我们有效识别甲流中的两种亚型.

本文仅对甲流中的两种亚型进行了研究分析,后面我们将研究分析甲流中的其他亚型以及乙型丙型流感病毒.

[1] Morens D, Folkers G, Fauci A 2004 *Nature* **430** 242

[2] Chen J M, Sun Y X, Liu S 2009 *Chinese Science Bulletin* **54** 1657 (in Chinese)[陈继明、孙映雪、刘 朔 2009 科学通报 **54** 1657]

[3] Webster R G, Bean W J, Gorman O T 1992 *Microbiol Rev.* **56** 152

[4] Shi X M, Shi L, Zhang J F 2010 *Chin. Phys.* B **19** 038701

[5] Peng C K, Buldyrev S, Goldberg A L, Havlin S, Sciortino F, Simons M, Stanley H E 1992 *Nature* **356** 168

[6] Voss R F 1992 *Phys. Rev. Lett.* **68** 3805

[7] Buldyrev S V, Goldberger A L, Havlin S, Peng C K, Simon M, Stanley H E 1993 *Phys. Rev.* E **47** 4514

[8] Buldyrev S V, Goldberger A L, Havlin S, Mantegna R N, Matsa ME, Peng C K, Simon M, Stanley H E 1995 *Phys. Rev.* E **51** 5084

[9] Tai Y Y, Li P C, Tseng H C 2006 *Physica* A **369** 688

[10] Luo L F, Lee W J, Jia L J, Ji F M, Tsai L 1998 *Phys. Rev.* E **58** 861

[11] Yu Z G, Chen G Y 2000 *Theor. Phys.* **33** 673

[12] Yu Z G, Anh V, Gong Z M, Long S C 2002 *Chin. Phys.* **11** 1313

[13] Lopes S R C, Nunes M A 2006 *Physica* A **361** 569

[14] Gao J, Xu Z Y 2009 *Chin. Phys.* B **18** 370

[15] Jeffrey H J 1990 *Nucleic Acid Res* **18** 2163

[16] Almeida Jonas, carrico Joao A, Maretzek António 2001 *Bioinformatics* **17** 429

[17] Beran J 1994 Statistics for long-memory Processes (New Work: Chapman Hall)

[18] Hosking J R M 1984 *Water Resour. Res.* **20** 1898

[19] Crato N, Ray B K 1996 *Journal of Forcasting* **15** 107

[20] Ljung G M, Box G E P 1978 *Biometrika* **65** 297

[21] Li W K, Mcleod A I 1986 *Biometrika* **73** 217

Long-memory ARFIMA model for DNA sequences of influenza A virus*

Liu Juan Gao Jie†

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

(Received 16 April 2010; revised manuscript received 4 August 2010)

Abstract

Influenza viruses are divided into three types: A, B and C. Among them, type A virus is the most virulent human pathogen and causes the most severe disease. In this paper, we propose a new time series model for influenza A virus DNA sequence, i. e. chaos game representation (CGR) radians series. The CGR coordinates are converted into a time series model, and a long-memory ARFIMA(p,d,q) model is introduced to simulate the time series model. We select randomly 10 H1N1 sequences and 10 H3N2 sequences in analysis. we find in these data a remarkably long-range correlation and fit the model reasonably by ARFIMA(p,d,q) model, and also find that we can use different ARFIMA models to identify the two kinds of sequences, i. e. ARFIMA(0,d,5) model and ARFIMA(1,d,1) model that can identify H1N1 and H3N2 respectively.

Keywords: influenza A virus, time series model, chaos game representation(CGR), ARFIMA(p,d,q) model

PACS: 87.10.Vg, 02.50.Fz

* Project supported by the Innovative Research Team of Jiangnan University (Grant No. 2008CX002) the Foundamental Research Founds for the Central Universities (Grant No. JUSRP21117).

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