文章编号:1673-9590(2014)05-0116-05

基于扩展 Burrows – Wheeler 算法重构禽流感病毒的进化树

夏阳,白凤兰,刘立伟
(大连交通大学 理学院,辽宁 大连 116028)*

摘 要:基于核苷酸的物理化学性质,利用扩展 Burrows-Wheeler 算法和 Burrows-Wheeler 相似性分布,对80种 H5NI 病毒的 DNA 序列的 HA 片段进行相似性比较,同时构建序列进化树,得到较好的结果. 通过分析禽流感病毒之间的相似度关系,可为研究禽流感区域蔓延的特点提供一定的理论依据.

关键词: 扩展 Burrows-Wheeler 算法;禽流感病毒;进化树
文献标识码:A
DOI:10.13291/j.cnki.djdxac.2014.05.028

0 引言

禽流感病毒 H5N1 可以通过基因重组的方式适应新的各种自然环境和宿主. 这种重组可以是两种病毒基因间的重组,也可以是三种及以上病毒基因间的重组,其结果是新的重组病毒很快能适应新的各种自然环境和宿主,同时能替代先前流行的病毒而处于主导地位. 根据生物序列进化和生物亲缘关系基因序列测序以及序列进化树分析可知人类和动物的流感均来源于禽流感病毒[1]. 自然环境条件下进化率与物种种群分布密度有直接的关系,毒株可以在边缘地方普遍存在,并在抗体不足的地区容易暴发[2-3]. 随着病毒的不断重组和变异,众多国家极为重视新型抗流感药物的研发工作. 其中,揭示禽流感病毒的基因序列之间的进化关系对揭示流感病毒复制机制具有重要科学意义,也可为研究禽流感区域蔓延的特点提供一定的理论依据.

在国内外,许多学者用理论分析方法研究了禽流感病毒的进化关系. 如美国科学家 HolmesEC 等[4]利用拼接基因序列方法对人流感的 H3N2 病毒进行了全基因组的理论研究,比较好地反映了病毒 H3N2 的进化关系及其重组情况,提供了一个较好、较全面地研究禽流感病毒的进化关系的理论方法. 为了探索近年来禽流感病毒的编码序列进化和重组问题. 本文在利用 Burrows-Wheeler 方法时考虑了核苷酸的物理化学性质,从核苷酸 A、T、C、G 的化学结构入手,将它们分类,记 Y 为嘧啶,R 为嘌呤;即 Y = {C, T}和 R = {A, G};类似的可以将这四个核苷酸基分为酮基和氨基两类:即 M = {A, C}和 K = {G, T}[5]. 提取 Burrows-Wheeler 相似性分布,描述 H5N1 病毒的 DNA 序列,得到一种新的分析比较 DNA 序列方法. 将通过此方法比较80种 H5N1 病毒的 DNA 序列的 HA 片段的相似性,得到相似性矩阵,同时构建 DNA 序列进化树,比较全面地分析禽流感 H5N1 病毒的 DNA 序列之间的进化关系.

1 基本原理

1.1 Burrows-Wheeler 算法基本原理

Burrows-Wheeler 算法,应用于数据压缩技术中,也可称作块排序压缩,简称 BWT. 1994 年,在利福尼亚州帕洛阿尔托的 DEC 系统研究中心,Michael Burrows 和 David Wheeler 发明了该算法.

利用 Burrows-Wheeler 算法把某一字符串转

* 收稿日期:2013-10-29
基金项目:国家自然科学基金资助项目(61003139)
作者简介:夏阳(1986 –),女,硕士研究生;
白凤兰(1963 –),女,教授,博士,主要从事生物分子信息学的研究
E-mail:xy8288@126.com.

换时,不改变该字符串本身,而只改变这个字符串的先后顺序. 如果该字符串中有子串多次出现,则通过 Burrows-Wheeler 方法转换过的字符串上就会有一些字符连续重复,这一点在压缩中起到很大作用. 通过该算法的使用,处理字符串中连续重复字符的技术(如 MTF 变换和游程编码)的编码会更容易被压缩[6].

1.2　扩展 Burrows-Wheeler 算法基本原理

设 D 为一个有序字母表,在 D 上定义一个有限字符序列(有时也称为单词) $u = d_1 d_2 \cdots d_n$. 全部定义在 D 上的序列构成一个集合,记为 D^*. 设 $a, b \in D^*$,如果存在 $m, n \in D^*$ 使得 $a = mn$ 以及 $b = mn$,则称 a 是 b 的一个循环位移,也可称 a 与 b 是共轭的,记为 $a \sim b$. 若有 $n \in D^*$,且 $n = m^k \Rightarrow n = m, k = 1$,则称 n 为素词. 因为 D 是一个有序的字母集合,所以在 D^* 中任意两个不同的单词都可以按一定的顺序比较先后,我们称 D^* 为全序集合.

设 $m \in D^*$,令 $m^u = mmmm\cdots$,则 m^u 是一个定义在 D 上的无限词. 为判断两个无限词的序关系,可采用字典顺序. 假设给定两个无限词 $\alpha = \alpha_1\alpha_2\alpha_3\cdots$ 以及 $\beta = \beta_1\beta_2\beta_3\cdots$, $\alpha <_{lex} \beta$ 意味着存在 j 使得 $\alpha_i = \beta_i, i = 1, 2, \cdots, j-1$, 且有 $\alpha_j = \beta_j$.

取定两个素词 $p, q \in D^*$,其中 $q \notin C_p$,利用 Burrows-Wheeler 变换,分别得到 p, q 对应的共轭等价类 C_p 和 C_q. 易知, $C_p \cap C_q = \varphi$. 令 $S_{p,q} = C_p \cup C_q$. $S_{p,q}$ 按照 $< \omega$ 序关系构成一个全序集. 用 0 和 1 分别标记 $S_{p,q}$ 中属于 C_p 和 C_q 的元素.

将 DNA 序列看成字母集 $D = \{A, C, G, T\}$ 上的一个词.

例如,令 $p = AATGGTACC, q = GAATCGGAT$,则

$C_p = \{AATGGTACC, ATGGTACCA, TGGTACCAA,$
$GGTACCAAT, GTACCAATG, TACCAATGG,$
$ACCAATGGT, CCAATGGTA, CAATGGTAC\}$

$C_q = \{GAATCGGAT, AATCGGATG, ATCGGATGA,$
$TCGGATGAA, CGGATGAAT, GGATGAATC,$
$GATGAATCG, ATGAATCGG, TGAATCGGA\}$

$S_{p,q} = \{AATCGGTAC, AATGGTACC, ACCAATGGT,$
$ATCGGATGA, ATGAATCGG, ATGGTACCA,$
$CAATGGTAC, CCAATGGTA, CGGATGAAT,$
$GAATCGGAT, GATGAATCG, GGATGAATC,$
$GGTACCAAT, GTACCAATG, TACCAATGG,$
$TCGGATGAA, TGAATCGGA, TGGTACCAA\}$

$char(p, q) = 100110001111000110$

1.3　BurrowsWheeler 相似性分布

记 $char(p, q) = 0^{k_1}1^{k_2}0^{k_3}1^{k_4}\cdots 0^{k_m}1^{k_{m+1}}$,且当 $i \neq 1, m+1$ 时 $k_i > 0$ 记 $t_n = \#\{i \mid k_i = n\}$, $s = t_1 + t_2 + \cdots$.

注意,这里仅有有限个 $t_n \neq 0$.

序列 p, q 的 Burrow-Wheeler 相似性分布定义为: $P\{S_{pq} = k\} = \frac{t_k}{s}, k = 1, 2, 3, \cdots$

例如, $p = AATGGTACC, q = GAATCGGAT$,则 $char(p, q) = 100110001111000110 = 1^1 0^2 1^2 0^3 1^4 0^3 1^2 0^1$; $t_1 = 2, t_2 = 3, t_3 = 2, t_4 = 1$; $s = 8$.

$$P\{S_{pq} = 1\} = \frac{1}{4}, P\{S_{pq} = 2\} = \frac{3}{8},$$

$$P\{S_{pq} = 3\} = \frac{1}{4}, P\{S_{pq} = 4\} = \frac{1}{8},$$

$$P\{S_{pq} = k\} = 0, k > 4.$$

利用 BWSD 的数学期望定义序列 p 与 q 之间的距离为:

$$D_M(p, q) = E(S_{p,q}) - 1 = \sum_{k \geqslant 1} k \cdot \frac{t_k}{s} - 1 \qquad (1)$$

$D_M(p, q)$ 具有如下性质:

(1) $D_M(p, q) \geqslant 0$,当且仅当 $p = q$ 时, $D_M(p, q) = 0$;

(2) $D_M(p, q) = D_M(q, p)$;

(3) $D_M(p, q) \leqslant D_M(p, w) + D_M(w, q)$.

规定: $D_M(p, q)$ 越小序列 p, q 越相似,否则越不相似.

2　构建进化树

2.1　80 种 H5NI 禽流感病毒的相似性分析

选取 80 种 H5NI 禽流感病毒的 DNA 序列作为研究数据. 这些病毒的 DNA 序列都是近年来在东南亚国家出现的禽流感病毒亚型的编码序列,其中,一部分是从人体分离出来的病毒序列. 为了让这些禽流感病毒具有普遍性和适应性,在选取数据时,其分别提取自不同宿主体内、不同时间出现以及来自不同地区的病毒,并从 NCBI 数据库中下载这些病毒的基因序列.

利用式(1)计算了 80 种 H5NI 禽流感病毒的 DNA 序列的 HA 片段之间的相似性距离: $D_M(p, q)$,分别得到按照 YR 和 MK 碱基顺序排序的距

离矩阵 D_{M-YR}，D_{M-MK}. 由于矩阵维数过大，这里我们不给出了.

利用本文的方法来比较序列的相似性，其特点是不直接比对序列的碱基，而是先把碱基序列转化成数字序列，然后再去考虑其序列之间的 Burrows-Wheeler 相似性分布. 这样两个序列的比较就转换成了序列对应的数字序列的比较，最后利用数字序列的不变量来量化序列之间的差异程度，所以计算非常简单、快速，不足之处就是在利用不变量来比较序列相似性时会丢失一些序列结构方面的信息.

从距离矩阵 D_{M-YR}，D_{M-MK} 来看序列的相似性具有区域特性，如韩国地区的鸡鸭鹅的 H5N1 病毒之间很相似，广东、香港和广西地区的相似，印尼地区的相似，泰国和越南地区的 H5N1 病毒之间也很相似，而且发生的年份也比较接近. 通过对 77 与 75 和 70 与 69 的比较分析，我们推断从泰国人体上分离出的病毒序列与从香港和广东人体上分离出来的病毒序列具有很高的相似性.

2.2 对 80 种 H5NI 禽流感病毒序列构建进化树

将距离矩阵 D_{M-YR}，D_{M-MK} 输入 MEGA 5.1 软件中，输出 80 种 H5NI 禽流感病毒 DNA 序列的 HA 片段的进化树，见图 1、2.

观察图 1、2 可知，有一定的时间和地域特征，1996 年一枝，2002 年一枝，2003 ~ 2005 年一枝，2006 年一枝，2007 ~ 2008 年一枝. 韩国地区的鸭和鸡的 H5N1 病毒在一个分支内，香港、广东和汕头地区的鹅、鸭和鸡的 H5N1 病毒在一个分支内，印尼地区的，泰国和越南地区的，福建和日本地区的鹅、鸭和鸡的 H5N1 病毒在一个分支内分别属于不同分支，所以在 H5N1 的进化过程中地理因素起到了很重要的作用. 2004 ~ 2007 年期间 H5N1 病毒引起爆发的中心地区是香港，观察进化树 1、2，可知 H5N1 病毒宿主中的中心宿主是鸡和鸭.

观察进化树 1、2 可知，H5N1 病毒序列之间的进化关系与病毒的宿主有着密不可分的关系，即来自不同时间、不同地区，但它们因为有着相同的宿主而在病毒序列进化树中形成各自的分支. 例如，(31,41)，(24,32)(30,60)，(57,8)，(46,71)，(55,75)，(51,47,76,52,7)，(30,60)，(68,26,25)，(11,10,37,49). 可以看到，同一时间和

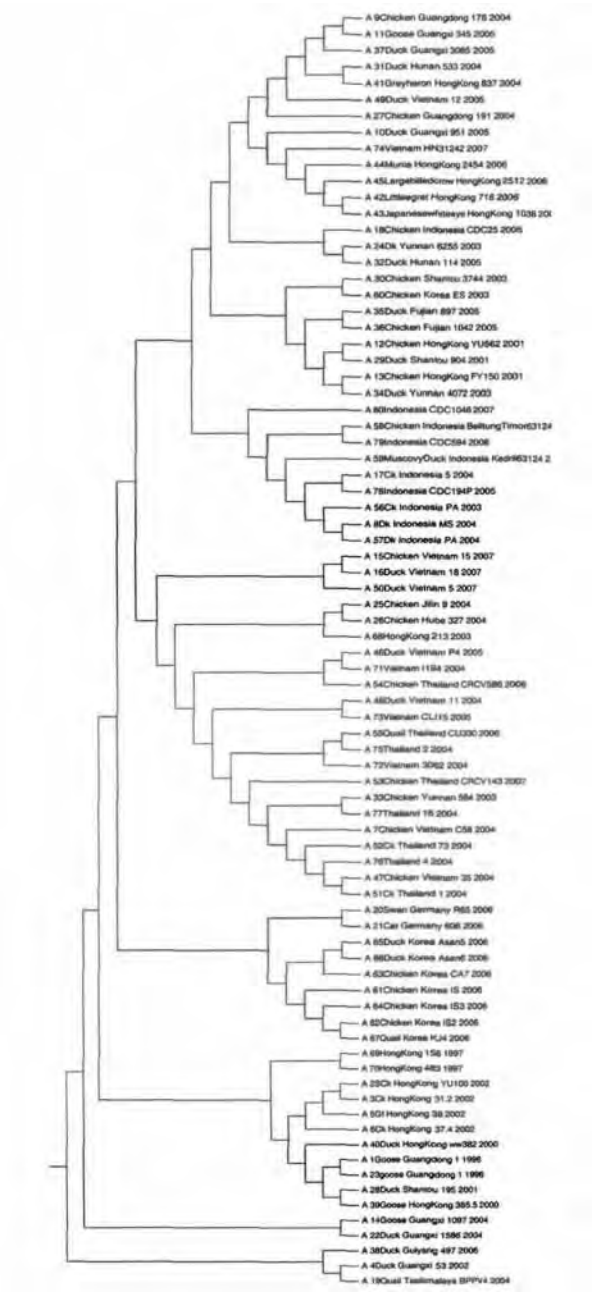

图 1 按照 YR 的碱基顺序构建的进化树

同一地区的病毒，尽管宿主不同，但它们也很相似. 如，(2,3,5,6)，(69,70)，(67,62,64,61)，(65,66)，(20,21)，(15,16,50) 等. 另外，不同地区、不同时间从白鹭、文鸟、乌鸦体内分离出来的病毒也很相似，从两个图中的小分支 (20,21) 推断天鹅和猫分离出来的 H5N1 病毒很相似，这和文献[10]结果一致.

从进化树 1、2 的小分支 (18,78)，(6,5)，(80,58,59)，(25,68,33,72)，(73,48)，(77,75,33)，(69,70)，(76,53,55,54)，(50,15,16,80)，(42,43,44,45,74) 中发现人感染的 H5N1 病毒片

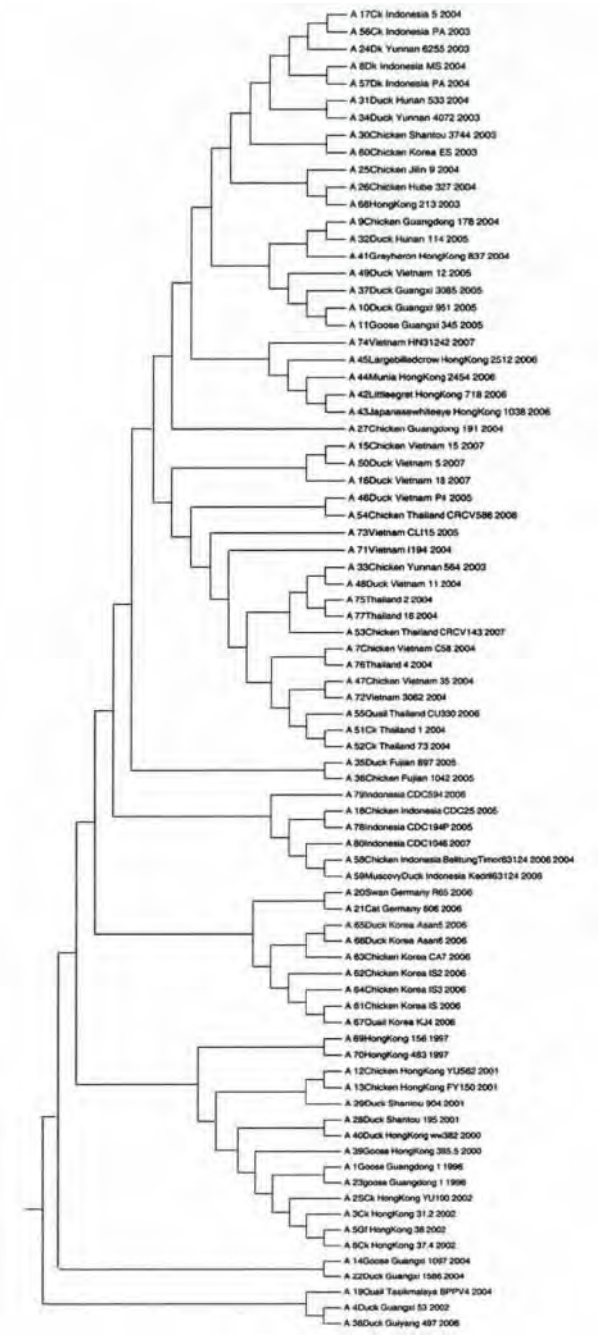

图 2　按照 MK 的碱基顺序构建的进化树

段 HA 与鸡和鸭感染的 H5N1 病毒片段 HA 非常相似. 所以推断人类感染的 H5N1 病毒可能是由于直接或间接接触禽类或家禽产品导致[7].

基于核苷酸的物理化学性质序列排列顺序的不同得到的三个进化树有不同之处,比较而言图 2 的结果比较符合已有的结果[9],说明核苷酸按酮基和氨基的顺序来提取特征来描述序列更接近事实,从而减少序列的生物信息丢失.

总的来说,本文构建的 H5N1 病毒片段 HA 序列的进化树中发现了病毒的地域蔓延和宿主的关键性作用,这和文献[10]结论一致. 说明本文对 H5N1 病毒片段 HA 序列的相似性分析是有效的并具有可行性.

3　结论

本文利用 Burrows-Wheeler 方法时考虑了核苷酸的物理化学性质,再根据 Burrows-Wheeler 相似性分布,得到一种新的分析生物序列方法,相对将碱基单纯的按照字典顺序进行分析比较,该算法可以得到相对高精度的相似性度量,说明核苷酸的物理化学性质对序列进化分析有不可忽视的作用. 通过此方法对 80 种 H5NI 病毒的 DNA 序列的 HA 片段进行相似性比较,同时构建序列进化树,并对其进行了详细分析,验证了我们方法的有效性和可行性,而且本文的方法优点是避免序列比对操作,时间复杂度不高. 另外,本方法对序列的长度没有任何规定. 对以后研究禽流感病毒之间的进化关系及研发禽流感病毒疫苗等具有一定的应用价值. 在碱基序列转化为数学对象的过程中,或多或少丢失了一些结构方面和生物信息,所以有可能不能真实的反映出禽流感病毒之间的进化关系. 这是本文的不足之处. 我们将在以后的工作中从多层次、多角度深入研究禽流感病毒.

参考文献:

[1] MEIJER A, WILBRINK M, DU RY VAN BH, et al. Highly pathogeni c avian inf luenza virus A (H7N7) infect ion of humans and human to human transmission during avian influenza outbreak in the Netherlands [C]. In: Kawaoka Y, ed. The Options for the Control of Influenza V. New York: Elsevier, 2004: 65-68.

[2] BURROWS M, WHEELER D J. A block-sorting lossless data compression algorithm [R]. Digital SRC Research Report, 1994.

[3] MANTACI S, RESTIVO A, ROSONE G, et al. An extension of the Burrows-Wheeler Transform [J]. Theoretical Computer Science, 2007, 387(3): 298-312.

[4] HOLMES EC, GHEDIN E, MILLER N, TAYLOR J, BAO Y. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses [J]. PLoS Biol, 2005, 3(9): 300.

[5] ZHANG R, ZHANG C T. Z-Curve, an intuitive tool for visualizing and analyzing the DNA sequences [J]. J. Biomol. str. Dyn, 1994 (11): 767-782.

[6] 田峰,饶妮妮,程煜,徐尚蕾. 基于BW算法的高采样率心电数据无损压缩[J]. 生物医学工程学杂志,2008,25(4):790-794.
[7] The World Health Organization Global Influenza Program Surveillance Network, emerging Infectious Diseases Evolution of H5N1 Avian Influenza Viruses in Asia [J]. Emerging Infectious Diseases, 2005, 11:1515-1521.
[8] ROBERT G, WEBSTER, MALIK PEIRIS, et al. H5N1 Outbreaks and Enzootic Influenza [J]. Emerging Infectious Diseasea, 2006, 12:3-8.
[9] LIU YANQIU, ZHANG YUSEN. A New Method for Analyzing H5N1 Avian Influenza Virus [J]. Journal of Mathematical Chemistry, 2010, 47:1129-1144.
[10] ROHM C. Characterization of a Novel Influenza Hemagglutinin, H15: Criter for Determination of Influenza A Subtypes [J]. Journal of General Virology, 1996, 217: 508-516.

Avian Influenza Virus for Phylogenetic Tree Reconstrction Based on Extension Burrows-Wheeler

XIA Yang, BAI Fenglan, LIU Liwei
(School of Mathematics and Physics, Dalian Jiaotong University, Dalian 116028, China)

Abstract: Based on physical and chemical properties of nucleotides, Burrows-Wheeler method and Burrows-Wheeler distribution similarity were used to compare the similarity of 80 kinds of HA fragments of H5NI virus DNA sequence, and the phylogenetic tree was built to obtain better results. Through the analysis of similarity relationships of the avian influenza virus, certain theoretical evidence can be provided for studying the characteristics of the virus spreading in bird flu area.

Keywords: extension Burrows-Wheeler; avian influenza virus; phylogenetic tree

(上接第94页)

Speed Control System Based on the Continuous Extrusion and Rolling of Neural Network PID

LIU Fudong, QI Wei, YUN Xinbing
(Engineering Research Center of Continuous Extrusion, Ministry of Education, Dalian Jiaotong University, Dalian 116028, China)

Abstract: Based on the characteristics of continuous extrusion and rolling speed control, the improved BP neural network is combined with PID control algorithm, and Siemens S7-200 PLC is adopted to realize the tandem mill of motor frequency and speed controls. By comparison with traditional PID and BP neural network PID algorithm, the improved BP neural network PID controller has shorter dynamic response time and good following characteristics, which indicates that the method has a good application value in the crowded and rolling speed control.

Keywords: continuous extrusion and rolling; PID; improved BP neural; speed control