·实验研究·

基于基因序列聚类的甲型流行性感冒病毒 H3 抗原变异规律研究

张文彤 姜庆五 蒋露芳 居丽雯

【摘要】 **目的** 利用聚类分析方法探讨全球甲型流行性感冒病毒(流感)H3 亚型抗原的变异规律。**方法** 下载 NCBI GenBank 和流感病毒数据库中全部的甲型流感病毒 RNA4 节段 H3 亚型基因序列,在 ClustalX 中进行序列对齐后,使用两阶段聚类法进行分析,并随后探讨各类的三间分布。**结果** 所有序列可被分为 10 类,其中 7 类主要为人流感病毒,人流感病毒和鸟类、其他哺乳动物的流感病毒被明确的分为不同类别,但和猪流感病毒则共存于数个类中。各类呈现出明显的时间分布和宿主分布规律,但并未发现地域分布规律。**结论** 由于受到人类免疫系统的选择压力,H3 抗原呈现出5-7 年出现一次较大变异的流行特征,且这一趋势随着近十年来流感疫苗的普遍使用而呈现加速趋势。同时,猪流感病毒和人流感病毒出现在同一类别中,两者的遗传距离较近,这为猪作为病毒重配的载体提供了新的佐证。

【关键词】 流行性感冒病毒;两阶段聚类法;抗原变异;生物信息学

Application of gene sequence cluster in research for H3 antigenic evolution of influenza A virus *ZHANG Wen-tong, JIANG Qing-wu, JIANG Lu-fang, JU Li-wen. Department of Health Statistic, School of Public Health, Fudan University, Shanghai* 200032, China

【**Abstract**】 **Objective** Gene sequence data were clustered to explore evolution lineages of H3 antigen of influenza A virus. **Methods** All data of H3 RNA sequence in NCBI Genbank and Influenza sequence database were downloaded and aligned in ClustalX while two step cluster method were applied to explore the data. **Results** All sequences were aggregated into ten clusters, while seven of them mainly were human virus. Human virus and avain/other mammal virus were separated into different clusters distinctively, but coexisted into same clusters with swine virus. Time and host distribution were very distinctive in these clusters, but no geographic distribution features were found. **Conclusion** With the interaction of human immunity system, H3 antigen mutated significantly every 5-7 years, and the speed of mutation had accelerated with the application of influenza vaccines in recent years. Meanwhile, human and swine influenza virus were not separated distinctly between clusters indicting that they had short inheritance distance. Result showed again that swine served as the mixer for antigenic recombination of different influenza virus.

【**Key words**】 Influenza A virus; Two-step cluster; Antigenic evolution; Bioinformatics

流行性感冒(流感)是当今各国流行病学研究的重点疾病之一,对甲型流感病毒变异规律的探讨一直没有停止,但目前的研究方向多基于实验室研究和连续动态监测,较少有研究从基因全序列遗传距离的角度进行分析。随着生物信息学技术的逐渐成熟,利用基因序列进行病毒变异规律的分析已成为可能。普林斯顿大学的 Plotkin 等[1]于 2002 年采用聚类分析方法对部分人流感病毒 HA 序列数据进行了分析,为流感抗原变异规律的研究提供了新的思路。但是,Plotkin 的研究有明显的不足之处;有鉴于此,我们在其研究思路的基础上,基于流感病毒的抗原基因序列,使用两阶段聚类这种新的聚类方法对流感病毒的抗原变异规律进行研究,以期能对流感病毒的变异规律有更深入的了解。

材料与方法

1. 基因序列的获取:本次研究所使用的基因序列来源于 NCBI GenBank 数据库(http://www.ncbi.nlm.nih.gov/)和美国洛斯阿莫斯国家实验室的流感病毒数据库(www.flu.lanl.gov)中截止 2003 年 11 月 1 日时所包含的全部甲型流感病毒 H3 抗

基金项目:国家自然科学基金资助项目(30400370)

作者单位:200032 上海,复旦大学公共卫生学院卫生统计与社会医学教研室(张文彤),流行病学教研室(姜庆五、蒋露芳、居丽雯)

原序列，这两个数据库集中了全球可供使用的所有流感病毒基因序列数据，并可提供免费下载。共计获得 H3 基因序列 1154 条。

2. 序列的预处理：血凝素抗原经细胞蛋白酶裂解后可分为 HA1 和 HA2 两部分，前者在免疫反应和病毒的抗原变异中非常重要，且所获得的大多数序列也只包括 HA1 片段基因，长度在 987 个碱基左右。因此研究中将只使用 HA1 的基因序列进行分析。在全部 1154 条序列中，有 154 条因长度过短，明显属于序列残片而删除；剩余的序列首先在 ClustalX 1.83版中进行序列对齐，随后截取这些序列对应于 HA1 基因的部分，最终进入聚类分析的序列共计有 1000 条，长度均为 990 个碱基（对齐后序列中部有三个碱基的间隙）。

3. 统计学分析：由于基因序列数据全部是以字符的方式进行记录，序列对齐后只包含五种字符：A、G、C、T 和间隔符“-”，属于无序分类资料。而传统的聚类方法对这种类型资料的利用并不充分，且无法自动判断出适宜的类别数。为此我们引入近年来发展起来的两阶段聚类法[2]，对序列数据进行分析。该方法正如其名称所示，在计算中共分两个阶段，第一阶段对原始记录进行分析，建立类别特征树；第二阶段则利用特征树进行系统聚类，并确定适宜的类别数。为了能够同时处理连续性变量和分类变量，两阶段聚类法采用对数似然值作为距离测量指标，两个类 j 和 s 间的距离由他们合并后对数似然值的下降程度来表示：

$$d_{(j,s)} = \xi_j + \xi_s - \xi_{(j,s)}$$

其中

$$\xi_v = -N_v\left(\sum_{k=1}^{K^A}\frac{1}{2}\log(\hat{\sigma}_k^2+\hat{\sigma}_{vk}^2)+\sum_{k=1}^{K^B}\hat{E}_{vk}\right)$$

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k}\frac{N_{vkl}}{N_v}\log\frac{N_{vkl}}{N_v}$$

在上式中，K^A 是所使用的连续性变量个数，K^B 是分类变量个数，L_k 是第 k 个分类变量的类别数，N_v 是类别 v 的记录数，N_{vkl} 是属于类别 v 的记录中第 k 个分类变量取值为 l 类的个数，$\hat{\sigma}_k^2$ 则表示第 k 个连续性变量的方差估计值。由于对数似然值可以同时对连续性变量和分类变量进行处理，这样就解决了对任意类型的变量进行聚类分析的问题。序列对齐后数据的两阶段聚类分析和描述工作均在 SPSS 12.0中完成。

结　果

1. 适宜类别数的确定：表 1 给出了样本被聚为 1～15 类时贝叶斯信息准则（Bayes' Information Criterion，BIC）等相关统计指标的具体数值，由 BIC 值可见，当类别数等于 6 时 BIC 值达到最小，随后逐渐增大，10 以上逐渐稳定下来。从类间距离比看，5、6 类时均较高，随后在 10 类时又出现一个高峰，显然，6 类或者 10 类都是较合适的类别数。为便于详细观察，这里将类别数定为较多的 10 类。

表1 两阶段聚类分析结果

类别数	BIC 值	BIC 改变量	BIC 改变率	最小类间距离比
1	302 791.949	-	-	-
2	226 234.418	-76 557.531	1.000	1.699
3	184 951.929	-41 282.490	0.539	2.080
4	169 883.775	-15 068.154	0.197	1.339
5	160 959.286	-8 924.488	0.117	1.884
6	160 539.210	-420.076	0.005	1.431
7	163 017.083	2 477.873	-0.032	1.130
8	166 268.819	3 251.736	-0.042	1.259
9	170 745.761	4 476.942	-0.058	1.059
10	175 484.157	4 738.397	-0.062	1.608
11	181 910.776	6 426.618	-0.084	1.068
12	188 514.340	6 603.564	-0.086	1.079
13	195 308.059	6 793.719	-0.089	1.053
14	202 222.508	6 914.449	-0.090	1.189
15	209 501.188	7 278.681	-0.095	1.001

2. 病毒类别的时间分布：表 2 给出的是全部序列的时间分布情况，可见 1980 年以前的序列数较少，大多数序列的采集时间集中在1985－1999 年间，因此在该时段内的聚类结果应当是较为准确的。为便于观察，这里使用箱式图作为各组序列年代变异规律的观察工具（图 1）。箱式图中间的粗线代表中位数，方框代表四分位数间距，两端的细线代表除去离群值外的全距，全距外的圈、点则代表离群值。从图中可见，这几类流感病毒的流行时间存在着明显的差异，大多数类的重叠时间也比较短，呈现出明显的流行株更替特征。相对而言，一直到 20 世纪 90 年代初，各类的流行时间还在5－10 年左右，而从 90 年代起，相应的 6、7、8、10 四类的流行时间跨度均在2－3 年左右，时间大大缩短。如果考虑分为 6 类的情形，则 10 类中 2、3、4 类的箱图合并，5、9 的合并，可见主要影响的是 1990 年前的分类情形，相应的时间分布特征并无大的变化。

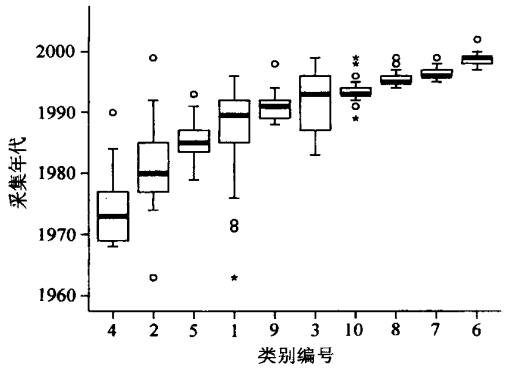

图1 各类别时间分布的箱式图

表2 样本序列的时间分布

年 代	频数	百分比(%)	累计百分比(%)
1960－1964	4	0.4	0.4
1965－1969	13	1.3	1.7
1970－1974	15	1.5	3.2
1975－1979	37	3.7	6.9
1980－1984	41	4.1	11.0
1985－1989	118	11.8	22.8
1990－1994	250	25.0	47.8
1995－1999	500	50.0	97.8
2000－	22	2.2	100.0
合 计	1000	100.0	－

3.病毒类别的宿主分布:从表 3 可见,这 10 类病毒有较明显的宿主分布特征,第一类的宿主完全是其他哺乳动物(马、海豹等),第二类则基本上以鸟类(鸭、水鸟等)为主。值得指出的是在3～9 类中,均出现了人和猪流感病毒被分在同一类中的情况,以第 4、8 两类最为明显。这说明对于人、猪流感病毒而言,宿主类型并未成为 H3 序列差异的重要特征,序列自身的变异要大于宿主种类间的变异,即有可能这两种不同宿主流感病毒的 H3 序列并无本质区别。

4.病毒类别的空间分布:表 4 为 10 类病毒在全球各地区的分布状况,可见除第 2、3 类因样本量较少而主要分布在中国和欧洲地区外,其余类别在各地均有出现,并无特别明显的地域分布规律。

讨 论

1.与经典病毒学分析结果的比较:本研究完全是用生物信息学方法对核酸进行序列分析,但得到的结论和经典病毒学方法并不矛盾,其结论还可以互为补充。甲型流感病毒的变异是引起新的流感大流行的主要因素,除因抗原漂移的点突变积累而引起抗原位点结构改变,从而不被机体特异性免疫识别外,多数学者认为由人和动物的流感病毒基因片段重组所引起的抗原转换也是变异的重要来源[3-6]。但是,由于流感病毒具有严格的宿主特异性,抗原如何进行重组一直是令人感兴趣的问题。自血清学追溯研究证实是猪型(H1N1)流感病毒导致了 1918 年的流感大流行以来,猪已被重点怀疑可能是不同宿主流感病毒的混合器。人流感病毒传染给猪目前已为大家所公认,近年来也已发现一些禽流

表3 各病毒类别的宿主分布

宿主种类	类别编号										合计
	1	2	3	4	5	6	7	8	9	10	
人类	0	0	3	34	89	249	141	77	102	165	860
鸟类	0	20	0	0	0	0	0	0	0	0	20
猪	0	2	19	15	2	5	8	10	2	0	63
其他哺乳动物	54	3	0	0	0	0	0	0	0	0	57
合 计	54	25	22	49	91	254	149	87	104	165	1000

表4 各病毒类别的地域分布

序列采集地点	类别编号										合计
	1	2	3	4	5	6	7	8	9	10	
无法确定	19	0	4	11	12	37	4	9	10	16	122
中国	2	11	1	17	14	26	30	9	28	33	171
北美	18	5	0	11	29	29	32	29	13	43	209
欧洲	8	1	17	7	8	91	32	15	21	51	251
亚洲(除中国外)	1	8	0	2	21	16	21	13	24	16	122
拉美	3	0	0	0	1	12	15	9	4	3	47
大洋洲	0	0	0	1	6	36	12	0	4	2	61
非洲	3	0	0	0	0	7	3	3	0	1	17
合 计	54	25	22	49	91	254	149	87	104	165	1000

感病毒也能传给猪[4]。在本研究的聚类结果中可以看到，人、鸟类和其他哺乳动物的 H3 序列差异较大，在类别上基本没有交叉，而猪和人的 H3 序列则被多次分入同一类中，有明显的交叉；且猪流感的 H3 序列在第 2 类中也和鸟、其他哺乳动物的病毒序列存在交叉。由于在各类内部，基因序列间的差异要小于不同类间的差异。这说明猪流感病毒和人流感病毒的 H3 抗原可能存在相当接近的亲缘关系，同时也可能与禽类和其他动物的流感病毒关系较近，该结果显然和血清学研究结果相当一致，共同支持了不同宿主的流感病毒可能在猪体内进行重配的假说。

2. 流行时间的变短原因：新流行株的出现一直是流感监测的重点内容[7]，在分析所得的 10 类中，1～3 类的宿主主要是其他生物，从箱式图可见，这三类的流行时间明显要长于另外 7 类，其四分位间距均接近 10 年；而其余 7 类的流行时间明显短于前者，除时间最早的第 4 类外，其四分位间距一般不超过 5 年。特别是在 1990 年以后的四类，可以看到其四分位间距基本上在 2 年左右。时间的分布规律充分体现出人流感病毒 H3 抗原的变异速度不仅要高于其他宿主的流感病毒，其速度在近 10 年还有加速的趋势。对此的解释是由于 H3 亚型在禽类、猪体内往往导致隐性感染，宿主的免疫反应较弱，而在人体主要引起显性感染，免疫反应较强，在人体免疫系统免疫选择压力的作用下，就会表现为流行株在数年后即自动消失，被新的流行株所替代。而近年来随着 WHO 推荐疫苗和各国卫生机构推荐疫苗的逐渐普及，新流行株遇到的免疫压力越来越大，很快就会因免疫屏障的形成而终止流行，这造成了相应毒株的流行时段越来越短。事实上，如果将各类别中毒株的名称和 WHO 所推荐的流感疫苗制备株名称相对应，就会发现两者正好完全重合，各类别所在的时间短正好就是相应毒株被推荐的时间，这也充分说明了疫苗使用和流行时间的缩短是有关联的。

3. 新聚类方法的作用：本次研究采用了一种新的聚类方法，除了能够充分利用信息，并自动判断最适宜的类别数外，其分析结果和 Plotkin 的结果出现了较大差异，最大的区别在于全部样本被较好的分入了几个大类，未出现许多小类。而在 Plotkin 的结果中，560 条序列共被聚成了 174 类，只有前 9 类的样本量大于 10。事实上，系统聚类法在聚类变量极多时往往会出现这一问题，容易将样本分出许多小类。本例中共使用了 990 个碱基位点，相当于聚类中有 990 个变量，因此这些小类的出现很有可能是方法不当所致的假象。另一方面，Plotkin 对数据的处理方式是将不相同的碱基记为距离 1，相同记为距离 0，这无疑在简化分析的同时丢弃了大量信息。由于聚类分析没有较好的效果判定方法，主要是看对结果的解释是否合理。综合判断之下，本次分析的结果应当更为合理。

参 考 文 献

1 Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. PNAS, 2002, 99: 6263.

2 SPSS Base 12 User's Guide. SPSS Inc. Chicago, Illinois, 2003. 391-399.

3 杨春，余佳. 甲型流感病毒抗原变异的机理. 国外医学病毒学分册，1999，6(3)：88.

4 闻玉梅，主编. 现代医学微生物学. 上海：上海医科大学出版社，1999. 1005-1020.

5 顾惠心，胡善联，杨志刚，等. 综合养鱼与人类甲型流感流行的关系. 中华流行病学杂志，1996，17：29-32.

6 李玉青，潘浩. 禽畜养殖与甲型流感抗体水平关系研究. 中国公共卫生，2003，19：45.

7 Bush RM, Bender CA, Subbarao K, et al. Predicting the evolution of human influenza A. Science, 1999, 286: 1921.

（收稿日期：2003-12-16）

（本文编辑：尹廉）