·实验研究·

H3N2 亚型人甲型流感病毒 HA1 序列进化正向选择位点研究

许慧琳 张文彤 赵耐青 姜庆五

【摘要】 **目的** 筛选 H3N2 亚型人甲型流感病毒红细胞凝聚素(HA1)序列中的正向选择位点，为进一步揭示流感病毒的变异规律。**方法** 从 NCBI GenBank 和流感数据库中下载甲型流感病毒 H3 亚型 RNA4 节段序列，用两步聚类法将全部序列分为 6 个年代类别，使用固定效应似然比检验模型分别筛选每一类中的正向选择位点，并使用两步聚类法归纳所筛选出位点的正向选择压力变化模式。**结果** 各年代类别筛选出的正向选择位点不完全相同，共有 50 个位点在进化过程中经历过正向选择，其中有 42 个位点属于 5 个抗原决定簇之一，A、B 抗原决定簇在各年代均有较多的位点入选，且承受较大的免疫选择压力。50 个位点可被归纳为 7 种变化模式，前 6 种模式的位点依次在某一个时间段存在正向选择，第 7 种模式则在多个年代中均存在正向选择。**结论** 分析结果提示流感病毒 HA1 序列在进化过程中的正向选择位点并非恒定，而是在不同年代中存在更替趋势。抗原决定簇 A、B 的变异在流感进化过程中起关键作用，8 个位于非抗原决定簇正向选择位点有可能是未被识别的抗原决定簇位点。

【关键词】 甲型流感病毒；亚型，H3N2；正向选择位点

Study on the detection of positive selected codons on HA1 sequence of human influenza A subtype H3N2

XU Hui-lin, ZHANG Wen-tong, ZHAO Nai-qing, JIANG Qing-wu. Department of Health Statistics and Social Medicine, School of Public Health, Fudan University, Shanghai 200032, China

Corresponding author: ZHANG Wen-tong, Email: wtzhang@shmu.edu.cn

【**Abstract**】 **Objective** To elucidate the evolution pattern of human influenza virus A H3 subtype by detecting positive selected codons in hemagglutinin gene. **Methods** All H3 sequences in NCBI GenBank and influenza sequence database were downloaded and two step cluster method was applied to divide sequences into six groups, which were corresponding to different period by turns. Fixed Effect Model was applied to detect positive selected codons in each group, and two step cluster method was then used again to summarize variation patterns of selective pressure among sites. **Results** Positive selected codons were different in groups corresponding different periods. 50 amino acid codons had been identified as positive selected sites in at least one time span. Among them, 42 codons belonged to one of the five known antigen-combining regions. A larger amount of sites as well as relatively higher selection pressure were identified in antibody combining regions A and B. Results showed that the 50 sites could be divided into seven different patterns. While other six patterns corresponding to positive selected codons at only one time span, the sites of the seventh pattern were under positive selection in several periods. **Conclusion** Positive selection codons in evolution of H3A1 strains were alternated in different time period whereas antibody combining regions A and B played more important roles in the evolution process. Other 8 identified codons out of the antibody combining regions might belong to unknown antigen regions.

【**Key words**】 Influenza virus type A; Subtype, H3N2; Positive selected codons

甲型流感一直都是人们所关心和研究的重点传染病之一，而流感病毒抗原，特别是红细胞凝集素(HA1)蛋白的高速变异使流感病毒新变异株能够不断逃避免疫屏障，成为流感预防的一大难题[1]，因此加强其变异规律的研究可作为流感防治的核心和切入点。目前认为抗原的高速变异是因为某些密码子位点承受着较大的免疫压力所致。密码子位点的突变可被分为同义突变和异义突变两种，如果突变引发的氨基酸变异有助于物种进化或生存，则异义突变数会显著高于同义突变数[2-4]，此时该密码子被称

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(许慧琳、张文彤、赵耐青)，流行病学教研室(姜庆五)

通讯作者：张文彤，Email：wtzhang@shmu.edu.cn

为正向选择位点，其异义替换率（d_N）与同义替换率（d_S）的比值（$\omega = d_N/d_S$）大于 1[5]，因此 ω 值可作为反映位点所承受免疫选择压力大小的重要指标。筛选流感病毒 HA1 序列进化中的正向选择位点已作为研究 HA1 进化的重要手段而被应用[2,6]，但由于使用的序列数量、年代范围及分析方法的限制，现有研究结论尚不能全面反映流感病毒进化中的正向选择历程。

随着生物信息学的发展，GenBank 等生物信息数据库已积累了大量的流感病毒序列数据，这些数据客观记录了流感病毒在历史上的变异过程。本课题组已对 1000 余条人 H3A1 序列通过聚类分析、进化树分析等方法进行了研究[7,8]，结果提示人 H3A1 序列呈现出单一主干的进化趋势，全部序列可以被聚为 6 类，各类间存在依次交替的基本趋势，分别对应了不同年代的进化树节段，且各类的进化模型相关参数也呈现出一定的变化规律。由于这些参数的变化反映出不同年代中抗原序列所承受的免疫选择压力存在差异，且历史上流感疫苗制备推荐株的抗原特征也各不相同，相应的变异在不同年代发生在不同位点上，不存在明显的单一变化规律，这提示由选择压力而导致的正向选择位点在不同年代可能并不相同。本研究将在原有研究的基础上，考虑分时间段筛选 H3A1 序列在进化过程中的正向选择位点，旨在从位点水平上多层次多方面地了解和分析流感病毒 HA1 的进化规律，为流感病毒的变异预测及疫苗株的选择提供基础理论依据。

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（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. lans. gov）中至 2006 年 1 月 1 日时所包含的人 H3A1 基因序列，在序列对齐、清理完重复序列后，用于分析的序列共 1665 条，均为 987 个核酸长度，329 个密码子位点。

2. 序列样本拆分：为确定适宜的年代分段，首先采用两步聚类对序列进行分析，将全部序列拆分成若干个年代类别。具体分析在 SPSS 13.0软件中实现[9]。

3. 正向位点筛选：正向选择位点的筛选在各个年代类别中分别进行，方法为固定效应似然比检验模型[10]，该模型推断正向选择位点主要包括两个过程：

（1）异义/同义替换率的估计：首先使用最大似然方法对序列构建进化树，在确定了序列进化过程中树的分支长度及替换率偏倚参数之后，再借助最大似然法使用密码子替换模型独立的对每一个密码子位点的异义/同义替换率参数进行估计。

（2）位点正向选择的似然比检验：对每个位点分别拟合异义、同义替换率相等的单参数模型 H0：$\alpha_s = \beta_s$ 和异义、同义替换率不相等的两参数模型 H1：$\alpha_s \neq \beta_s$，两模型的 2 倍对数似然值之差近似服从自由度为 1 的卡方分布，从而通过似然比检验来推断位点的异义、同义替换率是否有统计学意义。如似然比检验的 P 值小于显著性水平0.05，且该位点相应的异义/同义替换率比（ω）大于 1，则推断为正向选择位点。具体计算在 HYPHY 0.99 β 中实现[11]。

4. 正向位点 ω 值变化模式的确定：对在任何一个年代类别中曾被判断为正向位点的位点，使用其在各年代类别中计算出的 ω 值，采用两步聚类对其变化模式进行分析，将所有位点归纳为几种模式类型。具体分析在 SPSS 13.0软件中实现[9]。

结　果

1. 序列聚类结果：两步聚类的结果将全部序列分为 6 类，这些类别在时间上呈现出明确的更替规律。类别对应的进化树主干明显交替，而主干衍生出的分支上年代存在交替，聚类分析中处于变异交替状态的序列在聚类过程中被聚入相邻两类中的一类，所以相邻的类别存在年代的重叠。在随后的分析中我们将使用年代范围表示各类别（表 1）。这些类别的基本分布特征和前期研究结果基本一致[7]。

表1　H3A1 序列各类别的年代分布和样本量

序列分类	分离年代	序列数
1	1968 - 1980	70
2	1977 - 1991	114
3	1988 - 1999	289
4	1994 - 2002	214
5	1997 - 2003	514
6	2001 - 2005	464

2. 各年代类别中的正向选择位点：329 个密码子位点中共有 50 个在至少一个年代类别中被判断为正向选择位点（表 2），可见各年代类别中筛选出的正向选择位点数量不等，且具体的位点也并不完全相同，除个别位点外，绝大多数均只在一个年代类别中被选入。

表2 各时期中筛选出的正向选择位点

抗原决定簇	1968－1980 年		1977－1991 年		1988－1999 年		1994－2002 年		1997－2003 年		2001－2005 年	
	位点	ω 值	位点	ω 值	位点	ω 值	位点	ω 值	位点	ω 值	位点	ω 值
A	126	2.79	124	3.49	135	4.80	122	2.51	137	7.84	140	2.30
	137	4.76	137	4.14	138	5.59	133	3.02	142	2.45	144	2.68
	145	10.73	138	6.67	145	4.17	135	3.17	144	3.11	145	5.96
			145	3.49			137	2.48				
							142	3.50				
							145	3.87				
B	155	3.93	155	2.84	157	2.28	128	5.16	186	3.39	128	2.13
	159	2.93	156	5.77	159	2.12	157	4.83	194	3.38	156	2.98
	193	5.42	159	3.47	186	2.38	186	3.03			159	7.68
			186	5.17	193	4.58	190	3.66			186	2.20
			193	2.99	196	2.52	193	3.36			189	6.76
							194	13.33				
							198	2.43				
C					45	2.10	45	2.98	45	2.29	50	1.89
					276	3.01	50	2.42				
							275	2.58				
D	226	5.93	213	2.74	121	3.61	167	2.77	103	3.92	219	3.34
			248	3.15	201	2.80	219	2.37	167	2.27	226	5.37
					226	19.71	226	12.07	207	2.15	227	7.55
					248	2.28	229	3.25	226	2.14		
							246	4.68	229	6.02		
E	62	5.13			262	2.15			92	2.45		
NON[a]	9	3.03							3	2.39	3	2.24
	31	6.07							49	2.57	112	1.97
	164	3.26							220	9.30		
									291	1.96		

注：[a] 非抗原决定簇正向选择位点

3. 正向选择位点在抗原决定簇中的分布：全部 50 个正向选择位点中有 42 个位点分别属于已知的 5 个抗原决定簇之一[12-14]，结合抗原决定簇的位点数量可进一步揭示各决定簇位点发生正向选择变异的程度(表 3)。可见抗原决定区域 A、B 变异表现最活跃，在各年代均有位点选入，累计有 50% 以上的位点经历过正向选择，且由表 2 中位点的 ω 值可见，这两个决定簇的正向选择位点承受的选择压力均较大。尽管 D 抗原决定簇位点发生正向选择的比例明显低于 A、B 区域，但这一区域位点数最多，发生正向选择的位点绝对数量较大，且该决定簇中的入选位点数随着年代的推移而呈现出增加的趋势。C、E 抗原决定簇发生正向选择的数量及比例均较低，C 抗原决定簇在前两个年代类别中并未出现正向位点，在后三个年代类别中所筛选出的位点不仅少，而且 ω 值均较低。而 E 抗原决定簇仅在 1，3，5 类中分别筛选出了 1 个正向位点。

值得注意的是，筛选出的正向选择位点有 8 个不属于已知的 5 个抗原决定簇，且有的位点 ω 值还较高，这提示他们也可能在一定时期内承受着较高的进化选择压力。

表3 5 个抗原决定簇的正向选择位点发生情况

抗原决定簇	位点总数	正向位点数	正向选择次数	正向选择位点比例(%)
A	19	11	22	57.89
B	21	12	27	57.14
C	27	4	7	14.81
D	41	12	20	29.27
E	22	3	3	13.64

4. 50 个正向选择位点 ω 值的变化规律：由于筛选出的位点较多，这里使用 50 个正向选择位点在各年代类别中的 ω 值进行两步聚类，将其聚为 6 类。每类位点的 ω 值均为在某一个年代类别中明显较高，其余年代中则较低。但是考虑到各类中均有个别位点在多个年代类别中 ω 值均明显大于 1，因此这里将他们单独作为 1 类，最终将全部位点分为 7 类(表 4)。将年代类别与位点的聚类结果相结合，可以很清楚地看出，绝大多数正向位点都只是在流感病毒的某一年代类别内承受选择压力而发生正向

选择变化，而且在不同的年代中，承受正向选择压力的位点是不相同的，基本上呈现出批次替换的规律。除第 7 类的位点外，其余各类的位点均为在某一时间段内 ω 值远大于 1，明确属于正向位点，而在此时间段外 ω 值则在 1 上下波动，甚至于远小于 1，成为保守位点（图 1）。比较特殊的是第 7 类，这些位点不仅连续在几个年代类别中都经历正向选择变化，而且其相应的 ω 值都较大，显然是在不同的时间段上都承受着持续的选择压力（图 2）。

表4　正向选择位点 ω 值变化模式分类列表

分类	位　点　数
1	93,162,126,155,164
2	124,138,213,248
3	121,135,196,201,262,276
4	45,50,122,128,133,142,157,167,190,194,198,246,275
5	3,49,92,103,144,207,220,291
6	112,140,189,219,227
7	137,145,156,159,186,193,226,229

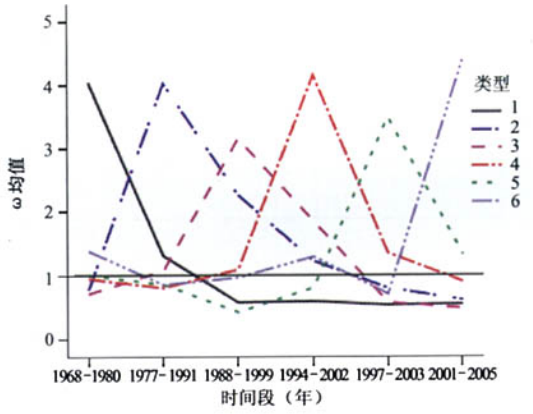

图1　1～6 类位点在各时间段的 ω 值均数

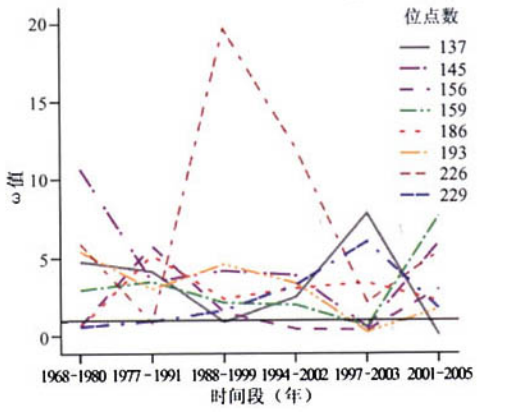

图2　第 7 类位点在各时间段的 ω 值

讨　论

正向选择位点的筛选一般都基于 ω 值进行。由于 H3N2 亚型流感病毒已在人间传播近 50 年，如果同一个位点在不同时间段上所承受的免疫压力并非恒定，而是一段时间内承受正向选择压力，其余时间处于中性进化状态（ω=1）或者抑制状态（ω<1），则分析中会因为整个时间跨度上突变情况的平均效应而不会被筛选为正向位点。本研究利用了聚类分析及进化树分析的研究结论[7,8]，将序列拆分为不同的年代类别分别进行正向选择位点筛选，从而避免了这些位点被淹没在大多数时期的非正向选择的状态中，应当能筛选出更多的此类位点，结果更加详尽。由结果可见，发生正向选择的位点 ω 值在各年代中变化各异，筛选出的 50 个正向选择位点除少数在各年代类别都发生正向选择变异外，大多数位点都只是在某一个年代类别中发生正向选择，即各年代中 HA1 序列的正向选择位点并不是完全相同，确实是在发生更替的。但也有少数位点在多个年代类别中均发生正向选择变化，其在病毒抗原变异中的作用可能更为关键，值得对其进行深入的生物学研究。

正向选择位点的发现一直是流感变异规律研究的重点之一，1999 年 Bush 等[2]利用当时收集到的 357 条人流感 H3HA1 抗原序列，采用直接计算各位点在主干上的异义/同义替换率的方法共筛选出 18 个正向选择位点。本研究的结果与之相比，可以发现所筛选的位点包括了除 158、197 外的其余 16 个位点。Bush 的分析使用1983－1997 年间采集的序列，这正好基本对应了本研究中1977－1991、1988－1999 年这两类，以及1994－2002 年这一类的前半段，而上述 16 个与 Bush 的结论相同的位点也正好是在这几个类别中选出，因此 Bush 的研究结论与本研究的结果并无大的矛盾，本研究使用的分析序列更多，得到的结果信息可能更加丰富。这在一定程度上也印证了流感病毒在进化中，不同年代的正向选择位点的确并不相同。Bush 研究与本研究结论不一致的 2 个位点我们认为可能是方法学差异所致。本研究采用的近两年才发展起来的固定效应似然比检验模型，能同时充分利用序列信息和树的拓扑结构信息，且避免了 Bush 方法中假定所有位点具有相同 ω 值而用每一个位点的 ω 值与序列平均的 ω 值进行比较来推断正向选择的缺陷，得到的结论较为可信[4]。

各抗原决定簇在病毒进化中发挥的作用并不完全相同，本研究的结果提示 A、B 决定簇在进化过程

的正向选择中表现最为活跃。在 HA1 蛋白的三维构象中，A、B 簇都位于蛋白质的头部，尤其是 B 簇直接位于抗原结构的最顶部，更容易与抗体接触而在选择压力的驱使下发生位点变异，因而抗原决定簇 A、B 上的位点承受的免疫压力更高，而这些正向选择位点发生变异的病毒株更也可能成为流感新的流行株[6]。C、D、E 3 个抗原决定簇正向选择发生率较低，可能是由于这 3 个决定簇在空间结构中的位置靠近内侧或基底部，位点的暴露程度较低，承受的选择压力也相对较小。但由于决定簇 D 位点多，尽管发生正向选择的比例显著低于 A、B 决定簇，但该簇发生过正向选择的位点绝对数仍不少，尤其在 80 年代末期以后，在每一个年代类别中发生正向选择的位点数明显较前增多，这一现象也说明了序列承受正向选择压力向 A、B 之外的抗原决定簇位点转移的多元化发展趋势，因此对于这些位点上发生较多突变的序列也应引起高度重视。在本研究中还筛选出了 8 个位于非抗原决定区域的正向选择位点，我们认为对此有两种解释，对于 3、9 等位于肽链起始段的位点，有可能对于 HA 蛋白质的功能、结构，以及抗原性的确定并不十分关键，因此其变异更容易被保留下来，因此更可能被判断为承受弱正向选择压力。而对于其余位点，我们认为有可能是未被发现的抗原决定位点，它们可能参与组成某一抗原决定簇，也可能在 5 个抗原决定簇之外构成新的抗原决定簇。

目前基于 ω 值的正向选择位点筛选方法发展很快，各种方法的优劣性尚无最终定论。本研究中所采用的是目前认为效果较好的固定效应似然比检验模型，它在一定程度上避免了计数模型及基于贝叶斯后验概率的随机效应模型的一些缺陷，理论上筛选的正向选择位点更加可靠[4,15,16]。但是，生物的进化规律各不相同，正向选择位点的筛选会涉及到生物进化中复杂的基因替换过程及大量的参数估计，针对流感病毒的进化规律，借助模拟研究方法对各种正向选择位点筛选方法在流感病毒进化研究中的适用性和结果准确性进行客观评价是绝对必要的，我们将对此进行深入研究。

参 考 文 献

[1] 金奇. 医学分子病毒学. 北京：科学出版社，2001：638-640.

[2] Bush RM, Fitch WM, Bender CA, et al. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol, 1999, 16: 1457-1465.

[3] Ina Y, Gojobori T. Statistical-analysis of nucleotide-sequences of the hemagglutinin gene of human influenza-A viruses. Proc Natl Acad Sci, 1994, 91: 8388-8392.

[4] Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol, 1999, 16: 1315-1328.

[5] Huges AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 1988, 335: 167-170.

[6] Bush RM, Bender CA, Subbarao K, et al. Predicting the evolution of human influenza A. Science, 1999, 286: 1921-1925.

[7] 张文彤，姜庆五，蒋露芳，等. 基于基因序列聚类的甲型流感病毒 H3 抗原变异规律研究. 中华流行病学杂志，2004，25：1046-1049.

[8] 张文彤，姜庆五. 全球历年人甲型流感病毒 H3A1 抗原的分子进化研究. 中华流行病学杂志，2005，26：843-847.

[9] 张文彤. SPSS 统计分析高级教程. 北京：高等教育出版社，2004：252-258.

[10] Sergei KP, Simon DW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol, 2005, 22: 1208-1222.

[11] Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics, 2005, 21: 676-679.

[12] Wilson D, Wilson I, Skehel J. Structure identification of the antibody-binding sites of Hong Kong influenza heamagglutinin and their involvement in antigenic variation. Nature, 1981, 289: 373-378.

[13] Wilson I, Cox N. Structure bases of immune recognition of influenza virus hemagglutinin. Annu Rev Immunol, 1990, 8: 732-737.

[14] http://www.bernstein-plus-sons.com/software/rasmol/.

[15] Suzuki Y. New methods for detecting positive selection at single amino acid sites. J Mol Evol, 2004, 59: 11-19.

[16] Nielsen R, Yang ZH. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics, 1998, 148: 929-936.

（收稿日期：2006-09-22）

（本文编辑：王多春）