·流感预防与监测·

全球历年人甲型流感病毒 H3A1 抗原的分子进化研究

张文彤 姜庆五

【摘要】 **目的** 应用生物信息学数据库和工具，对现有的 H3N2 亚型人甲型流感病毒全球分离株的 H3A1 抗原序列进化规律进行分析研究。**方法** 下载 NCBI Genbank 和流感病毒数据库中全部的甲型流感病毒 H3A1 序列，首先用两步聚类法进行样本拆分，随后分类绘制出完整的进化树。**结果** 人 H3A1序列呈现出单一主干的进化趋势，随着时间的推移，进化树结构和进化模型相关参数均呈现出一定的变化规律，关键变异株的出现则无明显的地域分布特征。**结论** 人流感病毒 H3A1 抗原的进化主要是病毒抗原漂移和人类免疫选择相互作用的结果，新变异株的出现并未出现明显的地域倾向性，中国华南地区不应当被认为是 H3 亚型新变异株的发源地。

【关键词】 甲型流感病毒；血凝素；生物信息学；两步聚类法；进化树

Phylogenetic analysis for H3A1 strain of all human influenza A virus *ZHANG Wen-tong, JIANG Qing-wu. Department of Health Statistic, School of Public Health, Fudan University, Shanghai* 200032, China

【**Abstract**】 **Objective** Influenza A virus remains an important pathogen which threatens humans. With the help of latest developed bioinformatics tools, all available human Influenza A virus H3A1 strains were explored to deeply understanding its evolution and variation rules. **Methods** All data of H3A1 sequence in NCBI Genbank and Influenza sequence database were downloaded and aligned in ClustalX with two step cluster method used to split the data and Bayesian phylogenetic tree analysis method applied to precisely construct phylogenetic tree for each clusters. **Results** Tree topology indicated that H3 strains evolved along a single evolution trunk and tree pattern and model parameter showed obvious variety tendency with time period. However, no geographic distribution features were found for key variation strains and big branch in trees. **Conclusion** The evolution of human H3 strains were mainly driven by the interaction of human immune barriers and antigenic drift of virus. Since the influenza subtype had already been spread in human population, south China should not be considered as the originated areas of new strains, hence it should be treated as equally as other places in the world.

【**Key words**】 Influenza A virus; Hemagglutinin; Bioinformatics; Two-step cluster; Phylogenetic tree

甲型流感是流行病学研究的重点疾病之一，其中 H3N2 亚型是目前流行的主要亚型，自 1968 年传入人群后，一直在人间持续流行和变异，给人类的生产、生活以及卫生防疫工作带来了繁重的负担。流感病毒持续流行的原因在于其血凝素抗原持续发生变异，逃避宿主免疫系统的识别与清除，从而能多次反复有效地突破人群的免疫屏障，引起新的流行。因此，加强流感病毒变异规律与流行的基础理论研究具有根本性的重要意义[1,2]。近年来，生物信息学在快速发展中积累了大量的测序数据，为流感的变异研究带来了新的机遇，国内外研究者也开始注意利用这些资料，结合统计模型和生物信息学技术对流感病毒的变异和进化规律加以研究。但是，与数据库中上千例可用的样本数相比，目前对这些资源的利用还很不充分，由于目前尚无较好的针对大样本的进化树方法，迄今最复杂的进化树也只利用了 357 条序列[3]，这使得分析结果的代表性受限。其次，除了更加准确地寻找进化树的拓扑结构外，目前的方法学进展还致力于在进化树模型中加入更多的参数，并通过对这些参数的估计来进一步深入的对序列进化规律加以刻画[4]。由于在早先的研究中已经发现聚类分析可以成功地对序列基本进化规律加以描述[5]，我们将进一步把聚类分析与最新的进

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)

化树分析方法相结合，在充分利用已有的全部 H3 序列测序结果的基础上，通过对全球历年所有 H3N2 亚型人甲型流感病毒分离株 H3A1 核酸序列进行进化分析，以从分子水平上多层次多方面地了解其变异特点和规律，为预防、控制和预报流感发生和流行提供理论依据。

资料与方法

1. 数据来源：使用的序列数据来源于 NCBI GenBank (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(http://www.flu.lanl.gov)中截止 2005 年 1 月 1 日时所包含的全部人甲型流感病毒 H3 抗原序列，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。序列下载完毕后首先在 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，最终进入分析的 H3A1 序列共计有 1214 条。

2. 聚类分析：由于本研究所用的样本量极大，远远超过了普通进化树分析方法能够承受的范围，因此首先考虑对样本进行拆分。由于在前期研究中我们已经成功地将序列聚成了若干类，且从结果解释可以发现，类别的划分实际上也反映了病毒进化和变异的规律[5]。另一方面，早期的进化树分析方法实际上就是聚类方法的衍生，如果聚类结果和进化树分枝结构间存在对应关系，则可以分类别进行分段拟合。从而大大简化进化树分析，在减少计算量的同时充分保证进化树结构的准确性。因此，这里将首先使用两步聚类法对总样本进行拆分。具体的两步聚类分析和描述工作均在 SPSS 12.0软件中完成[6]。

3. 进化树分析：进化树结构的准确性对本研究至关重要，综合考虑样本量大小与结果精度等要求，本研究中将采用基于 MCMC 抽样技术的 Bayes 方法来完成树结构的搜索和验证工作。由于除了搜索出进化树结构外，我们还希望能够进一步深入探讨序列进化规律是否存在变化，因此在进化模型中加入了相应的参数，称为 covarion-covariotide 模型[4]，它首先假设位点中存在不变异和变异这两个类别，当位点处于“关闭”状态时不允许发生变异，而处于“开放”时则按照模型中所指定的相应模型发生变异。位点在整个进化历程中会在这两类中发生转移，而参数 $s_{(off\to on)}$ 和 $s_{(on\to off)}$ 则表示相应位点由“关闭”转为“开放”和由“开放”转为“关闭”的速度。通过对这两个参数的变化规律进行分析，研究者就可以得到分析进化更详细的信息。

具体的进化树分析在 MrBayes 3.0b4版软件中完成[7]，为了进一步明确聚类结果和进化树结构间的关系，我们首先将相邻类别的序列合并分析，结果显示进化树的结构和聚类结果存在高度的一致性。在随后对各类别数据进行分析时，则依次使用上一个类别中的代表性序列作为树根，最早的类别则采用 1968 年香港株 A/HongKong/1/68 作为树根。按照作者所推荐的设定方式，均指定各参数的先验分布为均匀分布，马尔科夫链至少运行 200 万次，每隔 100 次进行抽样，并同时运行四条马尔科夫链，收敛诊断在基于 R 2.0的 BOA 软件中实现，结果表明全部类别均在 30 万次以后进入收敛状态。为保证结果的稳定性，最终只取了最后 50 万次的抽样结果(即 5000 个样本)进行汇总。

结　果

1. 聚类分析：表 1 给出了样本被聚为1～10 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 5 时 BIC 值达到最小，类间距离比也较高，因此我们将按照 5 类进行样本的拆分。这 5 类的时间与空间分布和前期研究结果完全一致[5]，在时间上呈现出明确的更替规律，但并无明显地域分布规律，因此相应结果不再列出。

表1 两步聚类分析中各统计指标的变化情况

类别数	BIC 值	BIC 改变量	BIC 改变率	最小类间距离比
1	223 711.404	–	–	–
2	181 913.051	-41 798.353	1.000	1.822
3	162 951.147	-18 961.904	0.454	1.397
4	151 885.374	-11 065.772	0.265	1.670
5	148 800.363	-3 085.011	0.074	1.439
6	149 350.528	550.165	-0.013	1.182
7	151 177.674	1 827.147	-0.044	1.529
8	155 425.912	4 248.237	-0.102	1.040
9	159 851.300	4 425.388	-0.106	1.141
10	164 819.350	4 968.051	-0.119	1.375

2. 进化树的基本特征：本研究所得到的进化树的各节点其后验概率均在0.5以上，其中靠近主干的节点，以及较大分枝的主干节点其后验概率大多在0.8以上，表明所得到的树结构高度可信。和同类研究的结果相同，H3A1 序列的进化树呈现出独特的

斜长型，进化树的结构和聚类结果呈现出了高度的一致性。相应的各个类别就正好按照出现的时间顺序依次排列，构成了进化树的一个节段。因篇幅有限，这里无法将完整的进化树加以呈现，因此我们只按类别时间顺序对进化树特征进行描述。

在最早的第 5 类中，从 1968 年的 A/HongKong/1/68 起，整个进化树就呈现出单一主干方向，如图 1 所示，其中序列名称由采集地点和年代组成，疫苗制备推荐株则在名称上加“★”表示。可见进化树在本类中并无较大的分枝出现，且呈现出每隔一定时期就会出现新的疫苗制备推荐株的特征，值得指出的是，这些推荐株均离主干并不太远。本类中序列的地域分布以欧洲为主，其余地区均只有几株。另外在本类的进化树最下方，出现了由 3 株序列构成的一个偏离主干较远的分枝，检索文献可知这是在 1983－1985 年首先出现于意大利的人 H3 株表面抗原节段与猪 H1A1 株内部节段的重组株回复传播给儿童序列[8,9]。由于该重组株只是偶尔感染人，并未构成人间大的流行，所以对此不再进行讨论。在随后的第 4 类和第 3 类中，进化树出现了一些新的趋势。首先，抗原制备推荐株开始逐渐出现在大分枝的末端，而不是接近于进化树主干；其次，这两个类别的进化主干方向上均形成了一两个较大的分枝，甚至还出现了抗原推荐株，但这些分枝却最终未能形成新的进化主干，而是在流行几年后最终消失。这似乎提示该类的主要进化方向最终被人类的免疫屏障所完全封死，而另辟蹊径的进化分枝则找到了突破点。从地域分布上看，这两个类别最初均应当起源于中国，随后播散到世界各地，整个流行的中、后期毒株主要采集自欧洲和北美，在中国反而并不多见。

进化树的形态在最后的第 2、1 类中又有了新的变化。这两个类别实际上均由4～5 个较大的分枝构成，这些分枝依次排列在进化树主干旁，规模也基本相近，显示出在这些年代中流感病毒似乎在尝试不同方向的进化途径。但是这些分枝中出现的疫苗推荐株要少得多，观察这些分枝的地域分布，其流行地点也无明显规律，虽然以欧、美地区为主，但世界各地都有发现。仔细比较这两类的差异，则会发现在最后的第 1 类中，除最下方采集自东亚地区（中国、韩国、日本等地）的毒株呈现出较明显的进化特征外，其余大多数分枝均呈现出奇怪的“扁平”形状，这似乎提示无论采取哪种变异取向，抗原的进化均受到了比较强的免疫屏障的阻碍。而最下方的东亚分枝似乎最终找到了摆脱阻碍的变异方式，得到了快速进化。该分枝连续出现了 2002、2004 年 2 株疫苗推荐株，因此应当是将来进化主干的起点。

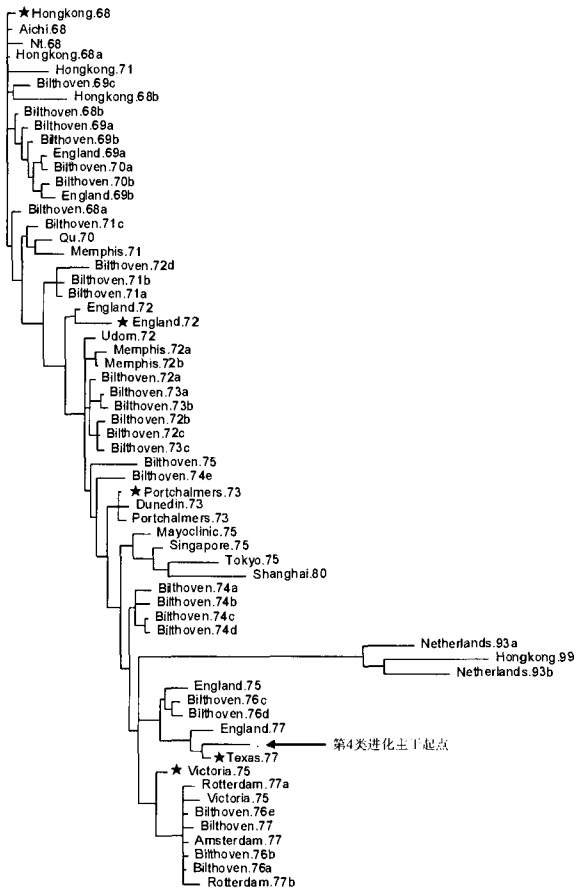

图1 流感病毒变异株的最早第 5 类所对应的进化树节段

3. 进化树各节段的参数估计：由于进化树的计算是按照分类的结果分别进行，因此可以进一步观察各模型参数，特别是 $s_{off\text{-}on}$ 和 $s_{on\text{-}off}$ 的估计值在各类中的变化规律，结果见表 2。首先，在所有的五个类别中，$s_{off\text{-}on}$ 都是远低于 $s_{on\text{-}off}$ 的，这说明位点中止变异的速度一直都要高于进入变异的速度。而按照各类别的时间顺序，这两个参数的大小也呈现出了明显的变化规律。在最早的第 5 类中，$s_{off\text{-}on}$ 为0.23，而 $s_{on\text{-}off}$ 则高达95.04，也就是说，在该类的进化历程中，虽然不断的有静止状态的位点转为可变异状态，但更多的则是处于可变异状态的位点进入静止状态，后者的频率远高于前者。在随后的第 4、3、2 类中，$s_{off\text{-}on}$ 和 $s_{on\text{-}off}$ 都有非常明显的下降，前者大约在0.05～0.08之间，而后者则降低为 10 左右。表明虽然是可变位点向静止位点的转换较多，但和以前相

比，这一转变的速度已经大大降低了。比较特殊的情况出现在最后的第 1 类中，$s_{off\text{-}on}$ 仍然保持在0.06的低水平上，而 $s_{on\text{-}off}$ 则高达66.23，变异位点被关闭的速度比以前大大上升。对此我们将在讨论中深入分析。

讨　论

1. 进化树算法的选择：目前进化树分析方法可以被分为距离阵方法、极大简约法和极大似然法(ML)三大类，其中 ML 法无疑是最精确，但计算量也是最大的一个，本研究样本量的限制使其实际无法使用[10]。近年来，基于马尔可夫链-蒙特卡罗(Markov chain Monte Carlo, MCMC)抽样技术的 Bayes 方法开始出现在进化树分析领域中。该方法在原理上可以和 ML 算法直接相结合，此时其结果的准确性完全能够与 ML 方法相比，且运算效率更高，根据 Huelsenbeck 等的研究，基于 MCMC 的 Bayes 算法其结果非常接近真值[11]。因此，我们选用 Bayes 进化树算法是较合适的选择。但 ML 方法还可以进一步完成似然比检验等工作，因此寻找高效的 ML 算法也将是本研究继续考虑的方向。

2. 分析结果的代表性：我们使用目前全部的 H3A1 序列进行分析，由结果可见，人甲型流感病毒 H3 亚型的进化路径表现为单一的进化主干，其高度传染性表现在进化树上，就是不存在明显的地域特征，在相同时期内，不同国家所采集的株系完全有可能是高度同源的。这充分说明对 IAV 进化规律的研究必须要从全球视角出发，仅取一国一地之数据，所得到的结果并不一定能反映该地区的序列进化规律。在条件许可的情况下，还是应当尽量利用全部的数据信息加以分析。

3. 对人 H3 序列基本进化规律的推测：综合进化树的时间、地域分布特征，以及相关参数在各类间的变化情况，我们可以推测流感 H3 序列的进化基本规律如下：在 1968 年 H3 株系刚跨宿主传播到人群中后，病毒的抗原序列需要加以调整，以更好的适应新的宿主环境。一旦相应的位点调整到位，就会停止变异。在模型参数上就表现为开始出现变异和停止变异的位点都很多，但后者的速度远高于前者。在这一时期中，由于人群中以前未出现相应亚型的流行，因此不存在免疫屏障，每几个位点变异的累积都可能形成新的流行株。在病毒的进化树结构上就表现为进化树形状比较规则，且每隔几年就会出现一个抗原推荐株。在进入第 4 类后，模型参数中的 $s_{off\text{-}on}$ 参数明显减小，$s_{on\text{-}off}$ 虽然仍然大于 $s_{off\text{-}on}$，但其数值也已经大大减小。这可以看成是病毒抗原结构为适应新宿主而进行的调整已基本结束，绝大部分位点已稳定下来。此时由静止状态进入可变异状态的位点应当是主要来源于位点的自然突变，而免疫屏障则对这种变异起到了筛选和促进的作用，如果这种变异恰好符合了免疫逃避的需求，则会形成新的流行株。由于位点进入变异的速度不高，因此虽然在免疫屏障的作用下 $s_{on\text{-}off}$ 是大于 $s_{off\text{-}on}$ 的，但是数值已经大大降低。在最后的第 1 类中，在参数 $s_{off\text{-}on}$ 大小保持基本不变的同时，$s_{on\text{-}off}$ 的数值却有明显升高。对此可能的一个解释是随着人群中免疫能力的逐渐积累，以及疫苗的不断普及，免疫屏障已经日渐完整。新变异株如果绕开免疫屏障，必须要经过多个位点变异的累积，而大多数突变株都在突变累积到这种程度之前被免疫屏障所阻止。因此进化树中就出现了多个中等规模的分枝，这些停止变异的分枝在模型参数上就表现为较高的 $s_{on\text{-}off}$ 值。显然，在经过几十年的持续流行后，疫苗的使用和人群免疫力的积累已经对 H3 亚型病毒的变异造成了很大的影响，其变异规律在近年内是否会发生较大变化将是非常值得关注的问题。

4. 中国在 H3 亚型流行和变异过程中所起的作用：长期以来，我国华南地区一直被怀疑是流感病毒新变异株的发源地，有学者认为该假说对于人群中已有亚型的抗原漂移也成立[12]，若的确如此，则在

表2　各类中开关参数的变化情况

类别标签	$S_{(off\to on)}$			$S_{(on\to off)}$		
	$\bar{x}$	σ^2	95% *CI*	$\bar{x}$	σ^2	95% *CI*
5	0.2321	6.9×10^{-3}	0.1861～0.2985	95.0416	10.2701	87.9005～99.7992
4	0.0833	3.1×10^{-4}	0.0721～0.0927	9.3425	1.3329	7.3754～11.8894
3	0.0571	1.8×10^{-4}	0.0480～0.0664	12.6050	2.0374	10.2184～15.7823
2	0.0632	1.7×10^{-4}	0.0556～0.0712	8.5353	0.8308	7.0104～10.4597
1	0.0606	9.0×10^{-5}	0.0554～0.0645	66.2267	2.7045	62.8118～69.1183

进化树中应当表现为重要的变异株应当主要来自于中国。但是在本研究所得到的进化树中，病毒的传播和变异并未出现明显的地域特征。其次，如果观察进化树中较大分枝的地域分布，则这些分枝的起源和流行地域也并无明显规律，在世界各地均有出现。最后，WHO 提供的抗原推荐株应当能够反映病毒变异的基本方向，在 H3 亚型近 40 年的人间流行中，WHO 一共筛选出了 23 株推荐株，虽然采集自中国的推荐株最多，共 9 株。但是如果注意到时间分布，则会发现中国主要是在1986－1995 年间一共出现了 7 株推荐株，特别是在 1987 年和 1989 年均采集了 2 株推荐株。在此时间段之外，除 1968 年跨宿主传播可能起源于中国外，就只有 2002 年又发现了 1 株推荐株。也就是说，在1970－1985 的 15 年间，虽然前后出现了 8 株推荐株，但均与中国无关。显然，以上结果并不能支持中国在 H3 亚型病毒抗原漂移中起重要作用的结论。

如果对这一问题进行深入分析，则会发现推论华南地区为流感病毒新变异株发源地的主要依据在于这个地区大量的人口，以及猪、鸭混养环境是形成流感重组株的理想环境。但从本研究的进化树结果可知，在传播到人间后，H3 亚型的进化主干基本上就是由抗原漂移通过人体免疫屏障的筛选和累积而形成，新旧变异株间有着明确的变异继承关系。虽然在数十年间曾多次出现过 H3 亚型跨宿主传播和重组事件[9,13,14]，但由于相应毒株的 H 抗原均来自于人株系，人群中已形成的免疫屏障直接阻止了相应毒株回传给人群，最终只在猪群、火鸡等其他宿主中造成了流行，并未影响人 H3 亚型的进化方向。显然，人株系 H3 抗原的变异主要受到的是人体免疫屏障的驱动，其他宿主的影响很小。此时华南地区的人禽混养环境就不应当影响到病毒的变异，因此同一亚型内的抗原漂移应当是和中国无明显关联的。考虑到新变异株的出现需要若干个位点的变异进行累积，我们提出如下假设：相比之下，各地人群的疫苗接种情况和既往流行史，以及基本卫生条件和医疗保健条件等才会真正影响到新变异株的出现。如果当地通过接种或者流行已经形成了较强的免疫屏障，则新变异株的出现会比较困难；反之，如果当地免疫屏障较弱，而卫生条件又较差的话，则可能更加容易形成新变异株。从历年来新推荐株的出现地点看，大多是在二三年中可能会在同一地区出现新推荐株，随后推荐株会转移到别的地区出现，这似乎也提示了这种可能性的存在。当然，对华南地区在抗原漂移中所起的作用这一问题更准确的解答方式，应当是将各种宿主种类的株系均放在一起进行进化树分析，以深入了解其相互影响和作用的情况，这将是我们下一步的研究内容。

参 考 文 献

1 Webby RJ, Webster RG. Are we ready for pandemic influenza? Science, 2003, 302: 1519-1522.

2 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.

3 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.

4 Tuffley C, Steel M. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci, 1998, 147: 63-91.

5 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.

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

7 Huelsenbeck JP, Ronquist F. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics, 2001, 17: 754-755.

8 Castrucci MR, Donatelli I, Sidoli L, et al. Genetic reassortment between avian and human influenza A viruses in Italian pigs. Virology, 1993, 193: 503-506.

9 Claas EC, Kawaoka Y, de Jong JC, et al. Infection of children with avian-human reassortant influenza virus from pigs in Europe. Virology, 1994, 204: 453-457.

10 Baxevanis AD, Ouellette BF. Bioinformatics: a practical guide to the analysis of genes and proteins. 2nd ed. John Wiley & Sons, Inc, 2001. 323-358.

11 Huelsenbeck JP, Ronquist F, Nielsen R, et al. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 2001, 294: 2310-2314.

12 Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. PNAS, 2002, 99: 6263-6268.

13 Zhou NN, Senne DA, Landgraf JS, et al. Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. J Virol, 1999, 73: 8851-8856.

14 Webby RJ, Swenson SL, Krauss SL, et al. Evolution of swine H3N2 influenza viruses in the United States. J Virol, 2000, 74: 8243-8251.

（收稿日期：2005-07-14）
（本文编辑：张林东）