(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 101847179 A
(43)申请公布日 2010.09.29

(21)申请号 201010147536.8

(22)申请日 2010.04.13

(71)申请人 中国疾病预防控制中心病毒病预防控制所
地址 100052 北京市宣武区迎新街 100 号流感室
申请人 中国科学院生物物理研究所

(72)发明人 舒跃龙 蒋太交 杜向军 蓝雨 吴爱平 董丽波 张烨 王大燕 彭友松

(74)专利代理机构 北京凯特来知识产权代理有限公司 11260
代理人 郑立明 赵镇勇

(51)Int.Cl.
G06F 19/00(2006.01)

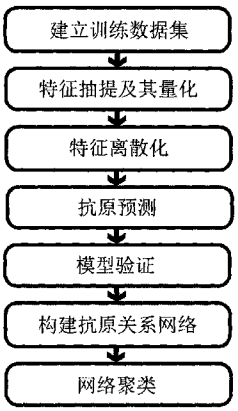
权利要求书 2 页 说明书 14 页 附图 1 页

(54)发明名称

通过模型预测流感抗原的方法及应用

(57)摘要

本发明公开了一种通过模型预测流感抗原的方法及应用，通过抽提影响流感抗原的 12 个特征：五个抗原决定簇氨基酸的突变个数、HA 蛋白氨基酸的五个理化特性、影响受体结合因素、糖基化位点改变的个数，氨基酸五个理化特性包括疏水性、体积变化、带电性、极性、可积表面积，并对 3681 对已知的抗原相似的病毒对和 1720 对抗原变异的病毒对的上述 12 个特征进行统计，建立一个抗原关系的预测模型，单纯从序列出发，就能给出病毒之间的抗原关系，简单、方便、灵敏度高。通过网络的方式能够形象的展示抗原进化的过程，用于揭示流感传播规律和筛选疫苗候选株等。



CN 101847179 A

1. 一种通过模型预测流感抗原的方法，其特征在于，包括通过以下表 1 和式 (1) 构建的模型预测两两病毒之间抗原相似还是抗原变异：

抽提影响流感抗原的 12 个特征：五个抗原决定簇氨基酸的突变个数、HA 蛋白氨基酸的五个理化特性、影响受体结合因素、糖基化位点改变的个数，所述 HA 蛋白氨基酸的五个理化特性包括疏水性、体积变化、带电性、极性、可积表面积；

对 3681 对已知抗原相似的病毒对和 1720 对抗原变异的病毒对的上述 12 个特征进行统计，得到：

表 1：

特征代号	抗原不变（$y_i=0$，抗原相似）		抗原改变（$y_i=1$，抗原变异）	
	特征改变小（$x_{ij}=0$）	特征改变大（$x_{ij}=1$）	特征改变小（$x_{ij}=0$）	特征改变大（$x_{ij}=1$）
1	2548	1133	1	1719
2	3663	18	1566	154
3	1685	1996	919	801
4	2060	1621	1463	257
5	1954	1727	990	730
6	770	2911	34	1686
7	2340	1341	180	1540
8	2468	1213	116	1604
9	2464	1217	213	1507
10	2121	1560	203	1517
11	2328	1353	165	1555
12	3493	188	1707	13

表 1 中的数据分别表示抗原相似病毒对的特征改变大和小的数量及抗原变异病毒对的特征改变大和小的数量；

$$P_{odds\ ratio}=\frac{1721}{3682}\times\left(\frac{3683}{1722}\right)^{12}\times\prod_{j=1}^{12}\frac{1+\sum_{y_i=1}\tilde{x}_{ij}}{1+\sum_{y_i=0}\tilde{x}_{ij}},\tilde{x}_{ij}=\begin{cases}x_{ij},if\ X_{j,new}=1\\1-x_{ij},if\ X_{j,new}=0\end{cases}\quad(1)$$

，式中：

$P_{odds\ ratio}$ 表示优胜率，$P_{odds\ ratio}<1$ 时被预测的病毒对抗原相似，$P_{odds\ ratio}>1$ 时被预测的病毒对抗原变异；

$X_{j,new}$ 表示待预测的病毒对的第 j 个特征改变情况，$X_{j,new}$ 以及 $P_{odds\ ratio}$ 通过以下方法计算：

首先，分别对所述影响流感抗原的 12 个特征进行量化，并分别取以下阈值 0、2、0、1、0、1.82、54.667、2.493、34.867、0.098、113.607、1；

然后，对被预测病毒对的 12 个特征分别进行比较计算，当二者的特征差异小于其对应的阈值时，$X_{j,new}=0$，$\sum_{y_i=0}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原相似的病毒对中特征改变小的数量，$\sum_{y_i=1}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原变异的病毒对中特征改变小的数量；当二者的差异大于或等于其对应的阈值时，$X_{j,new}=1$，$\sum_{y_i=0}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原相似的病毒对中特征改变大的数量，$\sum_{y_i=1}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原变异的病毒对中特征改变大的值。

2. 根据权利要求 1 所述的通过模型预测流感抗原的方法，其特征在于，所述的 3681 对

已知的抗原相似的病毒对和 1720 对抗原变异的病毒对通过以下方法得到：

已知 Smith 等人把 1968 年到 2003 年间的 253 株人 H3N2 流感病毒划分成 11 个抗原类；

对于这 253 株病毒，如果两个病毒处于同一抗原类，就认为它们是抗原相似株；如果这两个病毒处于不同的抗原类，就认为他们是抗原变异株，得到 31878 对两两病毒间的抗原关系；

选取两两病毒 HA1 蛋白序列差异数介于 1-9 的病毒对。

3. 根据权利要求 2 所述的通过模型预测流感抗原的方法，其特征在于，所述的阈值通过对所述的 3681 对已知的抗原相似的病毒对和 1720 对抗原变异的病毒对进行统计得到。

4. 一种权利要求 1、2 或 3 所述的通过模型预测流感抗原的方法的应用，其特征在于，用于构建抗原关系网络：

通过预测病毒对之间抗原相似还是抗原变异的关系，建立所有病毒之间的抗原关系网络，把每个病毒作为节点，把抗原相似的病毒之间给一个连线，构成所述抗原关系网络。

5. 根据权利要求 4 所述的通过模型预测流感抗原的方法的应用，其特征在于，包括对所述抗原关系网络进行聚类。

所述抗原关系网络的聚类包括抽提出所述抗原关系网络中的局部连接密度比较大的区域，作为抗原相似簇。

6. 根据权利要求 5 所述的通过模型预测流感抗原的方法的应用，其特征在于，所述抗原关系网络的聚类通过 MCL 方法。

7. 根据权利要求 6 所述的通过模型预测流感抗原的方法的应用，其特征在于，用于揭示流感的传播规律。

8. 根据权利要求 6 所述的通过模型预测流感抗原的方法的应用，其特征在于，用于按以下原则进行筛选疫苗候选株：

当有新的抗原相似簇出现，并且新的抗原相似簇所占比例不断增加，则选择该抗原相似簇作为疫苗候选株；

如果有多个新抗原相似簇同时满足上面条件，则选择变化更显著的抗原相似簇作为疫苗候选株。

通过模型预测流感抗原的方法及应用

技术领域

[0001] 本发明涉及一种流感抗原预测技术，尤其涉及一种通过模型预测流感抗原的方法及应用。

背景技术

[0002] 流感病毒是一种全球流行的病毒，它每年感染 300 ～ 500 万人，其中有 25 ～ 50 万人死亡，对人类社会造成巨大危害。流感病毒分 A、B、C 三个型，A 型和 B 型对人威胁较大，其中 A 型流感抗原变异频繁，对人类威胁最大。自 1968 年进入人群后，H3N2 亚型流感病毒在人群里占主导地位。H3N2 亚型流感病毒基因组包含 8 个片段，编码 11 个蛋白，其中 HA 跟 NA 是主要的表面抗原蛋白。相对于其它基因，HA 变异最快，使得抗原不断发生变化。注射疫苗是目前最有效防治流感的办法之一，由于流感病毒不断发生改变，所以必须不断更新疫苗成分。

[0003] 目前，世界卫生组织 (WHO) 通过与全球四个流感参比和研究合作中心及不同国家 / 地区的流感监测中心紧密合作，实时监测流感的抗原变化情况，并在每年二月（针对北半球）以及九月（针对南半球）通过评估全球流感流行情况推荐下一个流行季使用的疫苗株，指导疫苗的生产。但受人 H3N2 流感病毒的全球传播规律影响，疫苗株在不同地区的效果存在差异。对于源头地区，例如东亚、东南亚地区，因为新的抗原优势株在这个地区先出现并流行，使得现行推荐疫苗株对这个地区的保护性很差。最理想的情况是各个地区进行抗原监测，针对不同地区的差别分别推荐疫苗株。

[0004] 目前，使用 HI（血凝抑制反应）的方法对抗原进行检测，但这种方法费时、费力，而且有时候不够灵敏。

[0005] H3N2 病毒无休止地对人体免疫系统的逃避，使得其进化路径沿着一个主干行进，在进化树上表现为一条主干的进化模式，而其主干上的位点变化对其抗原性的进化起主要作用。其中，不同位点对抗原改变的贡献是不尽相同的，有的位点贡献大，而有的位点贡献相对小，但更多时候，抗原的改变是多个位点协同变化的结果。H3N2 病毒 HA 上存在五个抗原表位，是免疫系统抗体的主要识别区域，这些区域上的位点变化对于流感病毒抗原改变有显著的贡献。

[0006] 现有技术中，先找出跟抗原相关的位点，即所谓正选择位点，然后建立不同的位点模型来模拟和预测抗原变异。这些方法都有一定的预测能力，或多或少抓住了人 H3N2 流感病毒进化的一些规律。

[0007] 据目前的研究表明，影响抗原的所谓正选择位点是随时间变化的，即使是同一个位点，其结构背景不同，其变化的效果可能完全不一样。因此，这些基于位点的预测方法的缺点很明显 :对应时间段得到的规律只适用于对应时间段的数据，用到其他时间段上效果就会很差。

发明内容

[0008] 本发明的目的是提供一种简单、方便、灵敏度高的通过模型预测流感抗原的方法

及应用。

[0009] 本发明的目的是通过以下技术方案实现的：

[0010] 本发明通过模型预测流感抗原的方法，包括通过以下表 1 和式 (1) 构建的模型预测病毒对之间抗原相似还是抗原变异：

[0011] 抽提影响流感抗原的 12 个特征：五个抗原决定簇氨基酸的突变个数、HA 蛋白氨基酸的五个理化特性、影响受体结合因素、糖基化位点改变的个数，所述 HA 蛋白氨基酸的五个理化特性包括疏水性、体积变化、带电性、极性、可积表面积；

[0012] 对 3681 对已知抗原相似的病毒对和 1720 对抗原变异的病毒对的上述 12 个特征进行统计，得到：

[0013] 表 1：

[0014]

特征代号	抗原不变（$y_i=0$，抗原相似）		抗原改变（$y_i=1$，抗原变异）	
	特征改变小（$x_{ij}=0$）	特征改变大（$x_{ij}=1$）	特征改变小（$x_{ij}=0$）	特征改变大（$x_{ij}=1$）
1	2548	1133	1	1719
2	3663	18	1566	154
3	1685	1996	919	801
4	2060	1621	1463	257
5	1954	1727	990	730
6	770	2911	34	1686
7	2340	1341	180	1540
8	2468	1213	116	1604
9	2464	1217	213	1507
10	2121	1560	203	1517
11	2328	1353	165	1555
12	3493	188	1707	13

[0015] 表 1 中的数据分别表示抗原相似的病毒对的特征改变大和小的数量及抗原变异的病毒对的特征改变大和小的数量；

[0016] $$P_{odds\ ratio}=\frac{1721}{3682}\times\left(\frac{3683}{1722}\right)^{12}\times\prod_{j=1}^{12}\frac{1+\sum_{y_i=1}\tilde{x}_{ij}}{1+\sum_{y_i=0}\tilde{x}_{ij}},\tilde{x}_{ij}=\begin{cases}x_{ij},if\ X_{j,new}=1\\1-x_{ij},if\ X_{j,new}=0\end{cases}\quad(1)，$$

式中：

[0017] $P_{odds\ ratio}$ 表示优胜率，$P_{odds\ ratio}<1$ 时被预测的病毒对抗原相似，$P_{odds\ ratio}>1$ 时被预测的病毒对抗原变异；

[0018] $X_{j,new}$ 表示待预测的病毒对的第 j 个特征改变情况，$X_{j,new}$ 以及 $P_{odds\ ratio}$ 通过以下方法计算：

[0019] 首先，分别对所述影响流感抗原的 12 个特征进行量化，并分别取以下阈值 0、2、0、1、0、1.82、54.667、2.493、34.867、0.098、113.607、1；

[0020] 然后，对被预测病毒对的 12 个特征分别进行比较，当二者的特征差异小于其对应的阈值时，$X_{j,new}=0$，$\sum_{y_i=0}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原相似的病毒对中特征改变小的数量，

$\sum_{y_i=1}\tilde{x}_{ij}$对应表 1 第 j 个特征在抗原变异的病毒对中特征改变小的数量；当二者的差异大于或等于其对应的阈值时，$X_{j,new}=1$，$\sum_{y_i=0}\tilde{x}_{ij}$对应表 1 第 j 个特征在抗原相似的病毒对中特征改变大的数量，$\sum_{y_i=1}\tilde{x}_{ij}$对应表 1 第 j 个特征在抗原变异的病毒对中特征改变大的数量。

[0021] 本发明的上述通过模型预测流感抗原的方法的应用，用于构建抗原关系网络：

[0022] 通过预测病毒对之间抗原相似还是抗原变异的关系，建立所有病毒之间的抗原关系网络，把每个病毒作为节点，把抗原相似的病毒之间给一个连线，构成所述抗原关系网络。

[0023] 由上述本发明提供的技术方案可以看出，本发明所述的通过模型预测流感抗原的方法及应用，通过抽提一些反映抗体抗原相互作用破坏程度的特性，建立一个抗原关系的预测模型，单纯从序列出发，就能给出病毒之间的抗原关系，简单、方便、灵敏度高。通过网络的方式能够形象的展示抗原进化的过程。

附图说明

[0024] 图 1 为本发明中模型构建的技术路线示意图；

[0025] 图 2 为本发明中抗原关系预测结构模型示意图。

具体实施方式

[0026] 本发明的通过模型预测流感抗原的方法，其较佳的具体实施方式是，包括：

[0027] 通过以下表 1 和式 (1) 构建的模型预测病毒对之间抗原相似还是抗原变异：

[0028] 抽提影响流感抗原的 12 个特征：五个抗原决定簇氨基酸的突变个数、HA 蛋白氨基酸的五个理化特性、影响受体结合因素、糖基化位点改变的个数，所述 HA 蛋白氨基酸的五个理化特性包括疏水性、体积变化、带电性、极性、可积表面积；

[0029] 对 3681 对已知抗原相似的病毒对和 1720 对抗原变异的病毒对的上述 12 个特征进行统计，得到：

[0030] 表 1：

[0031]

特征代号	抗原不变（$y_i=0$，抗原相似）		抗原改变（$y_i=1$，抗原变异）	
	特征改变小（$x_{ij}=0$）	特征改变大（$x_{ij}=1$）	特征改变小（$x_{ij}=0$）	特征改变大（$x_{ij}=1$）
1	2548	1133	1	1719
2	3663	18	1566	154
3	1685	1996	919	801
4	2060	1621	1463	257
5	1954	1727	990	730
6	770	2911	34	1686
7	2340	1341	180	1540
8	2468	1213	116	1604
9	2464	1217	213	1507
10	2121	1560	203	1517
11	2328	1353	165	1555
12	3493	188	1707	13

[0032] 表 1 中的数据分别表示抗原相似的病毒对的特征改变大和小的数量及抗原变异的病毒对的特征改变大和小的数量；

[0033] $$P_{odds\ ratio}=\frac{1721}{3682}\times(\frac{3683}{1722})^{12}\times\prod_{j=1}^{12}\frac{1+\sum_{y_i=1}\tilde{x}_{ij}}{1+\sum_{y_i=0}\tilde{x}_{ij}},\tilde{x}_{ij}=\begin{cases}x_{ij},if\ X_{j,new}=1\\1-x_{ij},if\ X_{j,new}=0\end{cases}$$ （1），式中：

[0034] $P_{odds\ ratio}$ 表示优胜率，$P_{odds\ ratio}<1$ 时被预测的病毒对抗原相似，$P_{odds\ ratio}>1$ 时被预测的病毒对抗原变异；

[0035] $X_{j,new}$ 表示待预测的病毒对的第 j 个特征改变情况，$X_{j,new}$ 以及 $P_{odds\ ratio}$ 通过以下方法计算：

[0036] 首先，分别对所述影响流感抗原的 12 个特征进行量化，并分别取以下阈值 0、2、0、1、0、1.82、54.667、2.493、34.867、0.098、113.607、1；

[0037] 然后，对被预测病毒对的 12 个特征分别进行比较，当二者的特征差异小于其对应的阈值时，$X_{j,new}=0$，$\sum_{y_i=0}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原相似的病毒对中特征改变小的数量，$\sum_{y_i=1}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原变异的病毒对中特征改变小的数量；当二者的差异大于或等于其对应的阈值时，$X_{j,new}=1$，$\sum_{y_i=0}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原相似的病毒对中特征改变大的数量，$\sum_{y_i=1}\tilde{x}_{ij}$ 对应表 1 第 j 个特征在抗原变异的病毒对中特征改变大的数量。

[0038] 所述的 3681 对已知的抗原相似的病毒对和 1720 对抗原变异的病毒对可以通过以下方法得到（也可以通过其它的方法得到）：

[0039] 已知 Smith 等人把 1968 年到 2003 年间的 253 株人 H3N2 流感病毒划分成 11 个抗原类；

[0040] 对于这 253 株病毒，如果两个病毒处于同一抗原类，就认为它们是抗原相似株；如果这两个病毒处于不同的抗原类，就认为他们是抗原变异株，得到 31878 对两两病毒间的抗原关系；

[0041] 选取两两病毒 HA1 蛋白序列差异数介于 1-9 的病毒对。

[0042] 所述的阈值通过对所述的 3681 对已知的抗原相似的病毒对和 1720 对抗原变异的病毒对进行统计得到。

[0043] 本发明的上述的通过模型预测流感抗原的方法的应用，用于构建抗原关系网络：

[0044] 通过预测病毒对之间抗原相似还是抗原变异的关系，建立所有病毒之间的抗原关系网络，把每个病毒作为节点，把抗原相似的病毒之间给一个连线，构成所述抗原关系网络。

[0045] 还包括对所述抗原关系网络进行聚类。

[0046] 所述抗原关系网络的聚类包括抽提出所述抗原关系网络中的局部连接密度比较大的区域，作为抗原相似簇；

[0047] 所述抗原关系网络的聚类可以通过MCL方法(The Markov Cluster Algorithm，马尔可夫聚类算法)，也可以采用其它的方法；

[0048] 具体可以用于揭示流感传播规律。

[0049] 还可以用于按以下原则进行疫苗候选株的筛选：

[0050] 当有新的抗原相似簇出现，并且新的抗原相似簇所占比例不断增加，则选择该抗原相似簇作为疫苗株候选；

[0051] 如果有多个新抗原相似簇同时满足上面条件，则选择变化更显著的抗原相似簇作为疫苗候选株。

[0052] 本发明中的模型是通过以下方法得到：

[0053] 首先，构建训练数据集：

[0054] Smith 等人，把 1968 年到 2003 年间的 253 株人 H3N2 流感病毒划分成 11 个抗原类。通过如下原则构建训练模型需要的训练数据集。对于这 253 株病毒，如果两个病毒处于同一抗原类，就认为它们是抗原相似株；而如果这两个病毒处于不同的抗原类，就认为他们是抗原变异株，这样可以得到 31878 对两两病毒间的抗原关系。但这其中包含太多抗原变异病毒对的数据，会影响模型构建，因此选取两两病毒 HA1 蛋白序列差异数介于 1-9 的病毒对构建训练数据集，包括 3681 对抗原相似病毒对，以及 1720 对抗原变异病毒对。

[0055] 然后，进行特征选择：

[0056] 基于流感病毒抗原改变的结构本质抽提了 12 个特征用于构建抗原关系预测模型。这些特征包括：每个抗原决定簇的氨基酸突变个数（共五个抗原决定簇），五种氨基酸理化特性（疏水性、体积、带电性、极性、可积表面积），对受体结合影响，还有就是糖基化位点的改变。

[0057] 之后，对特征量化，并对特征离散：

[0058] 给定一个特定的病毒对，就可以通过比较它们的 HA1 氨基酸序列的差异，计算出上面提到的 12 个的特性的量化值。

[0059] 本发明中的模型最终只给出给定的两个病毒抗原是变了还是没变（及两个状态 0/1，分别代表抗原相似以及抗原变异），所有特征也离散成改变大小两个状态 (0/1，分别代表特性改变不能导致抗原改变以及能够导致抗原改变)。对于每一个特征，离散的原理就是找到一个阈值，使得以这个阈值为界对训练数据集中的病毒抗原关系对进行划分，划分的结果跟真实的抗原关系匹配最好。通过训练数据集学到的 1-12 特征的离散化阈值：0、2、0、1、0、1.82、54.667、2.493、34.867、0.098、113.607、1。

[0060] 最后，进行模型构建：

[0061] 通过构建 12 个特征的朴素贝叶斯模型（Naïve Bayes Model）来预测给定病毒对的抗原关系。假定选取的特征满足伯努利模型 (Bernoulli Model)，先检验分布满足正态分布，由贝叶斯理论，可以得到给定病毒对的抗原关系 $P_{odds\ ratio}$（抗原改变比上抗原不改变的比率）：

[0062] $$P_{odds\ ratio}=\frac{\frac{1+\sum_{y_i=1}1}{2+N}}{\frac{1+\sum_{y_i=0}1}{2+N}}\prod_{j=1}^{m}\frac{\frac{1+\sum_{y_i=1}\tilde{x}_{ij}}{2+\sum_{y_i=1}1}}{\frac{1+\sum_{y_i=0}\tilde{x}_{ij}}{2+\sum_{y_i=0}1}}=\frac{1+\sum_{y_i=1}1}{1+\sum_{y=0}1}\prod_{j=1}^{m}\frac{\frac{1+\sum_{y_i=1}\tilde{x}_{ij}}{2+\sum_{y_i=1}1}}{\frac{1+\sum_{y_i=0}\tilde{x}_{ij}}{2+\sum_{y_i=0}1}},\tilde{x}_{ij}=\begin{cases}x_{ij},if\ X_{j,new}=1\\1-x_{ij},if\ X_{j,new}=0\end{cases}$$

[0063] 其中 y_i 表示训练数据集中第 i 对病毒对的抗原关系 (0/1，分别表示抗原相似跟抗

原变异）。x_{ij} 表示训练数据集中第 i 对病毒第 j 个特性离散值 (0/1，分别代表特性改变不能导致抗原改变以及能够导致抗原改变）。m 表示我们抽提得到的 12 个特征 (m = 12)。计算针对整个训练数据进行，其实训练数据集给出的就是特性改变跟抗原改变的关系，也即训练集抽提如表 1 所示。

[0064] 可以得到 $\sum_{y_i=1} 1 = 1720$，$\sum_{y_i=0} 1 = 3681$，进而得到式 (1)；

[0065] 给定一对病毒，通过 12 个特征，如果 $P_{odds\ ratio} > 1$，抗原变异否则抗原相似。

[0066] 具体计算实例：

[0067] 给定一对病毒 A/Fujian/411/2002 跟 A/HongKong/1186/2003；

[0068] 其 HA1 的氨基酸差异包括 124(S- > N)，138(A- > S)，193(S- > N)，226(V- > I)，227(S- > P)，根据每个特征的阈值，可以得到每个特征的改变大小情况，即 $X_{new} = (X_{1,new}, \ldots, X_{12,new})$ 为 (1,0,0,0,0,0,0,0,0,0,1,0)

[0069] 然后根据表 1 和式 (1) 计算得到：

[0070] $$P_{odds\ ratio} = \frac{1721}{3682} \times \left(\frac{3683}{1722}\right)^{12} \times$$

[0071] $$\left(\frac{1720}{1134} \times \frac{1567}{3664} \times \frac{920}{1686} \times \frac{1464}{2061} \times \frac{991}{1955} \times \frac{35}{771} \times \frac{181}{2341} \times \frac{117}{2469} \times \frac{214}{2465} \times \frac{204}{2122} \times \frac{1556}{1354} \times \frac{1708}{3494}\right);$$

[0072] $$= 0.00042565940779$$

[0073] 因为 $P_{odds\ ratio} < 1$，因此我们预测 A/Fujian/411/2002 根 A/HongKong/1186/2003 抗原相似。

[0074] 抗原的变化本质上因为位点变化导致抗体抗原的相互作用发生变化，而单纯基于具体位点的模型显然不能反映这个本质。

[0075] 本发明从结构角度考虑，抽提一些反映抗体抗原相互作用破坏程度的特性，建立一个抗原关系的预测模型，单纯从序列出发，就能给出病毒之间的抗原关系。通过网络的方式能够形象的展示抗原进化的过程。

[0076] 通过预测抗原关系以及相关分析，可以得到中国大陆人 H3N2 流感病毒进化的规律，揭示优势抗原由南方到北方的传播规律。还可以更加细致分析亚洲不同地区流感的传播规律，揭示出亚洲热带、亚热带地区的流感传播的源头地位。能够有效监测抗原状态，进行疫苗候选株筛选。如应用到中国大陆地区，筛选出的疫苗候选株能够有效保护这个地区人群。而考虑到人 H3N2 流感病毒的全球传播趋势，这种基于起源地的抗原监测以及疫苗候选株筛选技术对流感防治意义重大。

[0077] 下面对本发明的原理和模型构建的过程进行详细的论述：

[0078] 具体如图 1 所示，包括：

[0079] 1、首先以 Smith 数据建立训练数据集，下载序列并抽提特征及量化，将特征离散化，建立抗原预测模型，验证模型，构建抗原关系网络，最后进行网络聚类。

[0080] 2、序列数据：

[0081] 从 GenBank 下载所有人 H3N2 流感病毒的 HA 蛋白质序列，截取其 HA1 区域，除去较短病毒序列（< 100aa）以及一些特殊序列（相对于所有病毒变化比较显著以及可能年代上标注错误），共得到 7297 条病毒蛋白质序列，其中 4711 条有月份信息，这些病毒序列按照其采集地划分到不同的区域。另外，基于国家流感中心的流感监测网络，收集并测序中国大

陆范围的932条病毒，其中506条有月份信息，加上从公共数据库中收集的中国大陆地区的序列数据，共得到中国大陆地区病毒序列1339条，其中705条有月份信息。另外根据病毒分离地点的不同对病毒的爆发区域进行划分，以秦岭淮河一线把中国划分成南北方。

[0082] 3、训练数据集：

[0083] Smith等，对1968年到2003年间的253株人H3N2流感病毒进行抗原测定并最终划分成11个抗原类，代表这段时间内全世界人H3N2流感病毒进化过程。

[0084] 按照如下的原则从文中抽提训练数据集：

[0085] 如果两个病毒处于同一抗原类，就认为它们抗原相似；而如果这两个病毒处于不同的抗原类，就认为他们抗原改变。同时，考虑到大于九个位点的氨基酸突变就会产生抗原改变，数据集中包含太多的这类数据会对模型造成影响，因此只保留小于等于九个位点突变的数据。最终从这套数据中抽提得到的训练数据集包含多对病毒的抗原关系，通过比较对应病毒对的HA1序列得到。

[0086] 4、特征抽提及其量化：

[0087] 抗原的改变本质上是抗体抗原相互作用的变化，因此按照经验以及数据分析抽提得到四组12个特性来反映抗原关系的改变：

[0088] 第一组特性包括五个特性，它们分别是流感HA上五个抗原决定簇的位点改变个数。这些特性广泛被人们所接受，主要反映了抗体结合区域的位点变化对抗原的影响；

[0089] 第二组特性也包含五个特性，这些特性主要从氨基酸变化对物理学直接相互作用破坏程度上来考虑，它们分别是疏水性、体积、带电性、极性和可积表面积；

[0090] 另外，受体结合区域及其周围区域位点的变化，将影响抗原的变化。这主要是两方面的叠加效应起作用：首先，受体结合区域及其周围区域的位点变化，将导致原来能结合到这个区域的抗体不能再结合，从而产生抗原变化；另外，抗体不能结合也为病毒更有效的结合宿主细胞表面的受体，为病毒的传播提供有利条件，这种适应性的优势将使对应抗原变异株更容易成为抗原优势株。

[0091] 基于以上分析，把影响受体结合作为第三组特性；

[0092] 糖基化位点的变化也将影响抗体与流感HA的相互作用，因而也将影响抗原的改变，把其作为第四组特性。

[0093] 为了表征每个特性对抗原影响的程度，首先对它们进行量化：

[0094] 第一组的五个特性的变化直接用对应抗原决定簇氨基酸改变个数来表示；

[0095] 第二组中不同氨基酸物理化学特性的变化，可以从AAindex（氨基酸指数）数据库中抽提。AAindex是一个代表各种理化和氨基酸及氨基酸对生化特性的数值指标数据库，其中需要的特性分别是代表疏水性的FASG890101、代表体积变化的GRAR740103、代表带电性的ZIMJ680104、代表极性的CHAM820101以及代表可积表面变化的JANJ780101；

[0096] 用距离受体结合区域的远近来度量第三组特性，但这要求首先要确定受体结合区域。应该说，从不同位点到受体结合区域的距离这个角度来讲，人H3N2流感病毒的HA的结构变化不大，因此用早期的H3N2结构(pdb 1MQN)为模板来计算。受体结合区域由三个结构元件组成：135～138的loop（环）、190～198的helix（螺旋）以及221～228的loop。以这三个结构元件出发，结合模板结构，确定131～138、155～160、186～196以及218～228为受体结合区域。任意位点对受体结合的影响用这个位点到受体结合区域的最短距离

来表示，但为了体现影响大小与距离大小的关系，用 HA 上所有位点距离受体结合区域的最远距离减去这个距离来表示，这样距离受体结合区域越近，其影响受体结合的数值就越大。

[0097] 以上第二组以及第三组特性的计算取两两病毒序列位点变化所导致特性变化的最大的三个值的平均值，之所以这样计算一定程度反映位点变化与抗原变化的一种关联，取平均值避免与第一组特性的重复；

[0098] 最后一组特性直接用糖基化位点改变个数来计算，糖基化位点的预测用 NetNGlyc 程序实现，用 0.5 作为阀值。

[0099] 5、特征的离散化：

[0100] 连续变量的过拟和是机器学习中经常遇到的问题，为了避免过拟和，这里对每个特性值进行离散化。前面抽提的训练数据集，其抗原状态已经被离散化：如果训练数据集中包含 N 对病毒，对于任意一对病毒 i(i = 1, ..., N)，其抗原关系用 y_i 表示，如果抗原相似 $y_i = 0$，反之 $y_i = 1$。对于每一个特性 j，其量化 s_{ij} 的离散化在这里就是找到一个合适的阀值，使得其对抗原关系的区分最好。如果用 N_1 表示抗原发生改变的病毒对数，用 N_0 表示抗原未发生改变的病毒对数，则 $N_0+N_1 = N$。对于任意一个特性 j，给定阀值 c，定义：

[0101] $$N_{j,11}^{(c)} = \{i : y_i = 1, s_{ij} > c\},$$

[0102] $$N_{j,10}^{(c)} = N_1 - N_{j,11}^{(c)},$$

[0103] $$N_{j,01}^{(c)} = \{i : y_i = 0, s_{ij} > c\},$$

[0104] $$N_{j,00}^{(c)} = N_0 - N_{j,01}^{(c)}。$$

[0105] 对于特性 j 最好的阀值通过下面公式得到：

[0106] $$c_j^* = \arg\max_c \sum_{p=0}^{1}\sum_{q=0}^{1} \frac{(N_{j,pq}^{(c)} - E_{j,pq}^{(c)})}{E_{j,pq}^{(c)}},$$

[0107] 其中

[0108] $$E_{j,pq}^{(c)} = \frac{N_p(N_{j,0q}^{(c)} + N_{j,1q}^{(c)})}{N}, p,q \in \{0,1\}。$$

[0109] 理论上，上面的操作是找出 N 个事例的 2×2 列联表的最显著卡方检验结果。按照上面方法计算得到的阀值，每个特性将被离散化成 0-1，分别表示特性改变没有造成抗原改变以及造成抗原改变，用 X 表示。

[0110] 6、抗原预测贝叶斯模型：

[0111] 如图 2 所示，Naïve Bayes Model（贝叶斯模型）在统计学习中广泛应用，其基本假设是每个特性之间是独立的。如果用 Y 表示抗原状态 (Y = 0 表示抗原相似，Y = 1 表示抗原变异，而用 $X_1, \ldots, X_m$ (m = 12) 表示每一个特性状态，应用 Bayes 定理：

[0112] $$P(Y \mid X_1, \ldots, X_m) = \frac{P(Y)\prod_{j=1}^{m} P(X_j \mid Y)}{P(X_1, \ldots, X_m)}。$$

[0113] 定义抗原改变的概率比上抗原不改变的概率为优胜率 (odds ratio)，它可以通过下面公式计算：

[0114] $$\frac{P(Y=1|X_1,...,X_m)}{P(Y=0|X_1,...,X_m)}=\frac{P(Y=1)}{P(Y=0)}\prod_{j=1}^{m}\frac{P(X_j|Y=1)}{P(X_j|Y=1)}$$。

[0115] 进一步假设 Y 以及给定 Y 的每一个 X_j 满足 Bernoulli models，比如：

[0116] $X_j|p_{0j}, Y=0 \sim \text{Bernoulli}(p_{0j})$，

[0117] $X_j|p_{1j}, Y=1 \sim \text{Bernoulli}(p_{1j})$，$j=1,\ldots,m$，

[0118] $Y|p_y \sim \text{Bernoulli}(p_u)$.

[0119] 如果认为 p_y，p_{0j} 以及 p_{1j} 的先验概率为均匀分布，定义训练数据集的抗原状态矢量为 $y=(y_1,\ldots,y_N)$ 以及离散化后的特性值矩阵 $X=(x_{ij})$，$i=1,\ldots,N$；$j=1,\ldots m$，给定训练数据集，p_y，p_{0j} 以及 p_{1j} 的后验概率可以很容易计算出来：

[0120] $$p_{0j}|X,y \sim Beta(1+\sum_{y_i=0}x_{ij},1+\sum_{y_i=0}(1-x_{ij})),$$

[0121] $$p_{1j}|X,y \sim Beta(1+\sum_{y_i=1}x_{ij},1+\sum_{y_i=1}(1-x_{ij})),$$

[0122] $$p_y|X,y \sim Beta(1+\sum_{y_i=1}1,1+\sum_{y_i=0}1).$$

[0123] 对于一个给定新的特性的观测量 $X_{new}=(X_{1,new},\ldots,X_{m,new})$，可以得到

[0124] $$P(X_{j,new}=1|Y_{new}=0,X,y)=E(P(X_{j,new}=1|Y_{new}=0,p_{0j})|X,y)$$

[0125] $$=E(p_{0j}|X,y)=\frac{1+\sum_{y_i=0}x_{ij}}{2+\sum_{y_i=0}1},$$

[0126] $$P(X_{j,new}=1|Y_{new}=1,X,y)=E(P(X_{j,new}=1|Y_{new}=1,p_{1j})|X,y)$$

[0127] $$=E(p_{1j}|X,y)=\frac{1+\sum_{y_i=1}x_{ij}}{2+\sum_{y_i=1}1},$$

[0128] $$P(Y_{new}=1|X,y)=E(P(Y_{new}=1|p_y)|X,y)$$

[0129] $$=E(p_y|X,y)=\frac{1+\sum_{y_i=1}1}{2+N}.$$

[0130] 而对于给定新的特性的观测量，其预测的优胜率 (odds ratio) 如下计算：

[0131] $$\frac{P(Y_{new}=1|X_{new},X,y)}{P(Y_{new}=0|X_{new},X,y)}=\frac{P(Y_{new}=1|X,y)}{P(Y_{new}=0|X,y)}\prod_{j=1}^{m}\frac{P(X_{j,new}|Y_{new}=1,X,y)}{P(X_{j,new}|Y_{new}=0,X,y)}$$

[0132] $$=\frac{1+\sum_{y_i=1}1}{1+\sum_{y_i=0}1}\prod_{j=1}^{m}\frac{2+\sum_{y_i=0}1}{2+\sum_{y_i=1}1}\times\frac{1+\sum_{y_i=1}\tilde{x}_{ij}}{1+\sum_{y_i=0}\tilde{x}_{ij}}\quad \tilde{x}_{ij}=\begin{cases}x_{ij}, if\ X_{j,new}=1\\ 1-x_{ij}, if\ X_{j,new}=0\end{cases}$$。

[0133] 如果优胜率大于 1，认为抗原发生了改变，反之抗原没有发生改变。

[0134] 7、模型验证：

[0135] 为了得到上述的抗原关系预测模型对训练数据集本身的预测能力，对训练数据集作 10-fold 交叉验证。把训练数据集随机分成十份，然后每次留出其中的一份作新的测试

数据集，而其余的九份为新的训练数据集，这样重复十次使得每一份都被作为测试数据集被预测一遍，得到预测准确率。同时，为了检验上述的抗原关系预测模型是否反映抗原变化的本质规律，进行前瞻性测试。从 Smith 文中的数据出发，按照年份信息把数据集分成不同时间段的数据集，分别用时间靠前的数据作训练数据集，而用时间靠后的数据作测试数据集。应该注意，基于每一次用到的新的训练数据集，都要重新进行特性的离散化，重新学习预测模型。

[0136]

Testing Dataset	Training Dataset From 1968 to						
	1973	1980	1985	1990	1995	2000	2003
CV	0.87	0.92	0.93	0.95	0.97	0.96	0.96
Remaining	0.93	0.92	0.93	0.96	0.96	0.98	/
1974-1980	0.96	0.97	0.97	0.97	0.93	0.93	0.93
1981-1985	0.92	0.93	0.98	0.97	0.95	0.94	0.94
1986-1990	0.92	0.93	0.91	0.96	0.95	0.95	0.95
1991-1995	0.91	0.91	0.91	0.97	0.97	0.97	0.97
1996-2000	0.95	0.95	0.96	0.95	0.95	0.95	0.95
2001-2003	0.98	0.98	0.99	0.98	0.98	0.98	0.98

[0137] 8、构建抗原关系网络：

[0138] 人 H3N2 流感病毒的进化过程是一个新抗原替换旧抗原不断反复的过程，本发明能够预测两两病毒的抗原关系，因此可以建立所有病毒之间的抗原关系网络。同时，人 H3N2 流感病毒的进化表现为从抗原相似的病毒中通过不断突变产生抗原变异株，基于此理解把每个病毒作为节点，而把抗原相似（优胜率 odds ratio <= 1）的病毒之间给一个连线，这样就构成一个抗原相关性网络。因为这个抗原相关性网络展示了人 H3N2 流感病毒通过不断抗原积累变化的进化过程，可以用这个网络来形象的反映抗原进化。用比较通用的 Cytoscape 来显示所有网络，并用 yFiles Organic Layout 来组织网络，这种显示方式能最大程度把网络的模块化给展示出来，反映出人 H3N2 流感病毒成簇进化的特征。

[0139] 9、网络聚类：

[0140] 为了抽提出抗原关系网络中的局部连接密度比较大的区域，也即认为的抗原相似簇，需要对得到的抗原关系网络进行聚类。网络聚类有很多种方法，但基于以下的考虑，选取 MCL 方法：

[0141] MCL 方法对每个病毒都进行分类，这符合任何病毒都可以划分抗原状态符合，虽然有一些病毒的抗原状态可能跟其他抗原状态有所区别，但这些抗原状态不是凭空出现，而是由其他一些主要抗原进化而来，从这个角度讲应该对每个病毒都给出抗原状态，这便于分析抗原进化过程。MCL 还能够利用边的权重，也就是可以利用病毒与病毒之间的抗原改变的优胜率作为权重来对抗原进行分类。

[0142] 抗原的进化是一种跟分子进化相比“更不连续”的过程，因此利用好边的权重将对于有效划分抗原类提供帮助。这样，加上权重，对上面得到的抗原相似性网络进行网络聚类，得到不同的网络模块。这些网络模块表现为抗原更相似，就把它们定义为不同抗原类。在分析人 H3N2 流感病毒整体上抗原进化规律以及传播规律时，我们采用优胜率的负对数

作为权重来进行网络聚类得到抗原类；而在疫苗候选株筛选时，因为需要跟踪抗原的细致动态变化，我们用优胜率的倒数作为权重。

[0143] 本发明可以有以下两方面的应用：

[0144] 一方面是，揭示流感传播规律：

[0145] 人 H3N2 流感病毒的进化是新抗原替换旧抗原的过程，在传播上表现为新抗原优势株由源头向其他区域传播的过程。能够用抗原关系网络来描述抗原的进化过程，并能够通过网络聚类得到不同的抗原类，这样就可以通过分析不同抗原类传播过程来得到传播规律。人 H3N2 流感病毒存在流行季的概念，即如果按照时间顺序，以月为单位，把每个时间点的病毒数列出来，可以看到在流行季病毒数比较多，而非流行季时间段病毒数相对来说就非常少。因此，考虑到不同流行季所监测病毒数以及测序病毒数的不均匀性，我们采用任意时间点的每个抗原类所包含的病毒数与其前后一个流行季时间段总病毒数的比例来描述病毒爆发以及抗原变化情况。用含有月份信息的数据，通过上面的处理，再比较不同地区新抗原出现的先后就可以确定其传播规律。

[0146] 中国人 H3N2 流感病毒进化：

[0147] 亚洲包括中国在内很可能是全球人 H3N2 流感病毒优势抗原的起源地，因而加强这个地区的流感抗原监测尤为重要。而很多研究也不约而同的指出，加强对亚洲热带、亚热带地区的流感监测，对全球流感的防治意义重大。中国大陆作为这个地区的大国，在流感的进化中扮演举足轻重的地位。而中国流感中心在中国流感监测中起领导地位，利用其健全的监测网络，可以详细描述人 H3N2 流感病毒在中国的抗原进化规律，以便更深入有效的理解 H3N2 在全球的进化与传播。

[0148] 中国国家流感中心建立了完善有效的人流感监测网络，即广泛分布于全国各个省份直辖市的哨点医院，每天进行采样，并依托分布于全国的流感网络实验室，上报国家流感中心，流感中心对流感的流行情况进行综合的分析与评价。基于这个监测网络，流感中心每年对人 H3N2 流感病毒的抗原状态进行有效监测，并结合病毒的序列分子进化分析，及时更新和推荐参考株，紧密与 WHO 沟通推荐疫苗候选株。流感中心的监测网络及其收集的数据对流感区域预防和防治起到十分重要的作用。

[0149] 为了从整体上有效描述人 H3N2 流感病毒在中国的进化规律，把 WHO 历年推荐的疫苗株加到中国大陆的序列库中，通过本发明发展的抗原关系预测模型对两两病毒抗原状态进行预测，并以此建立相应的抗原关系网络，进而对这个网络进行网络聚类，得到不同的抗原类。从进化树上看，中国的人 H3N2 流感病毒也满足主干式的进化模式，反映在抗原关系网络上，表现为抗原类的不断替换。分析中国流感的进化过程可以看到，中国国家流感中心推荐的参考株以及 WHO 的疫苗株很好的代表人 H3N2 流感病毒在中国的抗原进化过程，能够很好的覆盖抗原整个进化过程。

[0150] 但是，在抗原类 WI05（以 A/Wisconsin/67/2005 疫苗株为代表株）与 BR07（以 A/Brisbane/10/2007 疫苗株为代表株）之间，中国人 H3N2 流感病毒还存在一个独立的抗原类，而中国流感中心也推荐了相应的参考株用来表征这个抗原类，根据其参考株 A/Jiangxidonghu/312/2006，把这个抗原类命名为 JX06。JX06 在中国只持续了很短的时间 (06-07)，并很快被 BR07 抗原类取代。在中国，JX06 却在其流行期间占据 H3N2 的主导地位，并在抗原相关性网络中形成一个独立一簇。后续的分析表明，JX06 在世界其他地方并

没有流行，说明 JX06 在全球范围内并没有成为优势株。人 H3N2 流感病毒进化树的主干代表优势株的进化过程，如果把抗原流行株的抗原状态描述到进化树的主干上，JX06 出现在一个单独的分支上，这可能说明 JX06 只是一个抗原变异体，不具备很高的适应性，因而没有在全球流行。在人 H3N2 流感病毒的进化历史上，还存在一个类似的抗原类 BE89（以 A/Beijing/353/1989 疫苗株为代表株），在进化树上处于一个独立的分支，而在抗原关系图上也表现为游离于其他抗原类之外，但跟 JX06 不同的是 BE89 在全球范围内流行过，说明区域变异株的不同命运：如果适应性够好，则会像 BE89 一样全球流行；而如果适应性不是太好，则只能像 JX06 一样在局部区域短暂流行，并很快被其他优势抗原替代。JX06 的例子同时也说明，人流感抗原进化的复杂性，不同地区可能存在不同，因此对于不同地区，监测了解其抗原状态变化非常重要。

[0151] 中国人 H3N2 流感病毒传播规律及其与其他亚洲国家的关系

[0152] 上面提到了解人 H3N2 流感病毒区域进化规律对其防治的重要性，那人 H3N2 流感病毒在中国内部的传播规律如何？有什么特点？中国幅员辽阔，地大物博，地形多变，人口众多，横跨南北的疆域使得其几乎包含所有的气候类型。以秦岭淮河一线，把中国划分成南北方，而南北方在流感的流行规律上却也截然不同：北方属于典型的北半球气候，只有冬季一个流感流行季，在夏天几乎分离不到病毒；而南方则一年有多个流行季，全年流行。

[0153] 很多文章对不同地区的传播规律有过研究，John Paget et al 通过监测临床流感病例的峰值在不同地区的差异来看流感在欧洲的传播规律，而 Wladimir J.Alonso et al 利用类似的办法来研究流感在巴西的传播，但这些方法都依赖于详细的流行病学监测。而人 H3N2 流感病毒的进化就是优势抗原不断替代旧抗原的过程，换句话说就是优势抗原不断扩张的过程。既然本发明的方法从给定病毒序列的基础上，就能够有效描述出不同抗原类的进化过程，那本发明就可以直接看出优势抗原的出现在不同区域间有什么特征，从而研究 H3N2 的传播规律。把序列数据相对较多的 2002 到 2008 年这段时间内中国大陆南北方的抗原类进化过程按照月为时间单位描绘出来，可以看到，北方只有一个冬季流行季，而南方在一年则表现出多个流行季：很多时候是春季一个流行季夏季一个流行季。而不同抗原类在所分析的时间段里都表现出由南向北的传播规律，而且一般是优势抗原株在夏季流行季流行，并在随后的冬季流行季传到北方并引起流感流行。

[0154] 中国南方属于热带、亚热带气候，有很多研究者都指出热带、亚热带地区在流感进化中的特殊地位：因为流感在这些地区全年流行，因而更容易产生优势抗原株。从中国南北方传播的分析中也可以看到，中国南方在中国地区人 H3N2 流感病毒进化中有着重要地位：优势抗原类都是起源于南方，并在南方先流行，并在随后传到北方。诚然 Smith 把整个东亚、东南亚地区作为优势抗原的源头，但至少从中国分析的结果可以看出这个地区还是存在差异。

[0155] 承接对中国传播规律的分析以及亚洲不同地区之间关系的思考，用同样的方法来研究亚洲不同地区间的传播规律。分析可以看到，西亚、亚洲北部（蒙古）以及东亚的韩国、日本有着跟中国北方一致的只在冬季流行季流行的特征，而包括中国香港、澳门以及台湾在内的地区则跟中国南方的流行特点一致：全年流行，一年有多个流感流行季。另外东南亚、南亚地区同样有着跟中国南方一致的流感流行特征。这样可以根据流感的这种流行特征把亚洲地区划分成两个地区：一个是包括中国北方在内的温带地区，这个地区还包括

西亚、亚洲北部（蒙古）以及东亚的韩国、日本；另外一个地区是热带、亚热带地区，包括中国南部、中国香港、中国澳门、中国台湾、南亚以及东南亚。而分析抗原类的传播过程可以看到，人 H3N2 流感病毒在这段时间内都是由亚洲热带、亚热带地区传播到亚洲温带地区，也就是新的优势抗原类一般在亚洲热带、亚热带地区先流行，并在随后的冬季流行季到达亚洲温带地区。而优势抗原株在热带、亚热带地区的传播很复杂，不是由单一的国家或地区向其他地区传播，而是一个复杂的整体，在这个整体内，优势抗原不断出现，并传播到亚洲其他地区。

[0156] 另一方面是，筛选疫苗候选株：

[0157] 本发明的方法能有效描述抗原的进化过程，因此本发明可以通过监测抗原状态的变化来及时地筛选疫苗候选株。这里的抗原状态用某一个时间点的对应抗原类所包含的病毒数除以这个时间点所有病毒数的比例来表示，这样能够反映前后抗原比例的变化。中国作为流感新抗原株起源地之一，抗原很多时候都超前，使得现有的疫苗不能很好的保护中国的人群。这里以中国 2002 年到 2008 年数据的为例，用本发明的方法来筛选疫苗候选株，看筛选的疫苗候选株的保护效果。考虑到疫苗株制备需要至少 6 个月的时间，以及中国处于北半球，以对应冬季流行季（10 月到来年三月）的抗原状态变化来推荐下一个流行季的疫苗株。考虑到这种流行季相关的推荐方式，以及对应时间段的数据问题，以三个月为单位即季度为单位来分析。这样，以对应冬季流行季前两个季度的抗原状态变化来推荐来年流行季疫苗株，原则是：

[0158] 1）有新的抗原类出现，并且新的抗原类所占比例不断增加，

[0159] 2）如果有多个新抗原类同时满足上面条件，则选择变化更显著的抗原类作为疫苗候选株。

[0160] 基于中国监测数据的疫苗候选株筛选

[0161] 加深对人 H3N2 流感病毒进化规律的理解，特别是对其抗原进化规律的把握，了解其传播规律，最终的目的是为了能及时有效的对流感进行防治。现在最有效的防治方式还是疫苗，通过及时准确地推荐疫苗株，就能有效减少流感对于人类造成的伤害。WHO 通过全球流感监测网络，对全球流感进行抗原监测，再结合流行病学以及分子进化分析，适时推荐疫苗株。WHO 每年分两次分别对南北半球不同的流行季推荐疫苗株，指导疫苗的生产。但因为抗原监测方法的滞后性及灵敏度不够，加之人 H3N2 流感病毒区域传播所造成的抗原状态不同步问题，使得很多时间、很多地区存在疫苗株与流行株不匹配的问题。因为优势抗原更早的在起源地流行，这个问题在亚洲尤为突出。因此，对于起源地抗原状态的监测，以及基于此的疫苗株推荐就意义重大，这样可以一方面有效保护起源地地区的人群，另外对其他地区疫苗推荐以及流感防治具有十分重要的指导意义。这里以中国数据为例来筛选疫苗候选株。

[0162] 流感抗原进化就是优势抗原不断替代旧抗原的过程，优势抗原类一旦出现并在人群中流行开，那它将迅速成为流行株，替代原有的旧抗原类。基于这些认识，结合中国处于北半球的事实，从序列出发对疫苗候选株进行筛选：在北半球冬季流行季结束之前，如果有新的抗原类出现，并且其所占比例不断增加，那就使用这个新抗原类合适的病毒株为随后的流行季的疫苗候选株。应用到中国的数据上，分别推荐 02-03 到 07-08 流行季代表抗原类 FU02、FU02、CA04、WI05、WI05 以及 BR07 的疫苗株，而在中国地区，这段时间流行的优势

抗原类分别是 FU02、FU02、CA04、WI05、JX06 以及 BR07，其中 03-04 流行季流行的抗原类跟 FU02 抗原上非常相似，可以认为是 FU02。在这段时间，只有 06-07 流行季疫苗株跟流行株不匹配，而综观 WHO 推荐的疫苗株，则没有一个流行季匹配，而其对北半球其它地区也只有两个流行季能有效保护。这一方面说明现有方式推荐的疫苗株中国以及其他地区的保护不好，同时也说明通过本发明筛选疫苗候选株方法的的合理性与有效性。

[0163] 以中国的数据出发来筛选疫苗候选株，筛选的疫苗候选株能有效的保护中国地区的人群。但通过研究人 H3N2 流感病毒的传播规律知道，亚洲热带、亚热带地区作为优势抗原的起源地，优势抗原在这个地区流行比其他地区要更早流行，例如比澳洲、北美以及欧洲要早半个流行季甚至更长时间。这就使得针对优势抗原起源地的抗原监测以及疫苗候选株的筛选具有特殊的意义，因为基于这个地区筛选的疫苗候选株，不仅能有效保护这个地区的人群，对世界其他地区流感防治同样有指导意义。通过本发明的分析可以看到，如果优势抗原在亚洲的冬季流行季起源，那么在随后的夏季流行季传到澳洲（澳洲的冬季流行季），而下一个冬季流行季到达北美以及欧洲；而如果优势抗原是在亚洲夏季流行季起源，那么在同一个或再下一个夏季流行季（澳洲冬季流行）就有可能传到澳洲，并在接下来的冬季流行季到达北美以及欧洲使得可以有至少半个流行季 (3-6 个月）的预警期，可以根据起源地分别推荐不同的疫苗株，能对其它地区人 H3N2 流感病毒进行有效防治。

[0164] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。

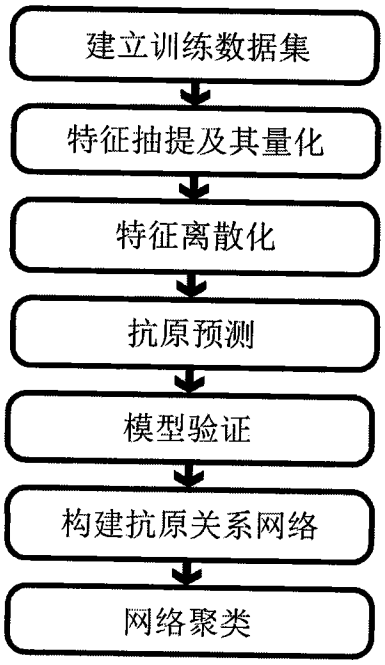
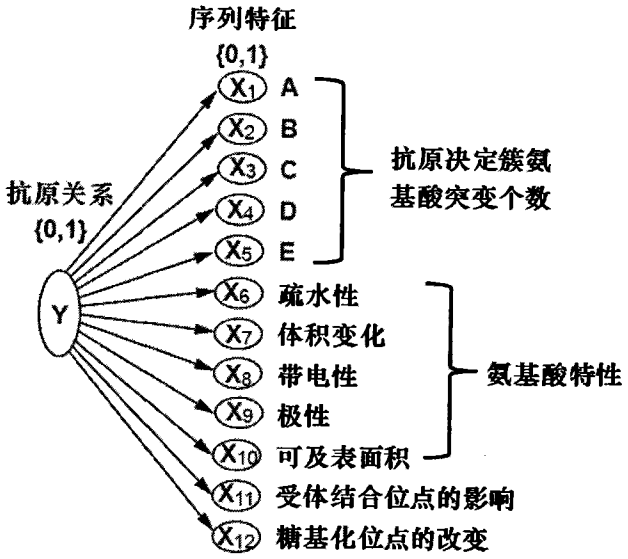


图 1



抗原关系预测贝叶斯数学结构模型

图 2