基于功率谱的流感病毒蛋白质序列结构分析

梁启浩，李阳，唐旭清*
（江南大学 理学院，无锡 214122）

摘要：基于经典HP模型，本文采用离散傅里叶变换获取蛋白质特征，利用分层聚类方法进行蛋白质序列的结构分析。其目的是将自动信号频谱分析技术与层次聚类方法相结合，并应用到蛋白质序列结构分析中。通过流感病毒HA和NA蛋白质序列的实验结果表明：应用该方法可得到非常好的分类结果。这些研究为基于大数据的蛋白质序列的自动信息提取和结构分析提供基础。

关键词：流感病毒；离散傅里叶变换；分层聚类；蛋白质序列
中图分类号：TP391；O29 **文献标识码**：A **文章编号**：1000-8721(2017)03-0313-07
DOI:10.13242/j.cnki.bingduxuebao.003152

人类基因组及其他模式的生物测序计划带来了海量的生物分子数据，从复杂的数据中挖掘有用信息是后基因组时代生物信息学的研究方向之一[1,2]，尤其是基于自动信息处理的数据挖掘技术。信号频谱分析技术正是基于自动信息处理，已被广泛应用于信息处理的各个领域，比如周期性分析、蛋白质编码区预测和基因识别等方面[3,4]。Berryman M J等[5]将信号处理与分析方法引入基因编码序列识别中。Hota等[6]基于快速离散傅里叶变换(Fast discrete fourier transform，DFT)和小波变换(Wavelet transform，WT)，从功率谱等信号处理方法的角度对基因识别进行了研究。Afreixo等[7]应用功率谱(Power spectrum，PS)探讨了核苷酸序列分布与频谱系数的关系。王其强等[8]基于功率谱将信号处理与分析方法应用于P53家族基因的三周期性特征分析。这些研究对于大数据中DNA序列数据的处理有重要的意义。

近来，聚类方法已广泛应用于大数据处理的各个方面，如分析蛋白质间的亲缘关系、提取蛋白质结构信息等[9,10]。谢佳新人[11]就不同暴发地点、不同宿主的H1N1型禽流感病毒，采用蛋白质序列进化距离进行聚类以构建进化树，对病毒株的分类及其序列变异性进行研究；Zhao等[12]应用不同的聚类算法在傅里叶变换(Fourier transform，FT)的基础上，进行了生物进化树的构建；Afreixo等[13]对一些物种聚类进行进化树研究；崔焕弟等[14]基于博卡病毒2基因组序列进行系统进化树分析；陆柔剑等[15]利用聚类分析方法构建进化树，研究中国首例输入性MERS-CoV与近几年其他国家一些地区的流行株的关系。这些研究表明基于生物进化树可以从本质上揭示蛋白序列的结构信息。

流感病毒属于正黏病毒科，包括甲、乙、丙三种类型。由于流感病毒的多样性及变异性，大多数人类对其缺乏相应的免疫力，且现有的流感疫苗无法起到有效的作用，从而流感病毒亚型的出现会造成流感的大流行，对人们生命健康产生极大的危害。因此迫切需要了解流感病毒蛋白的功能与变异特性[11]。现有的研究都是针对单病毒亚型及考虑血凝素(HA)蛋白进行流感病毒分析[16-18]，本文将在已有的HA和神经氨酸酶(NA)流感病毒蛋白序列数据基础上，基于信号频谱分析技术与层次聚类方法，构建流感病毒蛋白的分层结构(Hierarchical structure)，为基于大数据的蛋白质序列的自动信息提取和结构分析提供基础。

材料与方法

1 实验数据

流感病毒蛋白的基因组是由8个大小不等的可编码10个病毒蛋白的线状负链RNA片段组成[19,20]。这10个病毒蛋白为PB2、PB1、PA、HA、NP、NA、M1、M2、NS1、NS2，其中除了NS1、NS2为非结构蛋白外，其它均为结构蛋白。结构蛋白中的HA蛋白与NA蛋白为糖化蛋白，其它为非糖化蛋白。更重要的是，依据位于病毒外膜的血凝素(HA)和神经氨酸酶(NA)蛋白抗原性的不同，可将

收稿日期：2015-08-24；修回日期：2017-04-13
基金项目：国家自然科学基金(项目号：11371174)，题目：基于粒计算计算理论的生物网络结构建模、分析与算法研究；江苏省普通高校研究生科研创新计划资助项目(项目号：1145210232141170)
作者简介：梁启浩(1991-)，男，山西省介休市人，硕士研究生，主要研究方向为：生物信息学，E-mail：896055752@qq.com
*通讯作者：唐旭清(1963-)，男，安徽省望江人，教授，博士，主要研究方向为：智能计算，生态系统建模与仿真，生物信息学，E-mail：txq5139@jiangnan.edu.cn

流感病毒分为 16 个 H 亚型(H1～H16)和 10 个 N 亚型(N1～N10)[21,22]。因此,在流感病毒的暴发时 HA 和 NA 发挥直接且非常重要的作用,它们与流感的发生和流行最为密切。由此可知基于流感病毒 HA 和 NA 蛋白序列的分类可揭示流感病毒分子结构变异关系。

本文从 NCBI 网站中 Molecular Databases 的 Protein Sequence 下载了参考文献[23]中同时具有 HA 和 NA 两种蛋白的 29 条流感病毒进行研究(注:参考文献[23]中共采用 38 条流感病毒,但其中有 9 条不同时具有 HA 和 NA 两种蛋白)。具体的病毒名称如表 1 所示,其构成本文的研究实验数据。

表 1 29 条流感病毒及其对应的序号

Table 1 29 influenza viruses and corresponding serial number

Number	Virus name
1	[InfluenzaAvirus(A/turkey/Ontario/FAV110-4/2009(H1N1))]
2	[InfluenzaAvirus(A/mallard/NovaScotia/00088/2010(H1N1))]
3	[InfluenzaAvirus(A/thick-billed-murre/Canada/1871/2011(H1N1))]
4	[InfluenzaAvirus(A/pintail/Miyagi/1472/2008(H1N1))]
5	[InfluenzaAvirus(A/mallard/Korea/KNUYP09/2009(H1N1))]
6	[InfluenzaAvirus(A/mallard/Maryland/26/2003(H1N1))]
7	[InfluenzaAvirus(A/dunlin/Alaska/44421-660/2008(H1N1))]
8	[InfluenzaAvirus(A/mallard/Minnesota/Sg-00620/2008(H1N1))]
9	[InfluenzaAvirus(A/turkey/Virginia/4135/2014(H1N1))]
10	[InfluenzaAvirus(A/chicken/Yunnan/chuxiong01/2005(H5N1))]
11	[InfluenzaAvirus(A/chicken/Germany/R3234/2007(H5N1))]
12	[InfluenzaAvirus(A/domesticduck/Germany/R1772/2007(H5N1))]
13	[InfluenzaAvirus(A/wildbird/HongKong/07035-1/2011(H5N1))]
14	[InfluenzaAvirus(A/Chicken/HongKong/822.1/01(H5N1))]
15	[InfluenzaAvirus(A/Tokushima/chicken/Miyazaki/10/2011(H5N1))]
16	[InfluenzaAvirus(A/swine/Iowa/ chicken/Korea/es/2003(H5N1))]
17	[InfluenzaAvirus(A/mandarinduck/Korea/K10-483/2010(H5N1))]
18	[InfluenzaAvirus(A/turkey/VA/505477-18/2007(H5N1))]
19	[InfluenzaAvirus(A/Americanblackduck/NB/2538/2007(H7N3))]
20	[InfluenzaAvirus(A/Americanblackduck/ 02490/2007(H7N3))]
21	[InfluenzaAvirus(A/green-wingedteal/California//2007(H7N3))]
22	[InfluenzaAvirus(A/avian/DelawareBay/226/2006(H7N3))]
23	[InfluenzaAvirus(A/chicken/BritishColumbia/GSChuman_B/04(H7N3))]
24	[InfluenzaAvirus(A/chicken/Rizhao/713/2013(H7N9))]
25	[InfluenzaAvirus(A/duck/Jiangxi/3096/2009(H7N9))]
26	[InfluenzaAvirus(A/wildduck/Korea/SH19-47/2010(H7N9))]
27	[InfluenzaAvirus(A/mallard/Minnesota/AI09-3770/2009(H7N9))]
28	[InfluenzaAvirus(A/duck/HongKong/319/1978(H2N2))]
29	[InfluenzaAvirus(A/emperorgoose/Alaska/44297-260/2007(H2N2))]

2 方法

2.1 符号序列的数字表达 HP 模型

本文基于参考文献[23,24]考虑氨基酸的理化性质,在详细的 HP 模型中将 20 种氨基酸分成四大类,分别为极性亲水性(PQ),极性疏水性(PR),非极性亲水性(SQ)和非极性疏水性(SR),且 $PQ=\{G\}$,$PR=\{A,V,L,I,P,F\}$,$SR=\{W,M,Y\}$,$SQ=\{S,T,C,N,Q,K,R,H,D,E\}$。对含有 n 个氨基酸的蛋白质序列 $s=s_1s_2\cdots s_n$,其中 s_i,$i=1,2\cdots n$ 为组成此蛋白质序列的氨基酸,进行数据化定义:

$$\alpha_i=\begin{cases}0, & \text{若 } s_i\in PQ;\\ 1, & \text{若 } s_i\in PR;\\ 2, & \text{若 } s_i\in SQ;\\ 3, & \text{若 } s_i\in SR;\end{cases}$$

将任意一条蛋白质序列转化为一条由 0、1、2、3 构成的四元序列,记作:$X(s)=\alpha_1\alpha_2\cdots\alpha_n$。进而得到蛋白序列的指示序列:$u_{PQ}(n)$,$u_{PR}(n)$,$u_{SQ}(n)$,$u_{SR}(n)$。

2.2 基于功率谱的蛋白质特征向量提取

使用离散傅里叶变换,其显著的优点是使隐藏或潜伏在原始数据中的信息经周期性变换之后变得清晰[25,26]。下文的研究以极性亲水性(PQ)为例说

明。

利用离散的傅里叶变换及上述的指示序列,可以将蛋白质序列数据进行离散化[21]:

$$U_{PQ}(n)=\sum_{n=0}^{N-1}u_{PQ}(n)e^{-i\frac{2\pi nk}{N}},n=0,1,\cdots N-1$$

$$=\sum_{n=0}^{N-1}u_{PQ}(n)(\cos\frac{2\pi nk}{N}-i\sin\frac{2\pi nk}{N}),k=0,1,\cdots N-1$$

定义序列的功率谱

$P_{PQ}(k)=|U_{PQ}(k)|^2,k=0,1,\cdots N-1$,

同样的方法可以求得 $P_{PR}(k)$、$P_{SR}(k)$ 和 $P_{SQ}(k)$。原氨基酸序列的功率谱函数

$P(k)=P_{PQ}(k)+P_{PR}(k)+P_{SQ}(k)+P_{SR}(k)$, $k=0,1,\cdots N-1$.

通过功率谱构造向量,可分析生物序列的相似性。相似性则可以通过计算两向量之间的欧式距离得到。一般认为两向量欧式距离越小,两序列就越相似[27]。对于极性且亲水性(PQ)类氨基酸,在文献[23]中定义 j 阶矩为

$$M_j^{PQ}=\frac{1}{N_{PQ}{}^{j-1}(N-N_{PQ})^{j-1}}\sum_{k=1}^{\frac{N}{2}}(P_{PQ}(k))^j$$

类似地,可获得 M_j^{PR}、M_j^{SQ} 和 M_j^{SR}。因此通过 M_j^{PQ}、M_j^{PR}、M_j^{SQ}、M_j^{SR} 可以构建 12 维向量,也称为基于矩的蛋白质特征向量,即

$(M_1^{PQ},M_1^{PR},M_1^{SQ},M_1^{SR},M_2^{PQ},M_2^{PR},M_2^{SQ},M_2^{SR},M_3^{PQ},M_3^{PR},M_3^{SQ},M_3^{SR})$。

这样每一条蛋白质序列都会的得到一个 12 维的特征向量。特征向量间的距离按欧式距离进行计算。以下将在上面特征向量的基础上开展研究工作。

2.3　聚类

聚类分析是进行数据分析的一个基本方法,在数据挖掘、模式识别、生物信息学和统计学等领域都有广泛的研究与应用[28],也是探索或提取隐含在数据中的新规律和新知识的重要手段[29,30]。本文将采用基于距离的层次聚类方法获取系统的分层结构[31]。设 d 是有限集 $X=\{x_1,x_2,\cdots x_n\}$ 上一个距离,记

$D=\{d(x,y)\mid x,y\in X\}=\{d_0,d_1,\cdots,d_m\}$

其中 $d_0=0<d_1<\cdots<d_m$。其算法如下:

算法 A:

S1:输入 n 个样本,$i\Leftarrow 0$;

S2:构造 n 个类,每个类中只含有一个样本,记为 $X(d_i)=C=\{c_1,c_2,\cdots c_n\}$;

S3:$A\Leftarrow C,i\Leftarrow i+1,C\Leftarrow\varnothing$;

S4:$B\Leftarrow\varnothing$;

S5:对于任意的 $c_j\in A$,令 $B\Leftarrow B\cup\{c_j\}$,$A\Leftarrow A/c_j$;

S6:$\forall c_k\in A$,如果存在 $x_j\in c_j$,$y_k\in c_k$ 使得 $d(x_j,x_k)\leq d_i$,则 $B\Leftarrow B\cup\{c_k\}$,$A\Leftarrow A/c_k$;

S7:$C\Leftarrow\{B\}\cup C$;

S8:若 $A\neq\varphi$,则转 S4;

S9:$X(d_i)=C$,输出 $X(d_i)$;

S10:直到 $C\neq\{X\}$,否则转 S3;

S11:结束。

通过算法 A,可获得数据系统的分层结构。

结果分析与讨论

以下给出了采用本文方法对实验数据进行处理所得到的结果。由于流感病毒共编码 PB2、PB1、PA、HA、NP、NA、M1、M2、NS1、NS2 共 10 种病毒蛋白质,本文选取其中最重要的两种 HA 和 NA 进行研究。图 1 为结合 HA 与 NA 序列构造成 24 维特征向量得到的流感病毒蛋白质序列分层结构,图 2 和图 3 分别为采用 NA 序列与 HA 序列构造 12 维特征向量进行聚类得到的流感病毒蛋白质序列分层结构。

由图 1-3 可知:通过本文方法对流感病毒蛋白质 HA 与 NA 序列结合进行分类与仅以 HA 所进行的分类结果有较大差异,而与仅以 NA 所进行的分类结果差异较小。若将 29 条流感病毒蛋白质序列分为 5 类时,以 HA 与 NA 相结合构造蛋白质序列的特征向量进行聚类,分类结果为 H7N9,H7N3,H2N2,H1N1,H5N1(图 1);仅以 NA 构造蛋白质序列的特征向量进行聚类,分类结果为 H7N9,H7N3,H2N2,H1N1,H5N1(图 2),此与前述的结果一致。若将 29 条流感病毒蛋白质序列分为 4 类时,以 HA 与 NA 结合构造蛋白质序列的特征向量进行聚类的结果为:(H7N9,H7N3),H2N2,H1N1,H5N1(图 1);仅以 NA 构造蛋白质序列的特征向量进行聚类的结果为:H7N9,(H7N3,H2N2),H1N1,H5N1(图 2)。两种方法 H7N3 与不同的病毒聚在一起,其他各病毒序列所属类别均高度一致。事实上,H7N3 与 H7N9 病毒序列无论从发生的时间还是地点都较为接近且都属于 H7 亚型,而与 H2N2 相差较大。由此可见将 H7N3 与 H7N9 归为一类更为合理。而仅以 HA 构造蛋白质序列的特征向量进行聚类,只有分为 3 类时,分类结果为(H7N9,

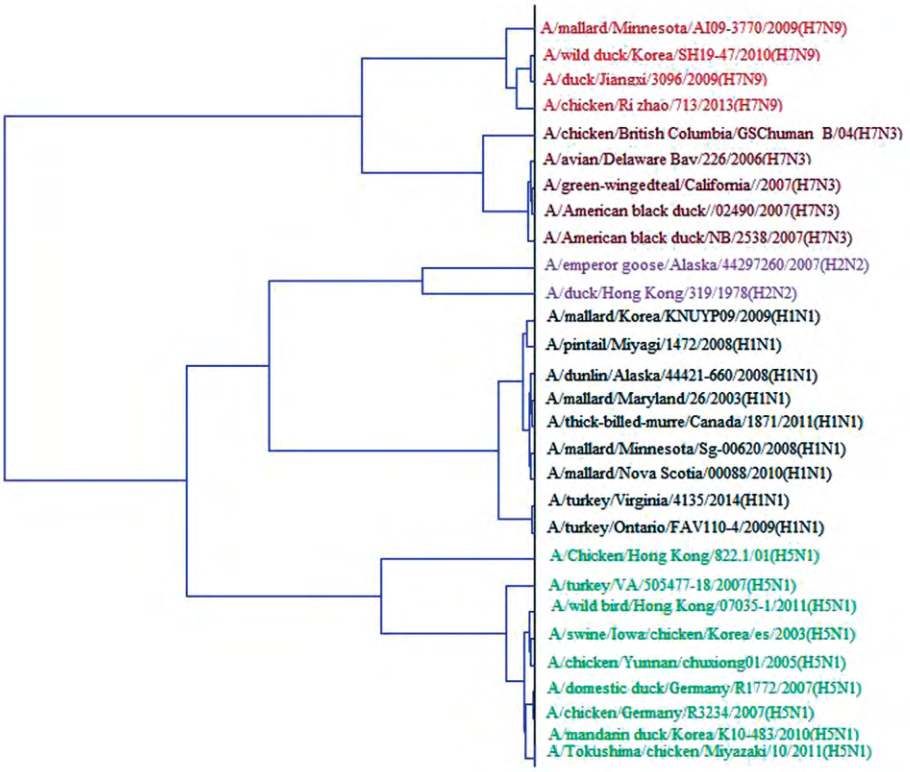

图 1　HA 与 NA 蛋白序列表达的流感病毒分层结构

Figure 1　Hierarchical structure of influenza viruses with sequences of HA and NA proteins

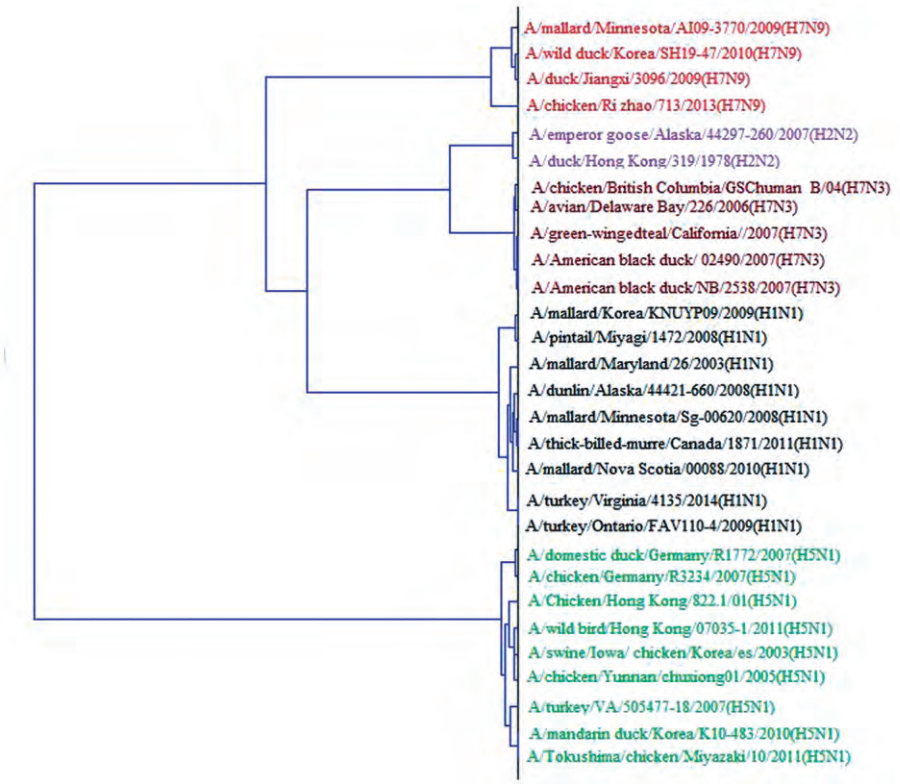

图 2　NA 蛋白序列表达的流感病毒分层结构

Figure 2　Hierarchical structure of influenza viruses with sequences of the NA protein

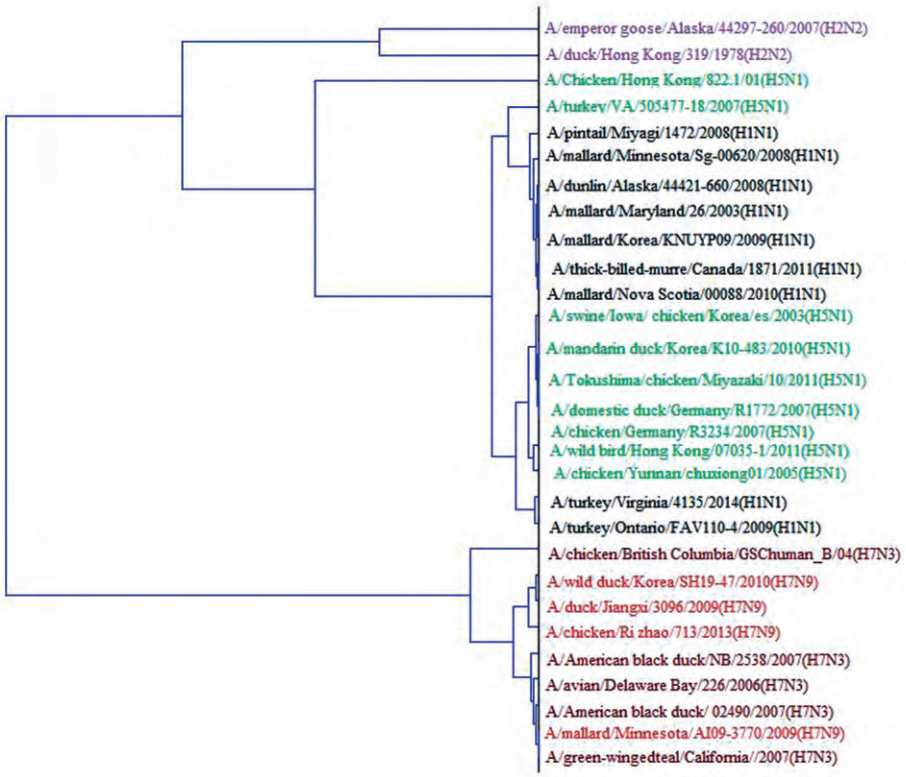

图 3　HA 蛋白序列表达的流感病毒分层结构

Figure 3　Hierarchical structure of influenza viruses with sequences of the HA protein

H7N3),H2N2,(H1N1,H5N1)(图 3)。将 H7 亚型、N1 亚型与 H2N2 分开,满足流感病毒的分类方式。因此对于给定的 29 种流感病毒,采用 HA 与 NA 相结合构造蛋白质序列的特征向量,所获得的分层结构能很好的反映其内在的结构特征。同时,实验结果表明我们方法的有效性。

由图 1 可以看出发生在同一宿主上的病毒很早就聚为一类,例如图 1 中 H7N9 的分类,虽然同为 H7N9,禽类宿主从 duck、chicken 到 mallard 都有涉及,从构建的进化树可以发现 A/duck/Jiangxi/3096/2009(H7N9)与 A/wild duck/Korea/SH19-47/2010(H7N9)尽管发生的时间与地点都不相同,但他们的宿主都是 duck,因此具有更为相似的序列结构,而 A/mallard/Minnesota/AI09-3770/2009(H7N9)宿主发生了变异,蛋白序列结构虽然相近但有了一定差异。病毒 A/chicken/Ri zhao/713/2013(H7N9)由于发生的时间和地点具有较大的跨度,关系较为疏远。同时发现,在同一宿主上可以检测到各种不同禽流感病毒,例如在 chicken 上有 H5N1、H7N3、H7N9,在 duck 上有 H5N1、H7N3、H7N9、H2N2,所以对被寄生多种病毒的宿主值得重点检测;从暴发地角度,从美洲(California)到亚洲(HongKong),地域分布较广,这是因为现代人员与货物流动较大,可能会成为流感暴发源头。总体上看:具有宿主相同、时间跨度相似、发生的地点相近以及相同名称的病毒蛋白更倾向于处于相同的分支,所以不同的年限、暴发地和宿主对流感病毒同源性有着重要的影响,这与参考文献[11,32]中的结论相一致。而相同的病毒名称,或者因为时间跨度、宿主的不同造成变异相似性远近有区分。所以需要对几类变异快的流感病毒亚型做进一步分析研究。

由于频谱信息的唯一性,它可以反映隐藏或潜伏在原始数据中的信息。本文所采用的 29 种流感病毒只以 HA 或 NA 进行分类时并不能反映序列中的信息,所以此时采用一种病毒蛋白进行研究时并不能完全代表此病毒,也得不到最优的结果。而通过 HA 和 NA 的结合能达到最优分辨并得到很好的分类结果。所以,HA 与 NA 能完全反映了这 29 种流感病毒的特征,不需要再使用其他的 8 种蛋白序列。但是在大数据的处理的过程中会用到 10

种蛋白其中的几种或者全部才能反映大数据中的全部信息，这是下一步我们需要进行的工作。

此外，用离散傅里叶变换构造特征向量来表征病毒蛋白序列，并结合其编码的蛋白可以完全包含序列所拥有的信息，且可用于自动特征信息提取。这正是本文研究的目的。

结 论

本文基于 DNA 序列在频率域上的表示方法，提出了基于功率谱的流感病毒蛋白质序列结构分析，将在 DNA 序列上的研究转换到蛋白质序列上。主要研究包括：

①在经典 HP 模型的基础上考虑以氨基酸的物理化学性质对蛋白质的氨基酸序列进行分类，继而在这种分类的基础上将序列数值化。

②在①的基础上，进行了离散傅里叶变换将序列离散化，根据定义计算序列的功率谱，在此基础上构造向量矩阵，并计算欧式距离。

③在②的基础上，采用层次聚类算法获取分层结构考察蛋白质序列的相似性。

在进行实验的过程中选取 29 种流感病毒序列，采用 12 维的特征向量分别表示 HA 与 NA 蛋白序列，进一步的将二者相结合，构造 24 维特征向量表征整条蛋白质序列，利用层次聚类的方法，对其进行分类，实验结果与已有的文献相吻合。因此将基于 DNA 序列在频率域上的特征提取上升到基于蛋白质的氨基酸序列在频率域上的特征提取方法，这为蛋白质序列研究提供更严谨的分类方法。这些为基于大数据的信息自动提取和结构分析提供研究路径，这也正是本文的创新之处。

参考文献：

[1] Ikeda H, Shin-ya K, Omura S. Genome mining of the Streptomyces avermitilis genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters [J]. J Ind Microbiol Biot, 2014, 41(2): 233-250.

[2] Marra M A, Jones S J M, Astell C R, et al. The genome sequence of the SARS-associated coronavirus [J]. Sci, 2003, 300(5624): 1399-1404.

[3] Anastassiou D. Frequency-domain analysis of biomolecular sequences[J]. Bioinformatics, 2000, 16(12): 1073-1081.

[4] Kotlar D, Lavner Y. Gene prediction by spectral rotation measure: a new method for identifying protein coding regions[J]. Genome Res, 2003, 13(8): 1930-1937.

[5] Berryman M J, Allison A, Wilkinson C R, Abbott D. Review of signal processing in genetics[J]. Fluct Noise Lett, 2005, 5(4): 13-35.

[6] Hota M K, Srivastava V K. Identification of protein-coding regions using modified Gabor-wavelet transform with signal boosting technique [J]. Int J Comput Biol Drug Des, 2010, 3(4): 259-270.

[7] Afreixo V, Ferreira P J S G, Santos D. Spectrum and symbol distribution of nucleotide sequences [J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 70(3): 031910.

[8] 王其强，谈承杰，晏寒冰，朱平. 基于碱基三周期性研究 P53 家族基因的特征[J]. 生物物理学报，2013，29(4)：296-309.

[9] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques [C] // KDD workshop on text mining. 2000, 400(1): 525-526.

[10] Hall L O. Exploring big data with scalable soft clustering [M]. Springer Berlin Heidelberg, 2013: 11-15.

[11] 谢佳新，殷建华，李淑华，鹿文英，韩一芳，韩磊，张宏伟，曹广文. 2009 年新型甲型 H1N1 流感病毒血凝素基因进化分析 [J]. 第二军医大学学报，2009，30(6)：613-617.

[12] Zhao B, Duan V, Yau S S T. A novel clustering method via nucleotide-based Fourier power spectrum analysis [J]. J Theor Biol, 2011, 279(1): 83-89.

[13] Afreixo V, Bastos C A C, Pinho A J, Garcia S P, Ferreira P J. Genome analysis with inter-nucleotide distances [J]. Bioinformatics, 2009, 25(23): 3064-3070.

[14] 崔换弟，金玉，谢广成，程卫霞，段招军. 人博卡病毒 2 基因组扩增及序列分析 [J]. 病毒学报，2014，30(3)：257-262.

[15] 陆柔剑，邹丽容，王延群，赵彦杰，周为民，武婕，王文玲，武桂珍，柯昌文，谭文杰. 中国首例输入性中东呼吸综合征冠状病毒结构基因与附属基因的序列分析 [J]. 病毒学报，2015，31(4)：333-340.

[16] Simth G J D, Vijaykrishna P V, Bahl J, Lycett S J, Worobey M, Pybus O G, MaS K, Cheung C L, Raghwani J, Bhatt S, Peiris J S M, Guan Y, Rambaut A. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic [J]. Nature, 2009, 459(7250): 1122-1125.

[17] Garten R J, Davis C T, Russell C A, et al. Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in human [J]. Science, 2009, 325(5937): 197-201.

[18] Plotkin J B, Dushoff J, Levin S A. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus[C]. PNAS, 2002, 99(9): 6263-6268.

[19] 王妮莎. 流感病毒 H1N1 亚型血凝素和神经氨酸酶的序列分析[D]. 广州,南方医科大学, 2010.

[20] 朱广蕊, 潘耀谦, 夏银可, 等. 甲型 H1N1 流感病毒致病机理研究进展[J]. 动物医学进展, 2011, 32(8): 70-74.

[21] Bhoumik P, Hughes A L. Reassortment of ancient neuraminidase and recent hemagglutinin in pandemic (H1N1) 2009 virus [J]. Emerg Infect Dis, 2010, 16 (11): 1748-1750.

[22] 修文琼, 中岛捷久, 信泽枝里. 禽流感病毒血凝素 HA 的可变性解析 [J]. 病毒学报, 2008, 24(1): 34-40.

[23] Hoang T, Yin C, Zheng H, Yu C, He R L, Yau S S T. A new method to cluster DNA sequences using Fourier power spectrum [J]. J Theor Biol, 2015, 372 (1): 135-145.

[24] Tang X Q, Zhu P, Cheng J X. The structural clustering and analysis ofmetric based on granular space [J]. Pattern Recog, 2010, 43(11): 3768-3786.

[25] Harris F J. On the use of windows for harmonic analysis with the discrete Fourier transform [C]. P IEEE, 1978, 66(1): 51-83.

[26] 赵剑, 许金涛, 顾凌榕. 蛋白质序列在频率域上的一种特征提取方法[J]. 南京工业大学学报(自然科学版), 2013, 35(6): 113-119.

[27] 赵静静, 齐斌, 王寒冰, 唐旭清. 基于矩阵图谱表达法的蛋白质序列的相似性分析 [J]. 计算机工程与应用, 2011, 47(7): 222-225.

[28] 陶华, 唐旭清. 蛋白质序列的聚类结构分析[J]. 生物信息学, 2012, 10(4): 269-273.

[29] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm [J]. Adv Neural Inform Proc Systems, 2002, 2: 849-856.

[30] Jain A K, Murty M N, Flynn P J. Data clustering: a review [C]. ACM Comput Surv (CSUR), 1999, 31 (3): 264-323.

[31] Tang X Q, Zhu P. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space [J]. IEEE T Fuzzy Syst, 2013, 21(5): 814-824.

[32] Mullick J, Cherian S S, Potdar V A, Chadha M S, Mishra A C. Evolutionary dynamics of the influenza A pandemic (H1N1) 2009 virus with emphasis on Indian isolates: Evidence for adaptive evolution in the HA gene [J]. Infect Genet Evol, 2011, 11(5): 997-1005.

Structural Analyses of the Protein Sequences of the Influenza Virus Based on the Power Spectrum

LIANG Qihao, LI Yang, TANG Xuqing*

(*School of Science, Jiangnan University, Wuxi* 214122, *China*)

Abstract: Based on the HP model and introduction of discrete Fourier transform spectroscopy to extractprotein features, the structural analyses of protein sequences using the hierarchical clustering method is described. We wished to combine automatic signal spectrum analyses with the hierarchical clustering method to analyze protein sequences. Test results on the protein sequences (HA and NA) of the influenza virus demonstrated a new way to obtain good classifications. These results provide a research path for the automatic extraction of information and structural analysis of big data.

Key words: Influenza virus; Discrete Fourier transform; Hierarchical clustering; Protein sequence

Funding: This work was supported by the National Natural Science Foundation of China (Grand No. 11371174) and Colleges and Universities in Jiangsu Province Plans to Graduates Research and Innovation (Grant No. 1145210232141170).

* *Corresponding author*: *TANG Xuqing*, *E-mail*: *txq*5139@*jiangnan.edu.cn*