Hans Journal of Computational Biology 计算生物学, 2018, 8(1), 15-23
Published Online March 2018 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2018.81003

Hans汉斯

Comparison and Analysis of H1N1 Hemagglutinin Protein Sequences Based on Fourier Power Spectrum

Hua Wang*, Fenglan Bai, Liwei Liu

College of Sciences, Dalian Jiaotong University, Dalian Liaoning
Email: *1123943421@qq.com

Received: Apr. 19th, 2018; accepted: May 11th, 2018; published: May 18th, 2018

Abstract

Based on the classical HP model, the H1N1 hemagglutinin protein sequence under different characteristics was converted into a digital sequence and the power spectrum of the corresponding sequence was calculated using a discrete Fourier transform. According to these power spectra, a mathematical moment function is established, and the digital sequence is converted into a multi-dimensional moment vector to obtain the corresponding feature vector of the protein sequence. Then using the middle distances between the feature vectors to compare and analyze the protein sequences, a good result was obtained. This method converts protein sequences of different lengths through power spectrum and moments into vectors of the same dimension, which makes it easier for us to compare and analyze biological sequences.

Keywords

Protein Sequence, Fourier Transform, Power Spectrum, Clustering

基于傅里叶功率谱的H1N1病毒血凝素蛋白质序列的比较分析

王　华*，白凤兰，刘立伟

大连交通大学理学院，辽宁　大连
Email: *1123943421@qq.com

收稿日期：2018年4月19日；录用日期：2018年5月11日；发布日期：2018年5月18日

*通讯作者。

文章引用：王华，白凤兰，刘立伟. 基于傅里叶功率谱的 H1N1 病毒血凝素蛋白质序列的比较分析[J]. 计算生物学, 2018, 8(1): 15-23. DOI: 10.12677/hjcb.2018.81003

摘 要

基于经典的HP模型，将不同特征下的H1N1病毒血凝素蛋白质序列转换为数字序列并且用离散傅里叶变换求出相应序列的功率谱。根据这些功率谱建立数学矩函数，并将数字序列转换为多维的矩向量，得到蛋白质序列对应的特征向量。再利用特征向量之间的中间距离对蛋白质序列进行聚类比较分析，得到了较好的结果。这一方法将不同长度的蛋白质序列通过功率谱和力矩将其转化为相同维数的向量，使我们更加容易比较分析生物序列。

关键词

蛋白质序列，傅里叶变换，功率谱，聚类

1. 引言

在过去几十年中，研究者已经提出了几种分类生物序列的方法。这些方法中的大多数是基于对齐的，其中通过使用选定的评分系统获得最佳比对。这些方法提供了生物序列的准确分类，并且已经开发并成功应用了几种算法[1] [2] [3]。然而，它的主要缺点是消耗时间长，这在需要进行快速聚类时(例如新的致命病毒)是不合适的[4]。以后，无对齐技术是一种趋势方法，它通常能在同一数据集上给出更快的分类[5] [6] [7] [8]。例如，k-mer 方法是最流行的无对准方法。为了测量两个序列的不同，收集两个生物序列中 k 个集合或长度为 k 的子序列，然后计算它们之间的进化距离[5] [9]。k-mer 方法给出与基于对准的方法相当的结果，同时计算速度更快[10]。离散傅里叶变换(DFT)是信号和图像处理中的强大工具。近年来，DFT 越来越多地用于信息处理的各个领域，如基因预测，蛋白质编码区和周期性分析[11] [12]。DNA 序列的 DFT 功率谱反映了该序列的核苷酸分布和周期性模式，并且已经应用于鉴定基因组序列中的蛋白质编码区[13] [14] [15]。

目前，聚类方法在处理大数据的各个方面扮演着越来越重要的角色，如分析蛋白质之间的相似性、提取蛋白质结构信息等[16] [17]。对于 H1N1 禽流感病毒，谢佳新等人[18]采用蛋白质序列进化距离进行聚类并构建进化树，在此基础上对病毒的变异性进化进行了研究；Zhao 等[19]在傅里叶变换的基础上应用不同的聚类法，对生物序列构建进化树；赵剑等人[20]在蛋白质的二位数字表达的基础上结合使用向量的傅里叶变换理论，提出高维共鸣识别法来判别蛋白质序列的相似性。李巍巍等人[21]在不同特征描述下对多条 H1N1 病毒血素蛋白质序列进行比较分析，在不影响表征蛋白质序列的前提下，用 16 维的特征向量代替已有的表征蛋白质序列的 40 维特征向量，大大减少了计算的复杂度。但是，此方法只考虑了物化性质而忽略了蛋白质序列的内部结构，得到的结果有一定的局限性。在本文中，在 20 种氨基酸以及氨基酸的四类理化性质(极性且亲水 *pq*，极性且疏水 *pr*，非极性且亲水 *sq* 和非极性且疏水 *sr*)两两连接所得的特征下用傅里叶变换将蛋白质的符号序列转换为数字序列，基于本文的方法得到蛋白质序列对应的特征向量，通过特征向量之间的中间距离对蛋白质序列进行相似性分析并聚类。

2. 材料

自流感病毒 H1N1 出现以来，世界各地的人们对其进行了研究，通过大量研究表明这种病毒是由禽流感、猪流感和人流感混合而成的。这种病毒的基因由 8 个长短不一的可编码的 10 个病毒蛋白的线状负链 RNA 片段组成。这 10 个病毒蛋白分别是 PB2、PB1、PA、HA、NP、NA、M1、M2、NS1、NS2，其中 NS1 和 NS2 为非结构蛋白外，其他均是结构蛋白[21]。

本文从 NCBI 网站中 Molecular Databases 的 Protein Sequence 下载了在 1902~2013 年全球 22,455 条 H1N1 型流感病毒中，选取了 31 条含有血凝素蛋白的蛋白质序列进行研究，如表 1。

3. 方法

在信号处理中，时域中的序列通常被转换成频域，使一些重要特征直观化。通过这种转换，没有信息丢失，而且一些隐藏的属性可以被揭示。

离散傅里叶变换是较最常见的转换方法之一。对于长度为 N 的信号 $f(n)$，$n=1,2,\cdots,N$。在频率 k 的信号的 DFT 为 $F(k)=\sum_{n=1}^{N} f(n)\mathrm{e}^{-i(2\pi/N)kn}$，$k=1,2,\cdots,N$。频率 k 处的信号的功率谱被定义为 $PS(k)=\left|F(k)\right|^2$，$k=1,2,\cdots,N$。

通过 DFT 功率谱将蛋白质序列转换成相应的数字序列之后，不同长度的数字序列之间进行相似性比较仍然很困难，解决这个问题的一个常用方法是矩向量，将不同长度的数字序列转换为相同维数的距向量，求出向量之间的中间距离矩阵，利用 SAS 软件建立基于距离矩阵的系统聚类树。我们将 PS-M 方法建立在蛋白质序列的不同属性上进行了比较。

Table 1. 31 influenza viruses and their corresponding serial numbers
表 1. 31 条流感病毒及其对应的序号

No.	Virus's name	No.	Virus's name
1	A/New York/4/1918	17	A/Oslo/868/2001
2	A/London/1/1919	18	A/swine/Iowa/H02NJ56391/2002
3	A/Fort Monmouth/1/1947	19	A/swine/Italy/151672-3/2003
4	A/Netherlands/001G1/1950	20	A/swine/North Carolina/00321/2004
5	A/Yamagishi/50	21	A/Massachusetts/6/2006
6	A/Kw/1/1957	22	A/swine/Kansas/01797/2007
7	A/Denver/1957	23	A/Brisbane/59/2007
8	A/swine/Hong Kong/1/1974	24	A/Kisii/5896/2008
9	A/swine/Hong Kong/59/1977	25	A/Tehran/2a/2008
10	A/USSR/90/1977	26	A/Thailand/CU-H1039/2009
11	A/mallard/Marquenterre/Z237/1983	27	A/Singapore/GP1022/2009
12	A/Memphis/12/1986	28	A/Japan/636/2009
13	A/Goroka/2/1990	29	A/Thailand/CU-H2717/2010
14	A/blue-wingedteal/Alberta/141/1992	30	A/swine/England/453/2006
15	A/Tokushima/20/1996	31	A/Shiraz/11/2013
16	A/swine/Hong Kong/5273/1999		

3.1. 符号序列的数字表达 HP 模型

3.1.1. 基于 20 种氨基酸的功率谱

对于一个长度为 N 的蛋白质序列 $S=s_1s_2\cdots s_N$，序列中 s_n $(n=1,2,\cdots,N)$属于一个有限的符号集合 $T=\{A_1,A_2,\cdots,A_{20}\}$，其中 A_t 为 20 种氨基酸中的一种，s_n 是 $A_1,A_2,\cdots,A_{20}$ 中的某个字母，符号序列 $s_1,s_2,\cdots,s_N$ 中符号 A_t 的指示函数为[22]：

$$u_{A_t}(n)=\begin{cases}1 & s_n=A_t\\ 0 & \text{其他}\end{cases}\qquad n=1,2,\cdots,N \tag{1}$$

例如，蛋白质序列 EVLVLWGVHHPPTGTDQQS，核苷酸 V 的相应指示剂序列是 $u_{A_{15}}=01010001000000000000$。

通过指示函数得到 20 个长度为 N 的二进制数列设为 $w_1,w_2,\cdots,w_{20}$，那么符号序列 $s_1,s_2,\cdots,s_N$ 可以表示为 $\sum_{t=1}^{20}u_{A_t}(1)w_t,\sum_{t=1}^{20}u_{A_t}(2)w_t,\cdots,\sum_{t=1}^{20}u_{A_t}(N)w_t$，记 $s(n)=\sum_{t=1}^{20}u_{A_t}(n)w_t$，$n=1,2,\cdots,N$。

因此蛋白质序列对应的离散傅里叶变换为:

$$\begin{aligned}S(k)&=\sum_{n=1}^{N}s(n)\mathrm{e}^{-\frac{i2\pi}{N}kn}\\&=\sum_{n=1}^{N}u_{A_1}(n)w_1\mathrm{e}^{-\frac{i2\pi}{N}kn}+\sum_{n=1}^{N}u_{A_2}(n)w_2\mathrm{e}^{-\frac{i2\pi}{N}kn}+\cdots+\sum_{n=1}^{N}u_{A_{20}}(n)w_{20}\mathrm{e}^{-\frac{i2\pi}{N}kn}\\&=U_{A_1}(k)w_1+U_{A_2}(k)w_2+\cdots+U_{A_{20}}(k)w_{20}\end{aligned} \tag{2}$$

其中 $U_{A_t}(n)$ 为 $u_{A_t}(n)$ 的离散傅里叶变换，即

$$U_{A_t}(k)=\sum_{n=1}^{N}u_{A_t}(n)\mathrm{e}^{-\frac{i2\pi}{N}kn}=\sum_{n=1}^{N}u_{A_t}(n)\left(\cos\frac{2\pi nk}{N}-i\sin\frac{2\pi nk}{N}\right),\quad t=1,2,\cdots,20,\quad k=1,2,\cdots,N \tag{3}$$

因此，得到数列的离散傅里叶变换为

$$S(k)=\sum_{t=1}^{20}U_{A_t}(k)w_t,\quad k=1,2,\cdots,N \tag{4}$$

定义 u_{A_t} 的功率谱为 $P_{A_t}(k)=\left|U_{A_t}(k)\right|^2$，$k=1,2,\cdots,N$，原蛋白质序列的功率谱函数为

$$P(k)=\sum_{t=1}^{20}P_{A_t}(k),\quad k=1,2,\cdots,N \tag{5}$$

3.1.2. 基于电荷和极性性质的功率谱

蛋白质序列的经典 HP 模型是以构成蛋白质序列的氨基酸的结构分类到物化特征间的对应关系为基础，将 20 种氨基酸分为 4 大类，分别是极性且亲水性(pq)极性且疏水性(pr)、非极性且亲水性(sq)和非极性且疏水性(sr)，$pq=\{G\}$，$pr=\{A,V,L,I,F,P\}$，$sq=\{S,C,N,E,T,Q,K,R,H\}$ 和 $sr=\{W,Y,M\}$。这也为蛋白质序列的结构与功能的研究提供了新思路，蛋白质序列的组成相似，进而推测出它们的结构和功能也相似，这就是经典 HP 模型的意义所在[23]。

经过分类之后，对任意一个长度为 N 的蛋白质序列 $S=s_1s_2\cdots s_N$，其中 s_i，$i=1,2,\cdots,N$ 为 20 种氨基酸中的某一种，进行数据化定义，以非极性氨基酸(NP)为例说明：

$$u_{NP}=\begin{cases}1, & s_i\in pq\\ 2, & s_i\in pr\\ -1, & s_i\in sq\\ -2, & s_i\in sr\end{cases} \tag{6}$$

显然，u_{pq} 是一长度为 N 的二进制的数列，将 20 个氨基酸一一对应于 4 个不同的向量 w_1,w_2,w_3,w_4。利用离散的傅里叶变换，可将指示函数得到的蛋白质序列数据离散化：

$$U_{pq}(n)=\sum_{n=1}^{N}u_{pq}(n)\mathrm{e}^{-i\frac{2\pi}{N}kn},\quad n=1,2,\cdots,N,\quad k=1,2,\cdots,N \tag{7}$$

序列的功率谱：$P_{pq}(k)=\left|U_{pq}(k)\right|^2$，$k=1,2,\cdots,N$，同样可以得到 $P_{pr}(k)$，$P_{UP}(k)$ 和 $P_{PP}(k)$。原蛋白质序列的功率谱为

$$P(k)=P_{pq}(k)+P_{pr}(k)+P_{sq}(k)+P_{sr}(k),\quad k=1,2,\cdots,N \tag{8}$$

定义 j 阶距[21]

$$M_j^{pq}=\frac{1}{N_{pq}^{j-1}\left(N-N_{pq}\right)^{j-1}}\sum_{k=1}^{N/2}\left(P_{pq}(k)\right)^j \tag{9}$$

同样可以求得 M_{pr}^{j}，M_{sq}^{j} 和 M_{sr}^{j}。我们的实验结果表明 $j=1,2,3,4,5$ 对于精确聚类来说是足够的。因此，每个蛋白质序列可以在 20 维欧氏空间中作为几何点来实现，即

$$\left(M_1^{pq},M_1^{pr},M_1^{sq},M_1^{sr},M_2^{pq},M_2^{pr},M_2^{sq},M_2^{sr},M_3^{pq},M_3^{pr},M_3^{sq},M_3^{sr},M_4^{pq},M_4^{pr},M_4^{sq},M_4^{sr},M_5^{pq},M_5^{pr},M_5^{sq},M_5^{sr}\right)$$

3.1.3. 聚类

聚类分析在数据分析领域应用甚广，如在数据挖掘、生物信息学和统计学等领域中扮演这非常重要的角色。聚类分析不仅可以达到物以类聚的效果，还可以探索和提取数据中隐含的新规律和新知识。本文将基于 Q 型系统聚类法，对所获得的数据进行聚类分析。设 n 个样本构成的有限集为 $X=\{x_1,x_2,\cdots,x_n\}$，$d=d(x_i,x_j)$ $(x_i,x_j\in X)$ 是任意两个样本之间的中间距离，记 $D=\{d(x_i,x_j)\,|\,x_i,x_j\in X\}=\{d_0,d_1,d_2,\cdots,d_m\}$，其中 $d_0=0<d_1<\cdots<d_m$。

4. 结果分析与讨论

根据公式(1)和公式(6)将 31 条含有血凝素蛋白的蛋白质序列转换为二进制序列和四元序列，利用离散的傅里叶变换及上述的二进制序列和四元序列，可将蛋白质序列数据离散化。由于不同长度的蛋白质序列通过傅里叶变换转换得到的数字序列的长度依然不同，使得分析蛋白质序列之间的相似性仍然很困难，为了解决这一难题，依据公式(9)将不同维数的特征向量转换维相同维数的特征向量，以此来达到蛋白质序列相似性分析的目的。如基于 20 种氨基酸的功率谱得到血凝素蛋白质的 20 维特征向量，表 2 是 6 种血凝素蛋白质序列的 20 种氨基酸的部分氨基酸数据(由于篇幅的问题这里不一一列举)。

应用 SAS 软件对 31 条血凝素蛋白质序列进行 Q 型系统聚类，根据上述的特征向量矩阵，先将各研究样本看成单独的一类，确定样本之间的‘距离’公式，再计算新样本与其他类之间的距离(本文采用中间距离法)，重复此过程，直到将所有的变量都找到各自的类别，最后通过 SAS 软件得到相应的聚类图定义，见图 1、图 2。

在流感病毒编码的 10 种病毒蛋白质中，本章选取了有血凝素蛋白质的病毒进行了研究。图 1 是依据 20 种氨基酸构造成 20 维特征向量得到的流感病毒蛋白质序列的聚类图，图 2 是依据氨基酸的四种理化

性质并通过数学力矩函数的思想构造了 20 维特征向量通过聚类得到流感病毒蛋白质序列的聚类图。例如，两者将 31 条 H1N1 病毒血凝素蛋白质序列分为不同类，图 1 是基于 20 种氨基酸对血凝素蛋白质序列进行聚类，分类结果为(1)，(2)，(3)，(5)，(13)，(15)，(25)，(26)，(28)，(29)，(8，16、20、22、30)，(4、6、7、10、12)，(17、21、23、24)，(11、14、19)，(9、18、27、31)；图 2 是基于氨基酸的四类理化性质对血凝素蛋白质序列进行聚类，分类结果为(1)，(2)，(3)，(5)，(13)，(15)，(25)，(26)，(29)，(8、17、21、23、30)，(16、20、22、24、28)，(4、6、10、11、9、12、14、19)，(7、18、27、31)。由图 1 和图 2 可知，基于蛋白质序列的不同特征属性，应用本文的方法对血凝素蛋白质序列进行分类的差异较小。将这两种分类结果与文献[24]进行比较发现，基于氨基酸的四种理化性质对血凝素蛋白质进行的分类结果更加接近文献[24]。第 9 条病毒蛋白质与第 4 条病毒蛋白质、第 6 条病毒蛋白质、第 10 条病毒蛋白质、

Table 2. Eigenvector data based on 20 amino acids
表 2. 基于 20 种氨基酸的特征向量数据

	A	W	C	D	E	F	G	H	I	Y	K	L
1	0.0602	0.1431	0.0000	0.0970	0.1431	0.2817	0.0740	0.1431	0.2817	0.0602	0.0970	0.0602
2	0.4183	0.7457	1.1064	0.8900	0.3857	0.7457	0.3139	1.4672	0.7457	0.3857	0.3857	0.2802
3	4.4910	22.004	17.650	5.5180	3.8000	5.5180	4.0310	17.650	3.9510	7.2050	3.8740	3.3590
4	2.6988	7.5539	5.2405	4.2231	2.1425	4.6752	2.0954	6.9369	2.4814	3.4097	2.1919	1.7558
5	2.7949	4.6770	6.6369	2.0944	1.8656	2.0944	2.1845	3.6220	1.1994	3.1533	1.9357	1.1276
6	2.6121	7.3109	5.0720	3.7295	2.1718	4.2946	2.0738	6.7138	2.7759	3.3002	2.4677	1.6996

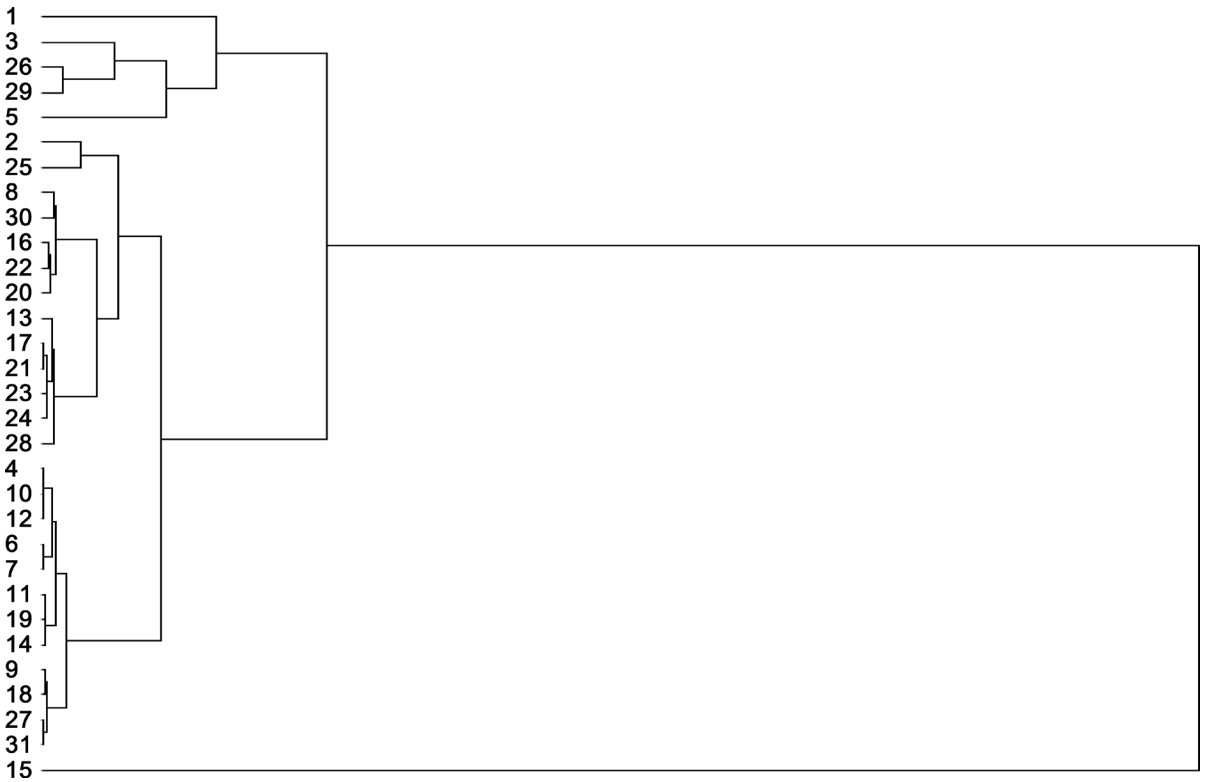

Figure 1. A clustering diagram based on the power spectrum of 20 amino acids
图 1. 基于 20 种氨基酸的功率谱的聚类图

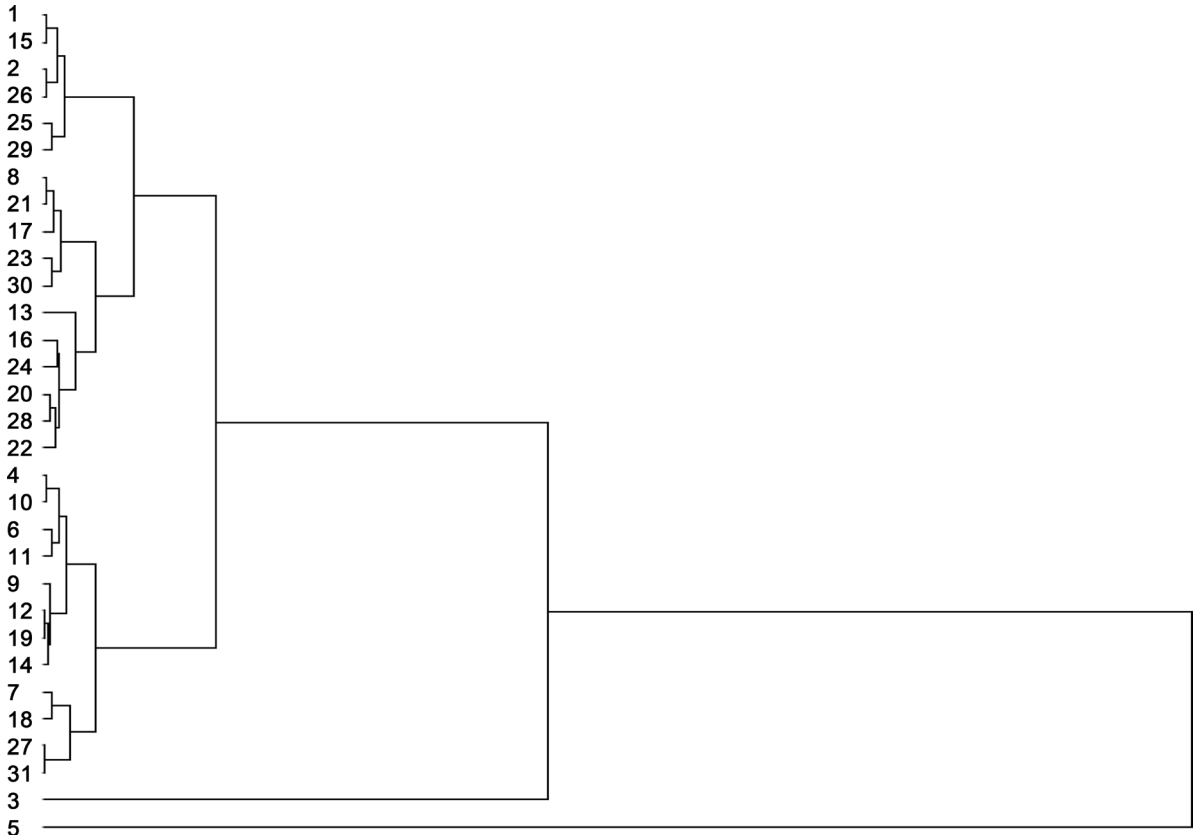

Figure 2. Clustering of power spectrum based on charge and polarity properties
图 2. 基于电荷和极性性质的功率谱的聚类图

第 11 条病毒蛋白质、第 12 条病毒蛋白质、第 14 条病毒蛋白质和第 19 条病毒蛋白质属于同一类，第 28 条病毒蛋白质与第 24 条病毒蛋白质属于同一类，以氨基酸的四种理化性质所进行的分类结果与文献[24]高度相似，而基于 20 种氨基酸所得到的结果并没有做到这一点。第 8 条病毒蛋白质与第 16 条病毒蛋白质、第 20 条病毒蛋白质和第 22 条病毒蛋白质属于同一类，以氨基酸的四种理化性质所进行的分类结果没有做到这一点，而第 7 条和第 30 条在不同特征下的聚类不一样。在本章中，我们只选取了蛋白质的 20 种氨基酸和四种理化性质对蛋白质序列进行研究，尽管得到的结论与文献[24]中的结果很相似，但是在大数据的处理过程中综合应用蛋白质的性质越多，对蛋白质的相似性比较越准确，这是我们在今后的研究中重点进行的工作。此外，傅里叶功率谱构造特征向量来表征蛋白质序列，并结合其数字编码的蛋白质可以完全包含序列的所有信息，可自动提取蛋白质序列特征信息，这正是本章内容研究的重点。

5. 总结

本章基于蛋白质二维数字表达结合高维共鸣识别法判别双序列蛋白质的相似性和在频率域上表示 DNA 序列的基础上，提出了应用傅里叶功率谱分析多个蛋白质序列的相似性。将 DNA 序列上传统的研究方法转换到研究蛋白质序列上，主要包括：在经典的 HP 模型之上，以 20 种氨基酸和氨基酸的四种理化性质为基础上将蛋白质序列数值化。在此基础上，通过离散的傅里叶变换将数字序列离散化，为了统一离散化序列的维数，再根据定义计算序列的功率谱，并构造向量矩阵，计算中间距离。在上述的基础上，采用系统聚类算法获取分层结构，构造聚类树讨论蛋白质序列的相似性。

本章选取了 31 条 H1N1 病毒血凝素蛋白质序列对提出的方法进行验证，在不同属性的基础上，经过

反复的验证，将多维的数字序列进行降维，最终我们采用 20 维的特征向量表征整条蛋白质序列，利用系统聚类算法，对 31 条蛋白质序列进行分类，实验结果与文献[24]高度吻合。因此将基于蛋白质二维数字表达结合高维共鸣识别法判别双序列蛋白质的相似性和在频率域上表示 DNA 序列方法的结合扩展到对多个蛋白质序列在频率域上的特征的提取，这为研究蛋白质序列提供了更为严谨的方法。这些对基于大数据和结构分析的研究具有积极的意义，将大大降低计算的复杂度。

致　谢

感谢基金项目：辽宁省教育厅科学研究一般项目(No. L2015093)对本论文的支持。同时，也要衷心的感谢本文中引用文章的作者。

参考文献

[1] Katoh, K., Misawa, K.K.-I. and Miyata, T. (2002) Mafft: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research*, **30**, 3059-3066. https://doi.org/10.1093/nar/gkf436

[2] Edgar, R.C. (2004) Muscle: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, **32**, 1792-1797. https://doi.org/10.1093/nar/gkh340

[3] Larkin, M.A., Blacksshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007) Clustal w and Clustal x Version 2.0. *Bioinformatics*, **23**, 2947-2948. https://doi.org/10.1093/bioinformatics/btm404

[4] Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., *et al.* (2003) The Genome Sequence of the Sars-Associated Coronavirus. *Science*, **300**, 1399-1404. https://doi.org/10.1126/science.1085953

[5] Vinga, S. and Almeida, J. (2003) Alignment-Free Sequence Comparison—A Review. *Bioinformatics*, **19**, 513-523. https://doi.org/10.1093/bioinformatics/btg005

[6] Yau, S.S.-T., Yu, C. and He, R. (2008) A Protein Map and Its Application. *DNA and Cell Biology*, **27**, 241-250. https://doi.org/10.1089/dna.2007.0676

[7] Yu, C., Deng, M. and Yau, S.S.-T. (2011) DNA Sequence Comparison by a Novel Probabilistic Method. *Information Sciences*, **18**, 1484-1492. https://doi.org/10.1016/j.ins.2010.12.010

[8] Yu, C., Hernandez, T., Zheng, H., Yau, S.-C., Huang, H.-H., He, R.L., Yang, J. and Yau, S.S.-T. (2013) Real Time Classification of Viruses in 12 Dimensions. *PloS One*, **8**, e64328. https://doi.org/10.1371/journal.pone.0064328

[9] Pandit, A. and Sinha, S. (2010) Using Genomic Signatures for HIV-1 Sub-Typing. *BMC Bioinformatics*, **11**, S26. https://doi.org/10.1186/1471-2105-11-S1-S26

[10] Blaisdell, B.E. (1989) Average Values of a Dissimilarity Measure Not Requiring Sequence Alignment for a Computer-Generated Model System. *Journal of Molecular Evolution*, **29**, 538-547. https://doi.org/10.1007/BF02602925

[11] Anastassiou, D. (2000) Frequency-Domain Analysis of Biomolecular Sequences. *Bioinformatics*, **16**, 1073-1081. https://doi.org/10.1093/bioinformatics/16.12.1073

[12] Kotlar, D. and Lavner, Y. (2003) Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein Coding Regions. *Genome Research*, **13**, 1930-1937. https://doi.org/10.1101/gr.1261703

[13] Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T., Kodo, Y., Mori, H. and Kanaya, S. (2002) Periodicity in Prokaryotic and Eukaryotic Genomes Identified by Power Spectrum Analysis. *Gene*, **300**, 203-211. https://doi.org/10.1016/S0378-1119(02)00850-8

[14] Yin, C. and Yau, S.S.-T. (2005) A Fourier Characteristic of Coding Sequences: Origins and a Non-Fourier Approximation. *Journal of Computational Biology*, **12**, 1153-1165. https://doi.org/10.1089/cmb.2005.12.1153

[15] Yin, C. and Yau, S.S.-T. (2007) Prediction of Protein Coding Regions by the 3-Case Periodicity Analysis of DNA Sequence. *Journal of Theoretical Biology*, **247**, 687-694. https://doi.org/10.1016/j.jtbi.2007.03.038

[16] Steinbach, M., Karypis, G. and Kumar, V. (2002) A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*, 1-20.

[17] Hall, L.O. (2013) Exploring Big Data with Scalable Soft Clustering. Springer, Berlin Heidelberg, 11-15. https://doi.org/10.1007/978-3-642-33042-1_2

[18] 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.

[19] Zhao, B., Duan, V. and Yau, S.S.T. (2011) A Novel Clustering Method via Nucleotid-Based Fourier Power Spectrum Analysis. *Journal of Theoretical Biology*, **279**, 83-89. https://doi.org/10.1016/j.jtbi.2011.03.029

[20] 赵剑, 阮越, 王嘉松. 数学结构的蛋白质二维数字表达及其应用[J]. 数据采用与处理, 2013, 28(11): 770-776.

[21] 梁启浩, 李阳, 唐旭清. 基于功率谱的流感病毒蛋白质序列结构分析[J]. 病毒学报, 2017, 33(3): 313-319.

[22] Hoang, T., Yin, C., Zheng, H., Yu, C., He, R.L. and Yan, S.S.T. (2015) A New Method to Cluster DNA Sequences using Fourier Power Spectrum. *Journal of Theoretical Biology*, **372**, 135-145. https://doi.org/10.1016/j.jtbi.2015.02.026

[23] 靳佩轩, 高洁. 流感病毒组成蛋白质序列的分析与预测[J]. 食品与生物技术学报, 2016, 35(4): 393-398.

[24] 李巍巍, 李阳, 唐旭清. 不同特征描述下 H1N1 病毒血凝素蛋白质序列的比较分析[J]. 生命科学研究, 2016, 20(2): 119-124.