DOI:10.12113/202209010

一种快速非比对的蛋白质序列相似性与进化分析方法

艾 亮,冯 杰*
(中央民族大学 理学院,北京 100081)

摘 要:本文提出了一种新的快速非比对的蛋白质序列相似性与进化分析方法。在刻画蛋白质序列特征时,首先将氨基酸的10种理化性质通过主成分分析浓缩为6个主成分,并且将每条蛋白质序列里的氨基酸数目作为权重对主成分得分值进行加权平均,然后再融合氨基酸的位置信息构成一个26维的蛋白质序列特征向量,最后利用欧式距离度量蛋白质序列间的相似性及进化关系。通过对3个蛋白质序列数据集的测试表明,本文提出的方法能将每条蛋白质序列准确聚类,并且简便快捷,说明了该方法的有效性。

关键词:蛋白质序列;主成分分析;相似性;系统进化树
中图分类号:Q516 **文献标志码**:A **文章编号**:1672-5565(2023)03-179-08

A fast alignment-free method for protein sequence similarity and evolution analysis

AI Liang, FENG Jie*
(*School of Science*, *Minzu University of China*, *Beijing* 100081, *China*)

Abstract: In this paper, we propose a new fast alignment-free method for protein sequence similarity and evolution analysis. First, 10 groups of physicochemical properties of amino acids are reduced to 6 principal components using principal component analysis, and the number of amino acids in each protein sequence is used as weights to the scores of the principal components. Then, the amino acid position information is fused to form a 26-dimension feature vector for each protein sequence. Finally, the Euclidean distance is used to measure the similarity and evolutionary distance between protein sequences. The test on three datasets shows that our method can cluster each protein sequence accurately, which illustrates the validity of our method.

Keywords: Protein sequences; Principal component analysis; Similarity; Phylogenetic trees

生物序列的相似性分析是生物信息学的重要研究方向之一。在早期研究中,通常采用多序列比对的方法对序列进行比较分析,许多算法现在已经非常成熟[1-3],例如使用较多的 ClustalW 算法。但多序列比对是基于同源序列片段间是邻接保守的假设,这与遗传重组相冲突,而且当样本量较大或序列长度较长时,比对算法的时间成本很高。因此,非比对方法[4]一经推出,立即受到研究人员的广泛关注。非比对方法不是具体比较基对,而是将序列看成是一个整体并将其转化为数值向量再进行分析比较,其优点是在计算机上计算迅速,且结果较准确。

蛋白质序列的比较分析方法大致分为两大类:图形表示方法和数值向量刻画方法。图形表示方法也称为可视化方法,其基本思想是建立一组映射,将氨基酸映射成平面或空间的点,然后将点连接起来得到空间曲线。进一步地,我们还可以从这些图形表示中提取生物序列的数值特征,利用这些数值特征进行序列分析[5-12]。数值向量刻画方法主要是将蛋白质序列转换为多维的数值向量,例如 Chou K. C.[13]和 Chen W.等[14]将氨基酸的20维频率向量与理化性质或者氨基酸之间的相互作用结合起来构建(20 + λ)维向量来表示蛋白质序列,其中 λ 指的是理化性质个数或者氨基酸相互之间作用的指标数。贾美多等[15]结合氨基酸的5-字母分类模型和序列

收稿日期:2022-09-10;**修回日期**:2022-10-25;**网络首发日期**:2022-11-11.
网络首发地址:https://kns.cnki.net/kcms/detail/23.1513.Q.20221110.1245.002.html
* **通信作者**:冯杰,女,讲师,研究方向:计算生物学和生物信息学. E-mail:fengjie0536@163.com.
引用格式:艾亮,冯杰.一种快速非比对的蛋白质序列相似性与进化分析方法[J].生物信息学,2023,21(3):179-186.
AI Liang, FENG Jie. A fast alignment-free method for protein sequence similarity and evolutionary analysis[J]. Chinese Journal of Bioinformatics, 2023, 21(3): 179-186.

的 k-字节模型,提取信息将序列转化为一个 30 维向量,之后利用欧氏距离求得蛋白质序列两两间的相对距离进而构建系统进化树。Xian-HuaXie 等[16]使用氨基酸随机和独立放置的序列分布图之间的相对偏差来定义序列间的差异。Li Y 等[17]结合氨基酸的概率、平均出现位置概率和两个相邻氨基酸的马尔科夫转移概率分布来构建蛋白质数值向量表示。Yongkun Li 等[18]将蛋白质序列中 20 种氨基酸的数目、平均位置和位置的正则化中心二阶矩结合起来构成 60 维数值向量来衡量病毒之间的相似性。Lily He 等[19]基于氨基酸的亲水性指数、极性需求和侧链的化学成分将氨基酸分成 8 类,之后融合蛋白质序列中这 8 类氨基酸的数量、平均位置和位置的二阶矩信息构建 24 维特征向量进行进化分析。朱臣臣等[20]选择 3 种氨基酸的理化性质绘制蛋白质序列的 3D 图形,再基于氨基酸的位置信息构建 20 个点集,分别求其转动惯量构建 23 维特征向量。Stephen S.-T. Yau 等[21]、Qi Dai 等[22]和 Yufeng Liu 等[23]统计序列中所有的长度为 k 的子串的频率,将这些数字组成向量,使用该向量刻画生物序列的特征。

蛋白质是由氨基酸组成的,已有研究表明氨基酸的物理化学性质对蛋白质序列分类和进化具有重要意义[24-25]。本文将氨基酸的 10 种理化性质通过主成分分析浓缩为 6 组主成分,对每条蛋白质序列,计算反映氨基酸理化性质的 6 组主成分得分均值,再融合 20 个氨基酸的位置信息构成一个 26 维的蛋白质序列特征向量,最后利用欧式距离度量蛋白质序列间的相似性并构造系统进化树。通过对三组蛋白质序列数据集的测试表明,本文的方法能将每条蛋白质序列准确聚类,结果与现有进化关系一致,说明了该方法的有效性。

1 蛋白质序列的特征向量构造

1.1 基于氨基酸理化性质的向量表示

蛋白质的基本单位是氨基酸,每种氨基酸都具有多种理化性质,氨基酸的理化性质对蛋白质的结构和功能起着重要的作用。本文考虑氨基酸的 10 种理化性质:解离常数($pK_a(NH_3^+)$ 和 pK_a)、等电点(pI)、相对分子质量(Mw)、旋光率($[a]_D(H_2O)$ 和 $[a]_D(HCl)$)、极性需求(Pr)、侧链的化学成分(Cc)、疏水值(Hb)和侧链质量(Scm),具体数值见表 1。

表 1 氨基酸的 10 种理化性质

Table 1 10 physicochemical properties of amino acids

氨基酸(aa)	pK_a (NH_3^+)	pI	Mw	$[a]_D$ (H_2O)	$[a]_D$ (HCl)	pK_a	Pr	Cc	Hb	Scm
A	9.69	6.02	89.06	1.8	14.6	0	7	0	0.62	15
R	9.04	10.76	174.4	12.5	27.6	1	9.1	0.65	-2.53	101
N	8.8	5.41	132.6	-5.3	33.2	0	10	1.33	-0.78	58
D	9.82	2.97	133.6	5	25.4	1	13	1.38	-0.9	59
C	10.78	5.02	121.12	-16.5	6.5	1	4.8	2.75	0.29	47
Q	9.13	5.65	146.08	6.3	31.8	0	8.6	0.89	-0.85	72
E	9.67	3.22	147.08	12	31.8	1	12.5	0.92	-0.74	73
G	9.6	5.97	75.05	0	0	0	7.9	0.74	0.48	1
H	9.17	7.59	155.09	-38.5	11.8	1	8.4	0.58	-0.4	82
I	9.68	6.02	131.11	12.4	39.5	0	4.9	0	1.38	57
L	9.6	5.98	131.11	-11	16	0	4.9	0	1.06	57
K	8.95	9.74	146.13	13.5	26	1	10.1	0.33	-1.5	73
M	9.21	5.75	149.15	-10	23.2	0	5.3	0	0.64	75
F	9.13	5.48	165.09	-34.5	-4.5	0	5	0	1.19	91
P	10.6	6.3	115.08	-86.2	-60.4	0	6.6	0.39	0.12	42
S	9.15	5.68	105.06	-7.5	15.1	0	7.5	1.42	-0.18	31
T	10.43	6.53	119.18	-28.5	15	0	6.6	0.71	-0.05	45
W	9.11	5.89	204.11	-33.7	2.8	0	5.2	0.13	0.81	130
Y	9.11	5.66	181.09	0	-10	1	5.4	0.2	0.26	107
V	9.39	5.97	117.09	5.6	28.3	0	5.6	0	1.08	43

为消除量纲不一的影响,先对 10 组氨基酸理化性质进行标准化处理,将其化为均值为 0,标准差为 1 的数据框。然后对该 20 × 10 的氨基酸理化性质矩阵进行主成分分析,将 10 组变量的信息压缩为几个综合变量,提取有效的主成分来表示 20 种天然氨基酸的理化性质。主成分分析结果见表 2。

表 2　重要主成分的贡献率

Table 2　Contribution of significant principal components

主成分	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
标准离差	1.789 8	1.558 3	1.313 8	1.054 2	0.817 5	0.673 2
贡献率	0.320 3	0.242 8	0.172 6	0.111 1	0.066 8	0.045 3
累计贡献率	0.320 3	0.563 2	0.735 8	0.846 9	0.913 8	0.959 1

由表 2 可以看到,前 6 个主成分的累计贡献率为 95.91%,远大于 85%,可以认为这 6 个主成分能代表原先 10 组理化性质的绝大部分信息。计算这 6 个主成分的得分,即把原来的 20 × 10 的氨基酸理化性质矩阵转化为 20 × 6 的主成分得分矩阵(见表 3)。

表 3　主成分得分矩阵

Table 3　Principal component score matrix

氨基酸(aa)	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
A	−1.568	0.953	−1.485	−0.843	−0.127	0.754
R	3.870	−0.649	0.720	−1.959	0.782	−0.211
N	1.103	0.965	−0.701	0.176	−1.042	−1.421
D	1.485	2.510	0.856	1.520	−0.909	0.658
C	−1.196	2.194	2.204	1.122	2.382	−0.664
Q	1.137	0.310	−0.748	0.261	−0.462	−0.906
E	1.888	1.715	0.388	1.652	−0.796	0.945
G	−1.785	1.831	−0.949	−1.052	−0.365	0.207
H	1.132	−0.788	1.278	−0.535	−0.125	0.144
I	−0.820	−0.467	−1.986	0.476	1.024	0.427
L	−1.292	−0.826	−0.974	0.085	0.361	0.273
K	2.898	−0.007	−0.062	−1.815	0.243	0.534
M	−0.359	−1.311	−1.054	0.401	0.030	−0.108
F	−1.024	−2.378	0.047	0.675	−0.477	−0.177
P	−3.623	−0.505	3.128	−1.387	−1.160	0.285
S	−0.520	1.322	−0.581	−0.329	−0.308	−1.162
T	−1.451	0.618	0.443	−0.544	0.430	−0.165
W	0.144	−3.341	0.524	1.204	−0.278	−0.638
Y	0.942	−1.975	0.817	0.924	0.437	0.861
V	−0.961	−0.170	−1.865	−0.030	0.360	0.363

对于任一长度为 n 的蛋白质序列 S,计算蛋白质序列 S 的各主成分平均值 $\overline{PC_j}$:

$$\overline{PC_j} = \sum_{i=1}^{20} \frac{n_i PC_{ij}}{n} \qquad j = 1,2,\cdots,6 \qquad (1)$$

其中 $\Omega = \{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$,$A_i(i = 1,2,\cdots,20)$ 表示 Ω 中的第 i 个氨基酸,n_i 表示序列 S 中氨基酸 A_i 的数量,PC_{ij} 表示氨基酸 A_i 的第 j 个主成分得分(见表 3)。对每条序列经计算可得到 6 维氨基酸主成分得分平均值向量 $(\overline{PC_1},\overline{PC_2},\cdots,\overline{PC_6})$。

例如对于蛋白质序列 MTMHTTMTTLTLTSL,$n_M = 3, n_T = 7, n_H = 1, n_L = 3, n_S = 1$,则 $\overline{PC_1} = \dfrac{-0.359 \times 3 - 1.451 \times 7 + 1.132 \times 1 - 1.292 \times 3 - 0.52 \times 1}{15} =$ −0.966 5,$\overline{PC_2},\cdots,\overline{PC_6}$ 类似可求得。由式(1)得知,氨基酸主成分得分平均值向量是以 20 种天然氨基

酸的数量为权重,对 6 个主成分进行加权平均而来,而 6 个主成分又是通过 10 组氨基酸的理化性质浓缩而来,因此,6 维氨基酸主成分得分平均值向量同时包含了氨基酸数量和理化性质信息。

1.2 基于氨基酸平均位置的向量表示

对于一条长度为 n 的蛋白质序列 $S=(s_1,s_2,\cdots,s_n)$,其中 $s_j \in \Omega, j=1,2,\cdots,n$,还可以基于每种氨基酸 $A_i(i=1,2,\cdots,20)$ 的平均位置[19]构造一个 20 维的特征向量,如下所示:

$$\mu_i=\frac{\sum_{j=1}^{n} j\times I(s_j=A_i)}{n_i} \tag{2}$$

其中 n_i 表示对应氨基酸的数量,$I(s_j=A_i)=\begin{cases}1, s_j=A_i\\ 0, s_j\neq A_i\end{cases}$,由此可以得到反映序列中氨基酸位置信息的 20 维氨基酸平均位置特征向量 $(\mu_1,\mu_2,\cdots,\mu_{20})$。例如对于蛋白质序列 MTMHTTMTTLTLTSL,由 $\Omega=\{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$ 知 $A_{13}=M$,计算它的平均位置,由于 $n_{13}=3$,有 $\mu_{13}=\frac{1\times1+3\times1+7\times1}{3}=\frac{11}{3}$,类似可求 μ_9、μ_{11}、μ_{16}、μ_{17},其它未出现在序列中的氨基酸的平均位置为 0。

1.3 蛋白质序列的特征向量

利用上述构建的两组特征向量,对每条蛋白质序列,将 6 维氨基酸主成分得分平均值向量和 20 维氨基酸平均位置向量结合起来可以得到一个 26 维蛋白质序列的数值化向量表示 $(\overline{PC_1},\overline{PC_2},\cdots,\overline{PC_6},\mu_1,\mu_2,\cdots,\mu_{20})$。由于两组特征的量纲不一,需将 26 维向量进行标准化处理。

2 蛋白质序列的相似性与进化分析

为验证本文所提方法的有效性,用三组蛋白质序列数据集[19-20]进行实验,利用欧氏距离计算两两蛋白质序列所对应的 26 维特征向量之间的距离,然后利用 UPGMA 算法(该算法已嵌入到 MEGA11 软件)构建生物系统进化树。

2.1 9 物种 ND5 蛋白质序列

9 个物种的 ND5 蛋白质序列信息在表 4 中给出。使用本文的方法,可以得到 9 物种 ND5 蛋白质序列的一个 9 × 26 特征矩阵,然后计算两两间的欧氏距离可以得到相似性距离矩阵,结果见表 5。

表 4 9 物种 ND5 蛋白质序列信息

Table 4 Information on 9 ND5 protein sequences

物种	英文名	登录号	长度/bp
人类	Human	AP_000649	603
普通黑猩猩	Common chimpanzee	NP_008196	603
侏儒黑猩猩	Pigmy chimpanzee	NP_008209	603
大猩猩	Gorilla	NP_008222	603
长须鲸	Fin whale	NP_006899	606
蓝鲸	Blue whale	NP_007066	606
小鼠	Mouse	NP_904338	607
大鼠	Rat	AP_004902	610
负鼠	Opossum	NP_007105	602

表 5 9 物种 ND5 蛋白质序列相似性距离矩阵

Table 5 The similarity/dissimilarity matrix of 9 ND5 protein sequences

Species	Human	C.chim	P.chim	Gorilla	F.whale	B.whale	Mouse	Rat
C.chim	3.146							
P.chim	4.035	**2.679**						
Gorilla	3.701	3.059	4.288					
F.whale	5.889	6.452	6.564	5.899				
B.whale	6.025	6.639	6.990	6.612	2.855			
Mouse	7.769	7.583	7.242	7.568	6.579	6.941		
Rat	9.701	9.304	8.939	9.808	8.830	8.997	5.886	
Opossum	9.057	8.066	7.618	8.250	8.817	8.709	8.137	**9.976**

观察表 5 可以看出,普通黑猩猩和侏儒黑猩猩的相似性距离最小,为 2.679,表示普通黑猩猩和侏儒黑猩猩间的亲缘关系最近;大鼠和负鼠的相似性距离最大,为 9.976,表示大鼠和负鼠间亲缘关系最远。同时,可以看到,人类、普通黑猩猩、侏儒黑猩猩和大猩猩这四个物种间的相似性距离比较小,说明它们的蛋白质序列相似性程度高,进化关系上较为接近;长须鲸和蓝鲸间相似性距离也很小,说明它们的进化关系接近;负鼠和其他八个物种的相似性距离都很大,表明在进化关系上与其它物种相比负鼠

相对比较独立。

进一步利用相似性距离矩阵构建物种进化树,结果如图 1 所示。通过观察发现 9 个物种被分成 4 个分支:第 1 个分支是侏儒黑猩猩、普通黑猩猩、人类和大猩猩,在这一分支中,侏儒黑猩猩和普通黑猩猩进化关系更近,其次是人类,而后是大猩猩,这与进化事实相符合;第 2 个分支是蓝鲸和长须鲸;第 3 分支为大鼠和小鼠;第 4 个分支为负鼠,与其他物种进化关系较远,单独成一个分支。从进化关系上看,侏儒黑猩猩、普通黑猩猩、人类和大猩猩都属于灵长目人科,蓝鲸和长须鲸都属于鲸目须鲸科,大鼠和小鼠都属于啮齿目鼠科,负鼠属于负鼠目负鼠科,本文的分析结果与实际进化关系相一致。

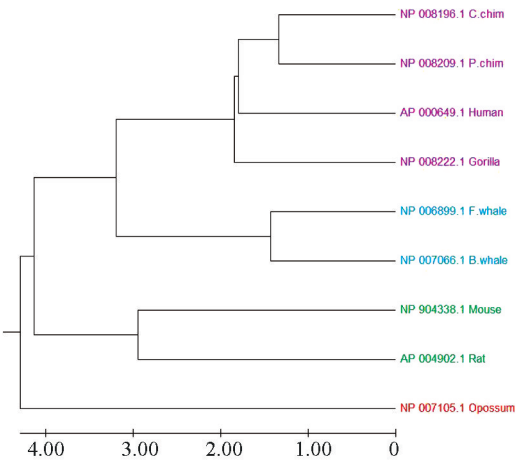

图 1　9 物种 ND5 蛋白质序列的进化树

Fig.1　The phylogenetic tree of 9 ND5 protein sequences

2.2　12 个杆状病毒蛋白质序列

12 个杆状病毒蛋白质序列信息见表 6,使用本文所提方法对其构建进化树,结果见图 2。由图 2 可以看到,Alphabaculovirus 病毒和 Betabaculovirus 病毒被分为两大分支,并且 Alphabaculovirus 中的 Group I 和 Group II 也都形成各自的分支,与实际的病毒进化关系一致。而文献[7]没有将 Alphabaculovirus 病毒和 Betabaculovirus 病毒形成两个大分支,并且 Group II 中的 6 个病毒不在一个分支,HzSNPV、HaSNPV、HearNPV 与 Betabaculovirus 病毒的进化距离要比与 Group II 中的其他病毒的距离要小,这与实际的进化关系不一致。文献[8]和[20]虽然将 Alphabaculovirus 病毒和 Betabaculovirus 病毒形成了两个大分支,但是 Group II 中的 6 个病毒并不在一个分支,Group I 中的 AcMNPV、BmNPV、RoMNPV 各自与 Group II 中的三个病毒形成分支。

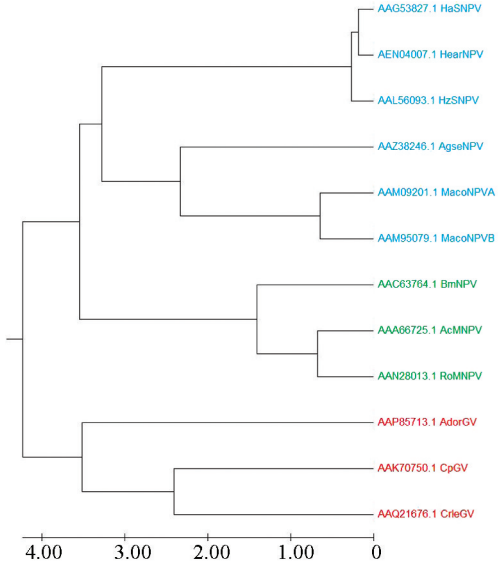

图 2　12 个杆状病毒蛋白质序列的进化树

Fig.2　The phylogenetic tree of 12 baculovirus protein sequences

表 6　12 个杆状病毒蛋白质序列信息

Table 6　Information on 12 Baculovirus protein sequences

属(组)	病毒名称	缩写	登录号	长度/bp
Alphabaculovirus (Group I NPVs)	Autographa californica MNPV	AcMNPV	AAA66725	1 221
	Bombyx mori NPV	BmNPV	AAC63764	1 222
	Rachiplusia ou MNPV	RoMNPV	AAN28013	1 221
	Helicoverpa armigera NPV	HearNPV	AEN04007	1 253
	Helicoverpa zea SNPV	HzSNPV	AAL56093	1 253
Alphabaculovirus (Group II NPVs)	Mamestra configurata NPVA	MacoNPVA	AAM09201	1 212
	Mamestra configurata NPVB	MacoNPVB	AAM95079	1 209
	Helicoverpa armigera SNPV	HaSNPV	AAG53827	1 253
	Agrotis segetum NPV	AgseNPV	AAZ38246	1 213
Betabaculovirus (GVs)	Adoxophyles orona GV	AdorGV	AAP85713	1 138
	Cydia pomonella GV	CpGV	AAK70750	1 131
	Cryptophlebia leucotreta GV	CrleGV	AAQ21676	1 128

2.3 35 个甲型流感病毒蛋白质序列

甲型流感病毒的一些亚型是根据 H(血凝素类型)的编号(H1 到 H18)和 N(神经氨酸酶类型)的编号(N1 到 N11)来标记的,最致命的甲流亚型是 H1N1、H2N2、H5N1、H7N3 和 H7N9,本文选取了 35 个与这些重要亚型相关的蛋白质序列。

使用我们的方法对该蛋白质序列数据集构建进化树,结果见图 3。由图 3 可知,五种最致命的甲型流感病毒亚型 H1N1、H2N2、H5N1、H7N3 和 H7N9 各自形成 5 个分支,35 个病毒都被正确聚类。相比之下,用 ClustalW 方法构建的进化树则有 3 个甲型流感病毒亚型聚类错误,如图 4 所示,其中 A/turkey/VA/505477 - 18/2007 (H5N1), A/turkey/Ontario/FAV110 - 4/2009 (H1N1) 和 A/turkey/Virginia/4135/2014(H1N1)没能正确被聚类。并且在同一台笔记本电脑下,ClustalW 方法完成多序列比对需要花费约7 s,而我们的方法将序列转化为特征向量只需0.17 s。

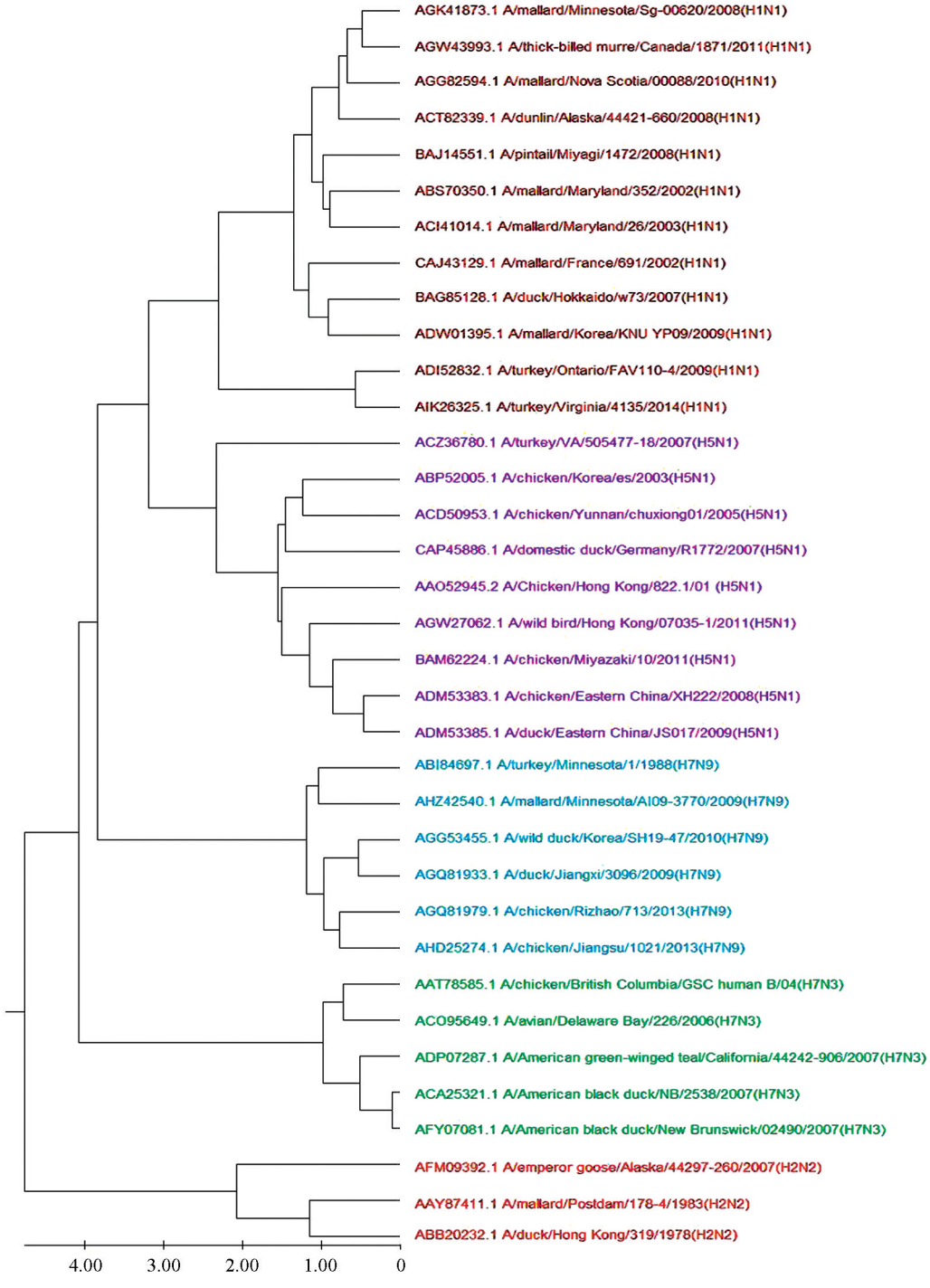

图 3 本文方法构建的 35 个甲型流感病毒蛋白质序列的进化树

Fig.3 The phylogenetic tree of 35 influenza A virus protein sequences constructed using our method

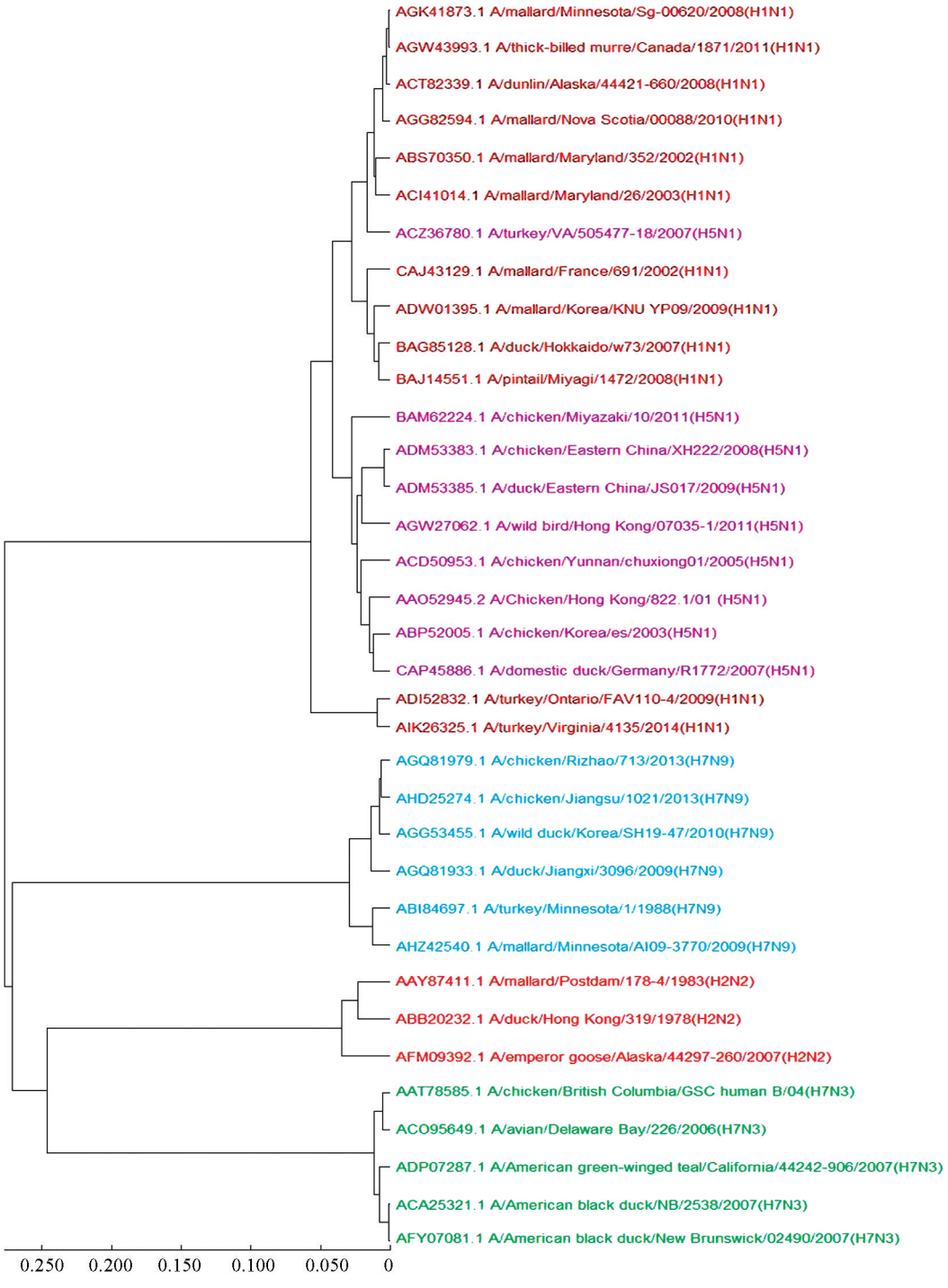

图 4　ClustalW 方法构建的 35 个甲型流感病毒蛋白质序列的进化树

Fig.4　The phylogenetic tree of 35 influenza A virus protein sequences constructed using ClustalW

3　总　结

新的非比对的蛋白质序列相似性分析的方法,将蛋白质序列转化为数值向量时,同时考虑了蛋白质序列中 20 种天然氨基酸的数量、理化性质和平均位置信息,最终将每条蛋白质序列都转化为唯一与之对应的 26 维特征向量。该新方法在 3 个数据集上均获得了准确的聚类结果,这说明该新方法在分析蛋白质序列的相似性方面是有效的。此外,该方法不需要复杂的计算,而且简便快捷。

参考文献(References)

[1] KATOH K, MISAWA K, KUMA K, et al. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform[J]. Nucleic Acids Research, 2002, 30(14): 3059-3066.DOI: 10.1093/nar/gkf436.

[2] THOMPSON J D, GIBSON T J, HIGGINS D G. Multiple

sequence alignment using clustalw and clustalx [J/OL]. Current Protocols in Bioinformatics, (2002-08-01)[2022-09-10]. https://doi.org/10.1002/0471250953.bi0203s00. DOI: 10.1002/0471250953.bi0203s00.

[3] ROBERT C, EDGAR. MUSCLE: Multiple sequence alignment with high accuracy and high throughput[J]. Nucleic Acids Research, 2004, 32(5): 1792-1797. DOI: 10.1093/nar/gkh340.

[4] VINGA S, ALMEIDA J. Alignment-free sequence comparison—a review[J]. Bioinformatics, 2003, 19: 513-523. DOI: 10.1093/bioinformatics/btg005.

[5] 张艳萍, 贺平安. 蛋白质序列的图形表示及其应用[J]. 浙江理工大学学报, 2010, 27(2): 308-314. DOI: 10.3969/j.issn.1673-3851.2010.02.029.
ZHANG Yanping, HE Ping-an. Graphical representation of protein sequences and its applications[J]. Journal of Zhejiang Sci-Tech University, 2010, 27(2): 308-314. DOI: 10.3969/j.issn.1673-3851.2010.02.029.

[6] 潘以红, 钱东, 朱平. 蛋白质序列图形变换及其相似性聚类分析[J]. 生命科学研究, 2018, 22(3): 191-200. DOI: 10.16605/j.cnki.1007-7847.2018.03.003.
PAN Yihong, QIAN Dong, ZHU Ping. Graphical transformation and similarity clustering analysis for protein sequences[J]. Life Science Research, 2018, 22(3): 191-200. DOI: 10.16605/j.cnki.1007-7847.2018.03.003.

[7] YAO Yuhua, YAN Shoujiang, XU Huimin, et al. Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation [J]. Evolutionary Bioinformatics, 2014,10(1):87-96. DOI: 10.4137/EBO.S14713.

[8] HOU Wenbing, PAN Qiuhui, HE Mingfeng. A new graphical representation of protein sequences and its applications [J]. Physica A Statistical Mechanics & Its Applications, 2016, 444: 996-1002. DOI: 10.1016/j.physa.2015.10.067.

[9] HE Pingan, XU Suning, DAI Qi, et al. A generalization of CGR representation for analyzing and comparing protein sequences[J]. International Journal of Quantum Chemistry, 2016, 116(6): 476-482. DOI: 10.1002/qua.25068.

[10] JIE L, KOEHL P. 3D representations of amino acids—applications to protein sequence comparison and classification [J]. Computational and Structural Biotechnology Journal, 2014, 11(18):47-58. DOI: 10.1016/j.csbj.2014.09.001.

[11] HU Hailong, LI Zhong, DONG Hongwei, et al. Graphical representation and similarity analysis of protein sequences based on fractal interpolation[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(1):182-192. DOI: 10.1109/TCBB.2015.2511731.

[12] AGATA C, DOROTA B W, PIOTR W, et al. 20D-dynamic representation of protein sequences [J]. Genomics, 2016, 107:16-23. DOI: 10.1016/j.ygeno.2015.12.003.

[13] CHOU K C. Some remarks on protein attribute prediction and pseudo amino acid composition[J]. Journal of Theoretical Biology, 2011, 273(1): 236-247. DOI: 10.1016/j.jtbi.2010.12.024.

[14] CHEN Wei, LIN Hao, CHOU Kuochen. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences[J]. Molecular BioSystems, 2015, 11(10): 2620-2634. DOI: 10.1039/c5mb00155b.

[15] 贾美多, 杨闫, 张盈盈, 等. 蛋白质序列基于k-字的数值刻画及应用[J]. 浙江农业学报, 2014, 26(6): 1635-1640. DOI: 10.3969/j.issn.1004-1524.2014.06.40.
JIA Meiduo, YANG Yan, ZHANG Yingying, et al. Numerical k-word based portrayal of protein sequences and applications [J]. Acta Agriculturae Zhejiangensis, 2014, 26(6): 1635-1640. DOI: 10.3969/j.issn.1004-1524.2014.06.40.

[16] XIE Xianhua, YU Zuguo, HAN Guosheng, et al. Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles[J]. Molecular Phylogenetics and Evolution, 2015, 89: 37-45. DOI: 10.1016/j.ympev.2015.04.008.

[17] LI Yushuang, SONG Tian, YANG Jiasheng, et al. An Alignment-Free Algorithm in Comparing the Similarity of Protein Sequences Based on Pseudo-Markov Transition Probabilities among Amino Acids[J]. PLoS ONE, 2016, 11(12): e0167430. DOI: 10.1371/journal.pone.0167430.

[18] LI Yongkun, TIAN Kun, YIN Changchuan, et al. Virus classification in 60-dimensional protein space[J]. Molecular Phylogenetics and Evolution, 2016, 99: 53-62. DOI: 10.1016/j.ympev.2016.03.009.

[19] HE L, LI Y, HE R L, et al. A novel alignment-free vector method to cluster protein sequences [J]. Journal of Theoretical Biology, 2017, 427: 41-52. DOI: 10.1016/j.jtbi.2017.06.002.

[20] 朱臣臣, 赵熙强. 基于氨基酸的理化性质和位置信息的蛋白质序列相似性分析方法[J]. 中国海洋大学学报, 2021, 51(增Ⅰ): 95-100. DOI: 10.16441/j.cnki.hdxb.20190110.
ZHU Chenchen, ZHAO Xiqiang. Similarity/dissimilarity analysis of protein sequence based on physicochemical properties and position information of amino acids[J]. Periodical of Ocean University of China (Natural Science Edition), 2021, 51(Sup.Ⅰ): 95-100. DOI: 10.16441/j.cnki.hdxb.20190110.

[21] YAU S S T, YU C, HE R. A protein map and its application[J]. DNA and Cell Biology, 2008, 27(5): 241-250. DOI: 10.1089/dna.2007.0676.

[22] DAI Qi, YANG Yanchun, WANG Tianming. Markov model plus k-word distributions: a synergy that produces novel statistical measure for sequence comparison [J]. Bioinformatics, 2008, 24(20): 2296-2302. DOI: 10.1093/bioinformatics/btn436.

[23] LIU Yufeng, ZENG Jianyang, GONG Haipeng. Improving the orientation-dependent statistical potential using a reference state[J]. Proteins, 2014, 82(10): 2383-2393. DOI: 10.1002/prot.24600.

[24] SALICHOS L, ROKAS A. Inferring ancient divergences requires genes with strong phylogenetic signals [J]. Nature, 2013,497:327-331. DOI:10.1038/nature12130.

[25] WIMLEY W C, WHITE S H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces [J]. Nature Structural Biology, 1996, 3: 842-848. DOI: 10.1038/nsb1096-842.