浙江理工大学学报(自然科学版),第 39 卷,第 4 期,2018 年 7 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 39, No. 4, Jul. 2018

DOI:10.3969/j.issn.1673-3851(n).2018.04.015

基于密码子特征的蛋白质序列图形表示

朱正阳,贺平安
(浙江理工大学理学院,杭州 310018)

摘　要:根据密码子碱基位置特征和氨基酸的疏水性指数值,20 种氨基酸被映射为 3 维空间中的向量,提出一种新的迭代函数关系将氨基酸序列转化为三维空间中的一条曲线,获得一种新的蛋白质图形表示方法。对蛋白质图形,利用闵可夫斯基距离刻画两个 3 维曲线之间的距离,依此推断蛋白序列之间的差异性和物种之间的进化关系。将该方法分别应用在 9 个物种的 ND5 蛋白和 12 个物种的 β-珠蛋白序列分析中,所得结果与 ClustalW 方法的结果以及其他文献中的结果对比后证明该文方法有效可行。

关键词:密码子特征;蛋白质序列;图形表示;相似性分析
中图分类号:O29　　**文献标志码**:A　　**文章编号**:1673-3851 (2018) 07-0474-10

0　引　言

基于 DNA 和蛋白质序列的相似性分析,可以推断出不同物种之间的进化关系。所以分析序列的相似性是生物信息研究的重要课题[1]。在序列的相似性分析研究中,最常用的是序列比对方法,但由于序列比对方法的计算复杂度大。近些年,许多非序列比对方法被提出来并得到使用,其中一种就是图形表示方法[2]。

在生物序列数据库中,DNA 和蛋白质序列是用字母序列表示的。在 DNA 数据库中,DNA 序列是由 A、C、G、T 四种字母组成;在蛋白质数据库中,蛋白质序列是由 A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W、Y 二十种字母组成[3]。图形表示方法是通过数学映射,把字母序列转化为空间图形,构建数学模型刻画图形之间的差异,进而推断物种之间的进化关系。图形表示法以其可视性好,计算简单,容易给出数值刻画等多方面的优点,受到研究者们的普遍关注[4-9]。

最初,Hamori 等[10]利用随机游走,将 DNA 序列转化为 3 维曲线(H-曲线)表示。此后,许多研究沿着这个思路,将序列转化为 1 维、2 维和 4 维曲线[11]。组成蛋白质序列的氨基酸有 20 种,与 DNA 序列的四种核苷酸相比,情况更为复杂,但蛋白质图形表示方法基本上都是 DNA 序列图形表示方法的推广,Rahman 等[11]和 Yu 等[12]所采用的方法均是 DNA 序列图形表示方法的简单拓展。对于映射关系,氨基酸的理化性质[6,9,13-18]和氨基酸的循环排列[2,5,19-22]常常是研究者们建立映射关系的依据。

通过氨基酸的映射关系,构造迭代函数,将序列中的氨基酸映射成空间中的一个点,顺次连接这些点,就构成了蛋白质序列的图形表示[23]。对于迭代关系式,Jeffrey[4]使用相同参数的迭代关系式图形表示方法,随后,许多研究者使用同参数的迭代关系式[3,21,23-26]。此外,Ma 等[8]首次提出异参数的迭代关系式,同样也取得不错的结果。结合 Jeffrey[4]和 Ma 等[8]的观点,He 等[27]提出广义混沌游戏表示方法。在把序列转化为曲线之后,数值刻画对于描述图形是必不可少的。图形表示只是提供视觉上的比

收稿日期:2017－11－19　　网络出版日期:2018－05－15
基金项目:国家自然科学基金项目(61772027)
作者简介:朱正阳(1991－),男,安徽定远,硕士研究生,主要从事生物信息学方面的研究。
通信作者:贺平安,E-mail:pinganhe@zstu.edu.cn

较,但是数值刻画能够量化不同曲线之间的差异性。例如,在 Yao 等[7,28]通过几何中心来描述曲线之间的差异,在 He 等[21]采用距离矩阵的特征值来描述曲线之间的差异等。

本文考虑密码子碱基位置特征和氨基酸的疏水性特征,将氨基酸映射成为一个三维向量。结合 Ma[8]和 He 等[27]结论,本文选取一种新的迭代关系式,获得新的 3 维图形表示方法。此外,为了比较曲线之间的差异,本文计算两个曲线之间的闵可夫斯基距离。应用新的 3 维图形表示方法到 9 个物种的 ND5 蛋白和 12 个物种的 β-珠蛋白序列上,分析它们之间的相似性并推断相应物种之间的进化关系。将相似性分析结果与经典的多重序列比对算法 ClustalW 方法得到的结果做相关性分析,验证本文的方法有效性。

1　蛋白质的 3 维图形表示方法

1.1　氨基酸的三维坐标

四种碱基 A、G、C、T 可以组成 64 种三联体密码子。其中,61 种密码子对应 20 种氨基酸,即通常所谓的遗传密码表。观察密码表可以看出,密码子的第一个位置和第二个位置的碱基几乎可以决定密码子翻译的是何种氨基酸。此外,密码子的第二个位置的碱基与密码子翻译的氨基酸的许多理化性质都有关联。所以,密码子的第二位碱基对于密码子的理化性质有着特殊的意义。由此,本文根据密码子第二位碱基的种类,将 64 种密码子分布在 2 维平面的四个象限内。

根据密码子第二位碱基确定三联体密码子所在象限,如图 1 所示,如果密码子的第二位碱基是 A,本文将第二位碱基是 A 的 16 种三联体密码子全部放置在第一象限,类似的,如果第二位碱基分别是 G、T、C,分别置于第二、第三和第四象限。选择第一个位置和第三个位置上碱基的映射关系如下:G→1、A→2、C→3、T→4。通过第一位和第三位碱基的映射关系,可以将 61 种密码子和 3 种终止密码子分布在 2 维平面内。例如密码子 TGA,中间碱基是 G,所以它放置在第二象限,又因为第一位置和第三位置上的碱基分别为 T 和 A,根据映射关系 T→4、A→2,所以密码子 TGA 的坐标为(−4,2)。根据这种规则,64 个三联体密码子分别被放置在一个二维平面内(图 2)。根据遗传密码表,将 64 个三联体密码子翻译成 20 个氨基酸与终止密码子(图 3)。

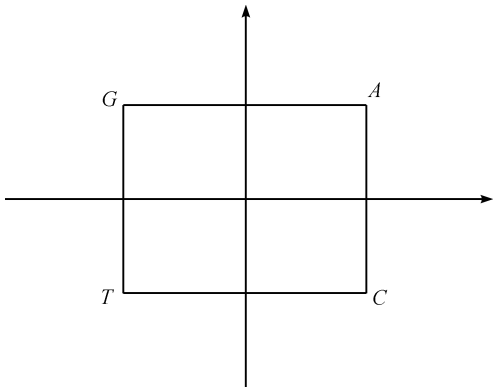

图 1　第二个位置碱基决定密码子象限图

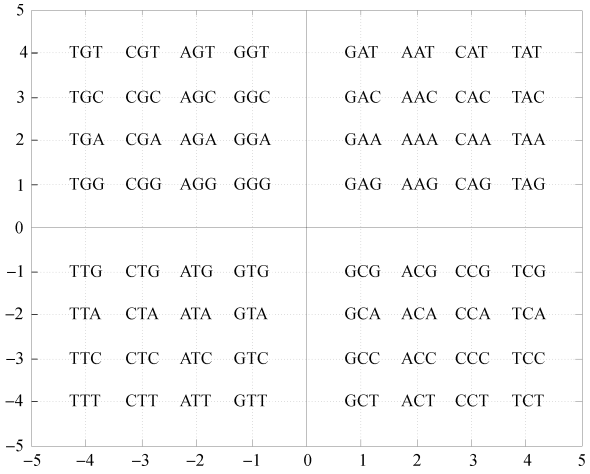

图 2　密码子平面分布图

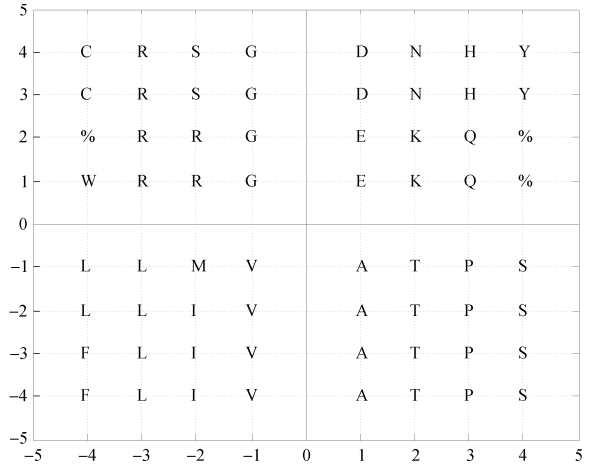

图 3　氨基酸平面分布图

对于图 3 中的氨基酸,取同一氨基酸坐标的平均值作为该氨基酸的坐标,记作(P_1,P_2)。此外,由于氨基酸疏水性在保持蛋白质的三级结构上起作用,将氨基酸的疏水性指数值[29]作为第 3 维坐标,记作 P_3,构建氨基酸的 3 维坐标,记作(P_1,P_2,P_3),如表 1 所示(第 3～5 列)。

为了消除坐标值来源不同对结果的影响,文章

对坐标值做了标准化处理，公式如下：

$$a'=2\times\frac{a-n_{\max}}{n_{\max}-n_{\min}}-1 \tag{1}$$

其中：a'是标准化之后的坐标值；a 是标准化之前的坐标值；$n_{\max}$和 $n_{\min}$为分别为坐标值中的最大值和最小值。根据式(1)可得新的 3 维坐标值，记作 X，Y，Z，见表 1(第 6～8 列)。20 个氨基酸的坐标对应的 20 个向量，如图 4 所示。

表 1 氨基酸的 3 维坐标

氨基酸	符号	P_1	P_2	P_3	X 坐标值	Y 坐标值	Z 坐标值
丙氨酸	A	1	−5/2	1.8	0.2500	−0.7143	0.4000
苯丙氨酸	F	−4	−7/2	2.8	−1.0000	−1.0000	0.6222
异亮氨酸	I	−2	−3	4.5	−0.5000	−0.8571	1.0000
亮氨酸	L	−10/3	−13/6	3.8	−0.8333	−0.6190	0.8444
蛋氨酸	M	−2	−1	1.9	−0.5000	−0.2857	0.4222
脯氨酸	P	3	−5/2	−1.6	0.7500	−0.7143	−0.3556
缬氨酸	V	−1	−5/2	4.2	−0.2500	−0.7143	0.9333
色氨酸	W	−4	1	−0.9	−1.0000	0.2857	−0.2000
半胱氨酸	C	−4	7/2	2.5	−1.0000	1.0000	0.5556
甘氨酸	G	−1	5/2	−0.4	−0.2500	0.7143	−0.0889
天冬酰胺	N	2	7/2	−3.5	0.5000	1.0000	−0.7778
谷氨酰胺	Q	3	3/2	−3.5	0.7500	0.4286	−0.7778
丝氨酸	S	2	−3/2	−0.8	0.5000	−0.4286	−0.1778
苏氨酸	T	2	−5/2	−0.7	0.5000	−0.7143	−0.1556
酪氨酸	Y	4	7/2	−1.3	1.0000	1.0000	−0.2889
组氨酸	H	3	7/2	−3.2	0.7500	1.0000	−0.7111
赖氨酸	K	2	3/2	−3.9	0.5000	0.4286	−0.8667
精氨酸	R	−8/3	13/6	−4.5	−0.6667	0.6190	−1.0000
谷氨酸	E	1	3/2	−3.5	0.2500	0.4286	−0.7778
天冬氨酸	D	1	7/2	−3.5	0.2500	1.0000	−0.7778

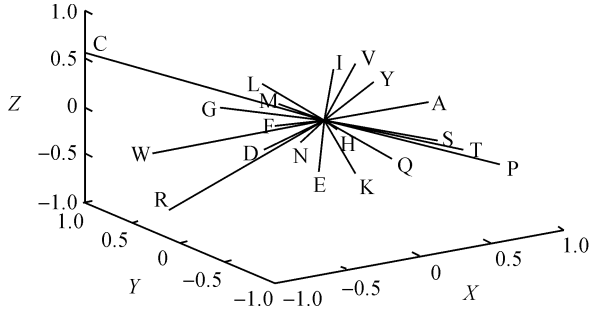

图 4 氨基酸对应的 3 维坐标向量

1.2 异参数迭代关系式

氨基酸是蛋白质的基本单元，通过氨基酸的 3 维坐标，笔者将蛋白质序列中的每个氨基酸转化为空间中的点，顺次连接这些点，得到蛋白质的 3 维图形表示。对于长为 n 的蛋白质序列 $s_1s_2s_3\cdots s_n$，每一个点(当 i 从 1 到 n)对应的坐标 $\boldsymbol{P}_i=(x_i,y_i,z_i)$，通过迭代关系式计算得出，通常迭代关系式的形式为：(本文选择 $\boldsymbol{P}_0=(0,0,0)$)

$$\begin{cases}x_i=\alpha x_{i-1}+\beta\boldsymbol{S}_i^1\\ y_i=\alpha y_{i-1}+\beta\boldsymbol{S}_i^2\\ z_i=\alpha z_{i-1}+\beta\boldsymbol{S}_i^3\end{cases} \tag{2}$$

其中：$\boldsymbol{S}_i^j(j=1,2,3)$表示第 i 个氨基酸的第 j 个坐标分量；$\alpha,\beta\in(0,1]$。异参数的迭代关系式在 *Ma* 等[8]首次提出，该文中选取参数 $\alpha=\frac{3}{4}$，$\beta=\frac{1}{2}$，并通过计算发现，β 为图形的压缩参数(Compression parameters)，仅影响图形的大小，而不影响图形的形状[8]，因此本文设置参数 $\beta=1$。此外，考虑异参数 $\alpha\neq\beta$，本文选取参数 $\alpha=\frac{4}{5}$，则迭代关系式为：

$$\begin{cases}x_i=\frac{4}{5}x_{i-1}+\boldsymbol{S}_i^1\\ y_i=\frac{4}{5}y_{i-1}+\boldsymbol{S}_i^2\\ z_i=\frac{4}{5}z_{i-1}+\boldsymbol{S}_i^3\end{cases} \tag{3}$$

通过式 3，氨基酸序列 $s_1s_2s_3\cdots s_n$ 被转化为 $\boldsymbol{P}_1$，$\boldsymbol{P}_2$，$\boldsymbol{P}_3$，$\cdots$，$\boldsymbol{P}_n$ 共 n 个三维空间中的点，顺次连接这 n 个点，获得氨基酸序列的三维曲线。

1.3 图形表示的数值刻画

本文选取闵可夫斯基距离表示不同曲线之间的距离。对每一个氨基酸序列对应的 3 维曲线，首先计算

曲线中每两个相邻点之间的差向量,再计算每两个曲线之间的闵可夫斯基距离。蛋白质序列 $s_1 s_2 \cdots s_{n_1}$ 和序列 $s'_1 s'_2 \cdots s'_{n_2}$($n_1$,$n_2$ 分别为序列的长度,假设 $n_1 > n_2$),首先计算三维曲线中前后两个点的差向量,计算公式为:

$$\begin{aligned} \boldsymbol{E}_i^1 &= (x_i - x_{i-1}, y_i - y_{i-1}, z_i - z_{i-1}) \\ &= (X_i, Y_i, Z_i), (i = 1 \cdots n_2), \\ \boldsymbol{E}_i^2 &= (x'_i - x'_{i-1}, y'_i - y'_{i-1}, z'_i - z'_{i-1}) \\ &= (x'_i, Y'_i, Z'_i), (i = 1 \cdots n_2). \end{aligned}$$

其次计算闵可夫斯基距离,计算公式为:

$$d_{1,2}(i) = \sqrt[m]{(X_i - X'_i)^m + (Y_i - Y'_i)^m + (Z_i - Z'_i)^m}, \quad i = 1, \cdots, n_2 \tag{4}$$

设 k 为剩余 $n_1 - n_2$ 个点之间的平均距离,所以两个序列的图形之间的距离公式为:

$$d_{1,2} = \sum_{i=1}^{n_2} d_{1,2}(i) + k \times (n_1 - n_2) \tag{5}$$

其中:在式(4)中,m 是正整数,对于两条曲线,每个 m 值都对应着一个距离,基于 $m \in [1,10]$ 中的 10 个正整数,比较不同 m 值对结果的影响,最终选取 $m = 1$ 时,即哈密顿距离;在式(5)中,由于 k 取 $\min d_{1,2}$,$\frac{\min d_{1,2}(i) + \max d_{1,2}(i)}{2}$ 和 $\max d_{1,2}(i)$ 值对距离矩阵没有影响,因此选取平均值 $k = \frac{\min d_{1,2}(i) + \max d_{1,2}(i)}{2}$,$i = n_2 + 1, \cdots, n_1$。

2　蛋白质序列的相似性分析

2.1　数据来源

本文将新的序列图形表示方法应用到 9 个物种的 ND5 蛋白(NADH dehydrogenase subunit 5)和 12 个物种的 β-珠蛋白上,蛋白相关的信息见表 2 和表 3。

表 2　9 个物种的 ND5 蛋白的相关信息

物种	ID 号(NCBI)	长度
人类	YP_003024036.1	603
大猩猩	NP_008222.1	603
倭黑猩猩	NP_008209.1	603
黑猩猩	NP_008196.1	603
鳍鲸	NP_006899.1	606
蓝鲸	NP_007066.1	606
大鼠	AP_004092.1	610
老鼠	NP_904338.1	607
负鼠	NP_007105.1	602

表 3　12 个物种的 β-珠蛋白的相关信息

物种	ID 号(NCBI)	长度
老鼠	NP_032246.2	147
牛	XP_002707739.2	147
鱼	NP_001117138.1	148
鸡	NP_990820.1	147
兔	NP_001075279.1	147
狒狒	NP_001162318.1	147
猪	NP_001138313.1	147
野马	NP_001157490.1	147
青蛙	NP_988859.1	147
人类	NP_000509.1	147
斑胸草雀	XP_002190521.1	147
狗	XP_534029.2	147

2.2　相似性分析

a) ND5 蛋白序列

根据表 2 中 9 个物种的 ND5 蛋白数据,氨基酸的映射关系和闵可夫斯基距离公式,计算出这 9 个物种 ND5 蛋白的距离矩阵,如表 4 所示。ClustalW 算法是目前被广泛应用的经典多重序列比对算法,为了说明本文方法的有效性,利用 Megalign 程序实现的 ClustalW 算法计算 9 个 ND5 蛋白的距离矩阵,结果如表 5 所示。

表 4　基于本文方法得到的 ND5 蛋白的距离矩阵

物种	大猩猩	倭黑猩猩	黑猩猩	鳍鲸	蓝鲸	大鼠	老鼠	负鼠
人类	146.6	103.6	96.7	428.7	427.5	551.2	546.8	850.3
大猩猩		130.4	131.1	442.5	442.5	538.5	527.7	863.9
倭黑猩猩			69.1	430.3	428.8	545.3	541.4	858.0
黑猩猩				428.5	428.2	541.7	538.5	858.7
鳍鲸					58.3	508.8	504.0	868.4
蓝鲸						505.2	494.5	873.1
大鼠							333.9	870.0
老鼠								824.4

表 5　ClustalW 算法得到的 ND5 蛋白序列的距离矩阵

物种	大猩猩	倭黑猩猩	黑猩猩	鳍鲸	蓝鲸	大鼠	老鼠	负鼠
人类	10.7	7.1	7.0	41.1	41.4	50.3	49.0	51.2
大猩猩		9.7	9.9	42.8	42.5	51.6	50.0	54.8
倭黑猩猩			5.2	40.2	40.2	50.3	49.0	50.6
黑猩猩				40.5	40.5	50.9	49.7	52.2
鳍鲸					3.6	46.0	47.5	53.8
蓝鲸						45.7	46.6	53.8
大鼠							25.7	54.8
老鼠								51.8

表 4 和表 5 数据表明，较大的元素均在最后一列，表示负鼠与其他 8 个物种的进化距离最远；最小的元素分别为 58.3(表 4)和 3.6(表 5)，均表示鳍鲸与蓝鲸的进化距离在这 9 个物种中是最接近的；在表 4 中，人类、大猩猩、倭黑猩猩、黑猩猩之间的距离明显小于与其他物种的距离，同样也能够在表 5 中得出这样的结论。根据两个表中数据的特点，大致可以将 9 个物种分为四类，分别为人类、大猩猩、倭黑猩猩、黑猩猩；鳍鲸、蓝鲸；大鼠、老鼠以及负鼠，两表中数据表示的进化关系基本一致。

此外，根据本文方法的结果(表 4)，运用生物信息学中构造进化树常用的 UPGMA 方法构造出进化树，结果如图 5 所示；根据 ClustalW 算法的结果(表 5)，构造出进化树，结果如图 6 所示。

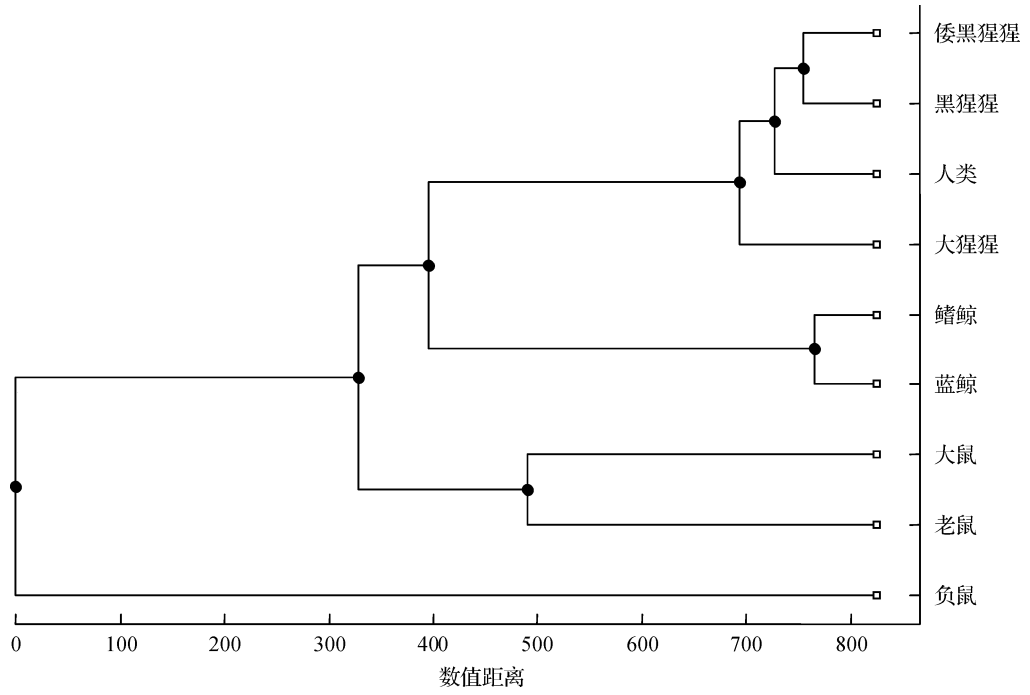

图 5 基于本文方法构建的 9 个物种的进化树

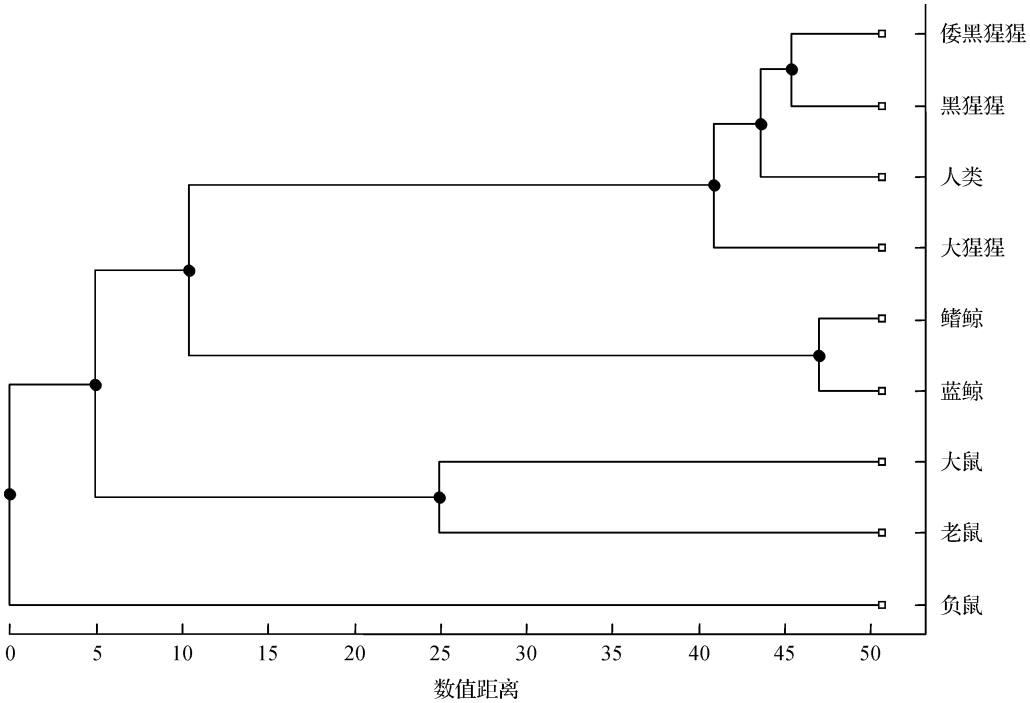

图 6 基于 ClustalW 方法构建的 9 个物种的进化树

由图 5 和图 6 可以看出,本文方法构造的进化树与 ClustalW 方法构造的进化树在结果上完全相同,且两个进化树的分支结构上也非常相似,鳍鲸和蓝鲸均在同一个分支,人类、大猩猩、倭黑猩猩和黑猩猩在同一个分支,老鼠和大鼠在同一个分支,表示在 9 个物种里,鳍鲸和蓝鲸的进化距离最近可以归为一类,人类、大猩猩、倭黑猩猩和黑猩猩归为一类,老鼠与大鼠归为一类,与这些类距离最远的是负鼠,进一步证明本文蛋白质序列相似性分析方法与序列比对方法 ClustalW 算法具有一致性,说明本文方法是可靠有效的。

为了更直观显示表 4 和表 5 数据的关系,利用相关性分析两种方法得到的距离矩阵,结果如图 7 所示,图中横坐标为 ClustalW 算法得到的距离值,纵坐标为本文方法得到的距离值。结果显示,两组数据的相关系数为 0.90,可见两种方法的结果具有很强的相关性。

本文进一步地分析两个距离矩阵每一行的相关系数,结果见表 6(第 2 列)。为了比较,本文计算了文献[14,21,28]中结果与 ClustalW 方法结果(表 5)的相关系数,结果见表 6(第 3—5 列)。观察表 6,对于所有物种,本文的结果与 ClustalW 算法结果的相关系数(第 1 行)明显高于文献[14,21,28]的结果与 ClustalW 结果的相关系数,同时,对于每个物种(第 2—8 行),相关系数同样明显高于文献[14,21,28]结果与 ClustalW 算法结果的相关系数,对比结果说明本文方法具有明显的优越性。

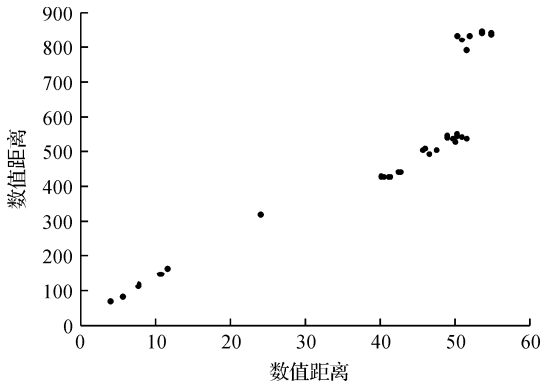

图 7　由表 4 和表 5 的数据作出散点图

表 6　本文结果及文献结果与 ClustalW 算法结果的相关系数(ND5)

相关系数	本文结果	文献[14]结果(表 3)	文献[21]结果(表 4)	文献[28]结果(表 4)
所有物种	0.90	0.69	0.73	0.69
人类	0.93	0.92	0.82	0.86
大猩猩	0.93	0.70	0.75	0.78
倭黑猩猩	0.93	0.92	0.80	0.80
黑猩猩	0.93	0.93	0.79	0.70
鳍鲸	0.89	0.60	0.83	0.75
蓝鲸	0.89	0.70	0.84	0.86
大鼠	0.77	0.69	0.53	0.81
老鼠	0.73	0.64	0.42	0.74
负鼠	0.60	−0.52	0.16	−0.51

b) β-珠蛋白序列

根据表 3 中 12 个物种的 β-珠蛋白数据,氨基酸的映射关系和闵可夫斯基距离公式,计算出这 12 个 β-珠蛋白的距离矩阵,结果如表 7 所示。本文利用 ClustalW 算法计算这 12 个蛋白的距离矩阵,结果如表 8 所示。

表 7　基于本文方法得到的 β-珠蛋白的距离矩阵

物种	牛	鱼	鸡	兔	狒狒	猪	野马	青蛙	人类	斑胸草雀	狗
老鼠	126.5	174.5	130.5	85.1	89.5	87.6	94.9	172.8	85.8	129.7	94.5
牛		212.7	159.6	119.2	119.6	120.7	120.1	201.8	113.1	156.4	120.4
鱼			177.5	198.0	194.0	192.2	193.6	204.1	193.8	172.2	198.3
鸡				101.9	102.1	102.4	116.2	168.5	110.9	50.1	111.0
兔					36.5	56.4	58.4	166.4	29.2	103.0	53.2
狒狒						60.8	60.2	173.1	27.2	103.5	41.7
猪							60.7	166.7	63.0	108.0	63.2
野马								183.4	58.2	114.2	65.7
青蛙									179.0	167.7	174.4
人类										108.3	43.5
斑胸草雀											106.9

表 8 基于 ClustalW 算法得到的 β-珠蛋白的距离矩阵

物种	牛	鱼	鸡	兔	狒狒	猪	野马	青蛙	人类	斑胸草雀	狗
老鼠	50.1	80.8	45.1	23.9	21.1	23.9	28.7	86.4	22.9	51.4	28.7
牛		103.6	65.6	46.4	47.6	47.6	47.6	103.6	45.1	70.4	51.4
鱼			67.2	94.6	86.4	82.6	88.4	90.4	82.6	75.4	86.4
鸡				42.7	38.2	43.9	41.6	75.4	39.3	17.6	38.2
兔					11.0	18.5	18.5	101.3	10.2	40.4	21.1
狒狒						19.3	18.5	96.8	5.7	43.9	15.9
猪							20.2	94.6	16.7	47.6	21.1
野马								96.8	18.5	42.7	24.8
青蛙									103.6	86.4	92.5
人类										43.9	15.1
斑胸草雀											38.2

观察表 7 和表 8 数据可以发现，两个表中，最小的元素分别为 27.2(表 7)和 5.7(表 8)，表示狒狒与人类的进化距离在这 12 个物种中是最接近的；观察两表中人类与其他物种的距离(倒数第 3 列和倒数第 2 行)，除了野马、猪和人类的距离，青蛙，鱼和人类的距离略有差异外，人类与其他物种的距离基本相同。同时观察到两表中鱼、青蛙同其他物种的距离均比较大，从表 7 和表 8 显示本文方法结果和 ClustalW 方法结果基本一致。

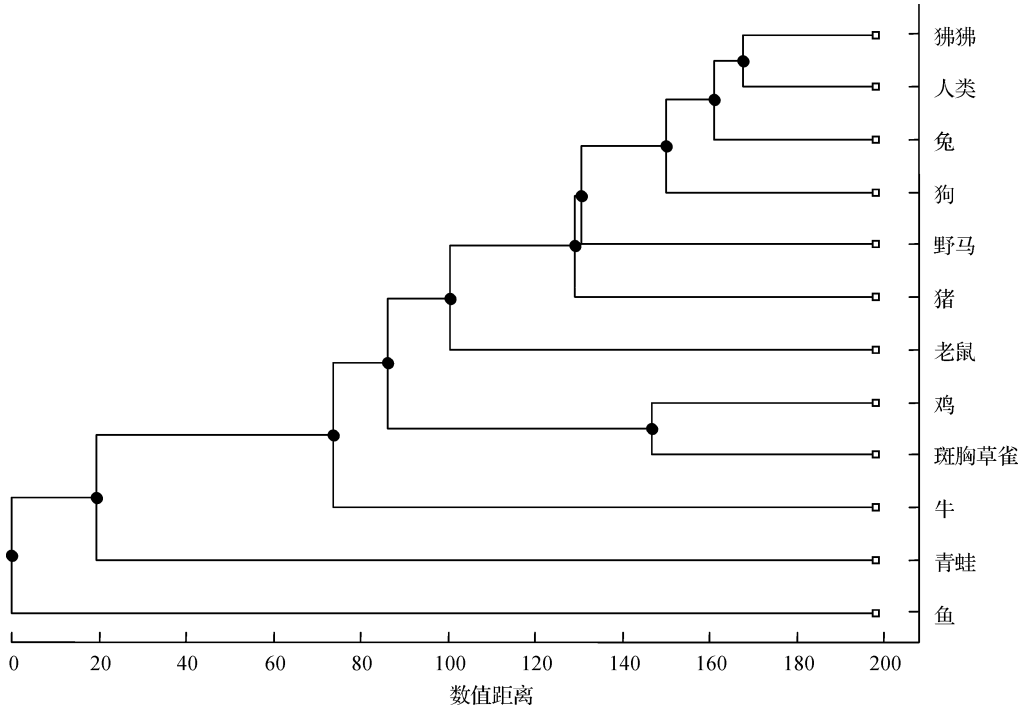

图 8 基于本文方法构建的 12 个物种的进化树

此外，基于本文方法的结果(表 7)构造出进化树如图 8 所示；基于 ClustalW 算法的结果(表 8)，我们构造出进化树如图 9 所示。对比图 8 和图 9 发现，除了野马和猪，鱼和青蛙的位置不同外，其余分支的结构完全相同。

分析本文结果与 ClustalW 算法结果的相关性，本文以 ClustalW 算法得到的距离值(表 8)为横坐标，以本文方法得到的距离值(表 7)为纵坐标，作出散点图如图 10 所示，从图中可以看出两表中数据呈正相关，此外，我们计算出两个距离矩阵的相关系数为 0.96，说明在 β-珠蛋白数据中，本研究结果同样与 ClustalW 的结果相关性很强，两种方法的结果具有一致性。

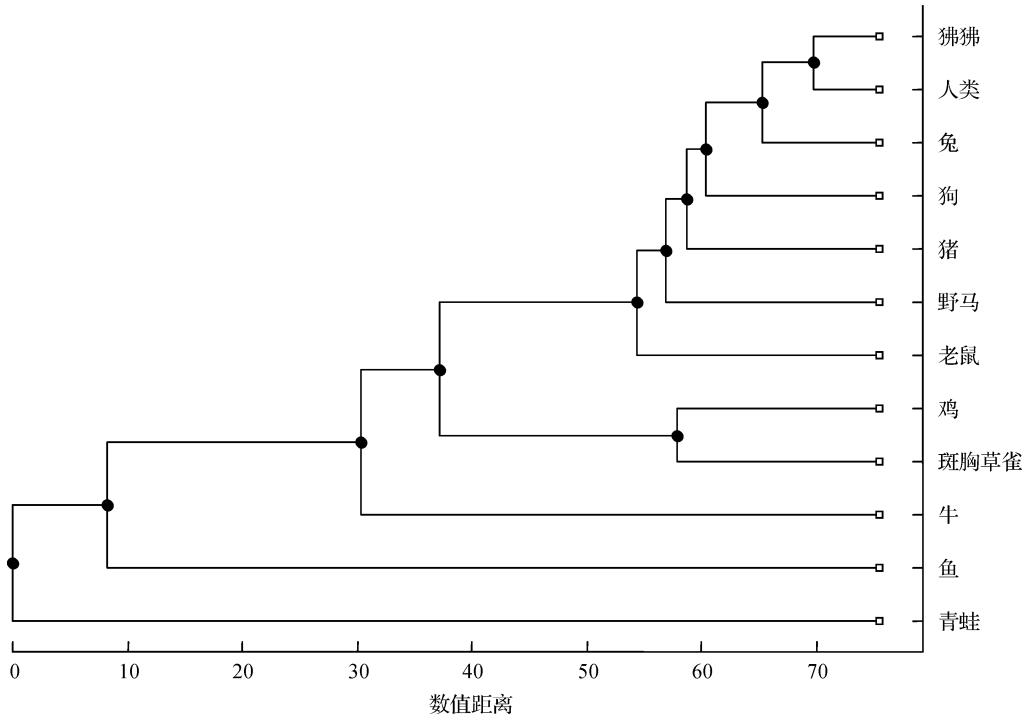

图 9　基于 ClustalW 算法结果构建的 12 个物种的进化树

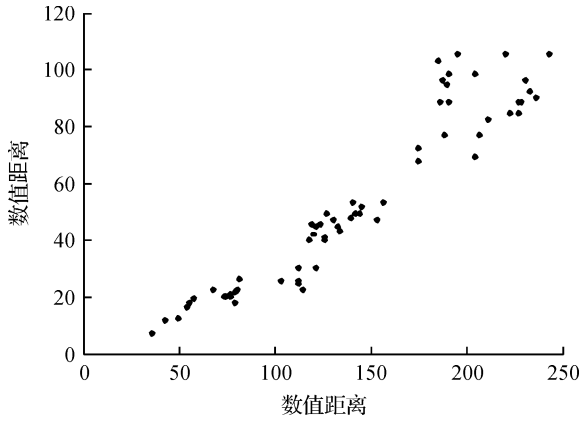

图 10　表 7 和表 8 的数据的相关性

本文进一步地分析两个距离矩阵每一行的相关系数,见表 9。结果表明,除了青蛙和鱼外,其他物种与 ClustalW 的结果相关系数均达到 0.96 以上。从 9 个物种的 ND5 蛋白和 12 个物种的 β-珠蛋白两组数据的相似性分析结果中,因此本文的方法是有效可行的。

表 9　不同物种与 ClustalW 结果的相关系数

物种	老鼠	牛	鱼	鸡	兔	狒狒	猪	野马	青蛙	人类	斑胸草雀	狗
相关系数	0.99	0.99	0.87	0.96	0.96	0.97	0.96	0.98	0.31	0.96	0.97	0.97

3　结　论

本文利用密码子的碱基位置特征与氨基酸疏水性特征,提出一种新的蛋白质 3 维图形表示方法,并选取闵可夫斯基距离描述图形的差异性。运用本文方法,分析 9 个物种 ND5 蛋白的相似性和 12 个物种 β-珠蛋白的相似性,并将相似性分析结果与 ClustalW 方法结果以及其他文献中方法的结果做比较,在 ND5 蛋白数据中,本文结果中与 ClustalW 算法结果的相关系数为 0.90;在 β-珠蛋白数据中,本文结果与 ClustalW 算法结果的相关系数为 0.96。比较结果说明本研究的方法是有效可行的。

参考文献:

[1] Wang J S, Yan M. Numerical Methods in Bioinformatics an Introduction[M]. Beijing: Science Press, 2013: 14-48.

[2] Randic M, Butina D, Zupan J. Novel 2-D graphical representation of proteins[J]. Chemical Physics Letters, 2006, 419(4/5/6): 528-532.

[3] He P A, Zhang Y P, Yao Y H, et al. The graphical representation of protein sequences based on the physicochemical properties and its applications[J]. Journal of Computational Chemistry, 2010, 31(11): 2136-2142.

[4] Jeffrey H J. Chaos game representation of gene structure[J]. Nucleic Acids Research, 1990, 18(8): 2163-2170.

[5] Randic M, Zupan J, Balaban A T. Unique graphical

representation of protein sequences based on nucleotide triplet codons[J]. Chemical Physics Letters,2004,397(1):247-252.

[6] Li C, Yu X Q, Liu Y, et al. 3-D maps and coupling numbers for protein sequences[J]. Physica A: Statistical Mechanics & Its Applications,2009,388(9):1967-1972.

[7] Yao Y H, Kong F, Dai Q, et al. A sequence-segmented method applied to the similarity analysis of long protein sequence[J]. MATCH Communications in Mathematical and in Computer Chemistry,2013,70(1):431-450.

[8] Ma T T, Liu Y X, Dai Q, et al. A graphical representation of protein based on a novel iterated function system[J]. Physica A: Statistical Mechanics and Its Applications, 2014,403(6):21-28.

[9] Qi Z H, Jin M Z, Li S L, et al. A protein mapping method based on physicochemical properties and dimension reduction[J]. Computers in Biology and Medicine,2015,57:1-7.

[10] Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences[J]. Journal of Biological Chemistry, 1983,258(2):1318-1327.

[11] Rahman R S, Rackovsky S. Protein sequence randomness and sequence/structure Corrections[J]. Biophysical Journal, 1995,68(4):1531-1539.

[12] Yu H J, Huang D S. Novel 20-D descriptors of protein sequences and it's applications in similarity analysis[J]. Chemical Physics Letters,2012,531:261-266.

[13] Aboel M I, Aboelkhier M M, Abbelwahaabm A. 3D graphical representation of protein sequences and their statistical characterization[J]. Physica A: Statistical Mechanics & Its Applications, 2010, 389(21): 4668-4676.

[14] Wen J, Zhang Y Y. A 2D graphical representation of protein sequence and its numerical characterization[J]. Chemical Physics Letters,2009,476(4-6):281-286.

[15] Yu C L, Cheng S Y, He R L, et al. Protein map: An alignment-free sequence comparison method based on various properties of amino acids[J]. Gene,2011,486(1-2):110-118.

[16] Zhao Y, Li X, Qi Z. Novel 2D graphic representation of protein sequence and its application[J]. Journal of Fiber Bioengineering & Informatics,2014,7(1):23-33.

[17] Liao B, Liao B Y, Lu X, et al. A novel graphical representation of protein sequences and its application[J]. Journal of Computational Chemistry, 2011, 32(12): 2539-2544.

[18] Yu J F, Sun X, Wang J H. A novel 2D graphical representation of protein sequence based on individual amino acid [J]. International Journal of Quantum Chemistry,2011,111(12):2835-2843.

[19] Bai F L, Wang T M. A 2-D graphical representation of protein sequences based on nucleotide triplet codons[J]. Chemical Physics Letters,2005,413(4):458-462.

[20] Liao B, Liao B Y, Sun X M, et al, A novel method for similarity Analysis and protein sub-cellular localization prediction [J]. Bioinformatics, 2010, 26(21): 2678-2683.

[21] He P A. A new graphical representation of similarity/dissimilarity studies of protein sequences[J]. SAR and QSAR in Environmental Research,2010,21(5/6):571-580.

[22] He P A, Li D, Zhang Y P, et al. A 3D graphical representation of protein sequences based on the gray code[J]. Journal of Theoretical Biology,2012,304(1): 801-807.

[23] Feng J, Wang T M. A 3D graphical representation of RNA secondary structures based on chaos game representation[J]. Chemical Physics Letters,2008,454(4-6):355-361.

[24] He P A, Yang J L, Li X F, et al, A novel descriptor for protein similarity analysis[J]. MATCH Communications in Mathematical & in Computer Chemistry, 2011, 65(2):445-458.

[25] Manikandakumar K, Gokulraj K, Muthukumaran S, et al. Graphical representation of protein sequences by CGR: Analysis of pentagon and hexagon structures[J]. Journal of Pharmacy Research,2013,13(6):764-771.

[26] Liu Y X, Li D, Lu K B, et al. P-H Curve, a graphical representation of protein sequences for similarities analysis [J]. MATCH Communications in Mathematical and in Computer Chemistry,2013,70(1):451-466.

[27] He P A, Xu S N, Dai Q, et al. A generalization of CGR representation for analyzing and comparing protein sequences[J]. International Journal of Quantum Chemistry, 2016,116(6):476-482.

[28] Yao Y H, Yan S J, Han J N, et al. A novel descriptor of protein sequences and its application[J]. Journal of Theoretical Biology,2014,347(1):109-117.

[29] Kyte J, Doolittle R F. A simple method for displaying the hydropathic character of a protein[J]. Journal of Molecular Biology,1982,157(1):105-132.

Graphical representation of protein sequences based on codon features

ZHU Zhengyang, *HE Pingan*

(School of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Based on the position characteristics of basic group of the codon and hydrophobicity value of amino acids, 20 kinds of amino acids were mapped to the vectors in the three-dimensional space. A new iterative function of different parameters was proposed to convert the amino acid sequence into a curve in three-dimensional space and a new graphical representation method of protein was obtained. For the protein graph, Minkowski distance was adopted to characterize the distance between two 3D curves. The difference of protein sequences and evolutionary relationship among species were thus inferred. This method was applied in sequence analysis of ND5 proteins of 9 species and β-globin proteins of 12 species. After the results gained were compared with the results of ClustalW method and the results in other literatures, the method proposed in this paper is feasible and effective.

Key words: codon characteristics; protein sequence; graphical representation; similarity analysis

(责任编辑: 唐志荣)