浙江理工大学学报，2019，41(2)：269-276
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2019.02.020

基于广义迭代函数的蛋白质分析及其应用

黄嘉禾，贺平安
（浙江理工大学理学院，杭州 310018）

摘　要：通过将蛋白质图形表示中的迭代函数推广到高维空间，提出了一种适用范围更加广泛的广义迭代函数，该函数能反映空间中坐标不同压缩比例与不同氨基酸残基理化性质参数。应用该广义迭代函数和氨基酸残基的某些理化指标，得到了一种新的蛋白质序列图形表示。研究结果表明，该图形表示反映了对应蛋白质序列的主要信息。应用该蛋白质图形表示方法和相应的数值刻画，比较了10种物种的ND5蛋白质序列相似性，并分析这些物种的进化关系。与经典的蛋白质多重序列比对ClustalW方法的比较结果表明，该方法对蛋白序列相似性比较和物种进化分析是可靠且有效的。

关键词：蛋白质序列；非序列比对；图形表示；广义迭代函数方法；相似性比较
中图分类号：O29　　**文献标志码**：A　　**文章编号**：1673-3851 (2019) 03-0269-08

Protein analysis based on generalized iterative function and its applications

HUANG Jiahe, *HE Pingan*
(School of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: By generalizing the iterative function in protein graphic representation to higher-dimensional space, a generalized iterative function with wider application scope was proposed, and this function can reflect different compression ratios of coordinates in space and different physicochemical properties of amino acid residues. A new graphical representation of protein sequences was gained by using this generalized iterative function and some physicochemical indexes of amino acid residues. The research results showed that the graphical representation reflected the main information of the corresponding protein sequences. The protein graphic representation method and corresponding numerical description were applied to compare the similarity of the ND5 protein sequences of 10 species and analyze the evolutionary relationships among these species. Compared with ClustalW method of the classical protein multiple sequence alignment, the results showed this method was reliable and effective for the comparison of protein sequence similarity and the evolutionary analysis of species.

Key words: protein sequence; non-sequence alignment; graphical representation; generalized iterative function system; similarity comparison

0　引　言

通过比较物种相关DNA和蛋白质序列的相似性，可以在一定程度上推断出这些物种之间的进化关系，同时也能预测新的蛋白质序列的结构与功能，所以生物序列的相似性比较是生物信息学中重要的

收稿日期：2018－08－21　　网络出版日期：2018－12－01
基金项目：国家自然科学基金项目(61772027)
作者简介：黄嘉禾(1992－　)，男，浙江兰溪人，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

基础研究课题。生物序列相似性比较的方法分为序列比对方法与非序列比对方法[1]。生物序列的图形表示方法作为一种非序列比对方法,具有易数学描述、计算简单和可用于处理大数据等优点,近年来一直受到广泛的关注。1983 年,Hamori 等[2]首次采用图形表示方法定量描述基因序列,从而为 DNA 序列的分析比较提供了一种可视化方法。自此,越来越多的图形表示方法被用来刻画和分析 DNA 序列。这些 DNA 序列图形表示方法,首先构建数学映射将 4 种核苷酸映射为 4 个基准向量;其次应用迭代函数将 DNA 序列转换为平面或空间中的曲线;最后通过对这些曲线进行数值刻画,从而定量地描述对应的 DNA 序列以供分析。DNA 序列图形表示方法中的迭代函数通常可表示为 $\boldsymbol{X}_i=\alpha\boldsymbol{X}_{i-1}+\beta\boldsymbol{S}_i$,其中:$\boldsymbol{X}_i$ 表示 DNA 序列中第 $\boldsymbol{i}$ 个核苷酸在空间中点的坐标;$\boldsymbol{S}_i$ 表示第 i 个核苷酸的基准向量;$\alpha>0$ 与 $\beta>0$ 是迭代函数中的参数。α 为前后两个核苷酸在空间中点的坐标压缩比例,β 为核苷酸基准向量的参数,并且迭代函数中不同的参数对应于不同的统计模型。例如,在文献[2-8]提出的方法中两个参数均取 1,该模型对应于等概率随机模型;在 Jeffry[9]提出的方法中两个参数均取 1/2,此模型对应于一类马尔科夫模型;在 Wu 等[10]提出的方法中两个参数均取 $1/k$(k 表示任意正整数)。

因为蛋白质序列是由 20 种氨基酸残基组成的,所以与 DNA 序列图形表示方法相比,蛋白质序列的图形表示方法更加复杂。近年来,许多学者提出了各种蛋白质的图形表示方法[11-29]用于定量地描述、分析和比较蛋白质序列。与 DNA 序列的图形表示方法类似,在蛋白质序列的图形表示方法中,首先确定 20 种氨基酸残基所对应的向量。例如,He 等[15]提出的方法中通过格雷编码表示氨基酸残基向量;Ma 等[21]、Zhao 等[24]与 Qi 等[25]提出的方法中通过氨基酸残基的理化性质表示氨基酸残基向量。对于迭代函数,除了上述取相同的参数外,Ma 等[21]提出了异参数迭代函数,即

$$\boldsymbol{X}_i=\alpha\boldsymbol{X}_{i-1}+\beta\boldsymbol{S}_i,\alpha\neq\beta \tag{1}$$

例如,在 Ma 等[21]提出的方法中取 $\alpha=\frac{3}{4}$ 和 $\beta=\frac{1}{2}$;在 He 等[22]提出的方法中取 $\alpha=\frac{2}{5}$ 和 $\beta=\frac{3}{5}$ 等。此外,Yao 等[17]、Zhao 等[24]与 Qi 等[25]提出了蛋白质序列的谱图表示方法,将每个蛋白质序列转换为二维曲线来刻画并分析蛋白质序列。该方法与前面的方法不同之处在于,蛋白质序列谱图表示方法中迭代函数的第一部分是刻画氨基酸残基在蛋白质序列的位置信息,另一部分迭代函数是刻画当前位置的氨基酸残基信息。

为了给出上述各种蛋白质图形表示方法迭代函数的统一形式,本文提出了一种更一般的广义迭代函数。首先,对氨基酸残基的理化指标进行归一化处理,选取 4 种相关性最弱的氨基酸理化指标并将每一种氨基酸残基映射为一个五维向量;然后,利用广义迭代函数得到了一种新的蛋白质图形表示方法,本文证明了该方法是非退化的,即这种图形表示方法能避免生物序列中主要信息的丢失;最后,为了说明该方法的可靠性和有效性,本文对 10 种物种的 ND5 蛋白质序列进行相似性比较和进化分析,并把所得到的相似性结果与经典的多重序列比对的 ClustalW 方法所得结果进行了比较。

1 广义迭代函数

蛋白质的图形表示方法通常是 DNA 序列图形表示方法的推广,因此大部分的蛋白质图形表示方法中的迭代函数仍然沿用公式 $\boldsymbol{X}_i=\alpha\boldsymbol{X}_{i-1}+\beta\boldsymbol{S}_i$。然而 Yao 等[17]、Zhao 等[24]与 Qi 等[25]提出了蛋白质的谱图表示,在他们的方法中迭代函数可表示为:

$$\begin{cases}\boldsymbol{X}_i^1=i\\ \boldsymbol{X}_i^2=S_i\end{cases} \tag{2}$$

因为在谱图表示中坐标的两个分量采用了两类不同的迭代函数,因此蛋白质序列的谱图表示中的迭代函数与一般的迭代函数 $\boldsymbol{X}_i=\alpha\boldsymbol{X}_{i-1}+\beta\boldsymbol{S}_i$ 是完全不同的。本文将常用的图形表示方法中的迭代函数推广到高维空间,提出了一种新的迭代函数:

$$\boldsymbol{X}_i=\boldsymbol{A}\boldsymbol{X}_{i-1}+\boldsymbol{B}\boldsymbol{S}_i \tag{3}$$

并称之为广义迭代函数,其中 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 分别表示 n($n\geqslant 1$)阶对角方阵。

容易看出在二维空间中,式(1)是 $\boldsymbol{X}_i=\boldsymbol{A}\boldsymbol{X}_{i-1}+\boldsymbol{B}\boldsymbol{S}_i$ 中取 $\boldsymbol{A}=\alpha\mathbf{I},\boldsymbol{B}=\beta\mathbf{I}$ 时的特殊情形,其中 $\mathbf{I}$ 为二阶单位矩阵;式(2)是 $\boldsymbol{X}_i=\boldsymbol{A}\boldsymbol{X}_{i-1}+\boldsymbol{B}\boldsymbol{S}_i$ 中取 $\boldsymbol{A}=\mathrm{diag}(1,0),\boldsymbol{B}=\mathrm{diag}(1,1)$ 时的特殊情形。

因为式(3)中 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 分别为 n 阶对角方阵,所以一方面不仅可以对两个氨基酸残基在空间中点的坐标选取不同的压缩比例,而且也可以对氨基酸残基不同的理化性质指标选取不同的参数;另一方面利用式(3)还可以把经典迭代函数与谱图迭代函数统一起来,只需要在一个高维空间中讨论。例如,在

Ma 等[21]的文章中应用 $\boldsymbol{X}_i=\frac{3}{4}\boldsymbol{X}_{i-1}+\frac{1}{2}\boldsymbol{S}_i$ 给出了蛋白质序列的二维图形表示；Yao 等[17]的文章中应用迭代函数 $\begin{cases}\boldsymbol{X}_i^1=i\\ \boldsymbol{X}_i^2=S_i\end{cases}$ 给出了蛋白质序列的二维谱图表示。如果在式(3)中取 $\boldsymbol{A}=\mathrm{diag}\left(\frac{3}{4},\frac{3}{4},1,0\right)$，$\boldsymbol{B}=\mathrm{diag}\left(\frac{1}{2},\frac{1}{2},1,1\right)$，则显然式(3)中前两个分量代表 Ma 等[21]的结果，后两个分量代表的是 Yao 等[17]的结果。

2　蛋白质的图形表示方法

基于广义迭代函数式(3)，结合氨基酸残基的某些理化性质，本文给出一种新的蛋白质图形表示方法来数值刻画蛋白质序列。

2.1　确定氨基酸残基坐标

蛋白质的基本单位为氨基酸残基，蛋白质的一级结构就是其氨基酸残基序列，蛋白质是由所含氨基酸残基的亲水性、疏水性、带正电、带负电等理化性质通过残基间的相互作用而折叠成立体的三级结构。因此氨基酸残基的理化性质对描述、分析和比较蛋白质序列、折叠、分类和功能等起到重要的作用[25]。在 AAindex 数据库中收录了 544 种不同的氨基酸残基理化指标[30]。文献[25]提出了分析蛋白质序列相对重要的 12 种氨基酸残基理化性质，它们分别为侧链化学组成(chemical composition of the side chain)、极性(polar requirement)、亲水性指数(hydropathy index)、等电点(isoelectric point)、分子体积(molecular volume)、极性值(polarity values)、芳香性指数(aromaticity)、脂化度指数(aliphaticity)、氢化作用指数(hydrogenation)、羟基化(hydroxythiolation)、pK_1(—COOH)和 pK_2($-NH_3^+$)。

首先考虑到不同理化性质之间指标数值差异较大，利用式(4)将文献[25]中的所有理化指标归一化到[0,1]区间上，则式(4)表示为：

$$a'=\frac{a-m}{M-m} \tag{4}$$

其中：a' 是归一化后所得到的指标，M 和 m 分别表示 20 种天然氨基酸残基的某种理化指标中的最大值和最小值。对归一化的数据进行相关性分析后，本文取相关性最小的亲水性指数、芳香性指数、pK_1 与 pK_2，并以这 4 种理化性质作为研究的基础，其中将亲水性指数记为 Hy，将芳香性指数记为 Ar。表 1 列出了 20 种天然氨基酸残基的 4 种理化指标及其归一化后的指标，其中 V_1、V_2、V_3 与 V_4 分别表示 Hy、Ar、pK_1 与 pK_2 利用式(4)归一化后的指标。

表 1　20 种氨基酸残基的理化性质指标与其坐标分量

氨基酸残基	Hy	Ar	pK_1	pK_2	V_1	V_2	V_3	V_4
A	1.800	−0.110	2.340	9.690	0.700	0.225	0.685	0.450
C	2.500	−0.184	1.710	10.700	0.778	0.130	0.000	1.000
D	−3.500	−0.285	2.090	9.820	0.111	0.000	0.413	0.515
E	−3.500	−0.067	2.190	9.670	0.111	0.280	0.522	0.439
F	2.800	0.438	1.830	9.130	0.811	0.929	0.130	0.167
G	−0.400	−0.073	2.340	9.600	0.456	0.273	0.685	0.404
H	−3.200	0.320	1.820	9.170	0.144	0.778	0.120	0.187
I	4.500	0.001	2.360	9.680	1.000	0.368	0.707	0.444
K	−3.900	0.049	2.180	8.950	0.067	0.429	0.511	0.076
L	3.800	−0.008	2.360	9.600	0.922	0.356	0.707	0.404
M	1.900	−0.041	2.280	9.210	0.711	0.314	0.620	0.207
N	−3.500	−0.136	2.020	8.800	0.111	0.192	0.337	0.000
P	−1.600	−0.016	1.990	10.600	0.322	0.346	0.304	0.909
Q	−3.500	−0.246	2.170	9.130	0.111	0.050	0.500	0.167
R	−4.500	0.079	2.170	9.040	0.000	0.468	0.500	0.121
S	−0.800	−0.153	2.210	9.150	0.411	0.170	0.544	0.177
T	−0.700	−0.208	2.630	10.430	0.422	0.099	1.000	0.823
V	4.200	−0.155	2.320	9.620	0.967	0.167	0.663	0.414
W	−0.900	0.493	2.380	9.390	0.400	1.000	0.728	0.298
Y	−1.300	0.381	2.200	9.110	0.356	0.856	0.533	0.157

对于每一个氨基酸残基,利用表 1 中归一化后所得的指标,将其映射为五维向量 $(1,V_1,V_2,V_3,V_4)^{\mathrm{T}}$。例如丙氨酸 A 的对应向量为 $(1.000,0.700,0.225,0.685,0.450)^{\mathrm{T}}$。

2.2 基于广义迭代函数图形表示

利用 2.1 给出的氨基酸残基坐标,给出一类新的蛋白质图形表示:对一条含有 N 个氨基酸残基的蛋白质序列 $S=s_1s_2\cdots s_N$,从坐标原点 $\boldsymbol{X}_0=(0,0,0,0,0)^{\mathrm{T}}$ 开始,当 $i=1,2,\cdots,N$ 时,基于广义迭代函数 $\boldsymbol{X}_i=\boldsymbol{A}\boldsymbol{X}_{i-1}+\boldsymbol{B}\boldsymbol{S}_i$ 得到了五维空间中的 N 个点,顺次连接这 $N+1$ 个点,得到该蛋白质序列在五维空间对应的曲线 $\boldsymbol{X}_0\boldsymbol{X}_1\cdots\boldsymbol{X}_N$。这里 $\boldsymbol{X}_i$ 表示第 i 个点的坐标,$\boldsymbol{S}_i$ 表示蛋白质序列的第 i 个位置的氨基酸残基对应的坐标;$\boldsymbol{A}=\mathrm{diag}(a_1,a_2,a_3,a_4,a_5)$,$\boldsymbol{B}=\mathrm{diag}(b_1,b_2,b_3,b_4,b_5)$ 是五阶对角矩阵,其中 $0<a_j\leqslant 1$ 且若 a_j 为 1,则 b_j 也为 1;否则,$b_j=1-a_j$。特别地本文取 $a_1=b_1=1$,它们反映了氨基酸残基的位置信息。另外,$a_j(j=2,3,4,5)$取不尽相同的值可以实现前后两个氨基酸残基在空间中点的坐标有不同的压缩比例;$b_j(j=2,3,4,5)$取不尽相同的值可以实现氨基酸残基不同的理化性质有不同的参数。

对于某蛋白质序列的图形表示,设对应的曲线为 $\boldsymbol{X}_0\boldsymbol{X}_1\cdots\boldsymbol{X}_N$。如果存在 $0\leqslant i,k\leqslant N$ 且 $i\neq k$,使得 $\boldsymbol{X}_i=\boldsymbol{X}_k$,则称该蛋白质序列的图形表示存在环[25]。在生物序列的图形表示中,为了避免生物序列中信息的丢失,通常要求得到的生物序列图形表示是非退化的,即图形与蛋白质序列一一对应且图形中不含有环[25]。

令 V 表示由所有蛋白质序列组成的集合,V_N 表示长度为 N 的蛋白质序列组成的集合,则 $V=\bigcup_{N=1}^{\infty}V_N$。令由上述迭代函数得到 $\boldsymbol{X}_0\boldsymbol{X}_1\cdots\boldsymbol{X}_N$ 是五维空间中的一条曲线,所有这样的曲线集合记为 W_N,且记 $W=\bigcup_{N=1}^{\infty}W_N$。现按照上述的广义迭代函数可以定义一个映射 $\Phi:V\to W$,它把每个蛋白质序列映射到它所对应的五维图形曲线。

性质 1 映射 $\Phi:V\to W$ 是一个单射,即蛋白质序列与其五维图形曲线是一一对应的。

证明:对于 V 中任意两条蛋白质序列 $V^1=s_1^1s_2^1\cdots s_N^1$ 与 $V^2=s_1^2s_2^2\cdots s_M^2$,如果 $N\neq M$,则显然 $\Phi(V^1):=X_0^1X_2^1\cdots X_N^1\neq\Phi(V^2):=X_0^2X_2^2\cdots X_M^2$,所以不妨假设 $N=M$,即

$$V^1=s_1^1s_2^1\cdots s_N^1,V^2=s_1^2s_2^2\cdots s_N^2。$$

记

$$i_0=\min\{i:s_i^1\neq s_i^2\},$$

则由式(3)知,$X_i^1=X_i^2,i=0,1,2,\cdots,i_0-1$;因为 $s_{i_0}^1\neq s_{i_0}^2$,所以 $V_{i_0}^1\neq V_{i_0}^2$ 且 $\boldsymbol{B}$ 是非奇异的,从而 $\boldsymbol{X}_{i_0}^1=\boldsymbol{A}\boldsymbol{X}_{i_0-1}^1+\boldsymbol{B}V_{i_0}^1\neq\boldsymbol{A}\boldsymbol{X}_{i_0-1}^2+\boldsymbol{B}V_{i_0}^2=\boldsymbol{X}_{i_0}^2$,故 $\Phi(V^1)\neq\Phi(V^2)$。因此映射 $\Phi:V\to W$ 是一个单射,从而 $\Phi:V\to\Phi(V)$ 是双射,即蛋白质序列与其五维图形曲线是一对一的。

性质 2 对任意的蛋白质序列,由广义迭代函数得到的五维图形表示不存在环。

证明:显然,由式(3)得到的任意一条蛋白质序列的五维图形表示不可能存在环,这是因为 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_k$ 的第一个分量表示蛋白质序列位置信息,所以由 $\boldsymbol{X}_i=\boldsymbol{X}_k$ 只能推出 $i=k$。

性质 1 与性质 2 表明,该图形表示是非退化的,从而不会丢失蛋白质序列的主要信息。

由于本文中矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 均为对角矩阵,有

$$\boldsymbol{X}_i=\boldsymbol{A}\boldsymbol{X}_{i-1}+\boldsymbol{B}\boldsymbol{S}_i=\boldsymbol{A}^i\boldsymbol{X}_0+\boldsymbol{B}(\boldsymbol{A}^{i-1}\boldsymbol{S}_1+\boldsymbol{A}^{i-2}\boldsymbol{S}_2+\cdots+\boldsymbol{A}\boldsymbol{S}_{i-1}+\boldsymbol{S}_i)。$$

因为 $\boldsymbol{X}_0=(0,0,0,0,0)^{\mathrm{T}}$,所以 $\boldsymbol{X}_i=\boldsymbol{B}(\boldsymbol{A}^{i-1}\boldsymbol{S}_1+\boldsymbol{A}^{i-2}\boldsymbol{S}_2+\cdots+\boldsymbol{A}\boldsymbol{S}_{i-1}+\boldsymbol{S}_i)$。记第 k 个氨基酸残基对应向量为

$$\boldsymbol{S}_k=(1,\boldsymbol{S}_2^{(k)},\boldsymbol{S}_3^{(k)},\boldsymbol{S}_4^{(k)},\boldsymbol{S}_5^{(k)})^{\mathrm{T}},$$

$$\boldsymbol{X}_k=(k,x_2^{(k)},x_3^{(k)},x_4^{(k)},x_5^{(k)})^{\mathrm{T}},\quad k=1,2,\cdots,$$

则图形表示中第 i 个点的坐标为:

$$\begin{pmatrix}i\\x_2^{(i)}\\x_3^{(i)}\\x_4^{(i)}\\x_5^{(i)}\end{pmatrix}=\mathrm{diag}(1,b_2,b_3,b_4,b_5)\left[\mathrm{diag}(1,a_2^{i-1},a_3^{i-1},a_4^{i-1},a_5^{i-1})\times\begin{pmatrix}1\\S_2^{(1)}\\S_3^{(1)}\\S_4^{(1)}\\S_5^{(1)}\end{pmatrix}+\mathrm{diag}(1,a_2^{i-2},a_3^{i-2},a_4^{i-2},a_5^{i-2})\times\begin{pmatrix}1\\S_2^{(2)}\\S_3^{(2)}\\S_4^{(2)}\\S_5^{(2)}\end{pmatrix}+\cdots+\mathrm{diag}(1,a_2,a_3,a_4,a_5)\times\left[\begin{pmatrix}1\\S_2^{(i-1)}\\S_3^{(i-1)}\\S_4^{(i-1)}\\S_5^{(i-1)}\end{pmatrix}+\begin{pmatrix}1\\S_2^{(i)}\\S_3^{(i)}\\S_4^{(i)}\\S_5^{(i)}\end{pmatrix}\right]\right]=\mathrm{diag}(1,b_2,b_3,b_4,b_5)$$

$$\left[\begin{pmatrix}1\\a_2^{i-1}S_2^{(1)}\\a_3^{i-1}S_3^{(1)}\\a_4^{i-1}S_4^{(1)}\\a_5^{i-1}S_5^{(1)}\end{pmatrix}+\cdots+\begin{pmatrix}1\\a_2S_2^{(i-1)}\\a_3S_3^{(i-1)}\\a_4S_4^{(i-1)}\\a_5S_5^{(i-1)}\end{pmatrix}+\begin{pmatrix}1\\S_2^{(i)}\\S_3^{(i)}\\S_4^{(i)}\\S_5^{(i)}\end{pmatrix}\right]=$$

$$\begin{pmatrix}i\\b_2(a_2^{i-1}S_2^{(1)}+\cdots+a_2S_2^{(i-1)}+S_2^{(i)})\\b_3(a_3^{i-1}S_3^{(1)}+\cdots+a_3S_3^{(i-1)}+S_3^{(i)})\\b_4(a_4^{i-1}S_4^{(1)}+\cdots+a_4S_4^{(i-1)}+S_4^{(i)})\\b_5(a_5^{i-1}S_5^{(1)}+\cdots+a_5S_5^{(i-1)}+S_5^{(i)})\end{pmatrix}。$$

因为对于 $j=1,2,3,4,5,0\leqslant S_j^{(k)}\leqslant 1$，所以当 $0<a_j<1$ 时，$0<b_j=1-a_j<1$，故

$$0\leqslant x_j^{(i)}\leqslant b_j(a_j^{i-1}+\cdots+a_j+1)=b_j\times\frac{1-a_j^i}{1-a_j}=(1-a_j)\times\frac{1-a_j^i}{1-a_j}=1-a_j^i<1,i=1,2,\cdots;$$

当 $a_j=1$ 时，$b_j=1$，故

$$0\leqslant x_j^{(i)}\leqslant i,i=1,2,\cdots。$$

上述推导给出了该图形表示中每个点各个坐标的取值范围。

3　图形表示在蛋白质相似性分析中的应用

为了说明上述方法的可靠性与有效性，本文利用提出的方法比较 10 种物种的 ND5 蛋白质序列（NADH dehydrogenates subunit 5）的相似性。10 种物种的 ND5 蛋白质的相关信息见表 2。

表 2　10 种物种的 ND5 蛋白质的相关信息

物种	物种简写	中文名称	ID 号(UniProtKB)	长度
Homo sapiens	Hom	人	P03915	603
Pantroglodytes	Pan	黑猩猩	Q35648	603
Rattusnorvegicus	Rat	大鼠	P11661	609
Musmusculus	Mus	小鼠	P03921	607
Didelphisvirginiana	Did	负鼠	P41309	602
Gallus gallus	Gal	鸡	P18940	605
Bostaurus	Bos	牛	Q85BD6	606
Canis lupus familiaris	Can	狗	Q9ZZ57	606
Drosophila melanogaster	Dro	果蝇	B6E0R2	572
Daniorerio	Dan	鱼	Q9MIY0	606

首先根据表 1 的结果给出 20 种氨基酸残基的五维坐标，对于迭代函数 $\boldsymbol{X}_i=\boldsymbol{A}\boldsymbol{X}_{i-1}+\boldsymbol{B}\boldsymbol{S}_i$，为得到这 10 种物种最优的比较结果，经过数学方法筛选后，本文取

$$\boldsymbol{A}=\mathrm{diag}\left(1,1,\frac{3}{4},1,\frac{3}{4}\right),\boldsymbol{B}=\mathrm{diag}\left(1,1,\frac{1}{4},1,\frac{1}{4}\right),$$

即广义迭代函数为：

$$\boldsymbol{X}_i=\mathrm{diag}\left(1,1,\frac{3}{4},1,\frac{3}{4}\right)\boldsymbol{X}_{i-1}+\mathrm{diag}\left(1,1,\frac{1}{4},1,\frac{1}{4}\right)\boldsymbol{S}_i \tag{5}$$

对于蛋白质序列 $S_1=s_1s_2\cdots s_M$ 和序列 $S_2=t_1t_2\cdots t_N$（这里 M 与 N 分别表示两条序列长度，不妨设 $M\geqslant N$）的图形表示 $\boldsymbol{X}_0\boldsymbol{X}_1\cdots\boldsymbol{X}_M$ 与 $\boldsymbol{Y}_0\boldsymbol{Y}_1\cdots\boldsymbol{Y}_N$。笔者刻画了它们的差异性。

首先计算每一个图形表示中相邻两个氨基酸残基位点的差向量。计算公式为：

$$\boldsymbol{C}_i^1=(x_1^{(i)}-x_1^{(i-1)},x_2^{(i)}-x_2^{(i-1)},x_3^{(i)}-x_3^{(i-1)},x_4^{(i)}-x_4^{(i-1)},x_5^{(i)}-x_5^{(i-1)}):=(1,U_2^{(i)},U_3^{(i)},U_4^{(i)},U_5^{(i)}),(i=1,2,\cdots,M),$$

$$\boldsymbol{C}_i^2=(y_1^{(i)}-y_1^{(i-1)},y_2^{(i)}-y_2^{(i-1)},y_3^{(i)}-y_3^{(i-1)},y_4^{(i)}-y_4^{(i-1)},y_5^{(i)}-y_5^{(i-1)}):=(1,V_2^{(i)},V_3^{(i)},V_4^{(i)},V_5^{(i)}),(i=1,2,\cdots,N),$$

其中：$x_j^{(i)}$ 表示第一条序列的第 i 个位点所对应向量的第 j 个分量，$y_j^{(i)}$ 表示第二条序列的第 i 个位点所对应向量的第 j 个分量（$j=1,2,3,4,5$）。并计算前 N 个向量之间夹角的余弦值为：

$$\cos\theta_i=\frac{\boldsymbol{C}_i^1\cdot\boldsymbol{C}_i^2}{\|\boldsymbol{C}_i^1\|\|\boldsymbol{C}_i^2\|},i=1,2,\cdots,N。$$

然后计算第 i 个位点之间距离为：

$$d_{1,2}^{(i)}=\frac{1-\cos\theta_i}{2}。$$

定义两个蛋白质序列 $S_1=s_1s_2\cdots s_M$ 和序列 $S_2=t_1t_2\cdots t_N$ 的距离为：

$$D_{1,2}=\sum_{i=1}^{N}d_{1,2}^{(i)}+K\times(M-N),$$

其中采用参数 K 来刻画不同长度序列之间的距离。由于位点之间距离 $d_{1,2}^{(i)}\in[0,1]$，在此取 $K=1$。

利用上述数值刻画方法,笔者计算表 2 中 10 种物种的 ND5 蛋白质之间的距离,结果见表 3。一般地,2 种物种之间距离越小说明这两个物种相似度越高。

表 3 采用本文方法所得的 10 种物种 ND5 蛋白质距离矩阵

物种	Pan	Rat	Mus	Did	Gal	Bos	Can	Dro	Dan
Hom	1.888	13.916	15.275	23.872	27.416	13.723	11.915	36.073	30.815
Pan		13.529	14.815	23.357	26.632	13.434	11.440	35.917	29.908
Rat			8.931	29.122	30.838	17.698	10.009	46.354	34.380
Mus				23.099	24.091	20.909	10.777	41.997	29.580
Did					26.376	24.341	22.149	40.186	35.210
Gal						25.572	24.613	43.300	38.657
Bos							7.789	42.330	30.340
Can								39.385	28.524
Dro									43.742

观察表 3 可以看出,人与黑猩猩、牛与狗、大鼠与小鼠的 ND5 蛋白质之间的距离相对较小(均小于 10),说明这些物种之间具有较高的相似性。并且上述 6 种物种的 ND5 蛋白质之间的距离均小于 20,说明它们之间具有一定的相似性。众所周知在表 2 中的 10 种物种中,上述 6 种物种均为哺乳类动物。

此外,在所有物种中,果蝇与其他物种之间的距离都较大(均大于 35),表明果蝇与其余 9 种物种的生物相似性较低。由生物学知识,在这 10 种物种中,果蝇属于昆虫。这个结果符合一般生物学上的规律。

进一步地,本文利用表 2 的结果,应用 UPGMA 方法构建了这 10 种物种的进化树,结果如图 1 所示。图 1 表明人与黑猩猩、牛与狗、大鼠与小鼠分属于同一分支,之后它们被分入一个大的分支。在这 10 种物种中,属于哺乳类动物有人、黑猩猩、大鼠、小鼠、牛、狗和负鼠。进化树显示与哺乳类动物相似度比较接近的依次为鸡、鱼和果蝇。以上结果符合生物学中一般的进化关系。这些结果表明该方法在分析 ND5 蛋白质序列进化关系上具有一定的可靠性与有效性。

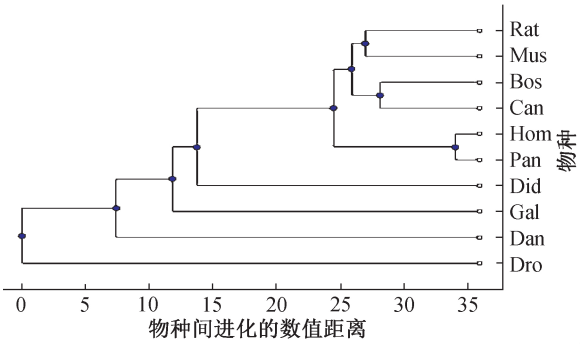

图 1 基于表 3 得到的 10 种物种的 ND5 蛋白质进化树

在生物信息学中,最常用的生物序列多重比对方法是 ClustalW 方法。为了进一步说明本文方法的可靠性,本文利用 MEGA 软件中的 ClustalW 方法计算了表 2 中 10 种物种的距离矩阵,见表 4。

表 4 采用 ClustalW 所得的 10 种物种 ND5 蛋白质距离矩阵

物种	Pan	Rat	Mus	Did	Gal	Bos	Can	Dro	Dan
Hom	0.055	0.396	0.409	0.436	0.560	0.324	0.349	1.097	0.579
Pan		0.396	0.406	0.439	0.577	0.314	0.339	1.091	0.579
Rat			0.190	0.442	0.541	0.334	0.362	1.097	0.591
Mus				0.439	0.529	0.347	0.372	1.129	0.582
Did					0.514	0.431	0.434	1.141	0.550
Gal						0.496	0.520	1.141	0.526
Bos							0.199	1.416	0.538
Can								1.157	0.550
Dro									1.113

观察表 4,能看出距离最小的依然是人与黑猩猩、牛与狗、大鼠与小鼠。果蝇与其他物种之间的距离是最大的。利用 UPGMA 方法与 Matlab 软件,笔者根据表 4 构建的 10 种物种的进化树与图 1 完全一致。这进一步说明本文的方法是可靠的和有效的。

此外,通过计算表 3 与表 4 结果的相关系数可以发现两个距离矩阵的整体相关系数为 0.9402。进一步地,逐行计算两个距离矩阵之间的相关系数,结果见表 5 第一列。同样地,逐行计算已有文

献的结果与表 4 结果之间的相关系数，结果　　见表 5。

表 5　本文方法、文献方法与 ClustalW 方法结果的相关系数

物种	本文	文献[24]	文献[27](表 3)	文献[27](表 4)	文献[18](表 3)	文献[18](表 4)	文献[28](表 5)
om	0.949	0.931	0.914	0.457	0.931	0.718	0.906
Pan	0.947	0.875	0.922	0.786	0.840	0.766	0.682
Rat	0.947	0.959	0.717	0.369	0.445	0.838	0.670
Mus	0.889	0.944	0.671	0.488	0.419	0.456	0.591
Did	0.929	0.776	−0.475	−0.204	−0.298	−0.433	−0.134
Gal	0.930	0.921	—	—	—	—	—
Bos	0.949	0.861	—	—	—	—	—
Can	0.953	0.923	—	—	—	—	—
Dro	0.955	0.873	—	—	—	—	—
Dna	0.963	0.940	—	—	—	—	—

观察表 5 可以看出，本文方法的结果与 ClustalW 的结果的相关性系数相对较大，说明它们之间具有很强的相关性。这表明本文方法与已有的蛋白质图形表示方法相比结果更加可靠。与 ClustalW 方法相比，本文方法的计算量较小，适用于更长蛋白质序列的相似性比较。综上所述，本文的方法对于比较生物序列的相似性是可靠的和有效的。

4　结　论

在分析经典图形表示和谱图表示中迭代函数基础上，本文提出了一种新的广义迭代函数 $\boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{X}_{i-1} + \boldsymbol{B}\boldsymbol{S}_i$。同时，根据氨基酸残基的 12 种理化性质，经过归一化及相关性分析等数学处理选出其中 4 种理化性质作为确定氨基酸残基坐标的依据。证明了形如 $\boldsymbol{X}_i = \mathrm{diag}(a_1,a_2,a_3,a_4,a_5)\boldsymbol{X}_{i-1} + \mathrm{diag}(b_1,b_2,b_3,b_4,b_5)\boldsymbol{S}_i$ 的广义迭代函数给出的图形表示是非退化的，从而表明该图形表示没有蛋白质序列信息的丢失。本文对 10 种物种的 ND5 蛋白质序列，应用广义迭代函数

$$X_i = \mathrm{diag}\left(1,1,\frac{3}{4},1,\frac{3}{4}\right)X_{i-1} + \mathrm{diag}\left(1,1,\frac{1}{4},1,\frac{1}{4}\right)S_i,$$

将每一个蛋白质序列转化为五维空间中的向量，通过计算两个蛋白质序列对应的向量之间夹角的余弦值得到两个图形间的距离，以此来刻画它们之的相似程度。利用相似性结果构建了 10 种物种的进化树，进化树的结果与生物物种进化规律一致。将本文结果、其它图形表示方法与 ClustalW 方法所得到的结果进行比较，发现本文的方法是可靠的和有效的。

本文将已有蛋白质图形表示方法中的迭代函数推广到高维空间，使得氨基酸残基不同理化性质对应不同的参数统计模型，改进了已有模型单一性的缺点。当更多氨基酸残基的理化指标融合进蛋白质的图形表示中，如何给出合理的迭代函数中的系数矩阵是未来研究的重点。

参考文献：

[1] Vinga S, Almeida J. Alignment-free sequence comparison-a review[J]. Bioinformatics, 2003, 19(4): 513-523.

[2] 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.

[3] Zhang C T, Zhang R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique[J]. Nucleic Acids Research, 1991, 19(22): 6313-6317.

[4] Zhang R, Zhang C T. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences [J]. Biological Structure Dynamics, 1994, 11(4): 767-782.

[5] Nandy A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes[J]. Current Science, 1994, 66: 309-314.

[6] Randic M, Vracko M, Nandy A, et al. On 3-D graphical representation of DNA primary sequences and their numerical characterization[J]. Journal of Chemical Information and Computer Sciences, 2000, 40(5): 1235-1244.

[7] Zhang Y, Liao B, Ding K. On 2D graphical representation of DNA sequence of nondegeneracy[J]. Chemical Physics Letters, 2005, 411(1): 28-32.

[8] Zhang Y, Liao B, Ding K. On 3D D-curves of DNA sequences[J]. Molecular Simulation, 2006, 32(1): 29-34.

[9] Jeffrey H J. Chaos game visualization of sequences[J]. Computers and Graphics, 1992, 16(1): 25-33.

[10] Wu D, Roberge J, Cork D J, et al. Computer visualization of long genomic sequences [C]. IEEE Conference on Visualization, 1993: 308-315.

[11] Randic M. 2-D graphical representation of proteins based on virtual genetic code[J]. SAR and QSAR in Environmental Research, 2004, 15(3): 147-157.

[12] Zhu X, Ping P, Qiu Y, et al. Similarities/dissimilarities analysis of protein sequences based on the appearance model [J]. Journal of Computational and Theoretical Nanoscience, 2017, 14(3): 1449-1460.

[13] Hu H, Li Z, Dong H. 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.

[14] Hou W, Pan Q, He M. A new graphical representation of protein sequences and its applications [J]. Physica A: Statistical Mechanics and Its Applications, 2016, 444: 996-1002.

[15] He P A, Li D, Zhang Y, et al. A 3D graphical representation of protein sequences based on the Gray code[J]. Journal of Theoretical Biology, 2012, 304 (1): 81-87.

[16] Randic M. 2-D Graphical representation of proteins based on physicochemical properties of amino acids[J]. Chemical Physics Letters, 2007, 444(1): 176-180.

[17] Yao Y H, Dai Q, Li C, et al. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation[J]. Journal of Computational Chemistry, 2010, 31(5): 1045-1052.

[18] Maaty M I A E, Abo-elkhier M M, Elwahaabm A A. 3D graphical representation of protein sequences and their statistical characterization [J]. Physica A: Statistical Mechanics and Its Applications, 2010, 389 (21): 4668-4676.

[19] Wu C, Gao R, Yang DM, et al. A novel model for protein sequence similarity analysis based on spectral radius [J].Journal of Theoretical Biology, 2018, 446.

[20] Randic M, Zupan J, Balaban A T, et al. Graphical representation of proteins[J]. Chemical Review, 2011, 111(2): 790-862.

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

[22] He P A, Xu S, 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.

[23] Yao Y H, Dai Q, Li C, et al. Analysis of similarity/dissimilarity of protein sequences [J]. Proteins-structure Function and Bioinformatics, 2010, 73 (4): 864-871.

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

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

[26] Qi Z H, Li K C, Ma J L, et al. Novel method of 3-dimensional graphical representation for proteins and its application [J/OL]. Evolutionary Bioinformatics Online, 2018, 14. (2018-06-12)[2018-08-21]. https://journals.sagepub.com/doi/10.1177/1176934318777755.

[27] Wen J, Zhang Y Y. A 2D graphical representation of protein sequence and its numerical characterization[J]. Chemical Physics Letters, 2009, 476 (4/5/6): 281-286.

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

[29] Nikhila K S, Nair VV. Protein sequence similarity analysis using computational techniques [J]. Materials Today, 2018, 724-731.

[30] Kawashima S, Pokapowski P, Pokarowska M, et al. AAindex: Amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2008, 36 (1): D202-D205.

(责任编辑:康 锋)