

一种基于混沌游走表示法的种系发生树构建*

陈勃¹, 季平²

(1. 福州大学 数学与计算机科学学院, 福州 350108; 2. 香港理工大学 工业与系统工程学系, 香港)

摘要: 针对生物信息学领域中种系发生树构建这一重要课题的需要, 利用 DNA 碱基序列的频度混沌游走表示法, 提出一种碱基序列自重复性的度量和一种序列间相关性的度量, 并由此出发, 提出了一种新的以此种相关性为依据的聚类方法。利用这样的方法, 通过 GenBank 中获取的线粒体 DNA 数据构建了一棵包含 20 个物种的种系发生树。实验结果验证了新提出的度量方法以及聚类方法在种系发生树构建问题上的有效性。此外, 由于这种方法使用碱基序列的图形表示法, 而非传统的串形表示法, 避免了建树过程中序列间联配的步骤。

关键词: 种系发生树; 混沌游走表示法; 聚类; 序列分析; 生物信息学

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)08-2956-05

doi:10.3969/j.issn.1001-3695.2012.08.040

Phylogenetic tree-building based on chaos game representation

CHEN Bo¹, JI Ping²

(1. College of Mathematics & Computer Science, Fuzhou University, Fuzhou 350108, China; 2. Dept. of Industrial & Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China)

Abstract: Phylogenetic tree-building is one of the most significant issues in bioinformatics. In order to develop a new tree-building method, this paper proposed new measures for the sequence repeatability and the sequence correlation based on the chaos game representation with frequencies (FCGR) of DNA sequences. Adopting such measures, then gave a new clustering approach for phylogenetic tree-building. The experiment obtained a phylogenetic tree of 20 species based on their mitochondrial DNA sequence data collected from GenBank. The result shows that the new measures and the clustering approach benefit to the phylogenetic tree-building. Moreover, such a tree-building avoids the alignment process since it is based on FCGR instead of the traditional representation of DNA sequences.

Key words: phylogenetic tree; chaos game representation (CGR); clustering; sequence analysis; bioinformatics

1 研究背景

种系发生学作为生物学中的一个重要部分, 旨在探究物种间、种群间、组织间或基因间的进化关系^[1]。用树型结构来表示生物家族的想法在比达尔文提出进化论更早的亚里士多德的时代就开始了, 生物学家们积累了大量的从形态学角度来构建物种间种系发生树的经验。而从生物信息学的角度, 研究者们则试图从物种的 DNA 碱基序列或蛋白质氨基酸序列出发来构建这样的树^[2~4]。这一课题的核心问题有: a) 理清物种间的家族条理; b) 推测各物种从它们共同祖先分支出来的时间^[5]。图 1 就是物种间种系发生树的一个简单例子。

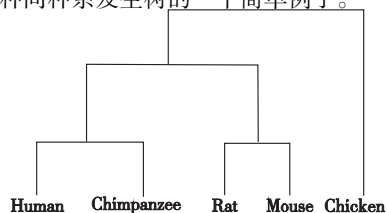


图1 一棵简单的种系发生树

不同于传统的串形表示法, 本文利用一种图形表示法, 即混沌游走表示法 (chaos game representation, CGR) 来表示 DNA 碱基序列^[6~8]。基于这种图形表示法, 本文提出一种新的序列自重复性度量和一种新的序列间相关性度量, 于是种系发生树就可以利用一种新的以上述度量为依据的聚类方法构建出来。

1.1 构建种系发生树的常见方法

构建种系发生树的方法可大致分为基于距离 (相关性) 的聚类方法、吝啬算法和极大似然算法。

在基于距离 (相关性) 的方法中, 每两个样本序列 (碱基序列或氨基酸序列) 之间的距离 (相关性) 被测算出来, 并且存放在一个矩阵当中。所谓距离 (相关性) 的度量, 根据算法的不同, 可以有多种定义。例如, Jukes-Cantor 模型^[9]中把距离定义为两个序列联配 (alignment) 后配对碱基数目的函数, 而 Hamming 距离^[10]则定义为序列联配中最小变异的步骤数。无论距离 (相关性) 是如何定义的, 当这些序列到序列的距离 (相关性) 被测得并记录下来后, 就可以利用某些聚类方法来实现种系发生树的构建。解决此问题常见的聚类方法包括非加权配

收稿日期: 2012-02-02; 修回日期: 2012-03-10 基金项目: 国家自然科学基金资助项目 (41076059); 福建省自然科学基金计划项目 (2012J05114); 福建省科技创新平台计划项目 (2009J1007)

作者简介: 陈勃 (1984-), 男, 福州人, 博士, 主要研究方向为生物信息学、计算机图形学 (bo.chen@fzu.edu.cn); 季平 (1961-), 男, 南京人, 副教授, 博士, 主要研究方向为生物信息学、优化算法。

对算术平均法 (unweighted pair-group method with arithmetic mean, UPGMA)^[11]、变形距离法 (transformed distance method)^[12]、邻近关系法 (neighbors-relation method)^[13] 和邻接法 (neighbor-joining method)^[14, 15]。

齐嵩算法^[15-17]的思想基于一个假设前提——进化是最有效的,即自然进化的总是只包含那些必须的步骤。这类算法尝试搜索所有可能的树结构,并在其中找出一棵“最高效的”树,这棵树的每个分叉处之下的分支具有最少的变异。目前已经有许多优秀的算法和软件包被开发出来,用于实现这一思想并找到近似的优化解,其中包括 Swofford 研发的 PAUP、Maddison 研发的 MacClade、Felsenstein 研发的 Phylip 以及 Goloboff 研发的 SPA。

在极大似然算法中,首先确定各种位点变异(转变)所发生的概率,而后对于每种可能的树结构,可以计算出沿这棵树获得目前观察到的这组碱基(或氨基酸)序列的概率,于是便可以从挑选出概率最大的那棵树作为结果。Cavalli-Sforza 等人^[18]最早对这种思想加以实现,之后又将其应用于氨基酸序列和碱基序列上^[19, 20]。由于计算的困难,此类方法在之前的研究中很少得到应用,但随着计算机技术的高速发展,此类方法也获得了研究者的重视^[15]。

虽然构建种系发生树的方法已有很多,但提高准确性和降低计算时间一直是研究者们关心的问题。值得注意的是,上述的大多数方法都依赖于序列联配(sequence alignment)的结果,而实现更高效的序列联配本身就是生物信息学领域的一项课题。于是,本项研究尝试通过其他手段,而非联配来估计序列间的关联性,从而实现种系发生树的构建。

1.2 频度混沌游走表示法

与传统的碱基序列串形表示法不同,混沌游走表示法是碱基序列的一种图形表示,由 Jeffrey 在 1990 年提出^[5]。这种表示法通过一系列迭代的过程构建一个含有分形模式的二维图像,以此来表示一个 DNA 的碱基序列。具体的过程是:首先在平面上创建一个方形区域,方形的四个顶点分别表示四种碱基,即 A、T、G 和 C,并将一个点放置在方形中央;然后逐位读取所要表示的碱基序列中的碱基,每读取一个碱基,就将方形中的点向该碱基对应的顶点移动,移动的距离为点的当前位置到目标顶点距离的一半。如此迭代地移动这个点直到序列终止,则这个移动的点在方形区域内留下的痕迹构成一个二维图像,此图像被认为是上述碱基序列的图形表示。这种方法被称做混沌游走表示法(CGR)(图2)。

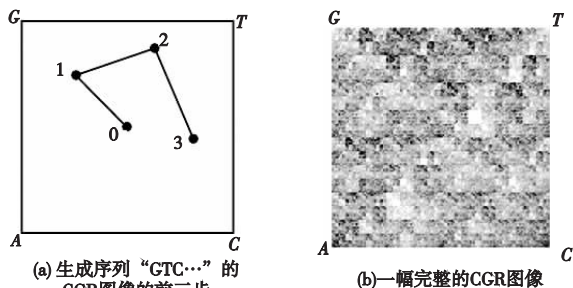


图2 CGR图像的生成

Jeffrey 指出,各种物种所携带的 DNA 碱基序列的 CGR 图像中含有其独特的分形模式^[5]。但由于此种方法最初只被用于可视化碱基序列,Jeffrey 并没有给出更详细的数学解释。在

这之后,研究者们开始尝试用各种理论来描述和研究 CGR。1993 年,Oliver 等人^[21]用信息学理论来分析 CGR 图像,并主要探讨了子序列的熵以及序列中的非随机因素。还有学者提出,CGR 图像实际上是状态的表示法。例如,Baldi 等人^[22]曾指出 CGR 图像是一个马尔可夫概率表,而若把所有可能的 CGR 图像组成的向量空间看做一个参考系,则所有可能的碱基序列都可以在其中找到它的位置,因此所有的序列模式都可以在此空间内得到表示。也有很多学者专注于对 CGR 图像多重分形特性的研究。例如,Gutiérrez 等人^[23]将 CGR 图像变换为一个一维的迭代函数系(iterated function system, IFS),然后研究其多重分形特性。Tavassoly 等人^[24]通过设定一些特殊的空间度量来研究 CGR 图像的多重分形特性,并以此为依据得出了一些关于物种划分的结论。仿真技术也同樣被应用在了对 CGR 图像的研究中。例如,Yu 等人^[25]在 2008 年利用一个递归迭代函数系(recurrent iterated function system, RIFS)进行的对 CGR 图像的仿真实验。

为了更好地表现序列中蕴涵的特殊属性,以适应不同的应用领域的需要,混沌游走表示法具有许多变种^[26-29]。例如,Randić 等人^[27]通过将方形区域的顶点放置到三维空间中,设计了一种新颖的用于蛋白质氨基酸序列的图形表示法。在 2008 年,Randić^[28]又将 CGR 精简称为一维的谱的形式,并进行了一些研究。

CGR 的另一个具有最重要意义的变种是频度混沌游走表示法(chaos game representation with frequencies, FCGR)。1999 年,Deschavanne 等人^[29]提出了这种变种用于便捷地表示序列中各个子序列所出现的频度,Almeida 等人^[30]在 2001 年将其命名为 FCGR。在 FCGR 中,原先的 CGR 图像被分割为 4^L 个更加细小的区域,每个区域被填充上不同的颜色(灰度)以表示落在该区域内的点迹的多少(图3)。在下文中,对于分辨率为 4^L 的 FCGR 图像,直接使用落在第 i 个小区域($i = 0, 1, \dots, 4^L - 1$)内的点迹个数 n_i 来表示为此区域填充的颜色值,而整个图像写做一个集合为

$$\text{FCGR}^L = \{n_i, 0 \leq i < 4^L\}$$

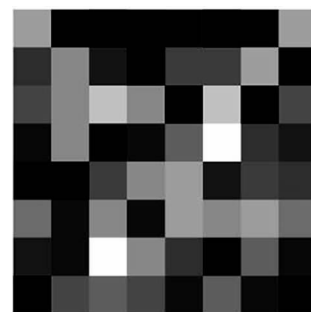


图3 一幅FCGR图像

Deschavanne 等人^[29]指出,每个细小的区域实际上表示了一个长度为 L 的子序列模式,而该区域的颜色正好可以表示这个模式在整个序列中出现的次数(频度)。近年来,基于 FCGR,研究者们设计了许多新的度量方法来考察序列间的距离和相关性。例如,Yu 等人^[31]通过提出基于 FCGR 的新度量方法而进行的序列相关性的研究,Wang 等人^[32]将一系列传统的基于串形表示法的距离计算方法移植到了基于 FCGR 图像的序列距离计算上来,提出了针对 FCGR 图像的 Euclid、Ham-

ming、Pearson 距离等距离形式,利用这些距离形式分别构建了含有 26 个物种的种系发生树,并对在不同距离形式下生成的树作了比较和讨论。可以看到,利用不同的距离(相关性)形式,可能得出不同结构的种系发生树,而目前没有哪种距离(相关性)形式可以保证得到绝对正确的结论,能够得到相对合理的结果的计算机方法则被认为是较有参考意义的。

2 方法介绍

2.1 序列自重复性

对于足够长的 DNA 碱基序列 s ,考察 $FCGR^L$ 中的元素 n_i ,即 s 中子序列模式 p_i (长度为 L)出现的次数(频度)。对于每个频度值 n ,以此频度出现在 s 中的模式的个数 μ 可以统计出来。于是可得到 n 到 μ 的函数(图 4)为

$$\mu(n) = \text{count}\{p_i | n_i = n, 0 \leq i < 4^L\} \quad (1)$$

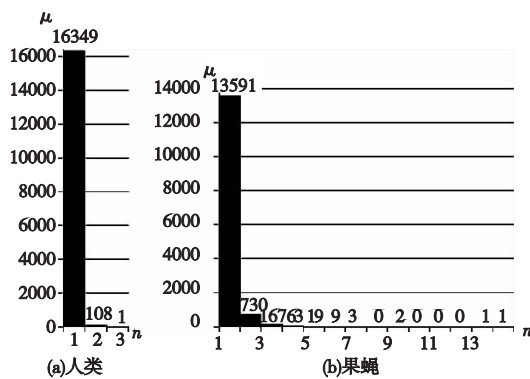


图 4 函数谱 $\mu(n)$ 的两个实例(取 $L=11$; $n=0$ 时的值未画出)

设随机变量 X 表示某个随机选取的模式在 s 中出现的次数。显然, X 的分布应当服从概率质量函数:

$$f(x) = \frac{\mu(x)}{4^L} \quad (2)$$

于是,对于序列 s ,定义 t 阶自重复性度量

$$R_t = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} f(x) = \frac{1}{4^L} \sum_{x=0}^{\infty} e^{tx} \mu(x) \quad (3)$$

这里,可将一阶自重复性度量 R_1 简称为序列 s 的自重复性度量 R

$$R = R_1 = E(e^X) \quad (4)$$

显然, R 实际上就是随机变量 e^X 的数学期望。一个序列如果具有较高的 R 值,则可说明此序列中具有较多高重复性的子序列模式。于是, R 可以用来表征给定序列的自重复性这一内在属性。

与此同时,还可以得到随机变量 e^X 的方差(本文中称做自重复性方差)为

$$V = \text{Var}(e^X) = E(e^{2X}) - E^2(e^X) = R_2 - R^2 \quad (5)$$

2.2 序列相关性

对于两条序列 s_1 和 s_2 ,分别具有分辨率为 4^L 的 FCGR 图像为

$$\begin{cases} FCGR_{s_1}^L = \{n_{i,s_1}, 0 \leq i < 4^L\} \\ FCGR_{s_2}^L = \{n_{i,s_2}, 0 \leq i < 4^L\} \end{cases} \quad (6)$$

则它们的自重复性度量分别为 $R(s_1)$ 和 $R(s_2)$,自重复性方差分别为 $V(s_1)$ 和 $V(s_2)$ 。设 X_1 和 X_2 是两个随机变量,分别表示某个随机选取的子序列模式 p (长度为 L)出现在 s_1 和 s_2 中

的次数,则根据前文的定义,有

$$\begin{aligned} R(s_1) &= E(e^{X_1}) & R(s_2) &= E(e^{X_2}) \\ V(s_1) &= \text{Var}(e^{X_1}) & V(s_2) &= \text{Var}(e^{X_2}) \end{aligned} \quad (7)$$

将 s_1 和 s_2 串连起来,构成复合序列 $s_1 + s_2$ 。若原序列 s_1 和 s_2 足够长,根据 FCGR 的特性(小区域的填充颜色值即为对应子序列模式在序列中出现的次数),则可直接通过将原先的两个 FCGR 图像叠加来近似计算复合序列 $s_1 + s_2$ 的 FCGR 图像:

$$FCGR_{s_1+s_2}^L = \{n_{i,s_1+s_2} \approx n_{i,s_1} + n_{i,s_2}, 0 \leq i < 4^L\} \quad (8)$$

并且也可以从中得到一个针对复合序列 $s_1 + s_2$ 的函数 $\mu(n)$ 。同样,用另一个随机变量 X' 来表示某个随机选取的模式出现在 $s_1 + s_2$ 中的次数,则可计算出该复合序列的自重复性度量为

$$R(s_1 + s_2) = E(e^{X'}) = \frac{1}{4^L} \sum_{x'=0}^{\infty} e^{x'} \mu(x') \quad (9)$$

若 s_1 和 s_2 足够长,则根据式(8)以及随机变量 X_1 、 X_2 和 X' 的定义可知

$$X' \approx X_1 + X_2 \quad (10)$$

于是有

$$E(e^{X_1+X_2}) = E(e^{X'}) = R(s_1 + s_2) \quad (11)$$

则随机变量 e^{X_1} 和 e^{X_2} 之间的线性相关系数可计算为

$$\begin{aligned} \rho(e^{X_1}, e^{X_2}) &= \frac{E(e^{X_1+X_2}) - E(e^{X_1})E(e^{X_2})}{\sqrt{\text{Var}(e^{X_1})\text{Var}(e^{X_2})}} = \\ &= \frac{R(s_1 + s_2) - R(s_1)R(s_2)}{\sqrt{V(s_1)V(s_2)}} \end{aligned} \quad (12)$$

从式(12)可以看出,这个系数可以仅从两个序列及其复合序列的 FCGR 图像中算得。该值体现了两个随机变量间的关联性,同时也从子序列模式的重复性这一角度表征了两个序列 s_1 和 s_2 之间的相似程度。本文将这个值称为序列 s_1 和 s_2 的相关性度量:

$$r(s_1, s_2) = \frac{R(s_1 + s_2) - R(s_1)R(s_2)}{\sqrt{V(s_1)V(s_2)}} \quad (13)$$

理论上说, $r(s_1, s_2)$ 的范围从 $-1 \sim 1$,而负值表明随机变量 e^{X_1} 和 e^{X_2} 负相关。但在实际中,由于实验序列的同源性,排除了负相关的情况,明显的负值并不会在实际应用中出现。于是,在实际中认为 r 值的范围从 $0 \sim 1$ (极微小的负值被认为是 0)。越相关的序列将得到越大的 r 值,而对于完全相同的两条序列,有 $r = 1$ 。

此外,将式(13)变换得到

$$R(s_1 + s_2) = R(s_1)R(s_2) + r(s_1, s_2) \sqrt{V(s_1)V(s_2)} \quad (14)$$

由式(14)可以看出,当将两条序列串连起来时,新序列的自重复性度量是两部分之和:一部分是原先两个序列的自重复性度量之积;而另一部分与原先两个序列间的相关性度量有关。这说明序列的自重复性度量也可以看做是其中子序列间相关性积累的结果。

2.3 聚类方法

利用前文中给出的序列自重复性度量和相关性度量,就可以通过一种新的聚类方法来构建种系发生树。其方法的步骤描述如下:

a)从数据库(如 GenBank)中取得所要聚类的多个物种的线粒体 DNA(或其他同源 DNA)的参考序列,将这些序列存放

于表 Q 中。对于表 Q 中的每个序列 s_i ,在待建立的种系发生树 T 中建立终端节点 N_i 。

b)根据表 Q 中序列的分辨率为 4^L 的 FCGR 图像,计算每两个序列间的相关性度量。

c)找出具有最大相关性度量的序列对 s_a 和 s_b 。在树 T 中建立 N_a 和 N_b 的父节点 $N_{(a,b)}$,并在表 Q 中用一个虚拟的父序列 $s_{(a,b)}$ 来替换序列 s_a 和 s_b 。此处,“虚拟”的含义在于,对这样的序列,无须考虑序列的串形形式,而只推测其图形形式,即 FCGR 图像(推测的方法见下文)。

d)若表 Q 中只剩下一条序列,则所有的序列已经聚类完毕,树 T 即为所求的种系发生树;否则返回步骤 b)。

在步骤 c)中需要推测两条序列 s_a 和 s_b 的父序列 $s_{(a,b)}$ 的 FCGR 图像。假设一个子序列模式 p_i 在 s_a 和 s_b 中分别出现 n_{i,s_a} 次和 n_{i,s_b} 次,则在本文中推测该模式在其父序列 $s_{(a,b)}$ 中出现 $(n_{i,s_a} + n_{i,s_b})/2$ 次。于是有 $s_{(a,b)}$ 的 FCGR 图像为

$$\text{FCGR}_{s_{(a,b)}}^L = \{n_{i,s_{(a,b)}}\} = \frac{n_{i,s_a} + n_{i,s_b}}{2}, \quad 0 \leq i < 4^L \quad (15)$$

3 实验与结果

为了实验构建一棵含有 20 个物种的种系发生树,本文从 GenBank 取得了这 20 个物种的线粒体 DNA 的参考序列(表 1)。而后,将每条序列转换为分辨率为 4^{11} (即 $L = 11$)的 FCGR 图像,并计算出每两条序列间的相关性度量(表 2)。

表 1 20 个实验物种的线粒体 DNA 资料

ID	GenBank 序列代码	物种名称	ID	GenBank 序列代码	物种名称
1	X52392	Chicken	11	X14848	Rat
2	D38113	Chimpanzee A	12	X54253	Round Worm
3	D38116	Chimpanzee B	13	X72004	Seal A
4	V00654	Cow	14	X63726	Seal B
5	X03240	Fruit Fly	15	X69067	Shrimp
6	L06178	Honey Bee	16	L29771	Trout
7	J01415	Human	17	J04815	Urchin A
8	M91245	Loach	18	X12631	Urchin B
9	L20934	Mosquito	19	X61145	Whale A
10	V00711	Mouse	20	X72204	Whale B

表 2 实验中每两条线粒体 DNA 参考序列间的相关性度量

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.042	0.044	0.048	0	0	0.042	0.042	0.002	0.046	0.045	0	0	0	0.003	0.04	0	0	0.053	0.052
2		0.64	0.084	0	0	0.402	0.039	0.014	0.068	0.074	0	0	0	0.006	0.036	0	0	0.082	0.085
3			0.083	0	0	0.397	0.042	0.019	0.069	0.073	0	0	0	0.005	0.035	0	0	0.078	0.08
4				0	0	0.083	0.038	0.009	0.088	0.089	0	0	0	0.011	0.038	0	0	0.138	0.138
5					0.391	0	0.028	0.285	0	0	0.999	0	0	0	0	0	0	0	0
6						0.009	0.111	0	0	0.391	0	0	0	0	0	0	0	0	0
7							0.04	0.013	0.066	0.072	0	0	0	0.004	0.04	0	0	0.082	0.082
8								0.012	0.043	0.041	0.028	0	0	0.005	0.095	0	0	0.04	0.043
9									0.009	0.007	0.285	0	0	0.017	0.001	0	0	0.006	0.005
10										0.172	0	0	0	0.012	0.033	0	0	0.076	0.089
11											0	0	0	0.007	0.033	0	0	0.082	0.081
12												0	0	0	0	0	0	0	0
13													0.91	0	0	0	0	0	0
14														0	0	0	0	0	0
15															0.006	0.083	0.083	0.006	0.005
16																0	0.038	0	0.041
17																	0.999	0	0
18																		0	0
19																			0.458

按照第 2 章中给出的聚类方法,这 20 个物种的一棵二叉种系发生树产生了(图 5)。在这个结果中,亲缘关系很接近的物种被十分清晰地体现出来,如猩猩(chimpanzee)和人类(human)、大鼠(rat)和小鼠(mouse)。由于对祖先序列 FCGR 图像的估计存在累积误差,在接近树根处(较高级别的节点)的祖先序列可能并不准确。而事实上,在接近树根处,序列间的相关性度量已经非常微小(图 5 中,节点 α 、 β 、 γ 和 δ 两两间的相关性度量都不大于 10^{-4}),不足以作为聚类的依据。因此,在高级别处可以放宽“二叉”这一限制,即当表 Q 中的序列两两间的相关性度量极小时(不大于 10^{-4}),将这些序列直接聚类。采用这种方案得到如图 6 所示的种系发生树。可以看到,在这棵树中,脊椎生物(vertebrates)与其他物种被区别开来。在脊椎生物中,除了海豹(seals)之外,哺乳动物(mammals)、鸟类(birds)、鱼类(fishes)区分十分清晰。同时,在低等级上,亲缘很近的物种仍然得到了很好的聚类。

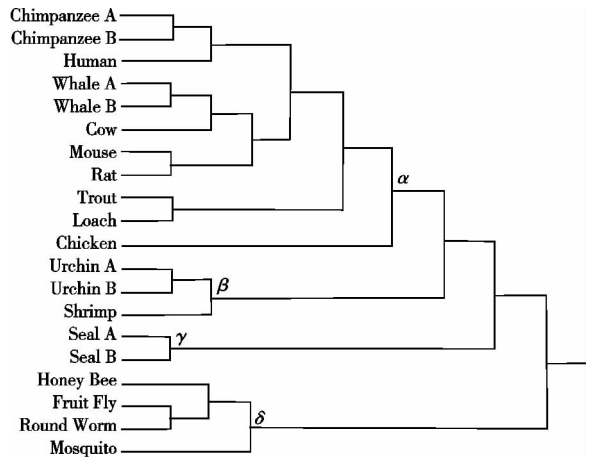


图 5 20 个实验物种的二叉种系发生树

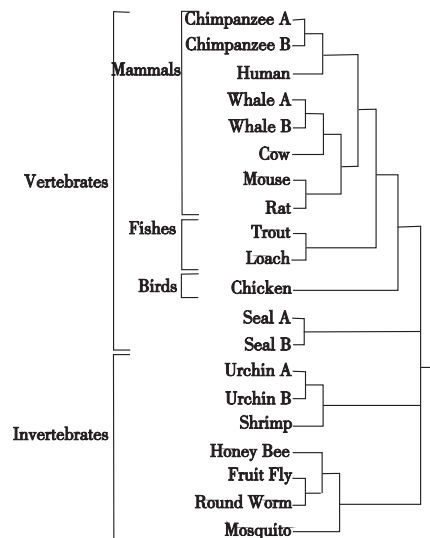


图 6 改进的种系发生树

与文献[32]中的一些建树结果相比较,本文得出的树(图 6)中的一些子结构与之相符合,而一些结构较之显得更加合理(从物种分类的角度)。另一方面,相比于第 1 章中介绍的许多基于序列串形表示的建树方法,本文提出的方法由于使用了序列的图形化表示,不依赖于序列联配的结果。因此,本文提出的序列自重复性度量、序列相关性度量以及基于这些度量的聚类方法可以作为无须联配的种系发生树构建的有效工具。

4 结束语

本文首先基于DNA碱基序列的图形表示法——频度混沌游走表示法(FCGR),提出了一种序列自重复性的度量。它可以看做是序列自身的属性,也可以看做是序列中子序列间相关性累加的结果。此外,本文还从序列自重复性的角度提出了一种序列间相关性的度量。这种度量可以在不需要进行序列间联配的前提下表示出同源序列间的相似程度。基于这种相关性度量,提出了一种新的聚类方法,用于种系发生树的构建。实验中利用此种方法,在不依赖于序列联配结果的情况下,得到了结构合理的种系发生树:对于近亲缘的物种,聚类结构十分清晰;在较大尺度的物种分类上,也得到了满意的结果。因此,相比于那些基于传统的序列串形表示法的度量和建树方法,本文提出的基于序列图形表示法的自重复性度量、序列相关性度量和聚类方法,更加适合无须联配的种系发生树构建。

参考文献:

- [1] LI W H. Molecular evolution[M]. Sunderland: Sinauer Associates Inc Publishers, 1997.
- [2] 耿立英, 张传生, 尹春光, 等. 依据锌指基因内含子序列构建羊、牛种系发生树的研究[J]. 黑龙江畜牧兽医, 2011(9): 9-11.
- [3] 刘建, 柳英, 王振海, 等. 宁夏地区人畜BDVp24基因系统发生树构建和序列分析[J]. 中华流行病学杂志, 2010, 31(2): 194-198.
- [4] 罗金, 刘光远, 田占成, 等. 基于18S rRNA基因序列的我国马梨形虫分类学地位分析[J]. 动物分类学报, 2011, 36(1): 99-103.
- [5] ABBAS A E, HOLMES S P. Bioinformatics and management science: some common tools and techniques[J]. Operations Research, 2004, 52(2): 165-190.
- [6] JEFFREY H. Chaos game representation of gene structure[J]. Nucleic Acids Research, 1990, 18(8): 2163-2170.
- [7] 高雷, 齐斌, 朱平. 基于混沌游走方法的Rh血型系统中RHD基因的分析[J]. 生命科学研究, 2009, 13(5): 408-412.
- [8] 张立婷, 管维红, 徐振源. 基于蛋白质CGR的线粒体蛋白质序列比对[J]. 计算机工程与应用, 2008, 44(13): 50-53.
- [9] JUKES T H, CANTOR C R. Evolution of protein molecules[M]//MUNRO H N. Mammalian Protein Metabolism. New York: Academic Press, 1969: 21-132.
- [10] HAMMING R W. Error detecting and error correcting codes[J]. Bell System Technical Journal, 1950, 29(2): 147-160.
- [11] SOKAL R R, MICHENER C D. A statistical method for evaluating systematic relationships[J]. University of Kansas Science Bulletin, 1958, 28: 1409-1438.
- [12] FARRIS J S. On the phenetic approach to vertebrate classification[M]//HECHT M K, GOODY P C, HECHT B M. Major Patterns in Vertebrate Evolution. New York: Plenum, 1977: 823-850.
- [13] SATTATH S, TVERSKY A. Additive similarity trees[J]. Psychometrika, 1977, 42(3): 319-345.
- [14] SAITOU N, NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees[J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.
- [15] 林昶东, 梁振, 郝宇, 等. 基于线粒体COI基因的风蝶亚科部分物种系统发育关系[J]. 动物分类学报, 2011, 36(3): 639-647.
- [16] ECK R V, DAYHOFF M O. Atlas of protein sequence and structure[M]. Washington DC: National Biomedical Research Foundation, 1966.
- [17] FITCH W M. On the problem of discovering the most parsimonious tree[J]. The American Naturalist, 1977, 111(978): 223-257.
- [18] CAVALLI-SFORZA L L, EDWARDS A W F. Phylogenetic analysis: models and estimation procedures[J]. American Journal of Human Genetics, 1967, 19(3): 233-257.
- [19] FELSENSTEIN J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters[J]. Systematic Zoology, 1973, 22(3): 240-249.
- [20] FELSENSTEIN J. Evolutionary trees from DNA sequences: a maximum likelihood approach[J]. Journal of Molecular Evolution, 1981, 17(6): 368-376.
- [21] OLIVER J L, BERNAOLA P. Entropic profiles of DNA sequences through chaos-game-derived images[J]. Journal of Theoretical Biology, 1993, 160(4): 457-470.
- [22] BALDI P, BRUNAK S. Bioinformatics: the machine learning approach[M]. Cambridge: MIT Press, 1998.
- [23] GUTIÉRREZ J M, RODRÍGUEZ M A, ABRAMSON G. Multifractal analysis of DNA sequences using a novel chaos-game representation[J]. Physica A: Statistical Mechanics and Its Applications, 2001, 300(1-2): 271-284.
- [24] TAVASSOLY I, TAVASSOLY O, RAD M S R, et al. Multifractal analysis of chaos game representation images of mitochondrial DNA[C]//Proc of Frontiers in the Convergence of Bioscience and Information Technologies. Washington DC: IEEE Computer Society, 2007: 224-229.
- [25] YU Zu-guo, SHI Long, XIAO Qian-jun, et al. Chaos game representation of genomes and their simulation by recurrent iterated function systems[C]//Proc of the 2nd International Conference on Bioinformatics and Biomedical Engineering. 2008: 41-46.
- [26] 赵静静, 齐斌, 王寒冰, 等. 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.
- [27] RANDI Ć M, ZUPAN J, BALABAN A T. Unique graphical representation of protein sequences based on nucleotide triplet codons[J]. Chemical Physics Letters, 2004, 397(1-3): 247-252.
- [28] RANDI Ć M. Another look at the chaos-game representation of DNA[J]. Chemical Physics Letters, 2008, 456(1-3): 84-88.
- [29] DESCHAVANNE P J, GIRON A, VILAIN J, et al. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences[J]. Molecular Biology and Evolution, 1999, 16(10): 1391-1399.
- [30] ALMEIDA J S, CARRICO J A, MARETZKE A, et al. Analysis of genomic sequences by chaos game representation[J]. Bioinformatics, 2001, 17(5): 429-437.
- [31] YU Zu-guo, ANH V, LAU K S. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses[J]. Journal of Theoretical Biology, 2004, 226(3): 341-348.
- [32] WANG Ying-wei, HILL K, SINGH S, et al. The spectrum of genomic signatures: from dinucleotides to chaos game representation[J]. Gene, 2005, 346: 173-185.