文章编号:1000-5013(2013)03-0308-05 **DOI**:10.11830/ISSN.1000-5013.2013.03.0308

DNA 序列碱基组合的频率矩阵及其应用

李玉双[1], 刘倩[1], 张昱[2]

(1. 燕山大学 理学院, 河北 秦皇岛 066004;
2. 石家庄邮电职业技术学院 计算机系, 河北 石家庄 050021)

摘要: 基于碱基组合在 DNA 序列中出现的频率,构造 11 个物种的 β-globin 基因第一个外显子的编码序列的频率矩阵. 借助矩阵 2-范数对 11 个物种进行相似性比较,并结合柱状图对物种之间的相似性进行分析. 研究结果表明:所构造的 DNA 序列频率矩阵不仅能够反映出 DNA 序列中碱基及碱基组合的含量分布,而且能够显示出序列碱基突变的情况.

关键词: DNA; 碱基组合; 频率矩阵; 相似性; 编码序列
中图分类号: Q 332 **文献标志码**: A

随着生物科学技术的迅猛发展,生物信息学越来越受到人们的重视,各种研究方法相继产生[1]. 近年来,数学模型被引入到该领域,对生物信息学本身而言,这是一次从量变到质变的飞跃. 众所周知,数学模型在生物序列和结构的比较中起到了很好的研究效果,在理论方面给出了很好的解释,如几何表示模型[2]、字统计模型[3]和马尔科夫模型[4]等. 隐马尔科夫模型在生物信息学的一系列问题都得到成功应用,如多序列比对[5]、基因识别[6]和蛋白质二级结构预测[7]等. 伴随生物研究中数学模型和算法的不断完善,产生了许多强有力的生物信息分析工具,如进化分析、聚类分析等,部分有效的分析工具极大地依赖于生物序列和结构的比较. 序列和结构的比较是最重要和最常用的原始操作,是许多其他复杂操作的基础. 序列的相似性分析是生物序列和结构比较中的一个重要问题. 从序列分析角度,判定两条序列同源与否的一个主要依据是探寻它们之间的相似性. 文献[8]提出了转移矩阵,将 DNA 序列看成是离散的马尔科夫链,分别以碱基 A,T,C 和 G 在序列中出现的次数作为基准来构造转移矩阵,进而刻画 11 个物种的 β-globin 基因第一个外显子编码序列的差别. 本文以序列的长度作为基准,基于碱基组合在 DNA 序列中出现的频率,构造了 DNA 序列的频率矩阵.

1 碱基组合的频率矩阵

1.1 频率矩阵的定义

给定长为 n 的生物序列 $l=l_1l_2l_3\cdots l_n$, $l_i\in S$, $S=\{\mathrm{A},\mathrm{T},\mathrm{C},\mathrm{G}\}$ 为碱基集合. 记 AA 在序列中出现的次数为 n_{AA},则定义 $P_{\mathrm{AA}}=n_{\mathrm{AA}}/n$. 同理,可分别定义 P_{AT}, P_{AC}, P_{AG}, P_{TA}, P_{TT}, P_{TC}, P_{TG}, P_{CA}, P_{CT}, P_{CC}, P_{CG}, P_{GA}, P_{GT}, P_{GC}, P_{GG}. 这里称 AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC, GG 为碱基组合. 定义该序列对应的频率矩阵 $\boldsymbol{P}$ 为

$$\boldsymbol{P}=\begin{bmatrix} P_{\mathrm{AA}} & P_{\mathrm{AT}} & P_{\mathrm{AC}} & P_{\mathrm{AG}} \\ P_{\mathrm{TA}} & P_{\mathrm{TT}} & P_{\mathrm{TC}} & P_{\mathrm{TG}} \\ P_{\mathrm{CA}} & P_{\mathrm{CT}} & P_{\mathrm{CC}} & P_{\mathrm{CG}} \\ P_{\mathrm{GA}} & P_{\mathrm{GT}} & P_{\mathrm{GC}} & P_{\mathrm{GG}} \end{bmatrix}.$$

由此可知,对应于文献[8]刻画的 11 个物种的 β-globin 基因第一个外显子编码序列,可以分别

收稿日期: 2012-10-17
通信作者: 李玉双(1980-),女,副教授,主要从事生物数学的研究. E-mail:liyushuang@yeah.net.
基金项目: 国家自然科学基金资助项目(11201409)

定义相应的频率矩阵，其碱基如表 1 所示．表 1 中：1～11 个物种分别是人类(human)，家山羊(goat)，负鼠目(opossum)，原鸡(gallus)，狐猴(lemur)，小鼠(mouse)，兔子(rabbit)，老鼠(rat)，大猩猩(gorilla)，牛科动物(bovine)，黑猩猩(chimpanzee)．

表 1　11 个物种的频率矩阵

Tab. 1　Frequency matrix of eleven species

碱基	human				goat			
	A	T	C	G	A	T	C	G
A	0.043 5	0.021 7	0.043 5	0.076 1	0.058 1	0.023 3	0.023 3	0.093 0
T	0.010 9	0.021 7	0.021 7	0.163 0	0	0.023 3	0.023 3	0.151 2
C	0.032 6	0.076 1	0.076 1	0.021 7	0.034 9	0.093 0	0.046 5	0.023 3
G	0.087 0	0.097 8	0.065 2	0.130 4	0.093 0	0.058 1	0.104 7	0.139 5

碱基	opossum				gallus			
	A	T	C	G	A	T	C	G
A	0.032 6	0.032 6	0.076 1	0.087 0	0.054 3	0.043 5	0.032 6	0.076 1
T	0.021 7	0.043 5	0.043 5	0.130 4	0	0	0.043 5	0.119 6
C	0.076 1	0.097 8	0.043 5	0	0.076 1	0.076 1	0.076 1	0.032 6
G	0.087 0	0.065 2	0.054 3	0.097 8	0.065 2	0.043 5	0.108 7	0.141 3

碱基	lemur				mouse			
	A	T	C	G	A	T	C	G
A	0.043 5	0.043 5	0.021 7	0.097 8	0.053 2	0.031 9	0.031 9	0.063 8
T	0.010 9	0.043 5	0.043 5	0.152 2	0	0.031 9	0.031 9	0.180 9
C	0.043 5	0.087 0	0.021 7	0.010 9	0.031 9	0.095 7	0.074 5	0.010 6
G	0.097 8	0.076 1	0.076 1	0.119 6	0.085 1	0.085 1	0.074 5	0.106 4

碱基	rabbit				rat			
	A	T	C	G	A	T	C	G
A	0.055 6	0.033 3	0.011 1	0.088 9	0.065 2	0.043 5	0.043 5	0.065 2
T	0	0.011 1	0.044 4	0.166 7	0.043 5	0.021 7	0	0.163 0
C	0.044 4	0.055 6	0.055 6	0.011 1	0.021 7	0.097 8	0.065 2	0.010 9
G	0.077 8	0.122 2	0.066 7	0.144 4	0.076 1	0.065 2	0.087 0	0.119 6

碱基	gorilla				bovine			
	A	T	C	G	A	T	C	G
A	0.043 0	0.021 5	0.043 0	0.075 3	0.058 1	0.023 3	0.023 3	0.093 0
T	0.010 8	0.021 5	0.021 5	0.161 30	0.046 5	0.011 6	0.151 2	
C	0.032 3	0.075 3	0.075 3	0.021 5	0.034 9	0.069 8	0.058 1	0.023 3
G	0.086 0	0.096 8	0.064 5	0.139 8	0.093 0	0.069 8	0.093 0	0.139 5

碱基	chimpanzee			
	A	T	C	G
A	0.047 6	0.028 6	0.038 1	0.076 2
T	0.019 0	0.028 6	0.028 6	0.152 4
C	0.038 1	0.066 7	0.066 7	0.019 0
G	0.076 2	0.104 8	0.057 1	0.142 9

从表 1 可以看到：11 个物种中 TG 出现的频率都是最高，其次是 GG，而 TA 和 CG 频率较低．这说明在 β-globin 基因的编码序列中 TG 和 GG 相对来说出现频繁，而 TA 和 CG 相对出现次数较少，有些物种甚至没有出现．从单个物种来说，opossum 和 gallus 又有些特殊的地方，例如 TG 中频率较其他物种偏低，CA 中频率较高．这说明了在 11 个物种的 β-globin 基因的编码序列中 opossum 和 gallus 有着特殊性．上述结果与代琦等[8]的结论基本一致．

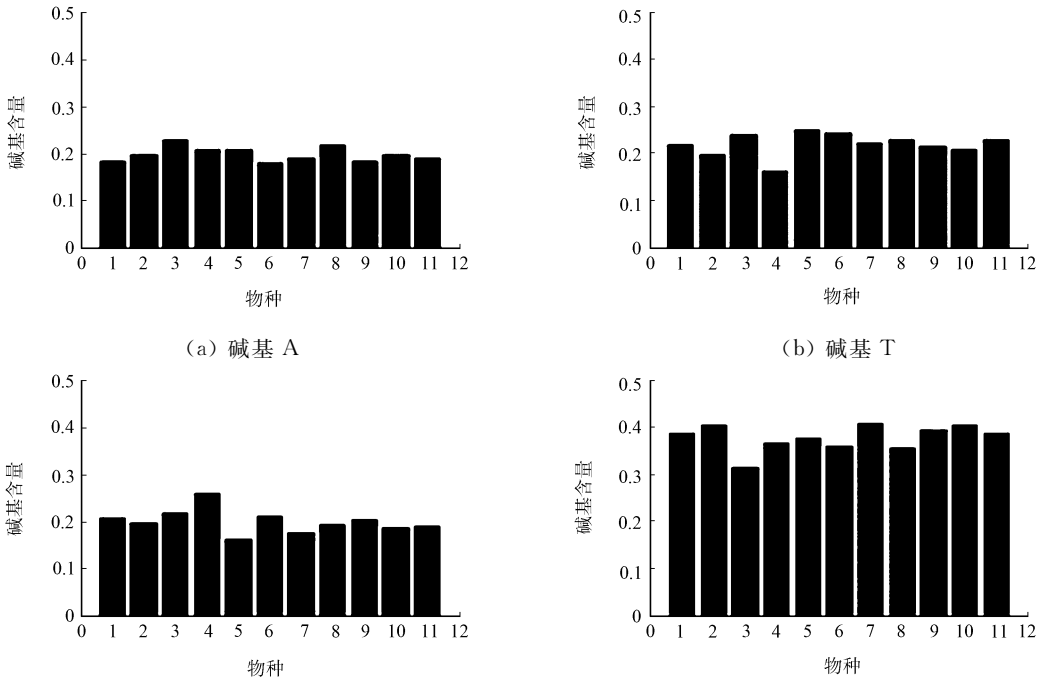
1.2　频率矩阵的性质

频率矩阵有如下 3 点性质：1) 各行元素之和能够反映出各个碱基在序列中的含量分布；2) 各列元素之和能够反映出碱基突变的分布情况；3) 所有元素之和为 $\frac{n-1}{n}$．

根据频率矩阵的性质1),可以计算出11个物种碱基含量的向量,即

human = (0.184 8,0.217 3,0.206 5,0.380 4),goat = (0.197 7,0.197 8,0.197 7,0.395 3),
opossum = (0.228 3,0.239 1,0.217 4,0.304 3),gallus = (0.206 5,0.163 1,0.260 9,0.358 7),
lemur = (0.206 5,0.250 1,0.163 1,0.369 6), mouse = (0.180 8,0.244 7,0.212 7,0.351 1),
rabbit = (0.188 9,0.222 2,0.166 7,0.411 1),rat = (0.217 4,0.228 2,0.195 6,0.347 9),
gorilla = (0.182 8,0.215 1,0.204 4,0.387 1), bovine = (0.197 7,0.209 3,0.186 1,0.395 3),
chimpanzee = (0.190 5,0.228 6,0.190 5,0.381 0).

对于序列的最后一个碱基,虽然它的含量不能通过上述向量中的对应值精确体现(由于计算的是碱基组合),但由于其他3个碱基的含量恰好就是向量中的对应值,所以能够很容易得到最后一个碱基的含量.如在human中,碱基A的含量是0.184 8,碱基T的含量是0.217 3,碱基C的含量是0.206 5,则碱基G的含量是0.391 4.图1为11个物种的碱基含量分布柱状图,可以更直观的展现碱基A,T,C,G在11个物种中的分布情况.



(a) 碱基A (b) 碱基T

(c) 碱基C (d) 碱基G

图1 碱基在11个物种中的分布图

Fig. 1 Distribution of nucleotide of eleven species

观察11个碱基含量向量及图1可以看出:11个物种序列中碱基G的含量都较高,碱基A的含量分布较为均匀;相比其他物种,gallus碱基G的含量明显偏低,lemur碱基C的含量偏低,opossum碱基G的含量偏低;human和gorilla的碱基含量几乎相等.众所周知,研究DNA序列的特殊区域能为基因组的组织结构和生物作用提供更加丰富的信息.这里借助碱基含量向量及图1可以很容易的得出特殊碱基组合的含量,如GC含量. GC含量为基因组提供了数量以及性质上的重要信息,GC含量高的DNA序列要比GC含量低的DNA序列更加稳定[9].

根据频率矩阵的性质2),可以计算出11个物种的碱基转移向量,即

human = (0.174 0,0.217 3,0.206 5,0.391 2), goat = (0.186 0,0.197 7,0.197 8,0.407 0),
opossum = (0.217 4,0.239 1,0.217 4,0.315 2), gallus = (0.195 6,0.163 1,0.260 9,0.369 6),
lemur = (0.195 7,0.250 1,0.163 0,0.380 5), mouse = (0.170 2,0.244 6,0.212 8,0.361 7),
rabbit = (0.177 8,0.222 2,0.177 8,0.411 1), rat = (0.206 5,0.228 2,0.195 7,0.358 0),
gorilla = (0.172 1,0.215 1,0.204 3,0.397 9), bovine = (0.186 0,0.209 4,0.186 0,0.407 0),
chimpanzee = (0.180 9,0.228 7,0.190 5,0.390 5).

通过比较碱基含量向量和碱基转移向量不难发现，每个物种的两个向量总有两个分量是相等的. 因为前者忽略了序列的最后一个碱基，后者忽略了序列的第一个碱基. 如果一个序列首尾碱基相同，则这个序列对应的两个向量一定相等. 从这个意义上来说，碱基转移向量也能够反映出各个碱基在序列中的含量分布. 此外，除首尾碱基相同的序列(注：这 11 个物种首尾碱基都不同)，不用计算通过比较两个向量就能确定每个物种中各个碱基的含量，如 human 的碱基转移向量的最后一个分量即为碱基 G 的含量 0.391 2，这与前面计算的结果一致(微小误差是由于计算时舍位引起的).

2　序列相似性分析

由于生物序列有其进化上的生物学意义，因此比较两条生物的相似性时，不能完全使用计算机科学中的模式匹配，常会借助“距离”来反映，如向量的欧氏距离、协方差距离、夹角距离等. 文中引入矩阵的 2-范数对 11 个物种进行相似性比较 .

设 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 为两个物种的频率矩阵，令 $\boldsymbol{Q}=|\boldsymbol{P}_1-\boldsymbol{P}_2|$，则 $\boldsymbol{Q}$ 的 2-范数计算公式为

$$\|\boldsymbol{Q}\|_2=\sqrt{\rho(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q})}.$$

利用 2-范数的计算公式来求两个物种的相似性大小，即求得的范数越小，代表两个物种所刻画的 DNA 序列越相似，两个物种越接近；反之，它们刻画的 DNA 序列差别越大. 利用 2-范数的计算公式和常用的欧式距离公式计算得到的 11 个物种的相似性矩阵，如表 2，3 所示.

表 2　由 2-范数算得的 11 个物种的相似性矩阵

Tab. 2　Similarity matrix of eleven species based on the 2-norm

物种	human	goat	opossum	gallus	lemur	mouse	rabbit	rat	gorilla	bovine	chimpanzee
human	0	0.069 5	0.089 8	0.090 7	0.075 6	0.047 0	0.060 4	0.063 1	0.009 7	0.057 6	0.029 7
goat		0	0.096 1	0.071 0	0.058 2	0.072 0	0.089 9	0.070 8	0.068 4	0.040 0	0.079 3
opossum			0	0.105 2	0.074 7	0.088 6	0.105 6	0.101 3	0.091 9	0.098 3	0.087 2
gallus				0	0.097 9	0.101 2	0.100 6	0.104 8	0.087 8	0.081 6	0.094 0
lemur					0	0.064 7	0.078 1	0.087 3	0.076 5	0.059 7	0.073 3
mouse						0	0.074 1	0.066 0	0.0517	0.070 4	0.061 6
rabbit							0	0.094 3	0.059 4	0.080 7	0.049 6
rat								0	0.064 8	0.072 5	0.071 9
gorilla									0	0.055 4	0.026 7
bovine										0	0.064 4
chimpanzee											0

表 3　由欧氏距离算得的 11 个物种的相似性矩阵

Tab. 3　Similarity matrix of eleven species based on the Euclidian distance

物种	human	goat	opossum	gallus	lemur	mouse	rabbit	rat	gorilla	bovine	chimpanzee
human	0	0.078 4	0.098 4	0.105 4	0.081 4	0.049 7	0.065 0	0.072 1	0.009 8	0.062 7	0.031 1
goat		0	0.109 2	0.081 8	0.063 4	0.076 3	0.092 9	0.077 2	0.074 0	0.040 3	0.083 2
opossum			0	0.116 1	0.082 9	0.100 6	0.127 7	0.108 8	0.101 3	0.109 6	0.098 9
gallus				0	0.109 8	0.112 0	0.116 2	0.109 8	0.103 9	0.094 0	0.105 5
lemur					0	0.078 2	0.084 8	0.093 5	0.082 8	0.066 2	0.078 8
mouse						0	0.083 2	0.068 9	0.055 4	0.073 3	0.067 4
rabbit							0	0.111 6	0.063 6	0.083 6	0.053 3
rat								0	0.073 9	0.079 5	0.081 1
gorilla									0	0.061 6	0.027 5
bovine										0	0.067 3
chimpanzee											0

比较表 2，3 可知：2-范数法要比常用的欧氏距离法好，但从整体上看两个方法求得的结果基本一致. 即 human 和 gorilla 相似性非常高，human 和 chimpanzee，gorilla 和 chimpanzee 相似性也很高，goat 和 bovine 相似性较高；相比之下，opossum 和其他物种相似性较低，这与 opossum 是与其他哺乳动物亲

缘较远的哺乳动物相符合;Gallus 和其他物种相似性也较低,这与 Gallus 是唯一的非哺乳动物相符合. 这些结论都与相关的文献结果一致[2,8].

3 结论

介绍一种利用 DNA 序列碱基组合的频率矩阵来刻画物种相似性的方法. 该矩阵的每一个分量都能够反映出对应碱基组合在序列中的含量分布情况,其行和能反映每个碱基在序列中的含量分布情况,列和能反映碱基突变的情况,而所有元素值之和为定值. 相较文献[8]中的转移矩阵,频率矩阵能够更好地从整体上反映出 DNA 序列中碱基以及碱基组合的含量分布,显示出序列碱基突变的情况.

文中引入矩阵的 2-范数对 11 个物种进行相似性比较,结果显示该方法要优于上述常用的距离分析方法. 频率矩阵的应用在物种的相似性比较方面得到了很好的体现,借助矩阵 2-范数和柱状图所得到的结果对物种的进化分析有一定的参考价值.

参考文献:

[1] 王勇献,王正华. 生物信息学导论:面向高性能计算的算法与应用[M]. 北京:清华大学出版社,2011:28-72.

[2] XIE Guo-sen, MO Zhong-xi. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications[J]. J Theor Biol, 2011, 269(1): 123-130.

[3] VINGA S, GOUVEIA-OLIVEIRA R, ALMEIDA J S. Comparative evaluation of word composition distances for the recognition of SCOP relationships[J]. Bioinformatics, 2004, 20(2): 206-215.

[4] PHAM T D, ZUEGG J. A probabilistic measure for alignment-free sequence comparison[J]. Bioinformatics, 2004, 20(18): 3455-3461.

[5] 罗泽举,宋丽红. 隐马尔可夫模型的多序列比对的研究[J]. 计算机工程与应用,2010,46(7):171-174.

[6] 丰月姣,贺兴时. 二阶隐马尔科夫模型在基因识别中的应用[J]. 佳木斯大学学报,2009,27(6):940-942.

[7] 石峰,莫忠息,张楚瑜. 隐马尔可夫模型-改进的预测蛋白质二级结构方法[J]. 生物数学学报,2004,19(2):233-237.

[8] 代琦. 生物序列、结构比较中若干数学模型研究及应用[D]. 大连:大连理工大学,2009:17-71.

[9] GAO F, ZHANG C T. GC-Profile: A web-based tool for visualizing and analyzing the variation of GC content in genomic sequences[J]. Nucleic Acids Res, 2006, 34: 686-691.

Frequency Matrix of Nucleotide Combination of DNA Sequences and Its Application

LI Yu-shuang[1], LIU Qian[1], ZHANG Yu[2]

(1. School of Science, Yanshan University, Qinhuangdao 066004, China;
2. Department of Computer, Shijiazhuang Post and Telecommunications Technical College, Shijiazhuang 050021, China)

Abstract: The frequency matrix of coding sequence of the first exon of β-globin gene of eleven species was proposed based on the frequencies of nucleotide combinations in the DNA sequences. The similarity of eleven species was compared with the aid of 2-norm of matrix. Moreover, the similarity analysis was spread among species with column graphs. The results showed that the frequency matrix of DNA sequences not only could reflect the content distribution of nucleotides and nucleotide combinations in DNA sequences, but also could display the mutations of sequences nucleotides.

Keywords: DNA; nucleotide combination; frequency matrix; similarity; coding sequence

(**责任编辑**: 陈志贤 **英文审校**: 刘源岗)