doi:10.3969/j.issn.1672-5565.2013.02.12

DNA 序列基于 k-字的数值刻画及其应用

刘俊宏,李 春*
(渤海大学数理学院,辽宁 锦州 121013)

摘 要:借助DNA序列中k-字的频数,将序列转化成一个340维向量,进而计算物种间的进化距离。作为应用:分别以15个物种的β球蛋白基因、13种汉坦病毒的S片段以及26个闭壳龟线粒体基因为例,构建系统发生树,所得结果与前人的结论一致,说明了该方法的有效性。

关键词:k-字;340维向量;系统发生树
中图分类号:Q523 **文献标识码**:A **文章编号**:1672-5565(2013)-02-142-04

A numerical characterization of DNA sequences based on k-words and its application

LIU Jun-hong, LI Chun*
(*Department of Mathematics, Bohai University, Liaoning Jinzhou* 121013, *China*)

Abstract: By means of the frequencies of k-words, the DNA sequence is transformed into a 340-D vector, and then the evolutional distance is obtained. The proposed measure is used to construct phylogenetic trees on three separate sets: the full β-globin genes of 15 species, the S segments of 13 hantaviruses and 26 Cuora mitochodriona genes. The results are consistent with those of previous analyses, which illustrates the utility of the approach.

Key words: k-words; 340-D Vector; Phylogenetic Trees

序列比较是生物信息学中最基本的操作,通过序列比较可以找到生物序列中的功能、结构和进化的信息。现在最重要的问题之一就是如何计算两条生物序列间的相似性距离。早期的工作中,序列比对和图形表示应用比较广泛[1-8]。序列比对依赖于打分函数,通过字符间的插入、删除和替换等编辑操作来定义序列之间的进化距离,但是打分函数中的空位罚分理论缺乏有效的依据,这促使大家寻找其他更为有效的方法来比较序列;图形表示则使用可视化的方法研究生物序列,将图形转化成矩阵的形式,提取不变量作为序列的描述子进行序列比较。本文利用DNA序列的k-字频数构造340维向量,计算序列间的相对距离,构建物种的系统发生树,并且结合实际数据来检验方法的有效性。

1 方法

1.1 k-字

DNA序列是由四种字符A、C、G、T构成的集合。我们将长度为k的字符串称为k-字,k-字的频数就是顺次出现在DNA序列中的字符串个数,k-字的频数包含着DNA序列的信息[9-11],本文中,k-字共取以下4种:

(1)1-字(4种):A、C、G、T;

(2)2-字(16种):AA、AC、AG、AT、…、TT

(3)3-字(64种):AAA、AAC、AAG、AAT、…、TTT

(4)4-字(256种):AAAA、AAAC、AAAG、AAAT、…、TTTT

收稿日期:2012-10-20;**修回日期**:2012-11-22.
基金项目:国家自然科学基金项目(11171042);辽宁省高等学校杰出青年学者成长计划(LJQ2011122)。
作者简介:刘俊宏,女,辽宁省辽阳市人,硕士生,研究方向:生物信息学。E-mail:w1988522@126.com.
*通讯作者:李春,男,教授,博士,主要研究方向为组合数学与生物信息学。E-mail:lchlmb@163.com.

总和为 340 个 k-字(k = 1、2、3、4)。

1.2　构造向量

给定一条长度为 n 的序列 $S = 1_1 a_2 a_3 \Lambda a_n$,统计 340 个 k-字出现的频数($k = 1$、2、3、4)。这里用 f_{ij} 表示长度为 i 的序列 $S = a_1 a_2 a_3 \cdots a_i$ 中第 j 个 k-字出现的频数,其中 $i = 1,2,\cdots,n$,$j = 1,2,\cdots,340$,为了消除长度的影响,将频数规范化得到 $V_{i,j}$:

$$v_{i,j} = \frac{f_{i,j}}{i}, i = 1,2,\cdots,n, j = 1,2,\cdots,340$$

任何一条生物序列能用下述的 340 维向量表示

$$v = (\sum_{i=1}^{n} v_{i,1}, \sum_{i=1}^{n} v_{i,2}, \cdots, \sum_{i=1}^{n} v_{i,340}) \quad (1)$$

以下面的序列片段为例

ATGCTGACTGCTGAGGAGAAGGCTGCCGT-CACCGCCTTTTGGGGCAAGGTCAAC,它对应的 340 维向量 v = (0.2,0.25,0.33,0.2,0.9,0.64,1.2,…,5.02)。

生物序列一旦转化成向量,序列之间的比较就转化成向量之间的比较。物种间的距离越小,这两条序列越相似。物种间距离的构造可以用两向量之间的欧氏距离和两向量之间的夹角余弦,两个物种的相对距离用下式来衡量:

$$dis(A,B) = [1 - d(A,B)] \times \cos(A,B) \quad (2)$$

2　应用

为了检测上面方法的有效性,我们利用公式(2)计算包括人类在内的 15 个物种 β 球蛋白基因(表 1)的距离。然后使用软件 MEGA4.0 中的非加权分组平均法(UPGMA)进行聚类,见图 1。

表 1　15 个物种的 β 球蛋白基因

Table 1　The β – globin genes of 15 species

No	Species	(AC) GenBank	Location	Length
1	Human	U01317	62 187 – 63 610	1 424
2	Chimpanzee	X02345	4 189 – 5 532	1 344
3	Gorilla	X61109	4 538 – 5 881	1 344
4	Lemur	M15734	154 – 1 595	1 442
5	Rat	X06701	310 – 1 505	1 196
6	Mouse	V00722	275 – 1 462	1 188
7	Goat	M15387	279 – 1 749	1 471
8	Sheep	DQ352470	238 – 1 708	1 471
9	Mouflon	DQ352468	238 – 1 706	1 469
10	Bovine	X00376	278 – 1 741	1 464
11	Rabbit	V00882	277 – 1 419	1 143
12	European hare	Y00347	1 485 – 2 620	1 136
13	Opossum	J03643	467 – 2 488	2 022
14	Gallus	V00409	465 – 1 810	1 346
15	Muscovy – duck	X15739	291 – 1 870	1 580

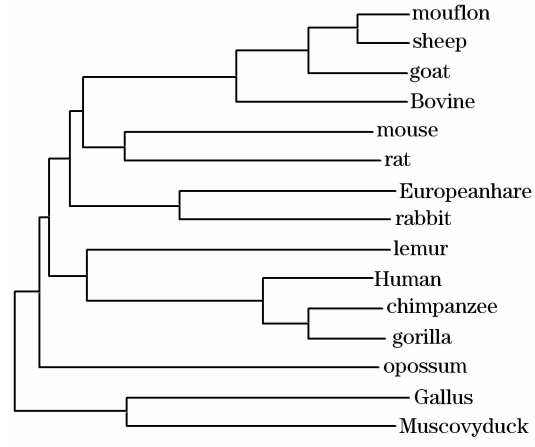

图 1　15 个物种 β 球蛋白基因的系统发生树

Fig. 1　The phylogenetic tree of the β – globin genes of 15 species

在图 1 中,可以看到 13 个哺乳动物被聚到一起,另两种非哺乳动物(Gallus 和 Muscovy duck)形成一个独立的分支。在哺乳动物的子树中发现 Sheep 和 mouflon 首先聚在一起,再与 goat 聚成一类,同时这三个物种与 Bovine 也很相似。同样地,Chimpanzee、Gorilla 与 Human 形成一个分支,同时它们与 Lemur 也有很大的相似度,因为它们都是灵长目动物,这和我们期望的结果一致。除此之外,Rabbit 和 European hare,rat 和 mouse 的相似度都明显高于同其他物种的相似度,这个结果与文献[7 – 8]、[12 – 17]的结论基本一致。

此外,还选择 13 种汉坦病毒的 S 片段构建系统发生树。汉坦病毒又名肾病综合征出血热病毒,是流行性出血热的病原体。它属于布尼亚病毒科汉坦病毒属,是一种有包膜分节段的负链 RNA 病毒。汉坦病毒感染人体后引发两种疾病:汉坦病毒肺综合征(HPS)和汉坦病毒肾综合征出血热(HFRS)。前者主要流行于美国,在阿根廷、巴西、巴拉圭等地也发现病例;中国是后者流行的主要地区,发病率高达 90% 以上。HFRS 的病原主要有四种:HTN、SEO、DOB 和 PUU,中国以 HTN 和 SEO 为主要流行株,我们选择长度为 1.6 ~ 2.0kb 的汉坦病毒的 S 片段构建系统发生树,并选择 Sotkamo 作为外群(表 2)。

利用同样的方法构建汉坦病毒 S 片段的系统发生树(图 2),能够发现,子树中 Z5、Z10、Z251 和 ZLS6-11、ZLS-12 聚合在同一分支上;ZT71 和 ZT10 先聚合在一起,再与 Z37 聚合,这三者的进化关系相对较近;GOU3 和 ZJ5 独立聚合成一个分支,然后与 ZT71、ZT10、Z37 聚合到同一分支。这与以前的研究结果是一致的[18]。

表 2 13 种汉坦病毒的 *S* 片段

Table 2 The *S* segments of 13 hantavirus

No	Strain	Type (AC)	Genbank	Location	Length
1	Z10	HTN	AF184987	Shengzhou	1 701
2	Z5	HTN	EF103195	Shengzhou	1 700
3	Z251	HTN	EF595840	Longquan	1 700
4	ZLS6 – 11	HTN	FJ753396	Lishui	1 700
5	ZLS – 12	HTN	FJ753398	Lishui	1 700
6	GOU3	SEO	AF184988	Jiande	1 759
7	ZJ5	SEO	FJ753400	Jiande	1 759
8	K24 – v2	SEO	AF288655	Xinchang	1 772
9	K24 – e7	SEO	AF288653	Xinchang	1 772
10	Z37	SEO	AF187082	Wenzhou	1 754
11	ZT71	SEO	AY750171	Tiantai	1 754
12	ZT10	SEO	AY766368	Tiantai	1 753
13	Sotkamo	PUU	X61035	Finland	1 830

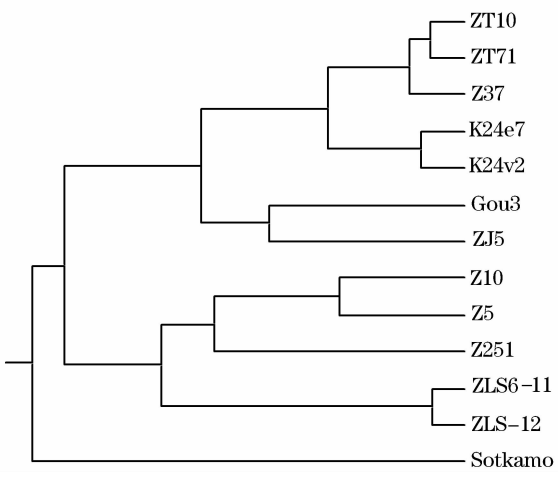

图 2 汉坦病毒 S 片段的系统发生树

Fig. 2 The phylogenetic tree of Hantavirus based on S segment

此外,利用上述的方法构建闭壳龟物种序列(表 3)的系统发生树(图 3)。闭壳龟是龟鳖目龟科一属的通称,又名呷蛇龟、壳蛇龟和亚洲箱龟,分布于亚洲的缅甸、泰国、越南、中国等东南亚地区。现已知中国有 9 种:黄缘闭壳龟、黄额闭壳龟、三线闭壳龟、云南闭壳龟、潘氏闭壳龟、马来闭壳龟、百色闭壳龟、周氏闭壳龟和金头闭壳龟,其中,云南闭壳龟是中国特有的闭壳龟种,仅产于云南昆明以及东川,非常稀少。黄缘闭壳龟分布最为广泛,栖息地位于海拔两千米以上的高地。从图 3 中发现,云南闭壳龟聚成单独的一类,周氏闭壳龟和潘氏、金头、三线闭壳龟聚为一类,它们与云南闭壳龟、黄缘壳龟形成三分叉关系。本文也不支持云南闭壳龟和黄缘闭壳龟相近的进化关系,这与之前的研究结果相同[19-20]。

表 3 闭壳龟线粒体 DNA

Table 3 Cuora mitochodriona DNA

No	Species	(AC) Genbank	Length
1	Aurocapitata A1	AY364606	892
2	Aurocapitata B1	AY572867	892
3	Chinemysreevesii A	AF348288	842
4	Chinemysnigricans A	AF348289	842
5	Mouhotii A1	AF348285	841
6	Galbinifrons bourreti B1	AY364618	892
7	Galbinifrons bourreti B2	AY364624	892
8	Galbinifrons galbinifrons B1	AY364615	892
9	Galbinifrons galbinifrons B2	AY364612	892
10	Galbinifrons picturata B1	AY364628	892
11	Galbinifrons picturata B2	AY364630	892
12	Mauremysmutica A	AF348279	892
13	Mccordi A	AY364608	892
14	Mouhotii A2	AF348286	842
15	Pani A	AY364607	892
16	Pani B	AY590461	892
17	Trifasciata A1	AF348297	842
18	Trifasciata A2	AF348296	841
19	Yunnanensis1	EF685042	893
20	Yunnanensis2	EF685041	905
21	Yunnanensis3	EF685043	890
22	Yunnanensis B	AY572868	892
23	Zhoui	EF685044	896
24	Zhoui B1	AY590462	887
25	Zhoui B2	AY572865	892
26	Zhoui B3	AY572866	892

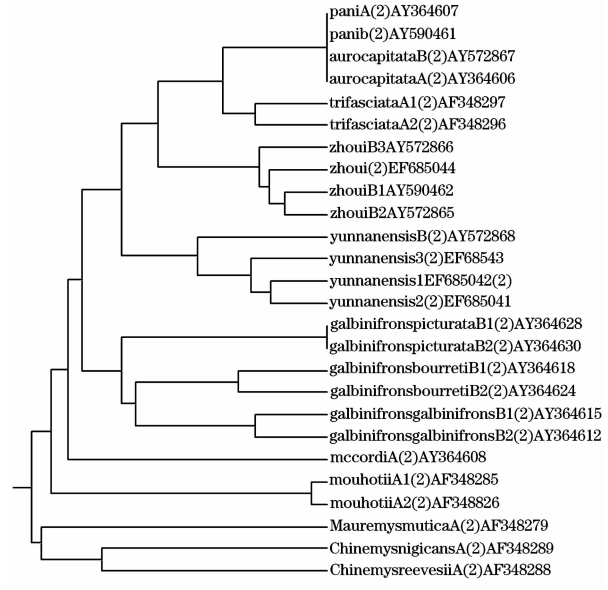

图 3 闭壳龟线粒体 DNA 的系统发生树

Fig. 3 The phylogenetic tree of Cuora mitochodriona DNA

3 结 语

k-字是指长度为 k 的字符串。本文通过考虑 DNA 序列中 4 个 1-字,16 个 2-字,64 个 3-字以及 256 个 4-字,构造了 DNA 序列的一种 340 维向量表示,并在此基础上给出了两条序列间的相对距离。

与大多数现有方法相比,我们的方法既不需要多重序列比对,也不涉及图形表示和复杂的高阶矩阵的计算。作为应用,我们分别构建了 15 个物种的 β 球蛋白基因,13 个汉坦病毒的 S 片段以及 26 个闭壳龟线粒体基因三组数据的系统发生树,所得的结果与已有文献一致。

参考文献(References)

[1] Zhang R, Zhang C T. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences[J]. J Biomol Struct Dyn., 1994, 11: 767-782.

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

[3] Randi M, Vracko M, Lers N, Plavsi D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation[J]. Chemical Physics Letters, 2003, 371:202-207.

[4] Leong P M, Morgenthaler S. Random walk and gap plots of DNA sequences[J]. Comput. Applic. Biosci., 1995, 12: 503-511.

[5] Li C, Tang N N, Wang J. Directed graphs of DNA sequences and their numerical characterization[J]. Journal of Theoretical Biology, 2006, 241: 173-177.

[6] 骆嘉伟,张惜珍. 一种新的基于 3D 图形的进化树构造方法[J]. 武汉理工大学学报, 2007, 29(4): 24-27.

[7] Randi M, Vracko M, Nandy A, Basak S C. On 3-D graphical representation of DNA primary sequence and their numerical characterization[J]. J. Chem. Inf. Comput. Sci., 2000, 40: 1235-1244.

[8] Yao Y H, Dai Q, Nan X Y. Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation [J]. J. Comput. Chem., 2008, 29(10): 1632-1639.

[9] 贾晓超,李培芳,罗辽复. 基因组中‘k 字’频数的分布[J]. 内蒙古大学学报, 2005, 36(3): 301-305.

[10] Zhang Y, Wang X H and Kang L. A K-mer scheme to predict piRNAs and characterize locust piRNAs[J]. Bioinformatics, 2011, 27(6): 771-776.

[11] Yu X Q, Zheng X Q, Meng L Y, Li C, Wang J. A Support Vector Machine Based Method to Predict Success for Polymerase Chain Reactions[J]. Combinatorial Chemistry & High Throughput Screening, 2012, 15(6): 486-491.

[12] Randi M, Zupan J, Novic M, Gute B D, Basak S C. Novel matrix invariants for characterization of changes of proteomics maps [J]. SAR QSAR Environ Res., 2002, 13(7-8): 689-703.

[13] Randi M, Guo X F and Basak S C. On the Characterization of DNA Primary Sequences by Triplet of Nucleic Acid Bases[J]. J. Chem. Inf. Comptu. Sci. 2001, 41: 619-626.

[14] He P-an, Wang J. Characteristic sequences for DNA primary sequence[J]. J. Chem. Inf. Comput. Sci., 2002, 42: 1080-1085.

[15] Li C, Wang J. Similarity analysis of DNA sequences based on the generalized LZ complexity of (0,1)-sequences[J]. J. Math. Chem., 2008, 43: 26-31.

[16] Liu N, Wang T. A weighted measure for the similarity analysis of DNA sequences[J]. J. Mol. Model, 2006, 12(6): 897-903.

[17] Li C, Ma H, Zhou Y, Wang X L, Zheng X Q. Similarity analysis of DNA sequences based on the weighted pseudo – entropy. Journal of computational chemistry[J], 2011, 32(4): 675-680.

[18] Yao P P, Zhu H P, Deng X Z. Xu F, Xie R H, Yao C H, Weng J Q, Zhang Y, Yang Z Q, Zhu Z Y. Molecular evolution analysis of hantaviruses in Zhejiang Province[J]. Chinese journal of virology, 2010, 26(6): 465-470.

[19] 何静, 周婷, 饶定齐, 张亚平. 云南闭壳龟(Cuora yunnanensis)的分子鉴定及进化地位研究[J]. 科学通报, 2007, 52(17): 2085-2088.

[20] 万全, 郑将臣, 程起群, 赵金良. 基于 12SrRNA 序列研究龟鳖类的系统进化特征[J]. 海洋渔业, 2010, 32(3): 264-275.