

基因组寡聚核苷酸频率组分差异的进化信息^{*}

熊远妍, 王军鹏, 蓝一杰, 文 明, 张尚宏

(中山大学基因工程教育部重点实验室//生物工程研究中心, 广东 广州 510275)

摘 要: 利用不同基因组全序列中寡聚核苷酸频率组分差异信息构建系统树, 并与传统方法的建树结果比较, 分析以寡聚核苷酸频率组分差异信息构建系统树的可行性及适用范围。在获得 194 种原核生物基因组全序列寡聚核苷酸频率数据的基础上, 依据寡聚核苷酸组分的保守性构建权重矩阵, 对寡聚核苷酸组分差异加权, 并尝试用绝对值距离、Lance 距离、欧氏距离三种方法计算距离矩阵, 再用距离法构建系统树。结果显示, 该方法对科以下分类单位的建树与传统方法的建树基本一致, 而对科以上的分类单位较难区分。首次利用寡聚核苷酸频率保守性进行加权, 通过非比对的算法用基因组全序列建树。

关键词: 基因组全序列; 寡聚核苷酸; 频率; 系统树; 信息分析; 数据发掘; 生物信息学

中图分类号: Q754 **文献标识码:** A **文章编号:** 0529-6579 (2008) 02-0084-05

随着越来越多的全基因组序列测序的完成, 出现了基于全基因组的分子系统学方法, 这些方法大多是整体分析比较基因组中整套或大部分的基因^[1], 比较基因的获得、丢失、转座等, 如基因种类 (gene content)^[2]、基因排列顺序 (gene order)^[3]等方法。已证实基于全基因组的方法至少在部分生物中是适用的^[4]。但是, 这些方法实际上并没有全部利用基因组全序列中的信息。

应用寡聚核苷酸频率差异构建系统树的方法, 以前由于基因组测序完成不多, 加上利用已有的基因组组分信息的建树结果不突出而没有被采用^[5]。另一方面, 寡聚核苷酸组分差异在比较基因组学中已有广泛的应用^[6-8], 如 G+C 含量常被看作是基因的一个特异信息^[9]。统计分析发现, 部分寡聚核苷酸的出现频率在不同基因组中存在着保守性^[10]。本文基于短的寡聚核苷酸具有保守性及物种特异性的假设, 以寡聚核苷酸组分变化波动构建权重矩阵, 并根据寡聚核苷酸组分差异计算物种间的遗传距离和构建系统树。从结果看出, 频率差异在科以下分类单位的分子系统学研究中适用。这一研究将有助于进一步利用基因组全序列中寡聚核苷酸的频率特征构建进化树, 并解译其进化信息。

1 材料和方法

1.1 数据来源

计算权值的 194 种原核生物的基因组全序列

(至 2005 年 8 月所有已测定全序列的原核生物物种)^[11]和用于建树的 77 条基因组全序列^[12]均从 NCBI (<http://www.ncbi.nlm.nih.gov/Genomes/>) 下载。后者包括有代表性的 63 条真细菌序列和 14 条古细菌序列^[12]。其中, 真细菌序列中又有 23 条同科物种的序列 (括号内为有关序列在图 1 中的缩写): *Buchnera aphidicola* str. APS (bucapa buc)、*B. aphidicola* str. Bp (bucapb buc)、*B. aphidicola* str. Sg (bucaps buc)、*Escherichia coli* CFT073 (esc-coc esc)、*E. coli* K12 (esc-cok esc)、*E. coli* O157: H7 (esc-c0 esc)、*E. coli* O157: H7 EDL933 (esc-coh esc)、*E. coli* UTI89 (esc-cou esc)、*E. coli* W3110 (esc-cow esc)、*Salmonella enterica* subsp. *enterica* serovar *Choleraesuis* str. SC - B67 (salens sal)、*S. enterica* subsp. *enterica* serovar *Paratyphi* A str. ATCC 9150 (salena sal)、*S. enterica* subsp. *enterica* serovar *Typhi* Ty2 (salent sal)、*S. enterica* subsp. *enterica* serovar *Typhi* str. CT18 (salenc sal)、*S. typhimurium* LT2 (salenl sal)、*Shigella boydii* Sb227 (shibob shi)、*S. dysenteriae* Sd197 (shidyd shi)、*S. flexneri* 2a str. 2457T (shifl2 shi)、*S. flexneri* 2a str. 301 (shifl3 shi)、*S. sonnei* Ss046 (shisos shi)、*Yersinia pestis* biovar *Medievalis* str. 91001 (yerpeb yer)、*Y. pestis* KIM (yerpek yer)、*Y. pestis* CO92 (yerpec yer)、*Y. pseudotuber-*

^{*} 收稿日期: 2007-09-14

基金项目: 国家自然科学基金资助项目 (30270752); 广东省自然科学基金资助项目 (031616); 国家 973 计划资助项目 (2005CB724600)

作者简介: 熊远妍 (1981 年生), 女, 博士研究生; 通讯联系人: 张尚宏; E-mail: lsszsh@mail.sysu.edu.cn

culosis IP 32953 (yerpsi yer)。

1.2 计算权值矩阵

寡聚核苷酸组分的权值获得依据如下：频率保守性大的寡聚核苷酸在物种间的频率差异可能更好地反映物种间的进化距离，因此加上较大的权值；反之，则加较小的权值。对于一种寡聚核苷酸，权值取上述 194 物种中该寡聚核苷酸频率的级差（最大值与最小值之差）、方差、卡方（频率观测值与理论值差异的卡方）^[10] 等的倒数或是倒数的平方， N 长度的寡聚核苷酸组分用不同的统计值可以得到不同的 4^N 维的权值矩阵。

寡聚核苷酸的字段长度取值范围在 2~6 之间，长度取值不宜太大，因为寡聚核苷酸特异性随长度的增加相对减少，与运用种属特异性信息差异建树的假设不相符，所以长度取了 2~6 进行计算分析。计算权值矩阵时对 194 个基因组序列分别统计 N ($1 < N \leq 6$) 长度的各种寡聚核苷酸频率，一寡聚核苷酸的频率取该寡聚核苷酸在基因组中出现的次数与基因组长度的比值。寡聚核苷酸频率信息为 4^N 矩阵，计算各组分频率的平均值、级差、方差、变异系数、卡方等，取相应的反映保守性的统计数作权值。

1.3 应用寡聚核苷酸频率组分差异计算距离矩阵

对用于建树的物种的每一基因组全序列，分别计算一定长度 N ($1 \sim 6$) 的各种寡聚核苷酸的频率，再以权值矩阵中对应的权值加权后作为代表寡聚核苷酸在该基因组中的信息，本研究直接将频率乘以权值后得到的值作为组分信息 x_i 。最后，得到 4^N 维频率组分信息矩阵，依据一定长度的寡聚核苷酸组分信息差异计算序列间的距离。本文计算三种距离，包括绝对值距离（Manhattan 距离）、欧氏距离和 Lance 距离，其中用得最多的是绝对值距离，即直接用组分差异的绝对值相加求和。设 i 和 j 为两基因组序列， d_{ij} 为基因组之间距离，寡聚核苷酸长度取 N ，共有 $n = 4^N$ 种寡聚核苷酸组分，则 d_{ij} 依据不同距离算法公式分别如下：

绝对值距离（Manhattan 距离）：

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|$$

欧氏距离：

$$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$$

Lance 距离：

$$d_{ij} = \frac{1}{n} \sum_{k=1}^n \frac{|x_{ik} - x_{jk}|}{x_{ik} - x_{jk}}$$

其中 x_{ik} 和 x_{jk} 分别表示物种 i 和 j 一种寡聚核苷酸的

组分信息。

1.4 构建系统树

得到距离矩阵后用于构建系统树的方法采用 Phylips 软件包里的 NJ 法^[13]。建树结果的拓扑结构优劣的判断主要是以核糖体 rRNA 基因比对后构建的系统树作为参照，同时对照从 NCBI 上公布的分类地位分析物种间的进化关系。

2 结 果

2.1 用于计算权值的统计量

表 1 显示了当 $N=2$ 时，用于计算各二核苷酸权值的主要统计参数。从统计结果来看，AA 和 TT 在基因组中频率的平均值较高；比较变异系数以及标准差，AC、AG、CA、CT、GA、GT、TC、TG 较其他 8 种组合稳定，变化波动较小。因此，这 8 种二核苷酸的频率在不同基因组中是较为保守的。这种频率保守的特性在部分三核苷酸和更高阶寡聚核苷酸中也存在。根据这些衡量不同寡聚核苷酸频率保守程度的参数，就可以进一步计算出相应的权值矩阵。

表 1 194 种原核生物基因组全序列中二核苷酸频率的统计

Tab. 1 Statistics of dinucleotide frequencies in 194 prokaryotic whole-genome sequences

二核苷酸	平均值 (%)	最大值 (%)	最小值 (%)	标准差	变异系数 (%)
AA	8.837604	17.32066	1.582243	4.055016	45.88366
AC	4.943659	6.866716	2.975966	0.643022	13.00701
AG	5.536866	7.873059	4.01773	0.749869	13.54319
AT	7.545879	14.42168	1.59203	2.891981	38.3253
CA	6.164954	7.84804	4.315466	0.744014	12.06844
CC	5.52609	15.02082	1.521166	2.541754	45.99552
CG	5.913428	18.1954	0.209034	4.286959	72.49533
CT	5.535939	8.208271	3.924078	0.763123	13.78489
GA	5.953901	8.842311	4.390572	0.808922	13.58642
GC	6.710828	15.52361	1.817691	3.562678	53.08849
GG	5.524357	14.99913	1.465212	2.546862	46.10242
GT	4.940368	6.851028	2.928956	0.638096	12.91596
TA	5.907548	13.563	0.965767	3.395791	57.48225
TC	5.959833	8.593161	4.374779	0.800733	13.43549
TG	6.154809	7.9365	4.295401	0.741376	12.04547
TT	8.843627	17.41653	1.590688	4.063708	45.9507

2.2 建树结果

当直接用基因组寡聚核苷酸频率差异建树时，结果的拓扑结构与参照树差异很大，但也有合理的地方。当只看小的分枝，即那些距离很近的同属物

种, 以及同种的不同菌株, 它们可以很好的聚集在一起。寡聚核苷酸的长度取不同的值时, 四核苷酸、五核苷酸、六核苷酸建树后的结果基本上差异不大, 因此 N 值取 2、3、4 比较有意义。在对寡聚核苷酸频率加权时用到了多种权值矩阵, 结果是加权并不会从根本上改变树的拓扑结构, 但有一定的微调作用。

为了增加科以上分类单位的差异, 还尝试引入了基因组大小这一指标, 和寡聚核苷酸频率一起用于计算距离矩阵。对它们的处理是用寡聚核苷酸频数代替其频率, 然后尝试用寡聚核苷酸频率加上其频数的对数作为基因组组分信息来建树。结果显示这一指标的引入对区分原核生物与真核生物有一定的作用^[12]。

科以下分类单位的建树结果都比较理想。有关材料是变形菌门 (Proteobacteria)、 γ -变形菌纲 (Gammaproteobacteria)、肠杆菌目 (Enterobacteriales)、肠杆菌科 (Enterobacteriaceae) 中的 5 个不同属 11 个不同种的 23 株细菌。用二核苷酸的频率不加或是加上不同权值矩阵作为组分信息, 分别以绝对值距离、欧氏距离、Lance 距离计算距离矩阵, 结果是几乎所有系统树的分枝都可以将关系很近的同属物种以及同种不同菌株聚集得很好 (图 1), 与参照树的拓扑结构整体上一致, 有些分枝甚至较传统的建树结果更优。用绝对值距离不加权时, N 取 3 和 4 的树形相同, 都比 N 取 2 要好; 用 lance 距离时, N 取 2、3、4 的结果基本一样, 树

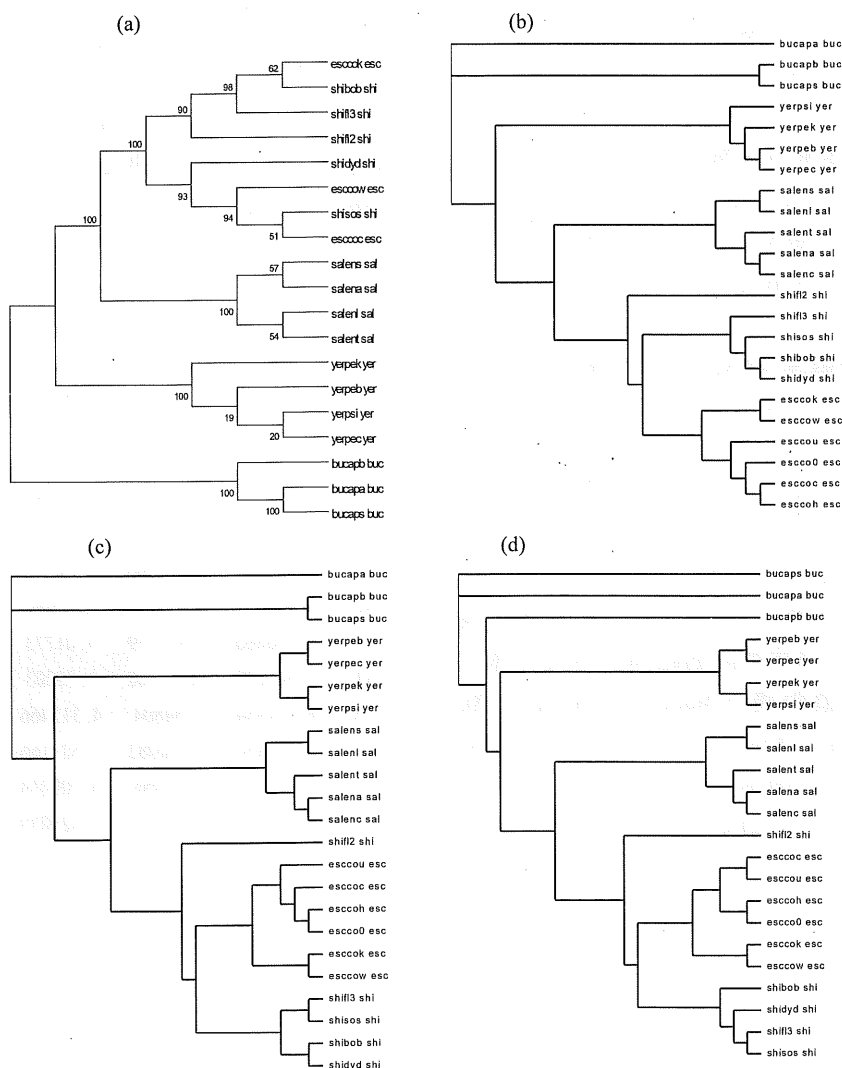


图 1 参照树和以寡聚核苷酸频率信息差异构建的系统树

Fig. 1 Reference tree and trees based on differences of oligonucleotide frequencies

(a) 用 NJ 法构建的 19 株细菌的参照树; (b) 用二核苷酸频率加上卡方倒数权值矩阵作为组分信息, Lance 距离计算距离矩阵的系统树; (c) 用三核苷酸频率作为组分信息, Lance 距离计算距离矩阵的系统树; (d) 用三核苷酸频率加上级差倒数权值矩阵作为组分信息, Lance 距离计算距离矩阵的系统树

形和参照树相似; 用欧氏距离不加权时, 结果相对较差, 但加权后也有很好的结果。因此, 寡聚核苷酸频率是分析同科不同属系统关系的良好标记。另一方面, 用寡聚核苷酸频数作为组分信息时, 系统树的结构明显不如用频率组分信息所得的结果, 且加权效果明显减弱。这进一步说明, 寡聚核苷酸频率信息是区分同科种属的标记, 加上其他参数可能使其作用减弱。

总的来说, 用寡聚核苷酸频率组分差异构建的系统树对科以上分类单位的区分与已有结果相差较大, 而对科以下亲缘关系较近的物种和同种不同菌株的区分与参照树很相似, 与 NCBI 上公布的分类也不矛盾。权值矩阵的引入对科以下分类单位的系统关系分析有精细的调节作用, 对科以上分类单位的作用则不明显。对科以下分类单位的系统分析用寡聚核苷酸频率作为组分信息较好, 权值微调的作用也比较明显的, 加上合适的权值, 系统树似乎可以更为合理。此外, N 的取值也不是要求很严格。3 种计算距离的方法中, Lance 距离相对较好。

3 讨论

应用寡聚核苷酸信息差异构建系统树的方法利用了基因组全序列的信息, 是真正意义上的基因组树, 且基于非比对的算法, 因此在很大程度上节约了建树的时间。以前由于基因组全序列比较少等原因, 并没有大量地对该方法进行考证。本研究首次通过以寡聚核苷酸频率保守性进行加权, 利用基因组全序列组分信息构建系统树。应用基因组全序列频率组分信息进行系统分析, 目前能得到比较合理结果的是亲缘关系较近的物种之间和种内菌株之间的比较, 而对科以上的分类单位, 建树结果则不理想。原因之一可能是基因组中的寡聚核苷酸组分信息有进化趋同现象, 从而频率差异较难很好地反映亲缘关系较远的物种之间的真实情况。因此, 对于分析科以上的分类单位, 仅仅是频率差异方面的信息量还不够。本文的结果只是比较系统树分枝的拓扑结构, 所以并没有做进一步的分枝长度处理。

本研究结果还进一步证实了基因组中寡聚核苷酸的出现频率具有一定的种属差异性, 从而说明这一信息用于进化分析是可行的。至于如何能更好地利用它, 则有待进一步的研究。另一方面, 本文的方法还可以考虑用更多的距离测度或相似性测度, 如离散二值距离、离散多值距离^[14]、相似系数 (Pearson 相关等)。选择不同类型的距离计算方法有助于得到更丰富和合理的结果。

参考文献:

- [1] WOLF Y I, ROGOZIN I B, KONDRASHOV A S, et al. Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context [J]. *Genome Res*, 2001, 11 (3): 356–372.
- [2] SNEL B, BORK P, HUYNEN M A. Genome phylogeny based on gene content [J]. *Nat Genet*, 1999, 21: 108–110.
- [3] WOLF Y I, ROGOZIN I B, GRISHIN N V, et al. Genome trees constructed using five different approaches suggest new major bacterial clades [J]. *BMC Evol Biol*, 2001, 1: 8.
- [4] HERNIOU E A, LUQUE T, CHEN X, et al. Use of whole genome sequence data to infer baculovirus phylogeny [J]. *J Virol*, 2001, 75 (17): 8117–8126.
- [5] LIN J, GERSTEIN M. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels [J]. *Genome Res*, 2000, 10 (6): 808–818.
- [6] KARLIN S, MRAZEK J, CAMPBELL A M. Compositional biases of bacterial genomes and evolutionary implications [J]. *J Bacteriol*, 1997, 179: 3899–3913.
- [7] ABE T, KANAYA S, KINOCHI M, et al. Informatics for unveiling hidden genome signatures [J]. *Genome Res*, 2003, 13: 693–702.
- [8] WANG Y, HILL K, SINGH S, et al. The spectrum of genomic signatures: From dinucleotides to chaosgame representation [J]. *Gene*, 2005, 346: 173–185.
- [9] KARLIN S, BURGE C. Dinucleotide relative abundance extremes: A genomic signature [J]. *Trends Genet*, 1995, 11: 283–290.
- [10] ZHANG S H, YANG J H. Conservation versus variation of dinucleotide frequencies across genomes: Evolutionary implications [J/OL]. *Genome Biol*, 2005, 6: 12 [2005-6-11]. <http://genomebiology.com/2005-06-11>.
- [11] 蓝一杰. 物种间基因组寡聚核苷酸的频率特征比较 [D]. 广州: 中山大学, 2006: 46–51.
- [12] 熊远妍. 基于基因组全序列的分子系统学方法研究 [D]. 广州: 中山大学, 2006: 12–16.
- [13] FELSENSTEIN J. PHYLIP: Phylogeny inference package (version 3.57) [CP]. Seattle, WA: University of Washington, 1995.
- [14] 张文军, 冯永军, 古德祥. 统计显著性标记的聚类分析算法与网络实现 [J]. *数值计算与计算机应用*, 2005, 3: 232–240.
ZHANG Wenjun, FENG Yongjun, GU Dexiang. Algorithm and implementation of a cluster analysis with statistical significance [J]. *Journal of Numerical Methods and Computer Applications*, 2005, 3: 232–240.

Evolutionary Information Contained in Oligonucleotide Frequency Differences Across Genomes

XIONG Yuan-yan, WANG Jun-peng, LAN Yi-jie, WEN Ming, ZHANG Shang-hong

(The Key Laboratory of Gene Engineering of Ministry of Education // Biotechnology Research Center,
Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Phylogenetic trees were constructed based on information from oligonucleotide frequency differences across genomes in this paper. Frequency data from 194 prokaryotic genomes were calculated and analyzed for their characteristics of conservation across genomes. Oligonucleotide frequency differences among genomes were weighted according to the characteristics of conservation. The weighted differences were then used as measures of genetic distances among genomes to construct phylogenetic trees. Compared with the traditional method, similar trees were obtained for classification units lower than family, while trees for classification units higher than family were different. The phylogenetic trees constructed in this study are true whole-genome trees. The approach used is alignment-free and specific in weighing the oligonucleotide frequency differences.

Key words: whole-genome sequence; oligonucleotides; frequency; phylogenetic tree; information analysis; data mining; bioinformatics

(上接第 83 页)

Ecological and Physiological Research on Arbuscular Mycorrhizas (AM) of Winter Cropping Italian Ryegrass in Different Areas of Guangdong

XIN Guo-rong, SUN Bin, LI Guo-xi, WU Jin, YANG Yu-jie, WANG Yu-tao, YANG Zhong-yi

(State Key Laboratory for Biological Control//School of Life Sciences, Sun Yat-sen University,
Guangzhou 510275, China)

Abstract: In this study, we investigated the formation rules, the nutritional and environmental conditions (moisture, pH value, soil temperature, soil nutrient, etc.) of the VA mycorrhizal in Italian ryegrass rhizosphere in different areas of Guangdong province. Our results showed that, the AMF colonization in Italian ryegrass root system is universal in different areas of Guangdong, but the AMF colonization in Italian ryegrass showed no regional characteristic. Besides the colonization of mycelium, a large number of AM structures of arbuscules and vesicles were also found. The total AM infection rates, the mycelium infection rates, the arbuscule infection rates and the vesicle infection rates all had significant negative correlations with soil total phosphorus and soil available phosphorus ($p < 0.01$). There were differences in the quantity of spores contained in Italian ryegrass rhizosphere in different areas of Guangdong. The number of the total AMF spores had significant positive correlations with the AM arbuscular infection rates and the AM vesicle infection rates ($p < 0.01$). It also had significant positive correlations with the soil moisture, organic matter, total nitrogen and soluble nitrogen ($p < 0.01$), and a significant negative correlation with the soil total phosphorus ($p < 0.05$). The differences of soil physical and chemical characters, regional and environmental factors may be the main factors inducing the differences of AMF quantity in different areas. The results may be applied to the management of paddy field and to the efficient fertilization.

Key words: *Lolium mutiflorum* L.; arbuscular colonization; quantity of AMF spores; physiology; ecology