计算分子生物学(网络版), 2012 年, 第 1 卷, 第 3 篇, 第 16–22 页
Jisuan Fenzi Shengwuxue (Online), 2012, Vol.1, No.3, 16–22
http://cmb.5th.sophiapublisher.com

研究论文
Research Article

原核生物基因组三核苷酸转移概率偏倚的物种特异性及致病关联性

章芬 [1], 黄庆生 [2], 严翠婷 [1], 吴建华 [1]

1 华南理工大学生物科学与工程学院, 广州, 510006;
2 中山大学生命科学学院, 广州, 510275
通讯作者: wujianhua@scut.edu.cn; 作者

计算分子生物学, 2012 年, 第 1 卷, 第 3 篇 doi: 10.5376/cmb.cn.2012.01.0003
收稿日期：2012 年 03 月 12 日
接受日期：2012 年 06 月 28 日
发表日期：2012 年 07 月 12 日

建议最佳引用格式：

引用格式(中文)：

章芬等, 2012, 原核生物基因组三核苷酸转移概率偏倚的物种特异性及致病关联性, 计算分子生物学(online) Vol.1 No.3 pp.16-22 (doi: 10.5376/cmb.cn.2012.01.0003)

引用格式(英文)：

Zhang et al., 2012, The Correlation between Species-specificity and Pathogenicity of Trinucleotide Transition Probability Bias in Prokaryotic Genomes, Jisuan Fenzi Shengwuxue (online) (Computational Molecular Biology) Vol.1 No.3 pp.16-22 (doi: 10.5376/cmb.cn.2012.01.0003)

摘 要 作为 DNA 序列的重要组成特征，基因组寡核苷酸使用模式及其偏倚的研究已被广泛应用于原核生物基因组的分析。然而，关于寡核苷酸使用模式的偏倚是否具有种群特异性并反映种群的功能这一问题，尚未阐明。我们基于一阶马尔可夫链模型，提出了一个度量寡核苷酸使用模式偏倚的新指标——基因组三核苷酸(trinucleotide, tri-)转移概率偏倚(transition probability bias, TPB)特征向量，或称之为三核苷酸转移概率最大偏倚分布，并分析比较了 727 条有代表性的原核生物基因组序列 tri-TPB 特征向量。结果表明，基因组 tri-TPB 特征向量具有物种特异性，亲缘关系越近的物种，它们的 tri-TPB 特征向量越相似；同种内之不同菌株具有几乎完全相同的 tri-TPB 特征向量，并且不依赖于基因组的 GC 含量；此外，基因组 tri-TPB 特征向量的相似性与菌株的致病性特征相关。本研究结果为基于全基因组寡核苷酸组成和分布信息的物种及其致病性进化分析提供了新的思路和方法。

关键词 原核生物; 三核苷酸转移概率偏倚; 分子进化; 物种特异性; 致病性

The Correlation between Species-specificity and Pathogenicity of Trinucleotide Transition Probability Bias in Prokaryotic Genomes

Zhang Fen [1], Huang Qingsheng [2], Yan Cuiting [1], Wu Jianhua [1]

1 School of Bioscience and Bioengineering, South China University of Technology, Guangzhou, 510006;
2 School of Life Sciences, Sun Yat-sen University, Guangzhou, 510275
Corresponding author: wujianhua@scut.edu.cn; Authors

Abstract As important characteristics of DNA sequence compositions, genomic oligonucleotide usage pattern and its bias study have been widely used in the analysis of prokaryotic genomes. Nevertheless, it remains unclear whether the bias of the genomic oligonucleotide usage pattern possesses species-specific properties of the genomes and reflects species functions or not. Based on a Markov chain model, a novel index — the characteristic vector of trinucleotide transition probability bias (tri-TPB), namely the distribution pattern of maximum trinucleotide transition probability bias, was proposed to measure the oligonucleotide usage pattern bias. 727 representative prokaryotic genomes were analyzed and compared their characteristic of tri-TPB vector. Our results showed that the closer the phylogenic relationship is, the more similar the characteristic of tri-TPB vectors is; especially, an almost identical characteristic vector tri-TPB pattern remains seen nearly in all genomes within the same species, is independent of genome GC contents. In addition, it was indicated that the similaritis of characteristic vectors of genomic tri-TPB patterns correlate closely with the pathogenicity of bacterial strains. The present results provide us a new perspective for the analysis of genome evolution and their pathogenicity evolution in genomic oligonucleotide composition and distribution.

Keywords Prokaryote; Trinucleotide transition probability bias; Molecular evolution; Species-specificity; Pathogenecity

无论是在原核生物，还是在真核及非细胞生物基因组中，短寡核苷酸含量及其分布受到人们的广泛关注和长期研究(Muto and Osawa, 1987; Karlin et al., 1993; Karlin et al., 1994; Karlin et al., 1997)。研究表明，密码子使用模式受到有效核糖体的选择、基因漂移以及偏倚突变等多种因素的调控，影响基因表达的效率与基因组内核酸的使用模式(Grantham et al., 1981; Bibb et al., 1984; Shah and Gilchrist, 2011)；生物有机体的二核苷酸相对丰度值是一种基因组标签，可以用来描述二核苷酸化学堆垛能、限制性内切酶的选择、物种特异性 DNA 修饰、复制和修复机制以及构象偏好等(Kariin and Burge, 1995; Karlin, 2001)；原核生物基因组的四核苷酸使用模式(tetranucleotide usage departure, TUD)具有物种特异性，用 TUD 构建的系统发育树含有一定的进化信息(Pride et al., 2003)。此外，许多神经系统相关疾病及肿瘤都与其基因组中的微卫星(三核苷酸的重复片段)的大量存在有关(Orr and Zoghbi, 2007; Haberman et al., 2008)。

我们知道，基因组的进化受到多重因素的影响，基于单个基因的分析已不足以全面了解相关物种的系统发育关系。尽管 SSU rRNA (small subunit rRNA)基因已被广泛应用于系统发育学的研究之中(Woese and Fox, 1977)，但由于作为基因组进化的重要动力源泉的平行转移基因(Ochman et al., 2000)的普遍存在，使得基于 SSU rRNA 基因以及其他蛋白编码基因获得的系统发育关系之间出现明显的不一致(Doolittle, 1999)；同时，SSU rRNA 基因具有高度保守性，这使得进化上远缘的 SSU rRNA 基因可能有非常相似的核酸组成，导致其在系统发育树上会被错误地聚集在一起(Hasegawa and Hashimoto, 1993)。

目前，基于马尔可夫链模型的分析方法(Phillips et al., 1987)是揭示短寡核苷酸在基因组中出现频率特征的一种有力工具。也许，寡核苷酸转移概率分布的局部偏倚，也就是转移概率矩阵相邻分量间的差异，不但是寡核苷酸在基因组中出现频率的差异的一种有意义的统计学度量，而且刻画了基因组中寡核苷酸成分的动力学稳定性，进而含有物种进化的信息。我们猜想，在基因组寡核苷酸成分的动力学稳定性之中，隐藏有 SSU rRNA 基因所无法揭示的物种间差异以及这些差异与菌株之特异生物学功能之间的关联。鉴于此，本文建议了一个新的度量基因组中寡核苷酸成分的局部偏倚或动力学稳定性的指标——基因组三核苷酸(trinucleotide, tri-)转移概率偏倚(transition probability bias, TPB)，以揭示隐藏于寡核苷酸相对丰度与密码子使用偏倚之中的有用信息。利用这一全新的指标，我们研究了来自古生菌、真细菌基因组及其质粒的 1 170 条 DNA 序列的 tri-TPB 特征向量或三核苷酸转移概率最大偏倚分布间的相似性，发现基因组 tri-TPB 特征向量具有极高的种群保守性和显著的致病关联性。

1 结果分析

1.1 tri-TPB 特征向量具有物种特异性

我们发现，种群内各基因组的 tri-TPB 特征向量是相似的。对迄今所发现的真细菌中两个最大的菌门(Gammaproteobacteria 和 Firmicute)而言，Gammaproteobacteria 菌门的两个 *H. pylori* 菌株(*H. pylori* 26695 和 *H. pylori* J99)有近乎相同的最大 tri-TPB 分布(R^2=0.996, 图 1A)，但它们与其远缘物种

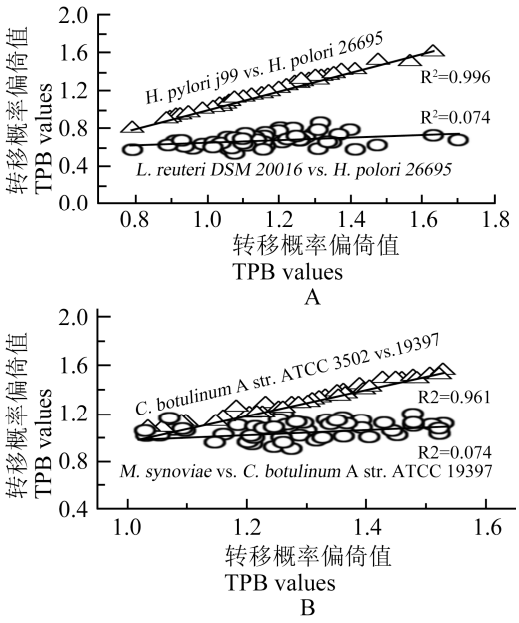

图 1 具有相同 GC 含量的细菌基因组 tri-TPB 特征向量间的线性相关性

注: A: Gammaproteobacteria 菌门的菌株 *H. pylori* 26695 及 *H. pylori* J99 和其远缘物种 *L. reuteri* DSM 20016 间的相似性; 这三个菌株基因组的 GC 含量相同, 都为 38%; B: Firmicute 菌门的菌株 *C. botulinum* A str. ATCC 19397, *C. botulinum* strain ATCC 3502 和其远缘物种 *M. synoviae* 间的相关性; 这三个菌株基因组的 GC 含量相同, 都为 28%

Figure 1 Linear correlations of characteristic tri-TPB vectors among some genomes with same GC content

Note: A: Similarities between Gammaproteobacteria *H. pylori* 26695, *H. pylori* strain J99 and an evolutionary distant strain *L. reuteri* DSM 20016. They contain the same GC content of 38%; B: Correlation among Firmicute bacterium *C. botulinum* A str. ATCC 19397, *C. botulinum* strain ATCC 3502 and evolutionary distant strain *M. synoviae*. All contain nearly the same GC content of 28%

L. reuteri DSM 20016 之间的相关性极少(图 1A)；对 Firmicute 菌门的两个 *C. botulinum* 菌株(*C. botulinum* A str. ATCC 19397 和 *C. botulinum* strain ATCC 3502)的 tri-TPB 特征向量而言，它们间的相关系数达到 0.961，但它们与其远缘物种 *M. synoviae* 的 tri-TPB 特征向量之间的相关系数值却只有 0.074 (图 1B)。这一 tri-TPB 特征向量的相似性，亦存在于具有多条染色体的菌株之中(图 2)。例如，*V. harveyi* 菌株的

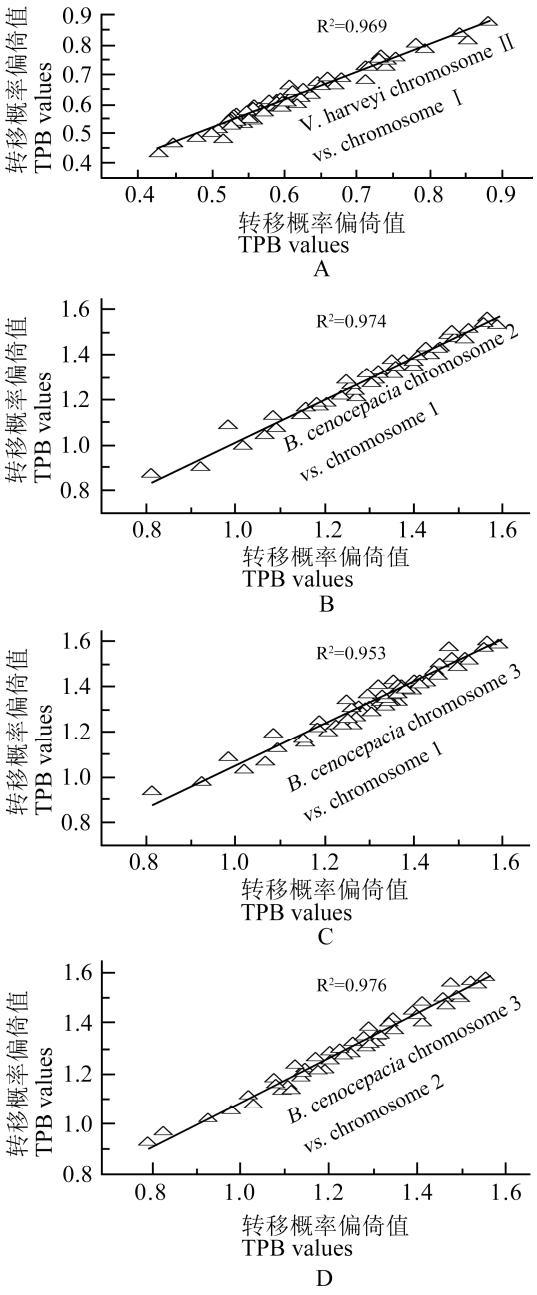

图 2 *V. harveyi* (A)和 *B. cenocepacia* (B, C, D)菌株不同染色体序列的特征 tri-TPB 向量的线性相关性分析

Figure 2 Linear correlation of characteristic tri-TPB vectors among different chromosomes of *V. harveyi* (A) and *B. cenocepacia* (B, C, and D)

条不同染色体基因的 tri-TPB 特征向量的相关系数达到了 0.969 (图 2A)，*B. cenocepacia* 菌株的三条染色体之间也呈现出极佳的线性相关性(图 2B; 图 2B C; 图 2B D)。上述结果与之前关于原核生物基因组 TUD 模式的研究结果相一致(Pride et al., 2003)。这提示，基因组的最大 tri-TPB 分布是物种特异性的，它可以刻画一个物种的特征。

我们的结果还表明，基因组 tri-TPB 特征分布不依赖于 GC 含量的大小。尽管 *H. pylori* 菌株的远缘物种 *L. reuteri* DSM 20016 具有和 *H. pylori* 菌株相同的 GC 含量(38%)，但它们的最大 tri-TPB 分布却不相似(图 1A)；Firmicute 菌门的两个菌株 *C. botulinum* A str. ATCC 19397 和 *C. botulinum* A str. ATCC 3502 与它们的远缘物种 *M. synoviae* 的 GC 含量均为 28%，但后者与前两者之间也无相似的最大 tri-TPB 分布(图 1B)；另外，对通过人工构建的一条与两个 *H. pylori* 菌株有着相同基因组大小和相同 GC 含量的随机 DNA 序列而言，它的 tri-TPB 特征向量与两个 *H. pylori* 菌株的 tri-TPB 特征向量几乎没有相关性(R^2<0.01, 数据未提供)。

1.2 原核生物 tri-TPB 特征向量间的相似性随着物种间进化距离的增加而减小

本研究展示了 *E. coli* str. K–12 substr. MG1655 与不同分类级别物种内的菌株 tri-TPB 特征分布向量间的相关性(图 3)。其中，A 组是 K–12 与同种内的其

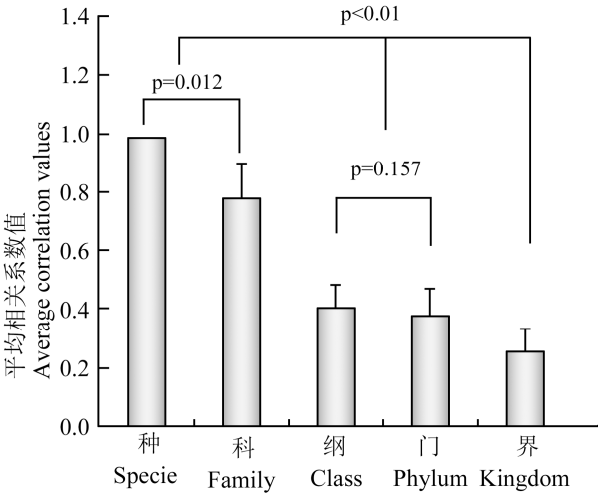

图3 *Escherichia coli* str. K–12 substr. MG1655分别与和它同种, 同科, 同纲, 同门以及同界的不同菌株间tri-TPB特征向量的平均相关程度

Figure 3 Average correlation coefficients of characteristic tri-TPB vectors among *E. coli* str. K–12 substr. MG1655 and other different strains within the same Species, Family, Class, Phylum, and Kingdom with K–12 respectively

它菌株的相关性；B 组是 K-12 与同科内的其它属的菌株的比较(不包含与 *Buchnera aphidicola* 属的菌株的比较)；C 组是 K-12 与同纲内的其它目的菌株的比较；D 组是 K-12 与同门内的其它纲的菌株的比较；E 组是 K-12 与其它门的菌株的比较；相关系数的值都是以均值±标准差的形式展示的(每组中数据点的个数分别为 8, 32, 102, 198 和 286)。除了纲和门两组无显著性差异(p=0.157)之外，其他各组间的比较都具有统计学意义上的显著性差异(p 值的变化范围为 0.01 到 0.012)。由于缺乏与 K-12 同属不同种以及同目不同科的物种数据，所以在图 4 中未含有这类比较结果。

结果表明，各 tri-TPB 特征向量间的平均相关系数值随着物种间进化距离的增加而减小，即：沿着从界、门、纲、科到种的进化路径，种群内各物种 tri-TPB 特征向量间的相似性逐渐增加。也就是说，在原核生物整体进化的水平上，分类学上亲缘关系越近的物种，它们的 tri-TPB 分布越相似。另一方面，与种间极小的差异不同，在界、门、纲和科内，各基因组 tri-TPB 特征向量间的相关系数的标准差是显著的(图 4)。这种在界、门、纲和科内的各基因组 tri-TPB 特征向量间的显著差异性，意味着原核生物基因组三核苷酸转移概率最大偏倚分布的多样性。

1.3 近缘物种的 tri-TPB 特征向量具有致病关联性

针对假单胞菌属(*Pseudomonas*)内具有不同致病性特征的菌群，我们分析比较了同一菌群内部及不同菌群之间的细菌基因组 tri-TPB 特征向量间的关系，所得结果如图 4 所示。其中，Ga 表示动物致病型菌群(Stover et al., 2000; Vodovar et al., 2006)，包含 *P. aeruginosa* PA7、*P. aeruginosa* UCBPP-PA14、*P. aeruginosa* LESB58、*P. aeruginosa* PAO1 和 *P. entomophila* L48 五个菌株；Gp 为植物致病型菌群(Feil et al., 2005)，含有 *P. syringae* pv. *syringae* B728a、*P. syringae* pv. *phaseolicola* 1448A 和 *P. syringae* pv. tomato str. DC3000 三个菌株；Gn 是非致病型菌群(Nelson et al., 2002)，由 *P. fluorescens* SBW25、*P. fluorescens* Pf-5、*P. fluorescens* Pf0-1、*P. putida* F1、*P. putida* KT2440、*P. putida* GB-1、*P. putida* S16、*P. putida* W619、*P. mendocina* NK-01 和 *P. mendocina* ymp 等 10 个菌株组成；符号“++”和“+”分别表示组间和组内比较，符号“-”表示未参与比较；利用 t-检验来估计相关数据的统计学差异的显著性。研究结果表明(图 4)，对动物致病型菌群，植物致病菌群和非致病型菌群而言，相同菌群内的 tri-TPB 特征向量具有很高的相似性，它们的相关系数大于 0.95；不同菌群的 tri-TPB 特征向量间的相似性要低于菌群内的相似性，且具有显著的统计学差异($p<0.01$)。基因组 tri-TPB 特征向量的差异，依赖于菌群的致病特性：最显着的差异可能存在于动物与植物致病型菌群中，非致病型与动物致病型菌群间的差异较小，而存在于非致病型与植物致病型菌群间的差异几乎可以不计。这表明：较之非致病与植物致病型菌群，动物致病型菌群具有更为特异的三核苷酸转移概率最大偏倚分布。

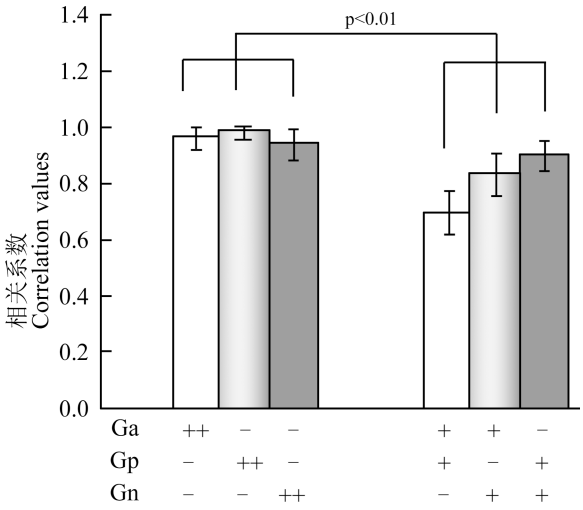

图 4 *Pseudomonas* 菌属内不同致病型菌株间 tri-TPB 特征向量相似性
注: 相关系数的结果以均值±标准差的形式来表示
Figure 4 Similarity of characteristic tri-TPB vectors among groups of *Pseudomonas* with different pathogenicity
Note: The data of correlation coefficient are presented as mean±SE

如上所述，假单胞菌属内菌株致病性的有无以及致病类型与菌株基因组 tri-TPB 特征向量的相似性之间存在相关性。这一菌株基因组特征与其致病性间的关联，通常不能从传统的系统发育分析中获得，并可能被传统分析方法所曲解。事实上，利用我们基于 16S rRNA 基因序列所构建的系统发育树，可以发现：动物致病型菌株 *P. entomophila* 与非致病型菌群 *P. putida* 处在同一分支，但很早以前就与动物致病型菌株 *P. aeruginosa* 发生了分歧(图 5)。

2 讨论

已有研究表明，细菌基因组中基因的多样性在很大程度上是来自于基因的平行转移，且这些转移

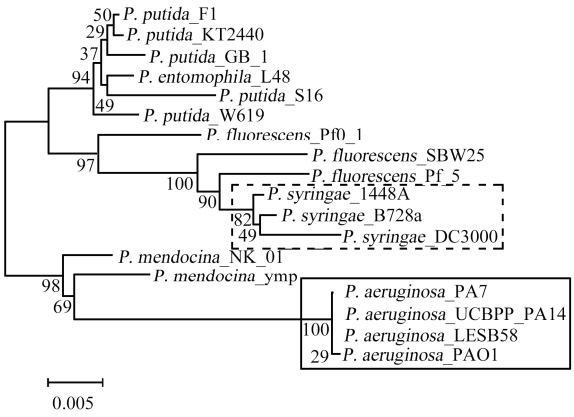

图 5 *Pseudomonas* 菌属内 18 个具有不同致病性特征的菌株的系统发育树

注: 实线框内为动物致病型菌株; 虚线框内为植物致病型菌株; 其它为非致病型菌株; 该系统发育树是基于 16S rRNA 序列, 依据最大组成似然距离矩阵, 采用邻接法构建而成; 树的节点处显示的数据是进行 500 次重复的自举检验得分

Figure 5 Phylogenetic tree of 18 *Pseudomonas* strains with different pathogenic features

Note: Animal pathogenic bacteria are within solid box; plant pathogenic bacteria are within dotted box; and others are avirulent bacteria. 16S rRNA sequences were subjected to neighbor-joining analysis using maximum composite likelihood distance matrices. Bootstrap values based on 500 replicates are represented at each node of phylogenetic tree

事件不仅存在于远缘物种的 DNA 序列之间(Ochman et al., 2000; Juhas et al., 2009), 同样也发生在近缘细菌和真核生物的同源基因之间(Hotopp et al., 2007)。这可能导致亲缘关系非常近的物种内的基因含量之显著差异。有趣的是, 我们的结果表明, 基因组特征 tri-TPB 向量在种群内具有保守性, 且这种保守性不依赖于基因组的 GC 含量的大小。这意味着: 基因的平行转移现象, 虽然会导致种群内基因的多样性, 但不会导致基因组 tri-TPB 特征向量的改变。尽管基因组 tri-TPB 特征向量间的相似性会随物种分类级别的升高而逐渐减小(图 4), 但却可能提供了一个研究分子遗传与进化的有用度量指标。

无论是在表型特征数据缺乏时所作的系统发育关系推断, 还是对现有表型特征数据所得结果间的比较, 基于核糖体 RNA 分子基因序列的系统发育分析均被视为一种可靠的方法。尽管如此, 这一方法在刻画物种间的差异时会出现盲区(Woese and Fox, 1977)。研究发现, 基因组中不稳定的基因通常与细胞表面信号转导或致病性等功能相关, 而不易发生转移的基因则参与到翻译或氨基酸的合成过程, 如编码核糖体 RNA 的基因(Lawrence, 1999)。这使得基于 SSU rRNA 等单个基因的传统系统发育分析方法很难察觉出物种中相对活跃的功能性基因的改变。因而, 系统发育分析难以回答菌株基因组特征是否可以表征菌株的致病特性。与之不同, 我们的研究表明, 近缘物种中具有相同致病性的菌株基因组 tri-TPB 特征向量间的相关性明显高于致病性不同的菌株, 这提示我们, 基因组 tri-TPB 向量也许还可以刻画菌株的致病特性。

可以认为, 基于基因组 tri-TPB 特征分布的分析, 也许是对其他基于全基因组分析工具的一个重要补充。与其他分析方法相比, 尽管基于基因组 tri-TPB 的分析也无法逾越大量平行转移基因和协同进化带来的物种进化分析障碍, 但却避免了复杂而耗时耗力的核苷酸或氨基酸序列比对。同时, 菌株基因组 tri-TPB 特征向量间的相似性与菌株致病性之间的相互关联, 可以为近缘物种致病性的进化分析以及相关疾病的预防和治疗提供新的思路和方法。

3 材料与方法

3.1 原核生物基因组及其质粒序列数据

本文的研究对象为 675 个原核物种的 727 条全基因组序列, 它们均下载自 NCBI 的 FTP 服务站点(ftp://ftp.ncbi.nih.gov/genomes/); 用于系统发育分析的 18 种原核生物的 16S rRNA 核苷酸序列均下载自 NCBI(http://www.ncbi.nlm.nih.gov/sites/gene/); 我们人工构建了一条长度为 5 Mbp, GC 含量为 50%的类似于大肠杆菌的完全随机序列, 并将其与 1 170 条完全测序成功的基因组序列进行比较, 检测分析本文分析方法的稳定性。

3.2 基因组转移概率偏倚(transition probability bias, TPB)及其特征向量

对于任意一条 DNA 序列, 将序列中长度为 k 的寡核苷酸片段记为 $\omega_1\omega_2\cdots\omega_k$, 其中 ω_s (s=1,…, k) 是四种碱基中的任意一种。将所有不同的 4^k 个长度为 k 的寡核苷酸序列中的第 i 个 $[\omega_1\omega_2\cdots\omega_k]_i$ 记为 A_i, 而将长度为 2k 的寡核苷酸序列 $[\omega_1\omega_2\cdots\omega_k]_i[\omega_1\omega_2\cdots\omega_k]_j$ 记为 A_iA_j, 其中 i 和 j 取 1 到 4^k 之间的整数。这样, 对于任意一条 DNA 序列, 基于马尔可夫链模型, 由寡核苷酸序列 A_i 过渡到 A_j 的转移概率 p_{ij} 可由下式计算:

$$p_{ji}=P(A_j|A_i)=\frac{P(A_iA_j)}{P(A_i)};\ i,j=1,2,\ \cdots,\ 4^k \qquad (1)$$

其中，$p_{ij}=P(A_iA_j)$和 $P(A_i)$分别是在长度为 k 和 2k 的读码框下观察得到的 k 阶寡核苷酸序列 A_i 和 2k 阶寡核苷酸短序列 A_iA_j 的出现频率。所有寡核苷酸的出现频数的计算都是基于 DNA 序列的正负两条链。

对于不同的 DNA 序列，其转移概率矩阵(transition probability matrix, TPM) (Van't Spijker et al., 2009)也是不同的。我们定义转移概率偏倚(transition probability bias, TPB)向量 $\Delta=\{\Delta_i\}$

$$\Delta_i=\sum_{j=1}^{n}\left|p_{j,i}-p_{j-1,i}\right|,\ i=1,2,...n \qquad (2)$$

其中 $n=4^k$。TPB 向量 Δ 是非唯一的，可以作为转移概率分布非均匀性的一种度量。对转移概率矩阵的每一行数据进行重排，可得 TPB 向量的 $M=4^k\times\ (4^k-1)\times\ldots\times2\times1$ 种不同表达形式 $\Delta^{(m)}=(\Delta_1^{(m)},\Delta_2^{(m)},\cdots,\Delta_n^{(m)})$ (m=1, 2,…,M)。为简化后续序列分析，我们引入 TPB 特征向量 Δ_c(它为所有可能的转移概率分布的拓扑结构中的一种, 表征转移概率最大偏移之分布)，其每一个元素对应相应 TPM 各行转移概率偏倚的最大值，也就是：

$\Delta_c=(d_1,d_2,\cdots,d_n)$，$d_j=\mathrm{Max}(\Delta_j^{(1)},\Delta_j^{(2)},\cdots,\Delta_j^{(M)})$，j=1, 2,…,n (3)

具有 TPB 特征向量 Δ_{1c} (x_1, x_2,…,x_n)和 Δ_{2c} (y_1, y_2,…,y_n)的两条不同 DNA 序列之间的相关性，由皮尔森(Pearson)相关系数 r 来衡量，公式如下：

$$r=\frac{\sum(x_j-\bar{x})(y_j-\bar{y})}{\sqrt{\sum(x_j-\bar{x})^2(y_j-\bar{y})^2}};\ \bar{x}=\sum_{j=1}^{n}x_j;\ \bar{y}=\sum_{j=1}^{n}y_j \qquad (4)$$

本研究中，所有寡核苷酸在基因组中出现频率的计算是利用 TCL 脚本程序完成，而 TPB 特征向量 Δ_c 的计算则是利用 MATLAB 程序完成。作为初步研究成果，我们这里仅讨论基因组 tri-TPB 特征向量或三核苷酸转移概率最大偏倚分布，tri-TPB 表示三核苷酸转移概率偏倚。

作者贡献

章芬负责实验设计、实验数据采集与分析及论文初稿写作；黄庆生负责编写程序，参与部分数据分析和讨论；严翠婷参与部分数据分析；吴建华负责研究方案与实验设计、数据分析、论文写作和修改。

致谢

本研究受到国家自然科学基金面上项目(10772069)、广东省工业攻关项目(2008B011000017)和广东省自然科学基金项目(S2011010005451)的资助。

参考文献

Bibb M.J., Findlay P.R., and Johnson M.W., 1984, The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences, Gene, 30(1-3): 157-166

Doolittle W.F., 1999, Phylogenetic classification and the universal tree, Science, 284(5423): 2124-2129

Feil H., Feil W.S., Chain P., Larimer F., DiBartolo G., Copeland A., Lykidis A., Trong S., Nolan M., Goltsman E., Thiel J., Malfatti S., Loper J.E., Lapidus A., Detter J.C., Land M., Richardson P.M., Kyrpides N.C., Ivanova N., Lindow S.E., 2005, Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. tomato DC3000, Proceedings of the National Academy of Sciences of the United States of America, 102(31): 11064-11069

Grantham R., Gautier C., Gouy M., Jacobzone M., and Mercier R., 1981, Codon catalog usage is a genome strategy modulated for gene expressivity, Nucleic Acids Research, 9(1): 213

Hasegawa M., and Hashimoto T., 1993, Ribosomal RNA trees misleading, Nature, 361(6407): 23

Haberman Y., Amariglio N., Rechavi G., and Eisenberg E., 2008, Trinucleotide repeats are prevalent among cancer-related genes, Trends in Genetics, 24(1): 14-18

Hotopp J.C.D., Clark M.E., Oliveira D.C.S.G., Foster J.M., Fischer P., Torres M.C.M., Giebel J.D., Kumar N., Ishmael N., Wang S., Ingram J., Nene R.V., Shepard J., Tomkins J., Richards S., Spiro D.J., Ghedin E., Slatko B.E., Tettelin H., and Werren J.H., 2007, Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes, Science, 317(5845): 1753-1756

Juhas M., van der Meer J.R., Gaillard M., Harding R.M., Hood D.W., and Crook D.W., 2009, Genomic islands: Tools of bacterial horizontal gene transfer and evolution, FEMS Microbiology Reviews, 33(2): 376-393

Kariin S., and Burge C., 1995, Dinucleotide relative abundance extremes: A genomic signature, Trends in Genetics, 11(7): 283-290

Karlin S., 2001, Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes, TRENDS in Microbiology, 9(7): 335-343

Karlin S., Blaisdell B.E., Sapolsky R.J., Cardon L., and Burge C., 1993, Assessments of DNA inhomogeneities in yeast chromosome III, Nucleic Acids Research, 21(3): 703-711

Karlin S., Mocarski E.S., and Schachtel G.A., 1994, Molecular evolution of herpesviruses: Genomic and protein sequence comparisons, Journal of Virology, 68(3): 1886-1902

Karlin S., Mrázek J., and Campbell A.M., 1997, Compositional biases of bacterial genomes and evolutionary implications, Journal of Bacteriology, 179(12): 3899-3913

Lawrence J.G., 1999, Gene transfer, speciation, and the evolution of bacterial genomes, Current Opinion in Microbiology, 2(5): 519-523

Muto A., and Osawa S., 1987, The guanine and cytosine content of genomic DNA and bacterial evolution, Proceedings of the National Academy of Sciences, 84(1): 166-169

Nelson K.E., Weinel C., Paulsen I.T., Dodson R.J., Hilbert H., Martins dos Santos V.A., Fouts D.E., Gill S.R., Pop M., Holmes M., Brinkac L., Beanan M., DeBoy R.T., Daugherty S., Kolonay J., Madupu R., Nelson W., White O., Peterson J., Khouri H., Hance I., Chris Lee P., Holtzapple E., Scanlan D., Tran K., Moazzez A., Utterback T., Rizzo M., Lee K., Kosack D., Moestl D., Wedler H., Lauber J., Stjepandic D., Hoheisel J., Straetz M., Heim S., Kiewitz C., Eisen J.A., Timmis K.N., Düsterhöft A., Tümmler B., Fraser C.M., 2002, Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440, Environmental Microbiology, 4(12): 799-808

Ochman H., Lawrence J.G., Groisman E.A., 2000, Lateral gene transfer and the nature of bacterial innovation, Nature, 405 (6784): 299-304.

Orr H.T., and Zoghbi H.Y., 2007, Trinucleotide repeat disorders, Annu. Rev. Neurosci., 30: 575-621

Phillips G.J., Arnold J., and Ivarie R., 1987, Mono-through hexanucleotide composition of the *Escherichia col*i genome: A Markov chain analysis, Nucleic Acids Research, 15(6): 2611-2626

Pride D.T., Meinersmann R.J., Wassenaar T.M., and Blaser M.J., 2003, Evolutionary implications of microbial genome tetranucleotide frequency biases, Genome Research, 13(2): 145-158

Shah P., and Gilchrist M.A., 2011, Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift, Proc. Natl. Acad. Sci. USA, 108(25): 10231-10236

Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik D.J., Lagrou M., Garber R.L., Goltry L., Tolentino E., Westbrock-Wadman S., Yuan Y., Brody L.L., Coulter S.N., Folger K.R., Kas A., Larbig K, Lim R., Smith K., Spencer D., Wong G.K., Wu Z., Paulsen I.T., Reizer J., Saier M.H., Hancock R.E., Lory S., and Olson M.V., 2000, Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen, Nature, 406(6799): 959-964

Van't Spijker A., Rodriguez J.M., Kreulen C.M., Bronkhorst E.M., Bartlett D.W., Creugers N.H., 2009, Prevalence of tooth wear in adults, Int. J. Prosthodont., 22(1): 35-42

Vodovar N., Vallenet D., Cruveiller S., Rouy Z., Barbe V., Acosta C., Cattolico L., Jubin C., Lajus A., Segurens B., Vacherie B., Wincker P., Weissenbach J., Lemaitre B., Médigue C., and Boccard F., 2006, Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium *Pseudomonas entomophila*, Nature Biotechnology, 24(6): 673-679

Woese C.R., and Fox G.E., 1977, Phylogenetic structure of the prokaryotic domain: The primary kingdoms, Proceedings of the National Academy of Sciences of the United States of America, 74(11): 5088-5090