计算分子生物学(网络版), 2012 年, 第 1 卷, 第 2 篇, 第 7-15 页
Jisuan Fenzi Shengwuxue (Online), 2012, Vol.1, No.2, 7-15
http://cmb.5th.sophiapublisher.com

数据分析

An Analysis

基于图像配准分析物种进化关系的新方法

严翠婷[1], 黄庆生[2], 章芬[1], 方颖[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: yfang@scut.edu.cn; 作者

计算分子生物学, 2012 年, 第 1 卷, 第 2 篇 doi: 10.5376/cmb.cn.2012.01.0002

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 09 日

建议最佳引用格式：

引用格式(中文)：

严翠婷等, 2012, 基于图像配准分析物种进化关系的新方法, 计算分子生物学(online) Vol.1 No.2 pp.7-15 (doi: 10.5376/cmb.cn.2012.01.0002)

引用格式(英文)：

Yan et al., 2012, A Novel Method for Evolution Analysis based on Image Registration, Jisuan Fenzi Shengwuxue (online) (Computational Molecular Biology) Vol.1 No.2 pp.7-15 (doi: 10.5376/cmb.cn.2012.01.0002)

摘 要 图像配准是图像处理的一个重要技术，可用于分析两幅图像之间的相似度。本文提出了一种基于图像配准分析物种进化关系的新方法：首先利用一阶马尔可夫链方法计算不同基因组序列的寡聚核苷酸转移概率矩阵；然后将转移概率矩阵转换为彩色图像矩阵，并绘制物种两两之间彩色图像矩阵的联合直方图；最后分析联合直方图点集的分布情况，引入直方图点集的散度公式，将其作为相似性测度的标准，从而鉴定物种亲缘关系的远近。100 种细菌全基因组的计算结果表明，相较于单基因法或基于基因组寡聚核苷酸频率组分差异信息的方法，本文提出的新方法具有更高的准确度和分辨力，它不仅能够很好地分辨科以下的分类单元，对科以上的分类单元同样具有较好的区分效果。该方法有望发展成为物种鉴定及系统发育推断的有效手段。

关键词 图像配准; 寡聚核苷酸转移概率矩阵; 联合直方图散度; 亲缘关系

A Novel Method for Evolution Analysis based on Image Registration

Yan Cuiting[1], Huang Qingsheng[2], Zhang Fen[1], Fang Ying[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: yfang@scut.edu.cn; Authors

Abstract Image registration is an important technique in image processing, which could be used to analyze the similarity between two images. Here, a novel method based on image is proposed to infer the evolutionary relatedness of various microbial organisms. Firstly, the oligonucleotide transition probability matrices of microbial genomes were calculated by applying 1st order Markov Chain Method. Secondly, each transition probability matrix was converted into a color image, and then combined with each other to depict as a joint histogram. Finally, the point set distribution of joint histogram was analyzed, and divergence formula was brought in and used as the similarity reference standard to determine the evolutionary relatedness of organisms. According to the study results of 100 bacterial genomes, our results suggest that this new method is more accurate and discriminable than the methods based on single gene or genomic oligonucleotide-based frequency component difference information methods. It not only can distinguish microbial organism under the family of taxonomy, but also has a better capability to distinguish microbial organism beyond the family of taxonomy. This method is expected to develop into effective device and apply to species identification and phylogeny inferring.

Keywords Image registration; Oligonucleotide transition probability matrix; Joint histogram divergence; Phylogenetic relationship

分析物种之间进化关系的传统方法是对同源基因进行多重序列比对(Wu and Eisen, 2008)，但是，这种基于单个或多个基因的分析方法存在着局限性，其可靠性依赖于所选择的基因是否能够真实反映物种的进化历史。并且，该方法的分辨力有限，它只对种以上的分类单元具有较高的分辨力。据 Mahoko 等人的研究发现，在只利用 16S rRNA 基因构建物种的系统发育树时，种内菌株虽然能够正确聚类，但是彼此之间的系统发育关系却无法确定(Takahashi et al., 2009)。另一种方法则是基于物种的

全基因组进行系统发育分析，它通常是比较物种基因组在 GC 含量或者寡聚核苷酸频率组分上的差异，进而分析物种之间亲缘关系的远近(Bohlin et al., 2008a; 2008b)。然而，这种方法只考虑了基因组中寡聚核苷酸的含量或相对丰度，却忽略了寡聚核苷酸的组成方式以及相邻寡聚核苷酸间的影响，因此也是片面的。而从适用性来看，熊远妍等(2008)人利用不同基因组中寡聚核苷酸频率组分差异的信息构建系统树，结果显示，该方法只有在分析科以下的分类单元时才能够得到比较合理的结果，而对科以上分类单元的分析结果则不理想。

鉴于此，我们提出了一个基于图像配准技术分析物种进化关系的新方法。为了充分考虑相邻寡聚核苷酸间的影响，在分析物种基因组时引入了马尔可夫链方法。我们假设基因组序列的延续是一个具有马尔科夫性质的离散时间随机过程，该过程中，序列中每一个寡聚核苷酸可以采取任何一种组合方式转移到下一个相邻的寡聚核苷酸，而这一步转移与之前的转移路径是无关的，其中与组合方式改变相关的概率叫做转移概率(Phillips et al., 1987)。利用一阶马尔可夫链方法分析基因组，可得到物种的寡聚核苷酸转移概率矩阵，该矩阵包含了基因组的全部信息，可用于推断物种的进化关系。接着在比较转移概率矩阵间的差异时，引入了图像配准技术(Pass and Zabih, 1999)，通过绘制转移概率矩阵间的联合直方图，并采用联合直方图散度分析其点集的分布情况(梅跃松等, 2007)。相较于前面提到的传统方法，新方法的适用范围更广，准确度和分辨力更高，它不仅适用于种内亲缘关系十分接近的物种间的鉴定，还可区分目以上亲缘关系较远的物种。

1 结果分析

1.1 联合直方图散度可真实反映物种的进化距离

对附录 1 中的 100 个物种全基因组进行分析，计算它们两两之间的联合直方图散度(HD)，结果发现，对于大部分的物种，分类单元的级别越高，即物种之间的进化距离越大，物种的联合直方图散度也越大。例如，*Streptococcus pneumoniae* TIGR4 (简写为 *S. pneu*_TI)与 *Streptococcus pneumoniae* D39 为同种的两个不同菌株，它们的 HD 为 49.027 3；而 *S. pneu*_TI 与 *Streptococcus gordonii* str. *Challis* substr. CH1 为同属不同种的两菌株，它们的 HD 则为 375.778 9；同样的，*S. pneu*_TI 与 *Lactococcus lactis* subsp. *Cremoris* MG1363、*Lactococcus salivarius* UCC118、*Staphylococcus epidermidis* ATCC RP62A、*Clostridium botulinum F* str. Langeland 及 *Aster yellows withches'-broom phytoplasma* AYWB 分别为同科不同属，同目不同科，同纲不同目，同门不同纲及同界不同门的两菌株，它们的 HD 分别为 1 488.71、2 015.58、2 339.58、3 054.30 及 4 088.10，呈现出一个递增的趋势。为了统计 HD 随着进化距离变化的总体趋势，按照“同种不同菌株”、“同属不同种”、“同科不同属”、“同目不同科”、“同纲不同目”、“同门不同纲”以及“同界不同门”的分类单元将所有物种两两之间的 HD 值分为 7 个小组，再计算各小组 HD 的平均值。各分组间进行两两 t 检验，组间显著性差异具有统计学意义($p<0.01$)。其联合直方图散度按生物分类级别的分布呈现单调递增的趋势：即随着分类级别的升高，HD 的平均值呈现出递增的趋势(图 1)。我们的计算结果说明联合直方图散度在鉴定物种进化关系方面是一个可靠的指标，它能够真实地反映物种的进化距离。

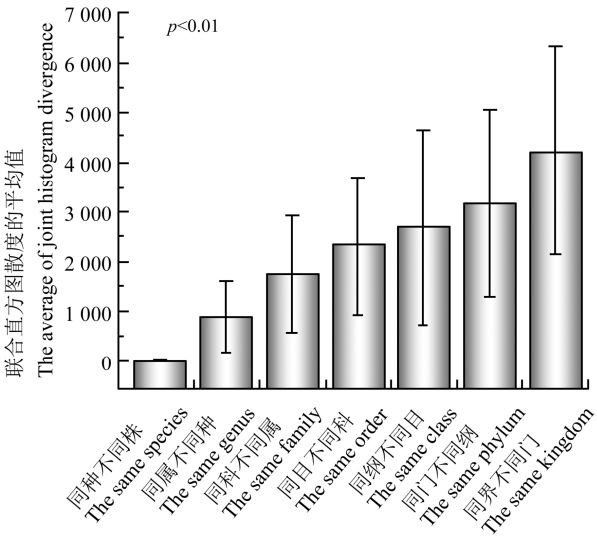


图 1 联合直方图散度平均值随分类单元级别的变化趋势
注: 横坐标表示不同级别的分类单元, 纵坐标表示各分组联合直方图散度的平均值; 各分组间通过 t 检验方法进行差异显著性检验($p<0.01$)
Figure 1 The distribution of the average of joint histogram divergence according to the rank of taxon
Note: The horizontal axis and the longitudinal axis represent the rank of taxon and the average of joint histogram divergence respectively; The differences between each taxon groups are statistically significant by t-test ($p<0.01$)

1.2 基于图像配准分析物种进化关系的新方法具有更高的分辨力

由于乳酸乳球菌与其他物种的种属关系比较清晰明了，在此以乳酸乳球菌为例对新方法的分辨力进行评估。例如 *Lactococcus lactis* subsp. *cremoris*

MG1363、*Lactococcus lactis* subsp. *cremoris* SK11 以及 *Lactococcus lactis* subsp. *lactis* Il1403 (分别简写为 *L. lac*_MG, *L. lac*_SK 与 *L. lac*_I1)为乳酸乳球菌的 3 个不同菌株，其中 *Lla* MG 与 *Lla* SK 属于同一个亚种 *Lactococcus lactis* subsp. *Cremoris*。一方面用欧几里德距离公式计算三者之间的距离，结果显示，同亚种的 *L. lac*_MG 与 *L. lac*_SK 的欧几里德距离为 0.043 8；而不同亚种的 *L. lac*_MG 与 *L. lac*_I1 的欧几里德距离为 0.062 4 (表 1)，即同亚种与不同亚种这两个分类单元之间的差异仅为 0.018 6。另一方面利用联合直方图散度比较上述三个菌株间的差异，结果发现，同亚种与不同亚种这两个分类单元之间的差异高达 51.691 4 (表 2)。该结果说明了联合直方图散度比欧几里德距离更能清晰分辨种内亲缘关系十分接近的物种。

另外，将 *L. lac*_MG 与 *Pediococcus pentosaceus* ATCC 25745、*Staphylococcus aureus* subsp. *aureus* USA300_FPR3757 及 *Alkaliphilus metalliredigens* QYMF (分别简写为 *P. pen*_AT, *S. aur*_U 及 *A. met*_QY) 进行比较，结果显示，*L. lac*_MG 与同目不同科的 *P. pen*_AT，同纲不同目的 *S. aur*_US 及同门不同纲的 *A.met*_Qy 间的欧几里德距离分别为 0.320 9、0.391 8 及 0.382 3，它们之间差异的平均值仅为 0.047 3；而 *L. lac*_MG 与 *P. pen*_AT，*S. aur*_US 及 *A. met*_Qy 间的联合直方图散度分别为 1 090.625 1，1 291.841 0 及 2 305.814 0，它们之间差异的平均值为 810.125 9，即联合直方图散度能够清晰分辨目以上亲缘关系比较远的物种。对金黄色葡萄球菌金黄亚种(*Staphylo coccus aureus* subsp. *aureus*)的 6 个菌株以及酿脓链球菌种(*Streptococcus pyogenes*)的 6 个菌株进行分析均能得到上述结论(表 3; 表 4)。综上所述，联合直方图散度比欧几里德距离在鉴定物种进化关系方面具有更高的分辨力。

1.3 联合直方图散度可作为单基因系统发育树构建过程的补充

以葡萄球菌属中的 5 个不同菌株为例，其中有 4 个菌株均属于金黄色葡萄球菌，另一个属于腐生性葡萄球菌。基于 16S rRNA 的核苷酸序列构建它们的系统发育树，结果显示，同为金黄色葡萄球菌的 4 个菌株，虽然都能够聚类在一起，但是它们之间的系统发育关系却不明确(图 2A)，也就是说单基因树无法鉴定出种内菌株的进化关系。而基于物种全基因组三核苷酸转移概率矩阵的联合直方图散度，不仅能够将同种的 4 个菌株聚类在一起，还能够很好地分辨它们之间的系统发育关系(图 2B)。因此，当单基因建树法无法分辨亲缘关系十分接近的物种时，可引入联合直方图散度这一参数，对其进行鉴定分析。换言之，联合直方图散度可作为单基因系统发育树构建过程的补充。在鉴定酿脓链球菌同种内的 6 个菌株时，同样能够得到上述结果(图 3)。

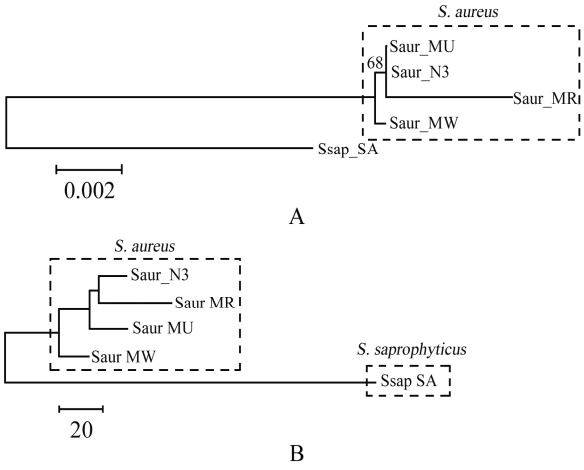


图 2 葡萄球菌属内 5 个菌株的系统发育树
注: A: 基于 16S rRNA 核苷酸序列构建的系统发育树; B: 基于基因组三核苷酸转移概率矩阵的联合直方图散度构建的系统发育树

Figure 2 The phylogeny tree of 5 strains of *Staphylococcus genus*
Note: A: The phylogeny tree based on 16S rRNA nucleotide sequence constructed by neighbor-joint method; B: The phylogeny tree based on joint histogram divergence of genomic trinucleotide transition probability matrix

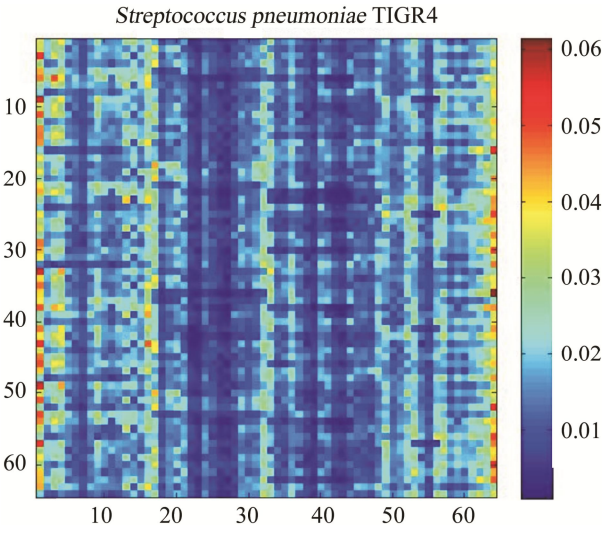


图 3 *Streptococcus pneumoniae* TIGR4 的色彩矩阵图
注: 横纵坐标分别表示转移前后的三核苷酸模式

Figure 3 The color matrix of *Streptococcus pneumonia* TIGR4
Note: The horizontal axis and the longitudinal axis represent the kinds of combination of trinucleotide respectively

表 1 细菌基因组三核苷酸转移概率矩阵间的欧几里德距离
Table 1 The Euclidian distance of trinucleotide transition probability matrices of bacterial genomes

	*L. lac*_MG	*L. lac*_SK	L. *lac*_I1	*S. pyo*_M1	*P. pen*_AT	*S. aur*_US	*A. met*_QY
*L. lac*_MG	0						
*L. lac*_SK	0.044	0					
*L. lac*_I1	0.062	0.073	0				
*S. pyo*_M1	0.291	0.295	0.308	0			
*P. pen*_AT	0.321	0.324	0.324	0.305	0		
*S. aur*_US	0.392	0.396	0.390	0.392	0.334	0	
*A. met*_QY	0.382	0.388	0.397	0.320	0.349	0.372	0

表 2 细菌基因组三核苷酸转移概率矩阵间的联合直方图散度
Table 2 The joint histogram divergence of trinucleotide transition probability matrices of bacterial genomes

	*L. lac*_MG	*L. lac*_SK	*L. lac*_I1	*S. pyo*_M1	*P. pen*_AT	*S. aur*_US	*A. met*_QY
*L. lac*_MG	0						
*L. lac*_SK	35.674	0					
*L. lac*_I1	87.366	89.496	0				
*S. pyo*_M1	1 020.933	1 043.886	1 202.584	0			
*P. pen*_AT	1 090.625	1 111.164	1 193.150	1 230.305	0		
*S. aur*_US	1 291.842	1 311.779	1 324.448	1 475.062	1 062.187	0	
*A. met*_QY	2 305.814	2 344.976	2 570.574	1 586.869	1 916.826	1 768.208	0

表 3 物种基因组三核苷酸转移概率矩阵间的欧几里德距离
Table 3 The Euclidian distance of trinucleotide transition probability matrices of bacterial genomes.

	*S. aur*_JH1	*S. epi*_AT	*S. pyo*_M1	*S. pyo*_M9	*S. aga*_26	*L. lac*_MG	*P. pen*_AT	*A. ore*_Oh
*S. aur*_JH1	0							
*S. epi*_AT	0.122	0						
*S. pyo*_M1	0.386	0.382	0					
*S. pyo*_M9	0.387	0.383	0.020	0				
*S. aga*_26	0.288	0.268	0.173	0.174	0			
*L. lac*_MG	0.389	0.379	0.291	0.292	0.267	0		
*P. pen*_AT	0.328	0.334	0.305	0.305	0.270	0.321	0	
*A. ore*_Oh	0.405	0.384	0.399	0.399	0.344	0.431	0.394	0

表 4 物种基因组三核苷酸转移概率矩阵间的联合直方图散度
Table 4 The joint histogram divergence of trinucleotide transition probability matrices of bacterial genomes.

	*S.aur*_JH1	*S.epi*_AT	*S.pyo*_M1	*S.pyo*_M9	*S.aga*_26	*L.lac*_MG	*P.pen*_AT	*A.ore*_Oh
*S. aur*_JH1	0							
*S. epi*_AT	476.659	0						
*S. pyo*_M1	1 460.841	1 771.400	0					
*S. pyo*_M9	1 465.781	1 767.048	26.995	0				
*S. aga*_26	1 062.527	932.606	549.594	541.027	0			
*L. lac*_MG	1 320.710	1 934.386	1 020.933	1 037.187	1 327.596	0		
*P. pen*_AT	1 043.516	1 377.373	1 230.305	1 231.175	1 094.646	1 090.625	0	
*A. ore*_Oh	1 839.600	1 758.309	1 938.365	1 934.658	1 516.067	2 265.351	2 028.414	0

2 讨论

通过寡聚核苷酸转移概率矩阵分析物种间的进化关系，利用了全基因组的信息，能够真实地反映物种的进化历史。本文提出的新方法，基于图像配准技术，利用联合直方图比较物种之间的差异，并引入联合直方图散度这一参数度量此差异，所得结果具有更高的准确度及分辨力。无论是种以内的近缘物种，还是目以上的远缘物种，该新方法都能够有效地衡量它们之间的进化距离，这是因为联合

直方图散度是基于对两个彩色图像矩阵间的灰度信息进行统计获得的，相较于单纯地计算寡聚核苷酸转移概率间的差值，新方法对物种基因组间的差异更为敏感，能更清晰地辨别不同物种间的进化距离。所以，这种基于图像配准分析物种进化关系的新方法，潜在有更广的适用性和更高的分辨力。

正如我们前面计算结果所证实的，对于大多数物种而言，联合直方图散度与物种间进化距离存在正相关关系：物种间的亲缘关系越远，它们之间的联合直方图散度就越大。然而，仍然存在某些特例，即个别亲缘关系较近的物种，其联合直方图散度反而大于亲缘关系较远的物种。究其原因，可能是微生物中基因水平转移的普遍存在，导致远缘物种间基因组的某些区段具有很高的相似性(Gogarten and Townsend, 2005)。另一个可能的原因是趋同进化现象的出现，导致不同的生物，甚至在进化上相距甚远的生物在基因组水平上产生了相似的变化(Amoutzias et al., 2004)。

基于图像配准分析物种进化关系的新方法不仅能够很好地分辨科以上的分类单元，与单基因建树法比较，对科以下的分类单元具有更好的区分效果。尽管目前这种用于物种进化分析的新方法还不够完善，对某些物种可能失效，但鉴于其对物种间差异的高度敏感性和分辨力，仍不失为物种鉴定及系统发育推断的一种有效手段和新型的辅助工具。例如可用于未知物种的鉴定(Tyagi et al., 2010)，通过与已知进化谱系的物种进行联合直方图分析，计算它们的联合直方图散度，从而确定该未知物种所属的分类单元。下一步的研究计划是完善并合理利用这种方法，使之真正发展为一个行之有效的物种鉴定的新手段。

3 材料与方法

3.1 基因组数据

本论文中，用于计算联合直方图散度的 100 种原核生物全基因组序列均下载自 NCBI (http://www.ncbi.nlm.nih.gov/sites/genome/)。这些物种的名称，NCBI 登录号，分类单元 ID 号以及进化谱系等信息见附录 1。进化谱系的分类单元从门到种，并对各物种的进化谱系按照一定的规律进行简写。例如，这 100 个物种分别属于硬壁菌门(*Firmicutes*)、软壁菌门(*Tenericutes*)和变形菌门(*Proteobacteria*)，可分别简写为 F.、T.和 P.；而硬壁菌门下又包含了杆菌纲(*Bacilli*)和梭菌纲(*Clostridia*)，又可分别简写为 F.1 和 F.2，依此类推。因此，物种“*Clostridium beijerinckii* NCIMB 8052”的进化谱系为 *Firmicutes* (Phylum1)-*Clostridia* (Class2)-*Clostridiales* (Order1)-*Clostridiaceae* (Family1)-*Clostridium* (Genus2)-*Clostridium beijerinckii* (Species2)，可简写为 F.2.1.1.2.2 (Sun et al., 2010; Qi et al., 2004)。

另外，在评估新方法中涉及到的 11 种原核生物 16S rRNA 核苷酸序列均下载自 NCBI (http://www.ncbi.nlm.nih.gov/sites/gene/)。它们的 Gene ID、Taxa ID 以及进化谱系等信息见表 5。

3.2 利用马尔可夫链方法计算基因组寡聚核苷酸转移概率矩阵

对于每一个基因组，分别计算长度为 n 的各种寡聚核苷酸转移到下一个相邻的长度为 n 的寡聚核苷酸的频率。具体算法是分别以 n 或 2n bp 大小的滑动窗口，每次移动 1 bp 的方法统计每一种长度为 n 或 2n 的寡聚核苷酸出现的频数，然后根据公式(1)计算得到它们的转移概率。由于细菌基因组的顺义链与反义链均可编码蛋白质，为了全面统计物种基因组包含的信息，我们计算了两条链的寡聚核苷酸转移概率矩阵，并对二者进行加和。最终得到了一个 $4^n \times 4^n$ 的转移概率矩阵。

寡聚核苷酸转移概率的计算公式：

$$p(\omega_{n+1}\cdots\omega_{2n}\,|\,\omega_1\cdots\omega_n)=\frac{f(\omega_1\cdots\omega_n\omega_{n+1}\cdots\omega_{2n})}{f(\omega_1\cdots\omega_n)} \quad (1)$$

其中，$f(\omega_1\cdots\omega_n)$ 为寡聚核苷酸 $\omega_1\cdots\omega_n$ 出现的频率；$f(\omega_1\cdots\omega_n\omega_{n+1}\cdots\omega_{2n})$ 为寡聚核苷酸 $\omega_1\cdots\omega_n\omega_{n+1}\cdots\omega_{2n}$ 出现的频率。$p(\omega_{n+1}\cdots\omega_{2n}\,|\,\omega_1\cdots\omega_n)$ 表示在 $\omega_1\cdots\omega_n$ 存在的情况下，$\omega_{n+1}\cdots\omega_{2n}$ 出现的概率，即 $\omega_1\cdots\omega_n$ 转移到 $\omega_{n+1}\cdots\omega_{2n}$ 的概率。

根据遗传信息传递过程的规律可知，基因组的转录、翻译都涉及到密码子的配对，而密码子是由三个单核苷酸组成的，每个密码子代表了一个氨基酸或者终止信号。为了能够将分析结果与遗传信息的传递过程结合起来，我们计算了每个物种基因组的三核苷酸转移概率矩阵，最终可分别得到一个 64×64 的转移概率矩阵。

3.3 通过图像配准技术比较各基因组三核苷酸转移概率矩阵间的差异

本文所采取的是基于灰度信息的图像配准方法。首先，将各物种基因组的三核苷酸转移概率矩阵转变为彩色图像矩阵。彩图矩阵中的每一个小方格对应着转移概率矩阵中相应位置的元素，根据元

表 5 各物种 16S rRNA 基因的相关信息 Gene ID, Taxa ID 以及进化谱系
Table 5 List of 16S rRNA related gene of organisms and evolutionary lineages used in this study

物种 Organism	基因 ID Gene ID	分类 ID Taxa ID	进化谱系 NCBI lineage	
			属名 Genus	种名 Species
Staphylococcus aureus subsp. *aureus* MRSA252 (*S. aur*_MR)	2861295	282458	Staphylococcus	*Staphylococcus aureus*
Staphylococcus aureus subsp. *aureus* Mu50 (*S. aur*_Mu)	1122189	158878	Staphylococcus	*Staphylococcus aureus*
Staphylococcus aureus subsp. *aureus* MW2 (*S. aur*_MW)	1004101	196620	Staphylococcus	*Staphylococcus aureus*
Staphylococcus aureus subsp. *aureus* N315 (*S. aur*_N3)	1124883	158879	Staphylococcus	*Staphylococcus aureus*
Staphylococcus saprophyticus subsp. *Saprophyticus* ATCC 15305 (*S. sap*_SA)	4036334	342451	Staphylococcus	*Staphylococcus saprophyticus*
Streptococcus pyogenes M1 GAS (*S. pyo*_M1)	2827784	160490	Streptococcus	*Streptococcus pyogenes*
Streptococcus pyogenes MGAS2096 (*S. pyo*_MG2)	4064804	370553	Streptococcus	*Streptococcus pyogenes*
Streptococcus pyogenes MGAS315 (*S. pyo*_MG3)	3284470	198466	Streptococcus	*Streptococcus pyogenes*
Streptococcus pyogenes MGAS5005 (*S. pyo*_MG5)	3571582	293653	Streptococcus	*Streptococcus pyogenes*
Streptococcus pyogenes MGAS9429 (*S. pyo*_MG9)	4061061	370551	Streptococcus	*Streptococcus pyogenes*
Streptococcus pyogenes SSI-1 (*S. pyo*_SSI)	1065158	193567	Streptococcus	*Streptococcus pyogenes*

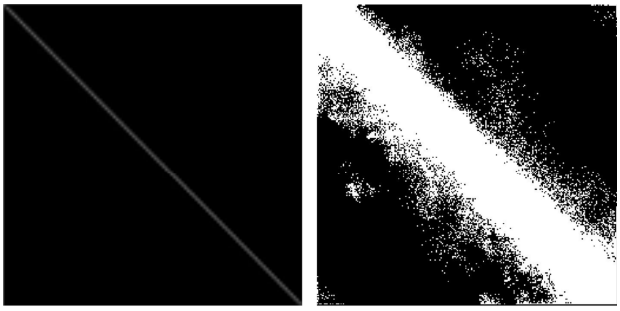
素值大小的不同，其对应小方格的颜色也不同(图 4)。然后，通过联合直方图分析彩色图像矩阵间的差异。联合直方图在使用颜色信息的同时，还引入了两幅图像像素的位置信息，它实际上统计了两幅图像对应像素的不同灰度组合出现的频数。具体算法是先定义一个 M×N 的矩阵 HIST [M, N]，其中 M 和 N 分别为图像 A 和图像 B 的灰度级数。然后，对于每一个像素 i∈A∩B，令 HIST [A(i),B(i)]+1，其中 A (i)和 B (i)分别为图像 *A* 和 *B* 在像素 i 处的灰度。这样最终统计出来的 HIST [M,N]就是图 A 和 B 的联合直方图矩阵(梅跃松等, 2007)。最后通过软件绘制出该联合直方图矩阵的图谱。因此，图 A 和图 B 越相似，它们的联合直方图谱就越趋近于 45°线；反之，则以 45°线为中心线，扩散程度越来越大(图 5)。

由此可见，联合直方图的扩散程度可作为两幅图像的相似性测度，我们将之定义为联合直方图散度。换言之，联合直方图散度可作为推断物种间进化关系的标准，物种亲缘关系越接近，基因组的相似性越高，计算得到的联合直方图散度就越小，反之则越大。

联合直方图散度的计算公式：

$$HD = \frac{\sum_{(i,j)} \left| j-i \right|^2 \cdot HIST\left[i,j\right]}{\sum_{(i,j)} HIST\left[i,j\right]} \qquad (2)$$

其中，i、j 分别代表矩阵 HIST 的行号和列号；HIST [i, j]代表矩阵 HIST 中第 i 行第 j 列的元素值。



A B

图 4 三核苷酸转移概率矩阵的联合直方图分析
注: A: 物种 *Streptococcus pneumoniae* TIGR4 自我比较的联合直方图; B: 属于同种不同菌株的 *Streptococcus pneumoniae* TIGR4 与 *Streptococcus pneumoniae* D39 的联合直方图
Figure 4 The joint histogram of trinucleotide transition probability matrix of organisms
Note: A: The joint histogram of both of *Streptococcus pneumonia* TIGR4 and itself; B: The joint histogram of *Streptococcus pneumonia* TIGR4 and *Streptococcus pneumonia* D39

公式(2)中，分子的几何意义是联合直方图上每一个点到 45°线的距离平方加权和；分母的几何意义是两幅图像重合部分像素的数量，其作用是去除相似性测度与两幅图像重叠度的关联性(梅跃松等, 2007)。彩色图像矩阵的绘制和联合直方图散度的计算均通过软件 MATLAB 完成。

3.4 基于图像配准分析物种进化关系方法的评估

依据一定长度的寡聚核苷酸组分差异计算序列间的距离，常用的是欧几里德距离公式，即直接用组分差异的平方相加求和。为了评估基于图像配准新方法的分辨力，我们将物种间三核苷酸转移概率矩阵的欧几里德距离与联合直方图散度进行比较，从而判断两者在表现序列间差异方面的优劣。

两个物种基因组寡聚核苷酸转移概率矩阵的欧几里德距离计算公式：

$$D = \sqrt{\sum_{i,j=1}^{N} (x_{i,j} - y_{i,j})^2} \qquad (3)$$

其中，N 为转移概率矩阵的总行数或总列数；$x_{i,j}$、$y_{i,j}$分别代表两个物种的转移概率矩阵中第 i 行第 j 列的元素值。

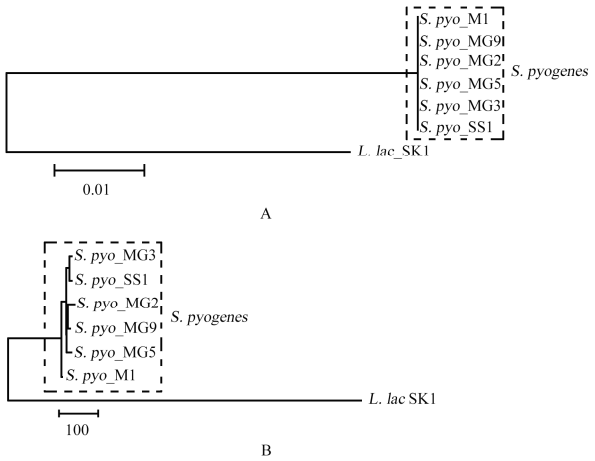


图 5 酿脓链球菌种内的 5 个菌株的系统发育树
注: A: 基于 16S rRNA 的核苷酸序列构建的系统发育树; B: 基于基因组三核苷酸转移概率矩阵的联合直方图散度构建的系统发育树
Figure 5 The phylogeny tree of 5 strains of *Streptococcus pyogenes*.
Note: A: The phylogeny tree based on 16S rRNA gene constructed by neighbor-joint method; B: The phylogeny tree based on joint histogram divergence of genomic trinucleotide transition probability matrix

另外，基于单基因的系统发育分析方法存在着局限性且分辨力有限，它只对种以上的分类单元具有较高的分辨力，而对种以下亲缘关系十分接近的物种则难以区分(Bohlin et al., 2008)。对于用单基因序列比对无法区分的物种，我们用联合直方图散度进行聚类分析，由此评估联合直方图散度在系统发育分析方面的优势。这里，我们先用 CLUSTAL X 对 11 个物种的 16S rRNA 基因进行多重序列比对，接着用邻接法构建单基因系统发育树；同时，通过构建这些物种的联合直方图散度矩阵，并利用 PHYLIP 软件进行聚类，推断它们的系统发育关系；最后，通过比较上述两种手段构建的系统发育树，对新方法进行评估。

作者贡献

严翠婷是研究的主要执行人，包括数据采集、分析和初稿写作；黄庆生编写程序，参与部分数据分析和讨论；章芬参与部分数据分析；方颖是项目负责人，指导整个实验设计、数据分析、论文写作和修改。

参考文献

Amoutzias G.D., Robertson D.L., Oliver S.G., and Bornberg-Bauer E., 2004, Convergent evolution of gene networks by single-gene duplications in higher eukaryotes, EMBO Reports, 5(3): 274-279

Bohlin J., Skjerve E., and Ussery D.W., 2008a, Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes, BMC Genomics, 9(1): 104

Bohlin J., Skjerve E., and Ussery W., 2008, Investigations of oligonucleotide usage variance within and between prokaryotes, PLoS Computational Biology, 4(4): 1-9

Gogarten J.P., and Townsend J.P., 2005, Horizontal gene transfer, genome innovation and evolution, Nature Reviews Microbiology, 3(9): 679-687

Mei Y.S., Yang S.X., and Mo B., 2007, Automatic image registration algorithm based on a novel similarity measurement, Yiqi Yibiao Xuebao (Chinese Journal of Scientific Instrument), 28(4): 336-339 (梅跃松, 杨树兴, 莫波, 2007, 一种基于新的相似性测度的自动图像配准算法, 仪器仪表学报, 28(4): 336-339)

Pass G., and Zabih R., 1999, Comparing images using joint histograms, Multimedia Systems, 7(3): 234-240

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

Qi J., Luo H., and Hao B.L., 2004, CVTree: A phylogenetic tree reconstruction tool based on whole genomes, Nucleic Acids Research, 32: W45-W47

Sun J.D., Xu Z., and Hao B.L., 2010, Whole-genome based archaea phylogeny and taxonomy: A composition vector approach, Chinese Science Bulletin, 55(22): 2323-2328

Takahashi M., Kryukov K., and Naruya S., 2009, Estimation of bacterial species phylogeny through oligonucleotide frequency distances, Genomics, 93(6): 525-533

Tyagi A., Bag S.K., Shukla V., Roy S., and Tuli R., 2010, Oligonucleotide frequencies of barcoding loci can discriminate sp ecies across kingdoms, Plos One, 5(8): 1-9

Wu M., and Eisen J.A., 2008, A simple, fast, and accurate method of phylogenomic inference, Genome Biology, 9: R151

Xiong Y.Y., Wang J.P., Lan Y.J., Wen M., and Zhang S.H., 2008, Evolutionary information of the diversity of oligonucleotide frequency of genomes, Zhongshan Daxue Xuebao (Zirankexue Ban) (Acta Scientiarum Naturalium Universitatis Sunyatsen), (2): 84-88 (熊远妍, 王军鹏, 蓝一杰, 文明, 张尚宏, 2008, 基因组寡聚核苷酸频率组分差异的进化信息, 中山大学学报: 自然科学版, (2): 84-88

附录

附录 1 各物种全基因组的相关信息

Supplement 1 List of genomes used in this study

物种 Organism	登录号 Accession	分类 ID TaxaID	进化谱系 NCBI Lineage
Staphylococcus aureus subsp. *aureus* JH1	NC_009632	359787	F.1.1.1.1.1
Staphylococcus aureus subsp. *aureus* JH9	NC_009487	359786	F.1.1.1.1.1
Staphylococcus aureus subsp. *aureus* Mu3	NC_009782	418127	F.1.1.1.1.1
Staphylococcus aureus subsp. *aureus* Mu50	NC_002758	158878	F.1.1.1.1.1
Staphylococcus aureus subsp. *aureus* USA300_FPR3757	NC_007793	451515	F.1.1.1.1.1
Staphylococcus aureus subsp. *aureus* USA300_TCH1516	NC_010079	451516	F.1.1.1.1.1
Staphylococcus epidermidis ATCC 12228	NC_004461	176280	F.1.1.1.1.2
Staphylococcus epidermidis RP62A	NC_002976	176279	F.1.1.1.1.2
Lactobacillus brevis ATCC 367	NC_008497	387344	F.1.2.1.1.1
Lactobacillus helveticus DPC 4571	NC_010080	405566	F.1.2.1.1.2
Lactobacillus johnsonii NCC 533	NC_005362	257314	F.1.2.1.1.3
Lactobacillus plantarum WCFS1	NC_004567	220668	F.1.2.1.1.4
Lactobacillus salivarius UCC118	NC_007929	362948	F.1.2.1.1.5
Pediococcus pentosaceus ATCC 25745	NC_008525	278197	F.1.2.1.2.1
Leuconostoc mesenteroides subsp. *mesenteroides* ATCC 8293	NC_008531	203120	F.1.2.2.1.1
Oenococcus oeni PSU−1	NC_008528	203123	F.1.2.2.2.1
Lactococcus lactis subsp. *cremoris* MG1363	NC_009004	416870	F.1.2.3.1.1
Lactococcus lactis subsp. *cremoris* SK11	NC_008527	272622	F.1.2.3.1.1
Lactococcus lactis subsp. *lactis* Il1403	NC_002662	272623	F.1.2.3.1.1
Streptococcus agalactiae 2603V/R	NC_004116	208435	F.1.2.3.2.1
Streptococcus agalactiae A909	NC_007432	205921	F.1.2.3.2.1
Streptococcus agalactiae NEM316	NC_004368	211110	F.1.2.3.2.1
Streptococcus gordonii str. *Challis substr.* CH1	NC_009785	467705	F.1.2.3.2.2
Streptococcus pneumoniae D39	NC_008533	373153	F.1.2.3.2.3
Streptococcus pneumoniae R6	NC_003098	171101	F.1.2.3.2.3
Streptococcus pneumoniae TIGR4	NC_003028	170187	F.1.2.3.2.3
Streptococcus pyogenes M1 GAS	NC_002737	160490	F.1.2.3.2.4
Streptococcus pyogenes MGAS2096	NC_008023	370553	F.1.2.3.2.4
Streptococcus pyogenes MGAS315	NC_004070	198466	F.1.2.3.2.4
Streptococcus pyogenes MGAS5005	NC_007297	293653	F.1.2.3.2.4
Streptococcus pyogenes MGAS9429	NC_008021	370551	F.1.2.3.2.4
Streptococcus pyogenes SSI−1	NC_004606	193567	F.1.2.3.2.4
Streptococcus suis 05ZYH33	NC_009442	391295	F.1.2.3.2.5
Streptococcus suis 98HAH33	NC_009443	391296	F.1.2.3.2.5
Streptococcus thermophilus CNRZ1066	NC_006449	299768	F.1.2.3.2.6
Streptococcus thermophilus LMD−9	NC_008532	322159	F.1.2.3.2.6
Streptococcus thermophilus LMG 18311	NC_006448	264199	F.1.2.3.2.6
Alkaliphilus metalliredigens QYMF	NC_009633	293826	F.2.1.1.1.1
Alkaliphilus oremlandii OhILAs	NC_009922	350688	F.2.1.1.1.2
Clostridium acetobutylicum ATCC 824	NC_003030	272562	F.2.1.1.2.1
Clostridium beijerinckii NCIMB 8052	NC_009617	290402	F.2.1.1.2.2
Clostridium botulinum A str. ATCC 19397	NC_009697	441770	F.2.1.1.2.3
Clostridium botulinum A str. ATCC 3502	NC_009495	413999	F.2.1.1.2.3
Clostridium botulinum A str. Hall	NC_009698	441771	F.2.1.1.2.3
Clostridium botulinum F str. Langeland	NC_009699	441772	F.2.1.1.2.3
Clostridium difficile 630	NC_009089	272563	F.2.1.1.2.4
Clostridium kluyveri DSM 555	NC_009706	431943	F.2.1.1.2.5
Clostridium novyi NT	NC_008593	386415	F.2.1.1.2.6
Clostridium perfringens ATCC 13124	NC_008261	195103	F.2.1.1.2.7
Clostridium perfringens SM101	NC_008262	289380	F.2.1.1.2.7

续附录 1
Continuing supplement 1

物种 Organism	登录号 Accession	分类 ID TaxaID	进化谱系 NCBI Lineage
Clostridium perfringens str. 13	NC_003366	195102	F.2.1.1.2.7
Clostridium tetani E88	NC_004557	212717	F.2.1.1.2.8
Bartonella bacilliformis KC583	NC_008783	360095	P.1.1.1.1.1
Bartonella henselae str. Houston−1	NC_005956	283166	P.1.1.1.1.2
Bartonella quintana str. Toulouse	NC_005955	283165	P.1.1.1.1.3
Bartonella tribocorum CIP 105476	NC_010161	382640	P.1.1.1.1.4
Bradyrhizobium sp. ORS278	NC_009445	114615	P.1.1.2.1.1
Mesorhizobium sp. BNC1	NC_008254	266779	P.1.1.3.1.1
Mesorhizobium loti MAFF303099	NC_002678	266835	P.1.1.3.2.1
Agrobacterium tumefaciens str. C58	NC_003304	176299	P.1.1.4.1.1
Dinoroseobacter shibae DFL 12	NC_009952	398580	P.1.2.1.1.1
Ruegeria pomeroyi DSS−3	NC_003911	246200	P.1.2.1.2.1
Anaplasma phagocytophilum HZ	NC_007797	212042	P.1.3.1.1.1
Anaplasma marginale str. St. Maries	NC_004842	234826	P.1.3.1.1.2
Ehrlichia canis str. Jake	NC_007354	269484	P.1.3.1.2.1
Ehrlichia chaffeensis str. Arkansas	NC_007799	205920	P.1.3.1.2.2
Ehrlichia ruminantium str. Welgevonden	NC_005295	254945	P.1.3.1.2.3
Acidovorax avenae subsp. *citrulli* AAC00−1	NC_008752	397945	P.2.1.1.1.1
Acidovorax sp. JS42	NC_008782	232721	P.2.1.1.1.2
Rhodoferax ferrireducens T118	NC_007908	338969	P.2.1.1.2.1
Polaromonas naphthalenivorans CJ2	NC_008781	365044	P.2.1.1.3.1
Polaromonas sp. JS666	NC_007948	296591	P.2.1.1.3.2
Verminephrobacter eiseniae EF01−2	NC_008786	391735	P.2.1.1.4.1
Aeromonas hydrophila subsp. *hydrophila* ATCC 7966	NC_008570	380703	P.3.1.1.1.1
Aeromonas salmonicida subsp. *salmonicida* A449	NC_009348	382245	P.3.1.1.1.2
Actinobacillus pleuropneumoniae L20	NC_009053	416269	P.3.2.1.1.1
Haemophilus ducreyi 35000HP	NC_002940	233412	P.3.2.1.2.1
Pasteurella multocida subsp. *multocida str.* Pm70	NC_002663	272843	P.3.2.1.3.1
Desulfovibrio vulgaris str. Hildenborough	NC_002937	882	P.4.1.1.1.1
Desulfovibrio vulgaris subsp. *vulgaris* DP4	NC_008751	391774	P.4.1.1.1.1
Geobacter metallireducens GS−15	NC_007517	269799	P.4.2.1.1.1
Geobacter sulfurreducens PCA	NC_002939	243231	P.4.2.1.1.2
Anaeromyxobacter dehalogenans 2CP-C	NC_007760	290397	P.4.3.1.1.1
Anaeromyxobacter sp. Fw109−5	NC_009675	404589	P.4.3.1.1.2
Campylobacter jejuni subsp. *jejuni* 81116	NC_009839	407148	P.5.1.1.1.1
Campylobacter jejuni RM1221	NC_003912	195099	P.5.1.1.1.1
Acholeplasma laidlawii PG−8A	NC_010163	441768	T.1.1.1.1.1
Aster yellows witches'-broom phytoplasma AYWB	NC_007716	322098	T.1.1.1.2.1
Mesoplasma florum L1	NC_006055	265311	T.1.2.1.1.1
Mycoplasma capricolum subsp. *capricolum* ATCC 27343	NC_007633	340047	T.1.3.1.1.1
Mycoplasma agalactiae PG2	NC_009497	347257	T.1.3.1.1.2
Mycoplasma gallisepticum str. R(low)	NC_004829	710127	T.1.3.1.1.3
Mycoplasma genitalium G37	NC_000908	243273	T.1.3.1.1.4
Mycoplasma hyopneumoniae 7448	NC_007332	262722	T.1.3.1.1.5
Mycoplasma mobile 163K	NC_006908	267748	T.1.3.1.1.6
Mycoplasma penetrans HF−2	NC_004432	272633	T.1.3.1.1.7
Mycoplasma pneumoniae M129	NC_000912	272634	T.1.3.1.1.8
Mycoplasma pulmonis UAB CTIP	NC_002771	272635	T.1.3.1.1.9
Mycoplasma synoviae 53	NC_007294	262723	T.1.3.1.1.10
Ureaplasma parvum serovar 3 str. ATCC 700970	NC_002162	273119	T.1.3.1.2.1