中国科学: 物理学 力学 天文学 2015 年 第 45 卷 第 5 期: 050501
SCIENTIA SINICA Physica, Mechanica & Astronomica phys.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

以多肽组分特异性和 GC 含量分类细菌的有效性分析

李静珂[①], 金涛[①*], 赵鸿[②]

① 陕西师范大学物理学与信息技术学院, 西安 710119;
② 厦门大学物理系, 厦门 361005
*联系人, E-mail: jintao@snnu.edu.cn

收稿日期: 2015-02-02; 接受日期: 2015-02-28; 网络出版日期: 2015-03-20
国家自然科学基金(批准号: 11147020)和中央高校基本科研业务费专项资金(编号: GK201102028)资助项目

摘要 本文研究了细菌的蛋白质多肽组分统计特征与基因组 GC(Guanine+Cytosine)含量的相关性, 发现当多肽长度较小时多肽组分特异性与 GC 含量存在着很强的关联; 随着多肽长度增加, 上述关联发生突变, 关联迅速丧失. 这一结果表明, 基于组分特异性确定细菌亲缘关系的方法的确给出了不同于 GC 含量的信息, 从而能实现有效分类.

关键词 种系基因组学, 非序列比对, 多肽组分矢量, GC 含量

PACS: 87.10.-e, 87.10.Vg, 87.14.ef, 87.14.gk

doi: 10.1360/SSPMA2015-00054

1 引言

传统生物亲缘关系的确定主要基于表型特征(如生存环境、外部形态、解剖生理和代谢途径等), 这对微生物来说就显得无能为力[1]. 分子生物学的发展, 特别是测序技术的不断完善, 使基于核苷酸或氨基酸序列确定生物亲缘关系成为可能. 其中, 最具代表性和影响力的是沃斯(Carl Richard Woese)及其合作者利用核糖体小亚基序列(SSU rRNA)发现了古细菌是介于真核生物和细菌的独立"域"这一事实[2–4]. 然而, 这种单基因序列比对的方法因其判定结果的不稳定性而备受质疑[5]. 随着不同物种全基因组测序的相继完成, 利用全基因组的序列特征(如共享基因数目[6]或其在染色体上的排序[7]、 基因含量[8,9]及保守基因对[10]等)来确定生物亲缘关系的方法不断发展, 逐渐形成了所谓的种系基因组学[11]. 但是这些方法本质上仍然需要序列比对, 与单基因序列比对方法一样易受基因组大小[12]、横向基因迁移(Horizontal Gene Transfer)[13]、平行基因缺失(Parallel Gene Loss)[14,15]及基因进化速度[16]等因素的影响, 导致判定指标特异性差、分辨率低而不能正确区分已测序物种的亲缘关系[17,18]. 另一方面, GC 含量特征是目前确定微生物亲缘关系的重要判据之一. 简单来说, GC 含量越接近的生物, 其亲缘关系也越近. GC 含量与基因识别标志有一定的关联, 有关基因识别标志(Genomic Signature)的研究发现, 四核苷酸(Tetranu-

引用格式: 李静珂, 金涛, 赵鸿. 以多肽组分特异性和 GC 含量分类细菌的有效性分析. 中国科学: 物理学 力学 天文学, 2015, 45: 050501
Li J K, Jin T, Zhao H. Validity of peptide composition and GC-content for classifying bacteria (in Chinese). Sci Sin-Phys Mech Astron, 2015, 45: 050501, doi: 10.1360/SSPMA2015-00054

cleotide)的频率分布特征与基因组的 GC 含量之间存在强关联[19,20]. 但是, GC 含量指标的分辨率太低而只能作为其他鉴定方法的辅助.

最近, 我国学者郝柏林课题组[21]提出利用全基因组或蛋白质组中"n 聚体"(即寡核苷酸或多肽)的组分特异性来确定生物亲缘关系. 这是一种非序列比对方法, 能准确鉴定目前综合不同方法确定出的微生物亲缘关系, 有效克服了原有方法的诸多限制, 引起了广泛关注[22–28]. n 聚体组分特异性的研究显示, 仅在 n 取值适当的情况下才能得到准确的生物亲缘关系. 例如, 当 n=5 或 6 时, 利用多肽组分矢量能够重建细菌的亲缘关系树; 而当 n 取较小的数值时, 则效果不好[21].

本文研究细菌的蛋白质多肽组分特异性与相应基因组 GC 含量的关联, 理解组分特异性分析方法与 GC 含量分析方法的不同之处, 揭示前者的优越性, 并对 n 的最佳取值提供新的注解. 实际上, 类似的相关性分析广泛存在于各个研究领域[29].

2 数据与方法

我们采用的数据源于 NCBI(National Center for Biotechnological Information)数据库截止 2014 年 9 月发布的 1564 种细菌的氨基酸序列(注: 同种细菌只取一个标准菌株所对应的氨基酸序列). 首先, 我们筛选其中 17 个门 99 种细菌的氨基酸序列为待研究数据集, 其亲缘关系如图 1 所示, 相关信息见附录中的表 a1 和 a2.

按照文献[21]提出的方法, 我们对第 i 种细菌统计所有氨基酸序列中长度为 n 的多肽 $\alpha_1\alpha_2...\alpha_n$ 出现的次数 $f_i(\alpha_1\alpha_2...\alpha_n)$, 得到观察频率

$$p_i(\alpha_1\alpha_2\ldots\alpha_n)=\frac{f_i(\alpha_1\alpha_2\ldots\alpha_n)}{l-n+1}, \tag{1}$$

其中, α_i 是 20 种氨基酸中的任意一种, 因此共有 20^n 种多肽; l 为氨基酸序列长度. 扣除由(n–2)阶马尔科夫过程估计的随机变异背景得

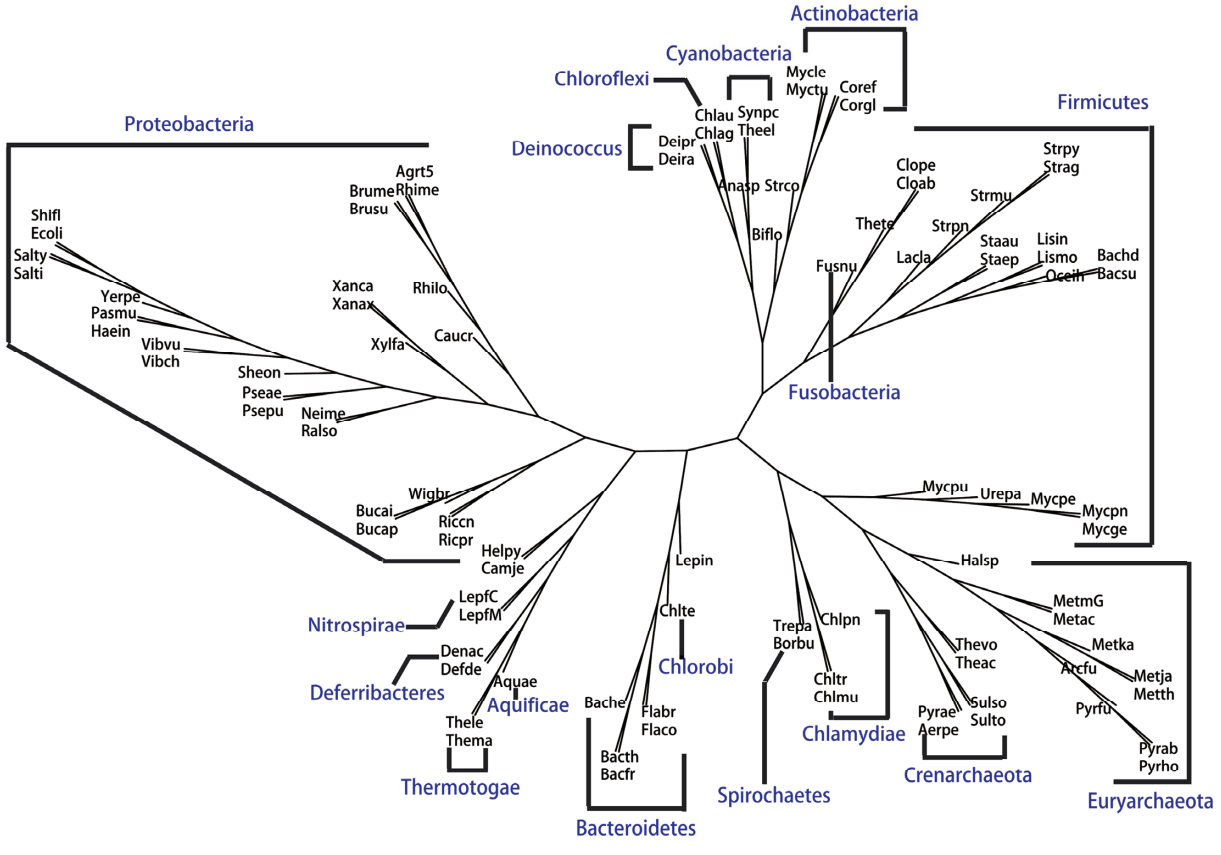

图 1 (网络版彩图)细菌亲缘关系树

Figure 1 (Color online) Phylogenetic tree of selected bacteria.

$$e_i(\alpha_1\alpha_2\ldots\alpha_n)=\frac{p_i(\alpha_1\alpha_2\ldots\alpha_n)}{b_i(\alpha_1\alpha_2\ldots\alpha_n)}-1, \quad (2)$$

其中

$$b_i(\alpha_1\alpha_2\ldots\alpha_n)=\frac{p_i(\alpha_2\alpha_3\ldots\alpha_n)\cdot p_i(\alpha_1\alpha_2\ldots\alpha_{n-1})}{p_i(\alpha_2\alpha_3\ldots\alpha_{n-1})}, \quad (3)$$

当 $b_i(\alpha_1\alpha_2...\alpha_n)=0$ 时, 我们同样令 $e_i(\alpha_1\alpha_2...\alpha_n)=0$. 这里 $p_i(\alpha_2\alpha_3...\alpha_{n-1})$是长度为 $n-2$ 多肽$\alpha_2\alpha_3...\alpha_{n-1}$ 的观察概率, 而 $p_i(\alpha_2\alpha_3...\alpha_n)$和 $p_i(\alpha_2\alpha_3...\alpha_{n-1})$是长度为 $n-1$ 的多肽$\alpha_2\alpha_3...\alpha_n$ 和$\alpha_1\alpha_2...\alpha_{n-1}$的观察概率. 对每种细菌 i, 把 20^n 个 $e_i(\alpha_1\alpha_2...\alpha_n)$按照相同的多肽组合顺序排成一列得到进化信息矢量 $\boldsymbol{E}_i$, 由此计算出细菌间亲缘关系距离

$$d_{ij}=\frac{(\boldsymbol{E}_i)^{\mathrm{T}}\cdot\boldsymbol{E}_j}{\|\boldsymbol{E}_i\|\cdot\|\boldsymbol{E}_j\|}, \quad (4)$$

d_{ij}越接近 1 说明细菌 i 和 j 的亲缘关系越近.

由文献[21]知, 上述扣除随机背景方法得到的多肽组分特异性对确定细菌亲缘关系效果最好, 并且 n 越大, $\boldsymbol{E}_i$ 的特异性越强, d_{ij} 分辨率也越高, 当 n=5 或 6 时就可以完全重建如图 1 所示的亲缘关系. 需要说明的是, 我们还计算了 n=2 时的 d_{ij}, 此时由最强马尔科夫过程来估计随机变异信息, 即 $b_i(\alpha_1\alpha_2)=p_i(\alpha_1)\cdot p_i(\alpha_2)$.

下面, 我们给出不同 n 值下多肽组分特异性(即进化信息矢量 $\boldsymbol{E}_i$)与基因组 GC 含量关联性计算的结果(图 2).

首先, 统计氨基酸序列所对应的核苷酸序列的 GC 含量 a_i, 并用

$$r_{ij}=\min\left\{\frac{a_i}{a_j},\frac{a_j}{a_i}\right\}, \quad (5)$$

即取 $\frac{a_i}{a_j}$ 和 $\frac{a_j}{a_i}$ 中较小的一个来度量细菌 i 和细菌 j 的亲缘关系. 由 GC 含量判据可知, r_{ij}越接近 1 表明两种细菌的亲缘关系越近. 即 r_{ij}和 d_{ij}具有相同含义, 都给出了细菌亲缘关系的相对距离. 这样, 我们就可以定义

$$c_i=\frac{\sum_{j=1}^{m}d_{ij}r_{ij}}{\sqrt{\sum_{j=1}^{m}d_{ij}^2}\cdot\sqrt{\sum_{j=1}^{m}r_{ij}^2}}, \quad (6)$$

来度量多肽组份特异性与 GC 含量的相关性. 显然, c_i 越接近 1, 说明矢量 $\boldsymbol{d}_i=(d_{i1}, d_{i2}, ..., d_{im})$和矢量 $\boldsymbol{r}_i=(r_{i1}, r_{i2}, ..., r_{im})$的方向越趋于一致, 表明由这两种距离预

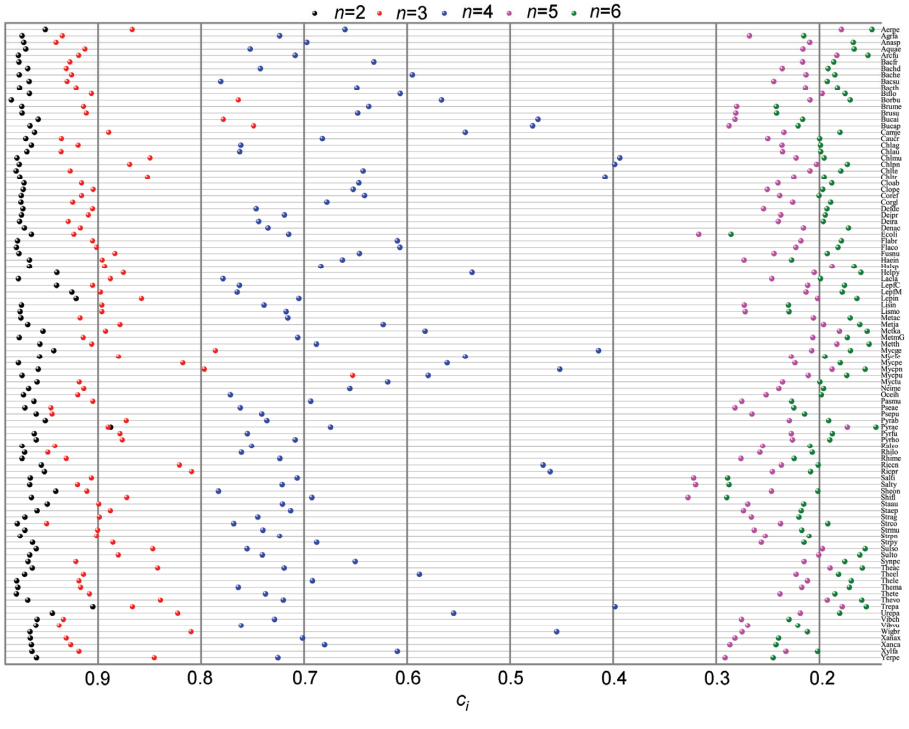

图 2 (网络版彩图)n 取不同值时 99 种细菌的 c_i

Figure 2 (Color online) The c_i of selected bacteria with different n.

测的细菌亲缘关系越趋于等价, 表明同种细菌的多肽组分特异性与其基因组 GC 含量的关联性越强.

3 结果与讨论

图 2 给出了上述 99 种细菌在 n 取不同数值时的 c_i. 可以发现, 当 n=2, 3, 4 时绝大多数的 c_i 都明显大于 0.5, 特别是当 n=2 或 3 时几乎所有的 c_i 都大于 0.8. 当 n=5 或 6 时所有的 c_i 都急剧减小到 0.3 以下.

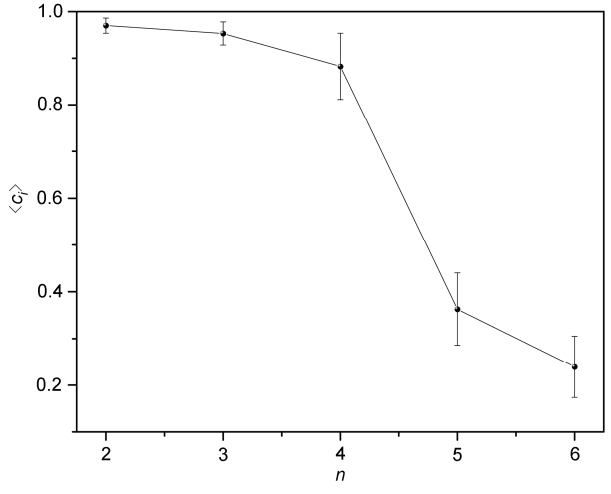

图 3 n 取不同值时的$\langle c_i \rangle$, 其中误差棒表示 c_i 的均方差

Figure 3 $\langle c_i \rangle$ for different n and error bar for standard deviation of c_i.

为了研究上述结果的普遍性, 同时为了能更清楚地显示 c_i 随 n 的变化规律, 我们计算了不同 n 值下所有 1564 种细菌的 c_i 平均值$\langle c_i \rangle$, 结果如图 3 所示. 可以看到, 在 n=4 和 5 时$\langle c_i \rangle$发生了一次突变.

以上结果表明: (1) 当多肽长度小于 4 时多肽组分特异性与 GC 含量之间具有很强的正关联. 意味着在这种情况下多肽组分统计所得到的信息等价于基因组 GC 含量的特异性信息. 由于郝柏林等人以前的研究表明在这一长度上 GC 含量和小 n 值下多肽组分特异性都能对"属"以上级别的亲缘关系做出准确且一致的判定, 表明 GC 含量和小 n 值多肽组分只能区分这个层次上的特异性. 由于和 GC 含量的等价性, 短多肽组分统计不能正确判定更精细层次上的细菌亲缘关系. 这一结果也暗示基因识别标志无法准确重构微生物亲缘关系, 因为它只等价于 GC 含量特异性. (2) 当多肽长度大于 4 时, 多肽组分特异性与 GC 含量之间的关联发生突变, 急剧降低到几乎无关联的程度. 在 n=5 或 6 的情况下, 多肽组分特异性能准确给出细菌到种株的分类树的事实表明, 这种新方法的确超越了 GC 含量分析方法和基因识别标志方法, 挖掘出了物种特异性特征. 我们还直接计算了寡核苷酸的组分特异性与相应基因组 GC 含量的关联, 得到一致的结论.

参考文献

1 Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev, 2001, 25: 39–67
2 Worse C R, Fox G E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA, 1977, 74: 5088–5090
3 Olsen G J, Woese C R. The wind of (evolutionary) change: Breathing new life into microbiology. J Bacteriol, 1994, 176: 1–6
4 Xie T, Ding D F. The third form of life-Advance in three boundary theory. Life Sci, 1997, 9: 233–236 [解涛, 丁达夫. 生命的第三界-三界学说的新发展. 生命科学, 1997, 9: 233–236]
5 Fox G E, Wisotzkey J D, Jurtshuk P. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol, 1992, 42: 166–170
6 Yang S, Doolittle R F, Bourne P E. Phylogeny determined by protein domain content. Proc Natl Acad Sci USA, 2005, 102: 373–378
7 Dandekar T, Snel B, Huynen M, et al. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci, 1998, 23: 324–328
8 Bansal A K, Meyer T E. Evolutionary analysis by whole genome comparisons. J Bacteriol, 2002, 184: 2260–2272
9 Snel B, Bork P, Huynen M A. Genome phylogeny based on gene content. Nat Genet, 1999, 21: 108–110
10 Wolf Y I, Rogozin I B, Grishin N V, et al. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol, 2001, 1: 8
11 Coenye T, Gevers D, Van de Peer Y, et al. Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev, 2005, 29: 147–167
12 Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. Genome Res, 1999, 9: 550–557
13 Philippe H, Douady C J. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol, 2003, 6: 498–505
14 Dutilh B E, Huynen M A, Bruno W J, et al. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J

Mol Evol, 2004, 58: 527–539
15 Lake J A, Rivera M C. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. Mol Biol Evol, 2004, 21: 681–690
16 Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool, 1978, 27: 401–410
17 Snel B, Huynen M A, Dutilh B E. Genome tree and the nature of genome evolution. Annu Rev Microbiol, 2005, 59: 191–209
18 Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet, 2005, 6: 361–375
19 Bohlin J. Genomic signatures in microbes-properties and applications. Sci World J, 2011, 11: 715–725
20 Bohlin J, Skjerve E. Examination of genome homogeneity in prokaryotes using genomic signatures. PLoS One, 2009, 4: e8113
21 Qi J, Wang B, Hao B. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. J Mol Evol, 2004, 58: 1–11
22 Gentles A J, Karlin S. Genome-scale compositional comparisons in Eukaryotes. Genome Res, 2001, 11: 540–546
23 Wu X, Wan X, Wu G, et al. Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbor-Joining method. Int J Bioinform Res Appl, 2006, 2: 219–248
24 Gao L, Qi J, Sun J D, et al. Prokaryote phylogeny meets taxonomy: An exhaustive comparasion of composition vector trees with systematic bacteriology (in Chinese). Sci China Ser C-Life Sci, 2007, 37: 389–401 [高雷, 戚继, 孙健东, 等. 原核生物系统发生学与分类学的一致性: 组份矢量树与原核生物分类系统的详尽比较. 中国科学 C 辑: 生命科学, 2007, 37: 389–401]
25 Sims G E, Jun S R, Wu G A, et al. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci USA, 2009, 106: 2677–2682
26 Jun S R, Sims G E, Wu G A, et al. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci USA, 2010, 107: 133–138
27 Hao B. CVTree: A whole-genome-based and alignment-free approach to microbial phylogeny. Int J Mod Phys Conf Ser, 2012, 9: 1–10
28 Li Q, Zuo G, Hao B L. Some mathematical problems inspired by the study of whole-genome-based phylogeny and taxonomy of Prokaryote. Sci Sin-Phys Mech Astron, 2014, 44: 1301–1310
29 Tang J. Cross-wavelet analysis of the radio flux of BL Lac object OJ 287. Sci Sin-Phys Mech Astron, 2014, 44: 865–871 [唐洁. BL Lac 天体 OJ 287 射电流量的交叉小波分析. 中国科学: 物理学 力学 天文学, 2014, 44: 865–871]

附录

表 a1 细菌的名字、缩写、索取号和 Berger's 代码

Table a1 Bacterium name, abbreviation, NCBI accession numbers and Berger's code

Species/strain	Abbrev	Accession No.	Bergey code
Aquifex aeolicus	Aquae	NC_000918	B1.1.1.1.1
Thermotoga lettingae TMO	Thele	NC_009828	B2.1.1.1.1
Thermotoga maritima	Thema	NC_000853	B2.1.1.1.1
Deinococcus proteolyticus MRP	Deipr	NC_015161	B4.1.1.1.1
Deinococcus radiodurans R1	Deira	NC_001263	B4.1.1.1.1
Chloroflexus aggregans DSM	Chlag	NC_011831	B6.1.1.1.1
Chloroflexus aurantiacus J	Chlau	NC_010175	B6.1.1.1.1
Leptospirillum ferriphilum ML	LepfM	NC_018649	B8.1.1.1.2
Leptospirillum ferrooxidans C2	LepfC	NC_017094	B8.1.1.1.2
Deferribacter desulfuricans SSM1	Defde	NC_013939	B9.1.1.1.1
Denitrovibrio acetiphilus DSM	Denac	NC_013943	B9.1.1.1.2
Thermosynechococcus elongatus BP	Theel	NC_004113	B10.l.1.1.13
synechocystis PCC6803	Synpc	NC_000911	B10.1.1.1.14
nostoc PCC7120	Anasp	NC_003272	B10.1.4.1.8
Chlorobium tepidum TLS	Chlte	NC_002932	B11.1.1.1.1
Rickettsia conorii	Riccn	NC_003103	B12.1.2.1.1
Rickettsia prowazekii	Ricpr	NC_000963	B12.1.2.1.1
Caulobacter crescentus CB15	Caucr	NC_002696	B12.1.5.1.1
Agrobacterium fabrum C58	Agrt5	NC_003062	B12.1.6.1.2
Sinorhizobium meliloti 1021	Rhime	NC_003047	B12.1.6.1.7
Brucella melitensis bv	Brume	NC_003317	B12.1.6.4.1

表 1(续)

Species/strain	Abbrev	Accession No.	Bergey code
Brucella suis 1330	Brusu	NC_004310	B12.1.6.4.1
Mesorhizobium loti	Rhilo	NC_002678	B12.1.6.5.6
Ralstonia solanacearum GMI1000	Ralso	NC_003295	B12.2.1.1.8
Neisseria meningitidis MC58	NeimeM	NC_003112	B12.2.4.1.1
Xanthomonas axonopodis citri 306	Xanax	NC_003919	B12.3.3.1.1
Xanthomonas campestris ATCC 33913	Xanca	NC_003902	B12.3.3.1.1
Xylella fastidiosa 9a5c	Xylfa	NC_002488	B12.3.3.1.12
Pseudomonas aeruginosa PA01	Pseae	NC_002516	B12.3.9.1.1
Pseudomonas putida KT2440	Psepu	NC_002947	B12.3.9.1.1
Shewanella oneidensis MR-1	Sheon	NC_004347	B12.3.10.1.14
Vibrio cholerae O1 biovar	Vibch	NC_002505	B12.3.11.1.1
Vibrio vulnificus CMCP6	Vibvu	NC_004459	B12.3.11.1.1
Buchnera aphidicola Sg	Bucap	NC_004061	B12.3.13.1.5
Buchnera sp. APS	Bucai	NC_002528	B12.3.13.1.5
Escherichia coli CFT073	EcoliC	NC_004431	B12.3.13.1.1
Salmonella enterica serovar Typhi CT18	Salti	NC_003198	B12.3.13.1.34
Salmonella enterica serovar Typhimurium	Salty	NC_003197	B12.3.13.1.34
Shigella flexneri 2a strain 301	Shifl	NC_004337	B12.3.13.1.37
Wigglesworthia brevipalpis	Wigbr	NC_004344	B12.3.13.1.41
Yersinia pestis strain C092	YerpeC	NC_003143	B12.3.13.1.43
Pasteurella multocida PM70	Pasmu	NC_002663	B12.3.14.1.1
Haemophilus influenzae Rd	Haein	NC_000907	B12.3.14.1.4
Campylobacter jejuni NCTC 11168 ATCC	Camje	NC_002163	B12.5.1.1.1
Helicobacter pylori 26695	Helpy	NC_000915	B12.5.1.2.1
Clostridium acetobutylicum ATCC824	Cloab	NC_003030	B13.1.1.1.1
Clostridium perfringens	Clope	NC_003366	B13.1.1.1.1
Thermoanaerobacter tengcongensis	Thete	NC_003869	B13.1.2.1.11
Mycoplasma genitalium G37	Mycge	NC_000908	B13.2.1.1.1
Mycoplasma penetrans	Mycpe	NC_004432	B13.2.1.1.1
Mycoplasma pneumoniae M129	Mycpn	NC_000912	B13.2.1.1.1
Mycoplasma pulmonis UAB CTIP	Mycpu	NC_002771	B13.2.1.1.1
Ureaplasma urealyticum	Urepa	NC_002162	B13.2.1.1.4
Oceanobacillus iheyensis	Oceih	NC_004193	B13.3.1.1.12
Bacillus halodurans	Bachd	NC_002570	B13.3.1.1.1
Bacillus subtilis 168	Bacsu	NC_000964	B13.3.1.1.1
Listeria innocua	Lisin	NC_003212	B13.3.1.4.1
Listeria monocytogenes EGD-e	Lismo	NC_003210	B13.3.1.4.1
Staphylococcus aureus Mu50	StaauM	NC_002758	B13.3.1.8.1
Staphylococcus epidermidis ATCC	Staep	NC_004461	B13.3.1.8.1
Streptococcus agalactiae 2603 V/R	StragV	NC_004116	B13.3.2.6.1
Streptococcus mutans UA159	Strmu	NC_004350	B13.3.2.6.1
Streptococcus pneumoniae R6	StrpnR	NC_003098	B13.3.2.6.1
Streptococcus pyogenes MGAS8232	Strpy8	NC_003485	B13.3.2.6.1
Lactococcus lactis sp. IL1403	Lacla	NC_002662	B13.3.2.6.2
Corynebacterium efficiens YS-314	Coref	NC_004369	B14.(1.5).(1.10).1.1
Corynebacterium glutamicum	Corgl	NC_003450	B14.(1.5).(1.10).1.1
Mycobacterium leprae TN	Mycle	NC_002677	B14.(1.5).(1.10).4.1
Mycobacterium tuberculosis CDC1551	MyctuC	NC_002755	B14.(1.5).(1.10).4.1
Streptomyces coelicolor A3	Strco	NC_003888	B14.(1.5).(1.14).1.1
Bifidobacterium longum NCC2705	Biflo	NC_004307	B14.(1.5).2.1.1
Chlamydia muridarum	Chlmu	NC_002620	B16.1.1.1.1
Chlamydia trachomatis	Chltr	NC_000117	B16.1.1.1.1

表 1(续)

Species/strain	Abbrev	Accession No.	Bergey code
Chlamydophila pneumoniae AR39	ChlpnA	NC_002179	B16.1.1.1.2
Borrelia burgdorferi B31	Borbu	NC_001318	B17.1.1.1.2
Treponema pallidum Nichols	Trepa	NC_000919	B17.1.1.1.9
Leptospira interrogans serovar lai	Lepin	NC_004342	B17.1.1.3.1
Bacteroides fragilis 638R	Bacfr6	NC_016776	B20.1.1.1.1
Bacteroides helcogenes P	Bache	NC_014933	B20.1.1.1.1
Bacteroides thetaiotaomicron VPI	Bacth	NC_004663	B20.1.1.1.1
Flavobacterium branchiophilum FL	Flabr	NC_016001	B20.2.1.1.1
Flavobacterium columnare ATCC	Flaco	NC_016510	B20.2.1.1.1
Fusobacterium nucleatum ATCC	Fusnu	NC_003454	B21.1.1.1.1

表 a2 古细菌的名字、缩写、索取号和 Berger's 代码

Table a2 Archaea name, abbreviation, NCBI accession numbers and Berger's code

Species/strain	Abbrev	Accession No.	Bergey code
Pyrobaculum aerophilum	Pyrae	NC_003364	A1.1.1.1.3
Aeropyrum pernix Kl	Aerpe	NC_000854	A1.1.3.1.3
Sulfolobus solfataricus	Sulso	NC_002754	A1.1.3.1.1
Sulfolobus tokodaii	Sulto	NC_003106	A1.1.3.1.1
Methanothermobacter thermautotrophicus	Metth	NC_000916	A2.1.1.1.4
Methanocaldococcus jannaschii	Metja	NC_000909	A2.2.1.2.1
Methanosarcina acetivorans C2A	Metac	NC_003552	A2.3.2.1.1
Methanosarcina mazei Goel	MetmG	NC_003901	A2.3.2.1.1
Halobacterium sp. NRC-1	Halsp	NC_002607	A2.4.1.1.1
Thermoplasma acidophilum	Theac	NC_002578	A2.5.1.1.1
Thermoplasma volcanium	Thevo	NC_002689	A2.5.1.1.1
Pyrococcus abyssi GE5	Pyrab	NC_000868	A2.6.1.1.3
Pyrococcus furiosus	Pyrfu	NC_003413	A2.6.1.1.3
Pyrococcus horikoshii	Pyrho	NC_000961	A2.6.1.1.3
Archaeoglobus fulgidus	Arcfu	NC_000917	A2.7.1.1.1
Methanopyrus kandleri AV19	Metka	NC_003551	A2.8.1.1.1

Validity of peptide composition and GC-content for classifying bacteria

LI JingKe[1], JIN Tao[1*] & ZHAO Hong[2]

[1] *School of Physics & Information Technology, Shaanxi Normal University, Xi'an 710119, China;*
[2] *Department of Physics, Xiamen University, Xiamen 361005, China*

In the past decades, a lot of methods have been proposed to construct Genome Tree. Among them, K-String Composition Approach which is Alignment-Free shows nonnegligible superiority. On the other hand, the species specificity of GC (Guanine+Cytosine)-content which actually is the lowest-order version of K-String Composition has been discovered for a long time, especially in bacteria. Unfortunately, its resolution is too poor to be applied to reconstruct phylogeny. Motivated by those facts, in this paper, relationship between composition vector of peptides and GC-content of corresponding DNA sequence is studied for bacteria. A strong correlation is uncovered for short peptides, and with the increase of peptide length the correlation exhibits an abrupt change, that is, tends to vanish quickly. These results indicate that the composition vector of longer peptide do contains more precise information of species specificity than that of GC-content, and therefore can effectively measure the genetic relationship of bacteria. Short peptides are obviously not competent.

phylogenomics, alignment-free phylogeny, composition vector, GC-content

PACS: 87.10.-e, 87.10.Vg, 87.14.ef, 87.14.gk

doi: 10.1360/SSPMA2015-00054