中国科学 C 辑: 生命科学
2007 年 第 37 卷 第 4 期: 389~401
http://www.scichina.com

《中国科学》杂志社
SCIENCE IN CHINA PRESS

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

高 雷[①②] 戚 继[①a)] 孙健冬[②③] 郝柏林[①②④*]

(① 中国科学院理论物理研究所, 北京 100080; ②复旦大学理论生命科学中心, 上海 200433; ③ 中国科学院北京基因组研究所, 北京 101300; ④ Santa Fe Institute, Santa Fe, NM87501, USA)

摘要 截至 2006 年 12 月 31 日, NCBI 共有 432 个原核生物的全基因组可供下载. 基于这些数据, 我们用组份矢量方法构建了原核生物的进化树. 最新的《伯杰系统细菌学手册》的在线大纲体现了细菌学家的分类系统. 我们对两者从各个分类单元、各个分支进行了详尽的比较. 组份矢量方法所得到的亲缘树和伯杰分类系统在整体结构和绝大多数的细微分支上都相当一致. 同时, 两者的多数不同之处也已经在一定程度上为生物学家所知, 从而为原核生物分类系统的修正提供了一定的启示. 本文重点阐述两者之间分歧之处的生物学含义, 而不再对居主导地位的相同之处做详细叙述.

关键词 组分矢量方法 伯杰分类 原核生物亲缘树 原核生物分类 系统发生 CVTree 16S rRNA

原核生物分类中存在的问题由来已久. 从林奈开始, 物种的分类一直主要基于它们的形态特征. 人们更加强调对相似物种的分组, 而忽略了寻求它们之间的进化关系. 直到Zuckerkandl和Pauling[1]提出使用保守的蛋白质序列作为进化"时钟", 才使得系统发生学和分类学深入到分子水平. 这种方法对各种真核生物取得了或多或少的成功. 几年之前, 为了重建所有现存物种之间的进化关系, 美国国家科学基金启动了"组装生命之树"(assembling the tree of life, AToL)计划 [2]. 然而, 具有讽刺意味的是, 不久之前发表在*Science*杂志上的一篇构建生命之树的文章 [3], 却缺少了占地球生物绝大多数的原核生物支.

原核生物的系统发生学研究进展缓慢并非偶然. 因对于原核生物可获得的形态特征太少, 难以使用传统方法重建其亲缘关系. 直到Woese等人 [4]提出使用小亚基核糖体RNA(原核生物的 16S rRNA)前, 一直没有适用于原核生物的分子钟. 16S rRNA亲缘树获得了巨大成功, 新版《伯杰系统细菌学手册》(后面简称《伯杰手册》)也在序言中申明: 他们的系统发生学工作更多地基于对小亚基核糖体RNA的序列分析, 而不是传统的表型特征 [5].

然而, 系统发生和分类并不是同义词. 恰当的分类应该在主要的大分支和细微的小分支上都与系统发生相一致. 另一方面, 一套合适的系统发生关系也应该真实反映物种的进化历史. 最初的《伯杰手册》是几代细菌学家系统分类的总结, 而该书的新版则更多地基于 16S rRNA亲缘树. 因此, 《伯杰手册》需要独立于 16S rRNA分析的系统发生研究的支持和检验. 此外, 一个必须面对的问题是单基因(如 16S rRNA)的进化能够在多大程度上反映物种的进化历

收稿日期: 2007-04-02; 接受日期: 2007-07-03
a) Center for Comparative Genomics and Bioinformatics, Penn State University, 310 Wartik Building, University Park, PA16802, USA
* 联系人, E-mail: hao@mail.itp.ac.cn

史? 事实上, 即便是某些细菌核糖体的操纵子也可能被来自其他物种的相应操纵子替代[6], 基于单个基因或少数几条基因序列的系统发生学怎么可能免于受基因的“横向转移”影响呢? 几年之前, 对于原核生物的蛋白质中是否存在系统发生的信号还存有疑问: “由蛋白质序列构建的进化树在细菌门的水平上几乎是噪声, 表明即使利用大规模数据, 或许也不可能用标准的基于序列的方法重建原核生物的系统发生树”[7].

另一方面, 自1995年以来公布出越来越多的原核生物全基因组, 导致了基于全基因组的系统发生研究热潮[8]. 然而, 多数此类方法仍然以序列联配为基础, 包含过多可调参数——在许多研究中都使用BLAST就是例证. 一些所谓“自动重建”方法[9]也需要在一定程度上手工鉴定基因. 更有甚者, 即使存在过一棵“系统发生树”, 它也早已湮没在漫长的进化史中. 因此, 对系统重建的评估不得不依赖于自恰性论据和统计检验, 如常用的自举法和刀切法. 这就导致原核生物的分子系统发生研究囿于自身, 而很少有人将所得亲缘树直接和生物学家的生命之树做详细的比较.

鉴于上述情况, 近年来我们发展了一个用全基因组数据来推断原核生物亲缘关系的新方法[10,11], 它不依赖于序列联配, 不含可选参数, 并建立了名为CVTree的服务网站[12]. 本文中, 我们同时用CVTree做为该方法及由该方法得到的进化树的简称. 更重要的是, 我们采用了新的方式来验证所得到的进化树. 我们把输入基因组数据的CVTree方法看作一种理论构造, 将它的输出直接和反映在《伯杰手册》中的“试验事实”做详尽比较. 本文就是详细比较的概要.

1 材料和方法

1.1 全基因组

从NCBI的ftp站[13]下载了截止到2006年12月31日的 432 个原核生物的全基因组数据. 这些序列的NCBI索取号都以NC打头. 我们使用其中后缀为.faa文件——即所有蛋白质产物序列的文件. 这些序列的优点是它们经过NCBI同一组人的审读, 因此更具有可比性. 此外, 我们增加了 8 个真核生物的基因组作为外类群. 《补充材料》[14]中列出了所有物种的名字、缩写和索取号. 我们排除了质粒和其他的染色体外序列.

1.2 CVTree 方法

在以前的文章中已经对CVTree方法做过比较详细的描述[10,11,12], 在此只做简短概述.

CVTree方法将每个物种用一个“组份矢量”表示. 该矢量的元素通过以下方法构成: 首先从物种所有的蛋白质序列中数出各种固定长度为K的短肽数目, 然后为了突出选择进化的作用, 减除“统计背景”对上述数目进行修正. 减除的步骤基于$(K-2)$阶的马可夫预测, 因此最小的$K = 3$. 通过计算组份矢量之间的距离, 可以得到相应物种之间的距离矩阵, 进而用Phylip[15]软件包中标准的邻接法程序生成亲缘树. 本文中重建了K从 3 到 6 的共 4 棵CVTree. 《补充材料》[14]中详细给出了这些树. 虽然短肽的长度K看起来像是参数, 但是它是CVTree方法分辨率的一种测度. 近几年以来, 我们从 72 个原核生物基因组开始到目前的 432 个基因组, 多次构建了原核生物的系统发生树. 从与生物学家的分类相一致的角度看, K=5~6 时得到的结果最好. 这一结果也和生物学家关于“六个氨基酸足以辨别一条蛋白质”[16]的观点相呼应. 关于CVTree方法的理论基础及其合理性仍然在进一步论证中, 可参看文献[17].

1.3 原核生物的分类系统

原核生物的分类系统并没有一个正式的标准. 但是, 《伯杰系统细菌学手册》[5]中的分类方式目前被微生学家广泛接受[18]. 该手册的最新版除了参考经典的原核生物表型和生化特征外, 更多的是基于16S rRNA的系统发生分析[18]. 伯杰分类系统仅仅使用可培养的典型种, 这在整个原核生物中只占了极小的一部分, 因此该分类系统是较为保守的.

为了方便表达, 在本文及其《补充材料》[14]中对《伯杰手册》或其在线大纲[19]中的谱系进行了简写, 比如B13.3.2.6.2 或$B_{13}C_3O_2F_6G_2$ 表示细菌第十三门(B XIII, Firmicutes), 第三纲(Class III, Bacilli), 第二目(Order II, Lactobacillales), 第六科(Family VI, Streptococcaceae), 第二属(Genus II, *Lactococcus*). 我们称这种简写为“伯杰编码(Bergey's code)”[10]. 必须注意, 这里的“伯杰编码”仅仅是一种方便简写, 它们可能会随着《伯杰手册》版本的不同而有所变化. 本文中的“伯杰编码”基于其在线大纲的 5.0 版[19].

NCBI在其网站上也提供了一套分类系统[20]. 虽

然NCBI特别声明其“命名和分类不具权威性”, 但是该分类系统的优点在于它的动态性和快速更新. 当NCBI的分类系统和伯杰分类系统不同但与CVTree结果相符时, 我们会参考NCBI的分类系统. 有时, 为了获得更多的信息, 我们也同时参考EBI[21]的分类列表. 在讨论中我们也会偶尔参考其他的分类系统, 比如《五界论》[22]等.

2　CVTree 和原核生物分类系统的比较

我们将CVTree的系统发生树和伯杰分类系统进行了详尽的比较: 从株和种直到纲和门, 逐个分类层次进行比较. 下一小节中是对古细菌的详细分析. 限于篇幅, 对细菌进行的类似分析将在《补充材料》[14]中给出, 本文仅仅在下文 2.3 小节中加以概述.

在将进化树的每个分支和细菌分类做比较时, 应当注意到分类的各个单元——如门、纲、目、科、属、种、株等——都是人类发明的. 只有分类单元的两个极端才更有意义, 如属中各个种的归属和所有低阶分类单元聚集成的最高的分类单元——门或纲[1)]. 因而, 本文的分析遵循两个指导原则. 在株和种的水平, 我们检验各个分类单元是否随着 K 的增加而逐渐聚集到一起, 我们称这个过程为“收敛”; 在高层分类单元上, 我们考察最高分类单元的从属成员是否形成单源支, 而不注重成员间的相互关系.

2.1　生命的三域论

由Carl Woese及合作者[23]提出的生命三域论是人类认识地球生命世界的重大进展. 是否能够清楚地区分生命的三域成为检验系统发生方法的试金石. 在表 1 中, 我们列出了CVTree方法将 440 个基因组分成三域的情况. 可以看出, 随着K的增加, “收敛”明显.

然而, 表中涉及一个内共生菌 *Carsonella ruddii*[24]需要特别说明, 请参看关于高阶分类单元位置的 2.3 中(7)小节.

2.2　古细菌分支的分析

我们对CVTree上由 31 个古细菌构成的分支进行了详细分析, 并以此说明进行比较的方法. 《补充材料》[14]中列出了这 31 个古细菌基因组的基本信息. 表 2 给出了它们分类的分布情况.

表 1　440 个基因组形成三域 a)

K=3	K=4	K=5	K=6
7E	8E	8E	8E
1E in B			
25A in B	25A in B	31A	31A
1A(Arcfu) in B	1A(Methj) in B		
1A(Naneq) in B	1A(Naneq) in B	400B	400B
4A in B	4A in B		

a) 其中 A=古细菌(*Archaea*), B=细菌(*Bacteria*), E=真核生物(*Eukarya*)

表 2　31 个古细菌分类的分布

门	纲	目	科	属	种
A1	1	3	4	4	7
A2	8	9	12	18	23
A3	1	1	1	1	1
合计	10	13	17	23	31

在总共 23 个属中, 有 5 个属包含至少 2 个种; 在 17 个科中, 有 4 个科包含至少 2 个属; 在 13 个目中, 有 3 个目包含至少 2 个科; 在 10 个纲中, 有 2 个纲包含至少 2 个目; 在 3 个门中, 只有一个门包含了至少 2 个纲. 这些数字是根据伯杰分类系统统计出来的. 到目前为止, 我们仅仅参考了伯杰的分类系统. 下面对 CVTree 与伯杰分类系统进行比较. 考察从 K=3~6 的 4 棵 CVTree 树, 在种的水平, 我们可以看到: (ⅰ) 有 2 个属 *Pyrobaculum* 和 *Thermoplasma* 包含 2 个种. 它们在 K 从 3~6 的树上总是分别自成一组. 我们用罗马数字表示同一个属内种的数目. 这两个属分别记为 *Pyrobaculum*(II)和 *Thermoplasma*(II); (ⅱ)包含 3 个种的属有 3 个: *Sulfolobus*, *Methanosarcina* 和 *Pyrococcus*. 它们在 K 从 3~6 的树上总是分别自成一组. 这 3 个属分别用 *Sulfolobus* (III), *Methanosarcina* (III)和 *Pyrococcus*(III)标记.

当一个分类单元包含三个或更多低一级分类单元时, 下级分类单元相互之间的关系也是值得仔细考察的. 例如: *Pyrococcus* 属对所有 4 个 K 都以(Pyrfu, (Pyrab, Pyrho))形式出现; *Sulfolobus* 属在 K =3, 5 和 6 时, 3 个种聚为(Sulso, (Sulac, Sulto)), 但在 K =4 时, 聚为 (Sulac, (Sulso, Sulto)); *Methanosarcina* 属在 K =4, 5 和 6 树上, 3 个种的顺序为(Metbf, (Metac, Metma)), 而在 K = 3 时, 则是(Metma, (Metac, Metbf)). 随着 K 的增加, 大致可以看到“收敛”趋势: 从树的拓

1) 达尔文在其《物种起源》一书中多次提到种、属和科, 但几乎没有提到更高的分类单元.

扑结构与标准的分类系统的一致性来看, 除了个别情况以外, K = 4 的树的拓扑结构与标准的分类系统的一致性比 K = 3 好, K = 6 和 K = 5 的树基本一致, 但 K = 6 的树略好一些. 本文中, 将在不同的分类单元上反复看到这种收敛性.

图 1 中给出在 K 从 3~6 的 CVTree 树上整个古细菌分支形成的属树, 这里使用了上文引入的对属的标记. 这里的 23 个叶节点分别相应于 23 个属, 见表 2. 包含两个属的科有 3 个: (ⅰ) *Methanobacteriaceae* 包含 *Methanosphaera* 和 *Methanobacterium*; (ⅱ) *Methanosarcinaceae* 包含 *Methanosarcina*(Ⅲ)和 *Methanococcoides*; (iii) *Thermococcaceae* 包含 *Thermococcus* 和 *Pyrococcus*(Ⅲ). 它们在 K 从 3 ~6 都收敛到一起.

A_2C_4 纲中唯一的科 $A_2C_4O_1F_1$ (Halobacteriaceae) 包含四个属: *Haloarcula* (Halma), *Natronomonas* (Natpd), *Halobacterium* (Halsa) 和 *Haloquadratum* (Halwd). 除了 K=4 有略微不同之外, 它们对所有的 K 收敛. 即使在 K =3 和 4, 该科跳出古细菌分支时, 也仍然保持为一个单系, 在表 1 中我们标记为“4A in B”. 虽然《伯杰手册》还没有列出 Halwd 种, 但是根据所有 4 棵 CVTree 树所显示, 它应该属于本科.

可以看出, K = 5 和 6 的CVTree树在种、属和科的水平上跟伯杰分类系统完全一致. 为了比较更高的分类单元, 在图 2 中, 我们重画了 K = 5 和 6 的树, 这里使用了前文所述的“伯杰编码”. 其中叶节点的相对拓扑关系和图 1 是相同的. 从图 1 和图 2 中我们可以很容易地看出在更高的分类单元上, 这两棵树和伯杰分类系统只有三处差异, (ⅰ) 一个表面看起来是跨门的差异: A2 (Euryarchaeota)门的A_2C_5 (Thermoplasmata)纲, 在K从 3 到 6 的树上, 总是和A_1 (Crenarchaeota)门稳定地聚在一起. 不过, 这样的聚合方式跟其他一些生物学家的分类系统一致, 如《五界论》[22]. 因此, CVTree的这个差异不是一个真正的问题. (ⅱ) Aerpe所处的位置使得$A_1C_1O_1$ 不能形成一个单源分支. (iii) Arcfu所处的位置使得A_2C_3 不能形

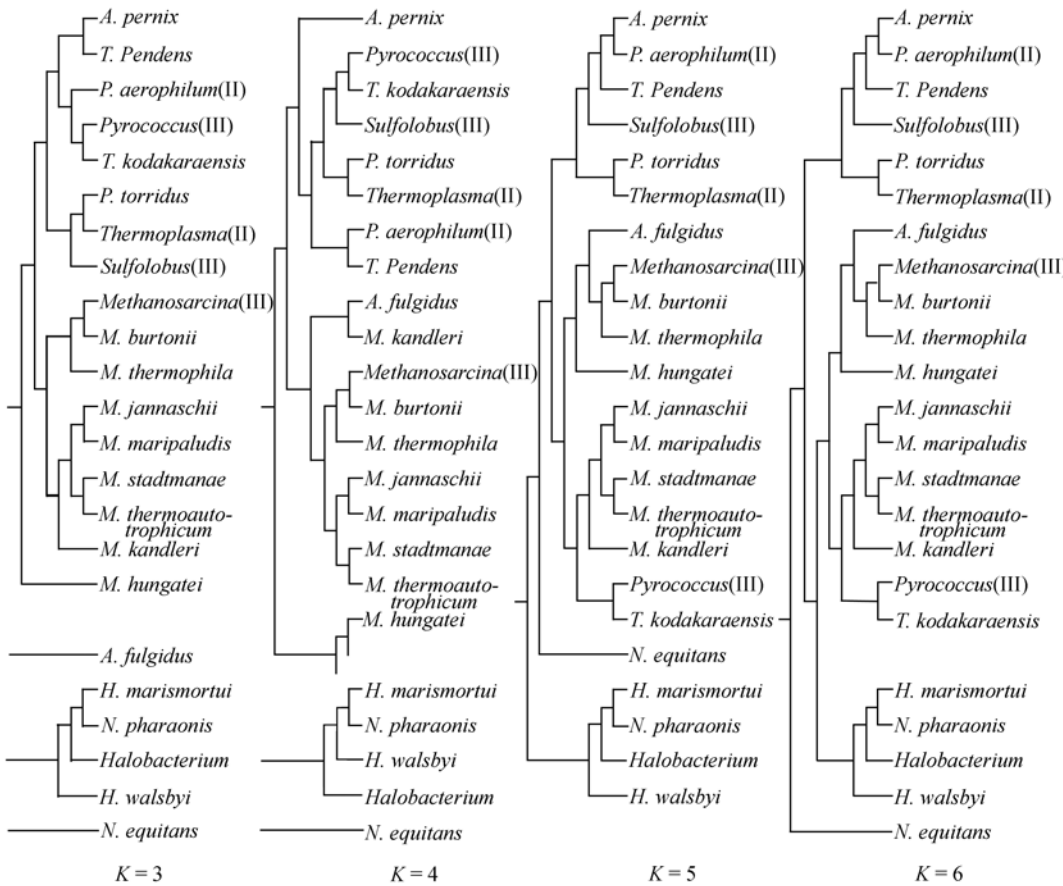

图 1 在由 440 个基因组重建的 K 从 3~6 的 CVTree 树上古细菌分支的属树

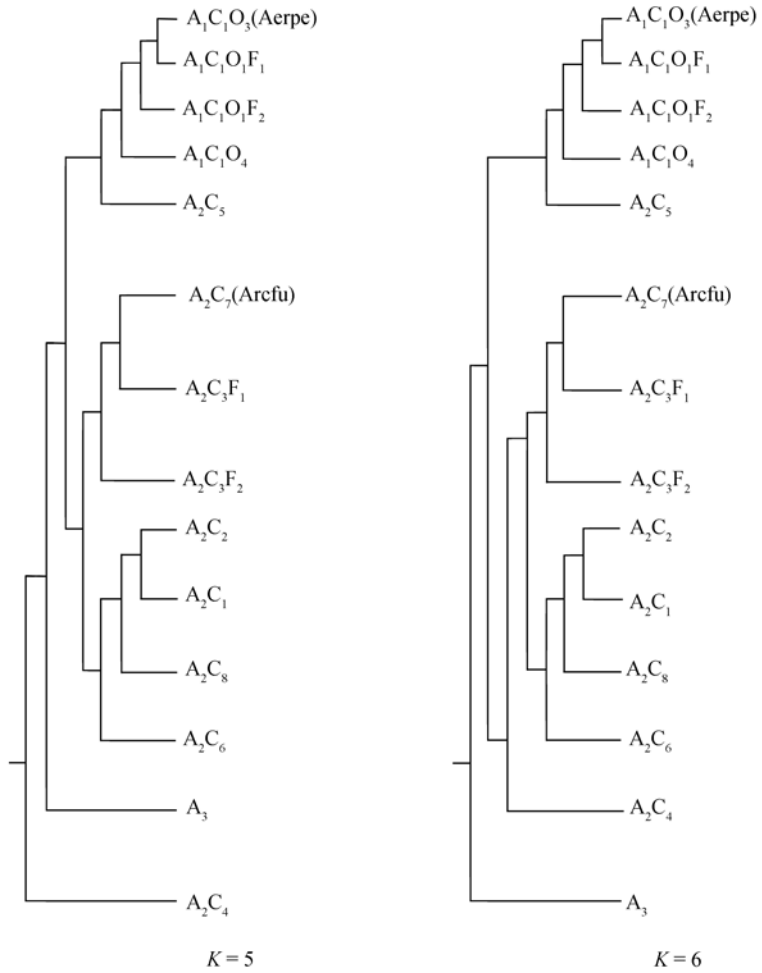

图 2　K = 5 和 6 时, 用伯杰编码表示的古细菌更高分类单元树

成一个单源分支. 后两个例子暗示应当对 Aerpe 和 Arcfu 的分类地位做出修正. 比如: 将 Aerpe 分到 $O_1F_?$, 将 Arcfu 分到 C_3 可以解决上述问题.

对 K 从 3 到 6 这 4 棵树进行的全面考察表明, 随着 K 的增加这些树有明显的收敛性. 在 K = 3 和 4 时, 31 个古细菌还不能构成一个单系; 但在 K = 5 和 6 时, 它们聚成了一个单源支. 新发现的 Nanoarchaeota 门, 目前只有一个代表物种 Naneq, 还没有被列入伯杰手册中. 它在 K=5 的 CVTree 树上没有形成一个单独的分支, 但是在 K=6 的 CVTree 树上独立成支. A_2C_4 (Halobacteria)纲在 4 棵树上都是一个单源分支, 在 K = 3 和 4 时一度插入了细菌域(表 1 中的 4A in B), 但在 K = 6 时最终回到了 A_2(Euryarchaeota)门, 因此, K = 6 的树支持将 Naneq 列为一个新的门.

总之, 对 31 个古细菌 CVTree 树和伯杰手册的详细比较分析显示仅仅 Aerpe 和 Arcfu 两个物种的分类存有疑问.

2.3　细菌分支的分析

表 3 列出了所有 401 个细菌基因组的分类分布情况. 我们对CVTree上由 401 个基因组构成的细菌分支同伯杰分类系统进行了详细的比较分析, 分析方法类似于上文对古细菌的分析. 下面仅仅进行概述. 对逐个分类单元的详细分析可以参看《补充材料》[14].

只有当一个分类单元包含至少两个低一层分类单元时, 在亲缘树上它才对应至少一个分支点. 如果一个分类单元包含至少三个低一层分类单元, 这些较低的分类单元在分类学中仅仅简单地并列在一起. 然而, 无论正确与否, 任何的系统发生树都会给出这些较低分类单元之间的顺序. 这为分类和系统发生

树的比较增加了一个新的角度, 因而可能导致新的进化知识. 表 4 列出了同一上层分类单元包含的低层分类单元的数目.

表 3 和表 4 中的数字满足一定的求和关系. 例如: 如果用 n_i^j 表示表 4 中第 j 行的第 i 个数字, 则求和 $a^j=\sum_{i\geq 1}n_i^j$ 与 $b^j=\sum_{i\geq 1}i\times n_i^j$ 满足关系: $a^{j-1}=b^j$, 其中 j 从 1 变到 6, 而 a^0 是所有细菌的数目, 本文中 $a^0=401$. 事实上, 上述这些 a^j 都出现在表 3 的最后一行中. 更多细节请参考《补充材料》[14].

表 4 的最后一列是至少包含两个低一层分类单元的分类单元总数. 也就是说, 有 54 个种至少包含两个株; 57 个属至少包含两个种; 34 个科至少包含两个属等. 总共有 184 个这样的分类单元. 特别应强调表 3 和表 4 中的数字完全来自伯杰分类, 而没有参考任何系统发生信息.

下面我们比较 K 从 3~6 的 CVTree 和分类学结果, 来检验上述 184 个分类单元. 我们对所有 184 个分类单元进行了详尽的分析, 结果表明其中有 138 个分类单元两者完全相同, 同时有 46 个分类单元存在差异. 在各分类水平上的差异情况请参看表 5. 较低的分类单元如种、属和科, 两者的一致性占主导. 相对而言在较高的分类单元(目、纲和门)上存在较多的差异. 事实上即使是分类学家们也常常对较高的分类单元存有争议. 从《补充材料》的详细描述可以看出这 46 个例子包含一些真正的小差异. 然而, 值得注意的一个事实是, 这 46 处差异中的大多数已经在某种程度上为生物学家所知晓, 或者正是不同分类学派的争议之处.

(1)平庸的差异: 在我们的数据集中各个门的物种数目并不均衡. B1, B2, B15 和B21 门都分别只有一个物种; 而B6, B11, B16, B19 和B20 门的物种数较少, 在所有较低的分类层次上, CVTree很容易同分类学保持一致. 所有这些门只有在涉及各门之间的相互位置时, 才会被提到. 这些平庸的差异只在《补充材料》[14]中进行描述. 下文我们只对五个门的比较结果做概述: 蓝细菌门(B10)、变形菌门(B12)、厚壁菌门(B13)、放线菌门(B14)和螺旋体门(B17).

表 3 401 个细菌基因组的分类分布

门	纲	目	科	属	种	株
B1 (Aquificae)	1	1	1	1	1	1
B2 (Thermotogae)	1	1	1	1	1	1
B4 (Deinococcus-Thermus)	1	2	2	2	3	4
B6 (Chloroflexi)	1	1	1	1	2	2
B10 (Cyanobacteria)	1	3	3	8	15	19
B11 (Chlorobi)	1	1	1	2	4	4
B12 (Proteobacteria)	5	33	53	99	157	208
B13 (Fermicutes)	3	7	14	22	58	96
B14 (Actinobacteria)	3	9	15	16	31	35
B15 (Planctomycetes)	1	1	1	1	1	1
B16 (Chlamydia)	1	1	2	3	7	11
B17 (Spirochaetes)	1	1	2	3	7	9
B19 (Acidobacteria)	1	1	1	2	2	2
B20 (Bacteroidetes)	3	3	5	5	6	7
B21 (Fusobacteria)	1	1	1	1	1	1
合计 15	25	66	103	167	296	401

表 4 同一上层分类单元包含的低一层分类单元的数目(细菌域)

i =	1	2	3	4	5	6	7	8	9	10	11	13	$\Sigma_{i>1}$
株/种	242	31	14	3	3			1	1		1		54
种/属	110	30	9	8	5		2		2	1			57
属/科	69	23	5	2	1	1	1				1		34
科/目	41	16	7	1	1								25
目/纲	15	3	2			2	2					1	10
纲/门	11		3		1								4
合计	488												184

表 5　CVTree 和伯杰分类的比较

分类层次	所含单元数目		*i*>1 情形与 CVTree 的比较	
	i=1	*i*>1	一致	不同
株/种	242	54	47	7
种/属	110	57	46	11
属/科	69	34	26	8
科/目	41	25	14	11
目/纲	15	10	5	5
纲/门	11	4	0	4
合计	488	184	138	46

(2) 蓝细菌门(B10, Cyanobacteria):表 6 列出了在株和种水平上 19 个蓝细菌在 CVTree 树上的收敛情况. 图 3 给出了 *K* 从 3 到 6 的四棵 CVTree. 可以看出, 在 *K* = 4, 5 和 6 时, 这 19 个细菌确实形成了一个单源分支, 这也验证了将它们放到同一门下的正确性.

表 6　蓝细菌(B10)门的 19 个细菌

伯杰分类				CVTree			
纲	目	科	属	种	株	聚类	*K*
C_1	1	F_1	G_7	1	1	Glovi	
			G_{11}	5	5	Prom9(II)	3,4,5,6
						Promt(II)	4,5,6
						Promm	
			G_{13}	3	8	Synja(II)	3,4,5,6
						Synp6(2)	3,4,5,6
						Synpx(4)	3,4,5,6
			G_{14}	1	1	Syny3	
			$G_?$	1	1	Synel	
	3	F_1	G_{17}	1	1	Triei	
	4	F_1	G_1	1	1	Anava	
			G_8	1	1	Anasp	

伯杰和 NCBI 的分类系统对蓝细菌的分类上存有分歧. 虽然两者都将蓝细菌归到了一个纲下, 然而在低层分类上, 两者的分类结果存有差异. 伯杰分类系统中, 在纲以下有 5 个未命名的亚分类单位. 而在 NCBI 的分类系统中, 有 7 个已命名的目, 其中: Chroococcales 目对应于亚分类单位 I, Oscillatorales 目对应于亚分类单位 III, Nostocales 目对应于亚分类单位 IV. 除此之外, NCBI 还有新的目, 如 Gloeobacterales 和 Prochlororales. CVTree 的结果可能有助于修正蓝细菌的分类.

本门的许多问题来自*Prochlorococcus*种. 这些最小的已知光合菌在 1980 年代后期被发现. 它们在海水中广泛存在, 在全球碳循环中起关键作用. 在伯杰分类系统中*P. marinus*属于Cyanobacteria纲亚分类单元I和科 1.1 下的形式属(FormGenus)XI, 它们都没有特别命名; 而在NCBI的分类系统中, 它属于Cyanobacteria纲Prochlorales目Prochlorococcaceae科, 也就是说, NCBI引入了一个全新的世系. 截止到 2006 年 12 月 31 日, 在公共数据库中共有*P. marinus* 的 5 个生态型[25, 26]的全基因数据, 参见表 7.

虽然这 5 个生态型的名字看起来像是同一菌种的不同菌株, 但是在 CVTree 中将它们当作不同的种来处理并不引起任何问题. 在 *K* = 3 到 6 的所有 CVTree 中(Promp, Prom9)形成一个稳定的小组, 这与它们的进化关系较近的高光适应性相吻合(表 7). 我们用 Prom9(II)标记它们. 在 *K* = 4 到 6 的 CVTree 上, (Proma, Promt)形成一个组, 同样与这两个生态型相近是一致的, 我们用 Promt(II)表示. 图 3 中使用了这些简写符号.

关于CVTree和伯杰分类系统在蓝细菌门的差异, 有以下几点: (ⅰ) *Synechococcus* sp.种的四个株(Synpx(4))没有归入相应的*Synechococcus*属, 而是和*Prochlorococcus*属的Promm种聚在了一起. 事实上这四个株与Promm的生活环境相同, 被看作是Promm的"潜在竞争者"[25]. *K*从 3 到 6, (Promm, Synpx(4))稳定地聚在一起. 因此, 将Synpx(4)列入Prochlorales目是合理的. (ⅱ) 虽然*Thermosynecho- coccus elongatus*还没有在伯杰手册的在线大纲中给出, 但是由于它在CVTree上和*Synechococcus elongatus*靠在一起, 因此可以据此将它的分类地位确定

表 7　*Prochlorococcus* 属的五个生态型

生态型	名称	缩写	备注
eMIT9312	*P. marinus* str. MIT 9312	Prom9	近表层, 适应高光照
eMED4	*P. marinus* MED4	Promp	同上
eSS120	*P. marinus* subsp. marinus marinus str. CCMP1375	Proma	深水, 适应低光照
eNATLA2	*P. marinus* NATL2A	Promt	同上
eMIT9313	*P. marinus* str. 9313	Promm	同上

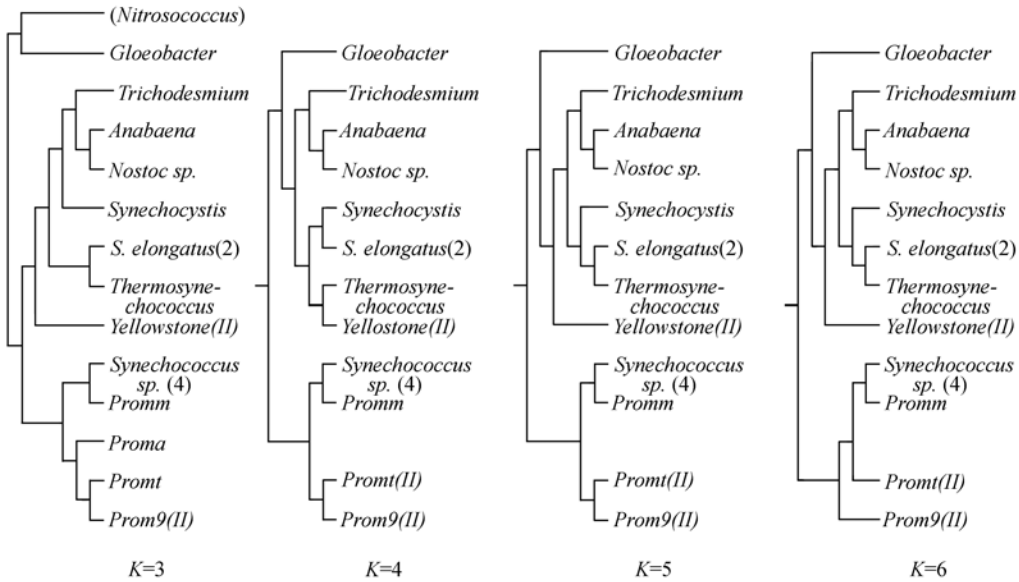

图 3 蓝细菌门 19 个细菌的收敛性

简写请参考表 7

到属: $B_{10}C_1O_1F_1G_?$. (iii)伯杰分类系统将 *Gloeobacter* 属和 *Prochlorococcus* 属放到了与 *Synechococcus* 属相同的科下, 这一点与 CVTree 的结果不同. 而 NCBI 的分类系统则将 *Prochlorococcus* 属放到了 Prochlorales 目下, 将 *Gloeobacter* 放到了 Gloeobacteriales 目下, 看起来比伯杰分类系统更加合理.

(3) 变形菌门(B12, Proteobacteria): 在我们的数据集中变形菌门共有 208 个全基因组. 伯杰分类系统将它们分成了 5 个纲或组. 下面我们将逐个讨论这些纲.

alpha 变形菌纲共有 55 个全基因组. 四棵 CVTree 在株、种、属和科的水平都充分收敛, 因此, 我们只需要讨论目水平的收敛性. 根据伯杰手册的在线大纲, 这 55 个细菌来自 6 个目, 其中有一个还没有在大纲中出现的新细菌: *Magnetococcus MC-1* (Magmc). 图 4 是在目的水平下本纲的四棵 CVTree. 这也是体现 CVTree 收敛性的另一个好例子. 6 个目在 K = 4, 5 和 6 时, 形成了一个单源分支. 同时, 这些树也暗示应当将 Magmc 放到 alpha 纲下的一个新的目中.

Beta 纲的 30 个基因组在株、种、属和科的水平完全收敛. 在 K = 4 时, 只有一个目跳到了外面, K = 3 时, 则更为分散, 但在 K = 5 和 6 时, Beta 纲的 6 个目形成了一个单源分支. 图 5 展示了随着 K 的增加的各个目是如何收敛的.

在我们的数据集中Gamma纲共有 101 个全基因组. 除了 2 个之外, 其它 99 个完全收敛到一个单源支. 详细的分析请参考《补充材料》, 在此重点讨论 16S rRNA树和CVTree树的一个共同特点. Woese及合作者在 16S rRNA树的研究中观察到: “Beta组实际上是在Gamma中较早分化独立的一个分支, Beta和Gamma作为一个整体与Alpha成姊妹关系”[27]. 在所有的CVTree上也同样如此. 然而, 我们现在对此有更进一步的认识. 从以肠道菌科(Enterobacteriaceae)为代表的肠道菌目(Enterobacteriales)的分开可以清楚看出Beta纲是怎样将Gamma纲分成“上”“下”两部分的. 该科是被研究得最多的科之一. 在所有四棵CVTree树上, 该科被Beta纲分成了两个组, 这两个组中细菌的基因组长度明显不同. “上”组包含了来自 7 个属的 28 个基因组, 除了内部关系的细微变化之外, 它们总是形成一个单源分支. 表 8 列出了此组各个属中最小基因组的长度.

肠道菌科“下”组的 8 种细菌全部是昆虫的细胞内共生菌. 由于特殊的生活环境, 它们的基因组经历了“还原进化”(reductive evolution), 因此基因组较小. 表 9 列出了每个属中最大的基因组. 可见, CVTree 和 16S rRNA 的一个共同问题是: 它们不能由在树上的位置来区分较早进化的基因组和经过还原进化的基因组.

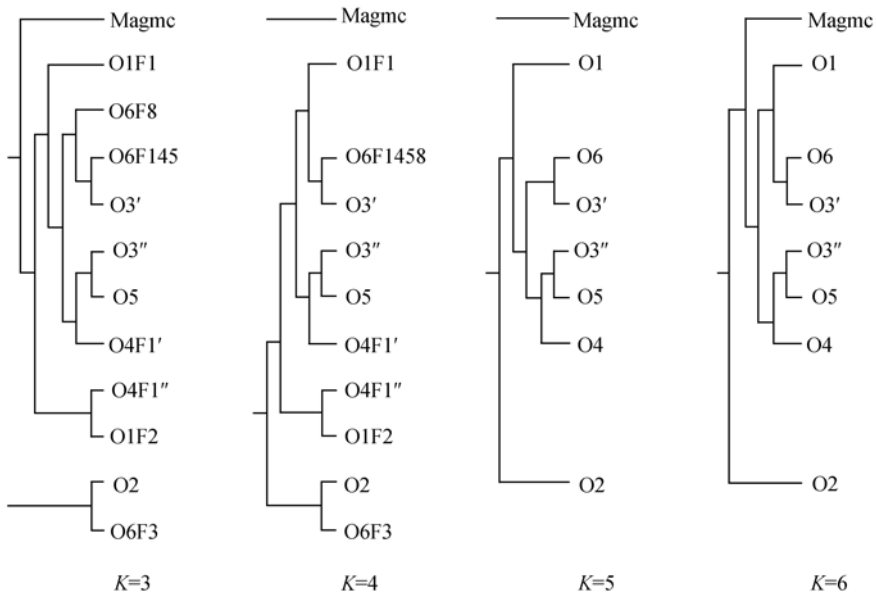

图 4　alpha 变形菌纲中各个目随着 *K* 增加的收敛性. 其中在 *K* = 4~6 时, 伯杰手册定义的 6 个目形成了一个单源分支. 新测序的 *Magnetococcus MC-1* 简写为 Magmc

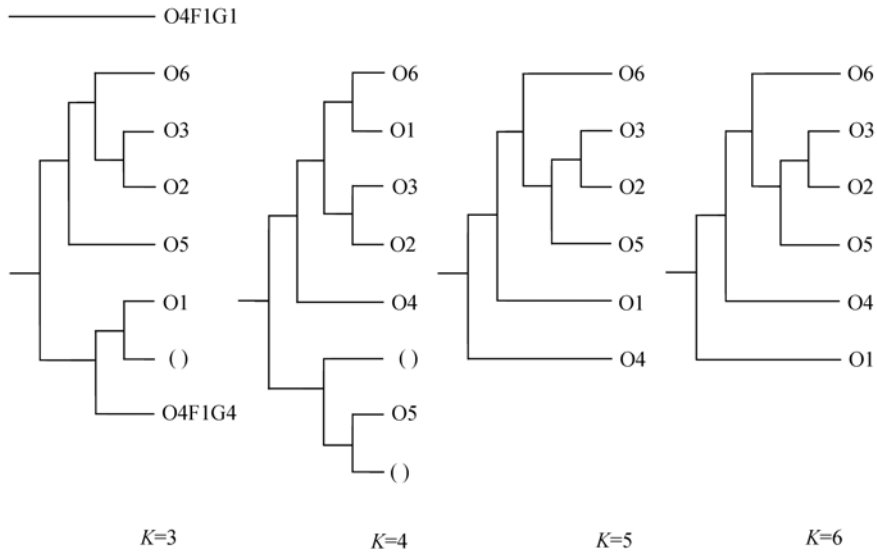

图 5　Beta 变形菌纲 6 个目的收敛性. 在 *K* = 5 和 6 时亲缘树的单源结构也表明了系统发生和分类的完美一致性. *K* = 3 和 4 的亲缘树中的小括号代表来自 Gamma 纲的变形菌

13 个Delta纲基因组有 10 个完全聚在一起, 形成一个单源支. 但是在所有 4 棵CVTree树上Bdellovibrionales目唯一的代表菌Bdeba和Myxococcales目的 2 个细菌Anade, Myxxd都跳到了这个分支的外面. 对于Bdeba需要更多同目的基因组被测序出来以后才能做进一步的分析. 而关于Anade和Myxxd, 我们注意到Myxococcales的分类地位是一个由来已久的问题. 多年以前, 甚至有人怀疑该目的物种根本不属于细菌 [28]. 这种情形在图 6 中用Delta(13-3)表示.

至于 Epsilon 纲, 唯一的差异是来自 Gamma 纲的 T. *denitrificans* 在所有 4 棵树上都插入了该纲. 这一点正好支持伯杰手册在线大纲中类似的观察, 参看

表 8 肠道菌科的“上”组中各属的最小基因组长度

种	缩写	基因组大小/bp	基因数目
Escherichia coli	EcoliK	4 639 675	4 237
Erwinia carotovora	Erwct	5 064 019	4 472
Photorhabdus luminescens	Pholl	5 688 987	4 683
Salmonella entrica	Salpa	4 585 229	4 093
Shigella dysenteriae	Shids	4 369 232	4 274
Sodalis glossinidius	Sodgl	4 171 146	2 432
Yersinia pestis	Yerpn	4 534 590	3 981

表 9 肠道菌科的“下”组中各属的最大基因组长度

种	缩写	基因组大小/bp	基因数目
Baumannia cicadellinicola	Bauch	686 194	595
C. Blochmannia pennsylvanicus	Blopb	791 654	610
Buchnera aphidicola	Bucap	641 454	546
Wigglesworthia brevipalpis	Wigbr	697 724	611

2.3 中第(7)小节.

(4) 厚壁菌门(B13, Firmicutes): 在本文数据集中, 厚壁菌门共有 96 个物种的全基因组. 厚壁菌非常多样化. “Hugenholz认为在厚壁菌门中至少有 4 个其他的门.”(伯杰手册在线大纲 [19]第二页脚注 3). 不过, 在K = 5 和 6 的CVTree上, 该门三个纲中的两个纲——Bacilli和Mollicutes——形成了两个相近但各自独立的单源支. Clostridia纲在K = 5 时分成了两部分, 但在K = 6 时聚为一整支并插入了Bacilli纲中.

厚壁菌门有几个菌种包含多个菌株的全基因组数据. 例如, 金黄葡萄球菌(*Staphylococcus aureus*)有 9 个株, 化脓性链球菌(*Streptococcus pyogenes*)有 11 个株. 由于CVTree方法的分辨能力甚高, 因而我们就有机会研究同一个种中各个株之间的进化关系. 事实上, 这些株之间的关系稳定而且收敛(参看《补充材料》[14]). 由于在包括厚壁菌门在内的所有门中, 绝大多数株都是收敛的, 因此, 个别不收敛的株值得我们特别关注. 例如, 蜡样芽孢杆菌(*Bacillus cereus*)的 3 个株和肺炎支原体(*Chlamydophila pneumoniae*)的 4 个株, 随着K的变化, 它们之间的相互关系也有所不同. 这是否是由株的快速变化引起需要进一步的研究.

另外, 伯杰手册在线大纲的第三版(Rel.3, July 2002)在变形菌门(B12)内引入了一个新属 *Oceanobacillus*. 然而, 在所有的 CVTree 上, 它一直和厚壁菌门的(B13)杆菌科(Bacillaceae)聚在一起. 然而伯杰在线大纲已经从第四版(Rel.4, Oct. 2003)起将其移到了 B13 门. 因此, 这个例子可以视为对 CVTree 方法的有力支持.

(5) 放线菌门(B14, Actinobacteria): 伯杰系统生物学手册的第一版(1986)中, 放线菌作为一个目分在厚壁菌门下. 在第二版中它们被提升成一个门. 在我们的数据集中该门共有 35 个全基因组, 其中 33 个在株、种、属和科的水平均有很好的收敛性, 同时也证明了将放线菌单列为一个新门的正确性 [29]. 关于两个例外的放线菌, 请参看《补充材料》.

有些例子表明, 在一定的条件下需要保持分类和系统发生的不同. *Mycobacterium bovis*种总是和两株*M. tuberculosis*菌(人类肺结核的病原体)聚在一起, 就像是*M. tuberculosis*种的一个新的株一样. 从基因组序列分析来看, *M. bovis*的长度只比*M. tuberculosis* 短了约 1%, 它们之间的序列相似性超过了 99.95%, 不过它们的感染模式显著不同 [30]. 这两个种临床表现的差异可能是由基因表达的不同所引起的. 目前所有以序列为基础的系统发生分析方法还没有考虑到这个因素. 大肠杆菌(*Escherichia Coli*)和志贺氏杆菌(*Shigella*)的株之间也存在相似的情形, 详见《补充材料》. 因此, 尽管这些种之间有较近的亲缘关系, 但也有理由把它们置于不同的分类地位.

(6) 螺旋体门(B17, Spirochaetes):关于螺旋体门有一个显著的现象: 在 K = 3 到 6 的所有 CVTree 上, 该门的两个科 Spirochaetaceae 和 Leptospiraceae, 从未聚在一起过. 事实上, 就像伯杰手册第一版(1984)所指出的那样: “*Treponema* 和 *Leptospira* 被指定到同一个目原因是由于它们都拥有螺旋状形态”. 因此, 它们更可能应划分到两个不同的门.

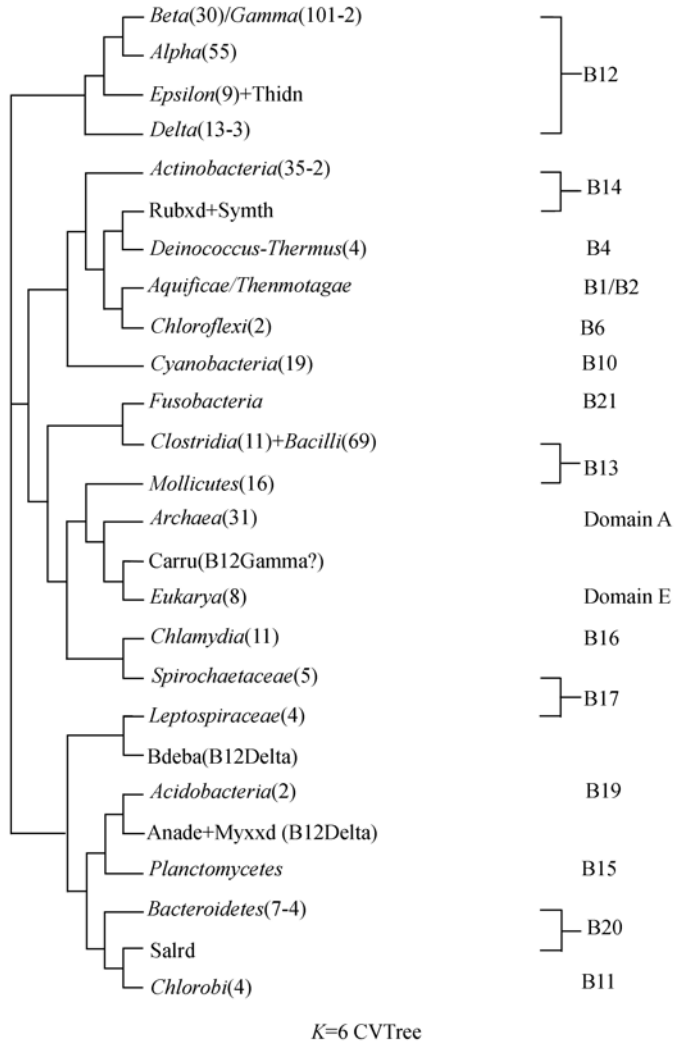

图 6　K=6 时最高分类单元的 CVTree

每个分类单元的名字代表了一个单源支, 括号中是该分支包含的细菌数目. 例如, Gamma(101-2)表示该支涵盖了 101 个 Gamma 变形菌中的 99 个, 两个例外 Carru 和 Thidn, 出现在树的其它位置. 这是一棵无根树, 而且分支没有标度

(7) 更高分类单元的定位:高层分类如何定位在分类学家之间一直都存有争议. 而原核生物就更是如此 [31]. 甚至是原核生物的种的概念也常常被质疑, 直到最近的 2006 年还有人对此提出异议 [32]. 鉴于这种情形, CVTree的结果就显得令人鼓舞: 432 个全基因组中的绝大多数都分别聚到了几个单源支, 并且这些分支分别对应于生物学家的分类. 图 6 是K = 6 时高层分类单元收敛最好的CVTree, K = 5 的CVTree只小有差异, 可参看《补充材料》[14].

如图 6 所示, 在 432 个全基因组中, 仅有 8 个例外: (ⅰ) 有两个例外来自Gamma变形菌纲. 其中一个是*Carsonella ruddii*[24](Carru), 这是一种细菌的内共生菌, 它的基因组高度退化, 只有 182 个编码蛋白、约 160 kb长, 远小于已知最小的可以独立生活的细菌(参看文献 [33]和它引发的讨论). 在对可以独立生活的原核生物的系统发生研究中, 它应当被舍弃. 我们在工作中保留了它, 然而它除了自身的位置存在问题之外, 并不能对树的整体结构产生多大影响. 另一个是*Thiomicrospira denitrificans* (Thidn), 事实上它不应该简单地算作例外. 在K =3 到 6 的CVTree上, 它一直稳定地插入了Epsilon组. 而伯杰手册的在线大纲已经注意到了这一点: 第 87 页的第 229 脚注指出"*T. denitrificans*的位置存在问题, 因为它属于Epsilon变形菌". (ⅱ)有两个来自放线菌门Actinobac- teria(35-2)

的例外, Rubxd和Symth, 但它们和其它 33 个放线菌形成的分支相距很近. (iii) 来自Delta(13-3)纲的三个例外已经在前面的 2.3 的第(3)小节中讨论过. (iv) 来自拟杆菌门Bacteroidetes(7-1)的唯一例外, Salrd, 仍然和其它 6 个拟杆菌同处于一个更大的分支下.

应当承认, 432 个基因组中只有 8 个例外, 表明系统发生和分类之间已经达到相当高的一致性. 除此之外, 在高于门的分组中也显示出某些特点: (ⅰ) 厚壁菌门的柔膜菌纲(Mollicutes)应当作为一个新的门. 其它两个纲在 K = 6 时聚在一起, 可以作为另一个门; (ⅱ) 钩端螺旋体科应从螺旋体门分离出来, 形成一个新门; (iii) B1, B2 和 B6 门在 K = 5 和 6 时, 形成了一个大的分支, 这需要等待包含更多基因组的进一步研究.

3 讨论和结论

CVTree 和越来越以 16S rRNA 分析为基础的伯杰分类系统之间存在着惊人的一致性. 这是一个值得特别关注的现象. CVTree 方法和 16S rRNA 分析可以说是使用了两种完全不同的数据和方法来推断系统发生信息. 然而, 它们在绝大多数分类单元的分支与聚类中互相支持, 因而为原核生物物种之间自然边界的划分提供了可靠的框架.

在CVTree和 16S rRNA树之间只有很少的差异. 例如, 根据rRNA的分析*Methanopyrus kandleri*没有加入其他已知的产甲烷菌(methanogenes)中 [34], 但是根据CVTree的结果, 它应当属于产甲烷菌.

几年前, 全基因组亲缘树还“不能分辨细菌的主要分支”[35], 今天CVTree方法从株直到纲和门的高分辨能力就真正是一个相当大的进步. 然而, 全基因组的使用有利有弊. 其优点是由于没有进行序列和基因的挑选, 因此最大程度地避免了在推断系统发生信息时的主观性和偏好性. 但是其缺点是因全基因组的获取不易, 限制了研究的范围. 目前, 新版伯杰手册已经包含了超过 6250 个原核生物物种. 在伯杰手册中单单是变形菌门就有 72 个科 425 个属和 1875 个种. 到 2006 年 12 月 31 日为止, 变形菌门中已有来自 53 科和 61 属的 123 个种有全基因组. 伯杰手册的第二版中最终可望增加几千个新的分类单元. 新版伯杰手册可能对CVTree方法建议的分类修正在更广阔的尺度上做出检验. 因此, 几年后一项类似于本文的更大规模的研究将为原核生物系统发生树提供一个骨架, 并为CVTree方法的预测能力提供进一步的测试.

到目前为止, 我们只是使用了 CVTree 方法的“定性”结果, 主要是树的拓扑结构. 然而组份矢量包含了更多的信息. 怎样利用其他的信息并进一步验证 CVTree 方法, 已经提上研究议程.

补充材料

《补充材料》[14]包含: 本文所使用的所有基因组的列表, 以及它们的简称, NCBI索取号和伯杰编码; K = 3 到 6 的四棵原始的CVTree; CVTree和生物学家分类系统逐个分类单元的详尽比较.

致谢 感谢来自复旦大学的支持. 感谢在 ATCC 和 Tim Lilburn 博士及 George Garrity 教授通过电话参与的讨论; 同时感谢在 UGA 和 Whitman 教授的讨论, 并感谢姜成林教授就放线菌的通信.

参 考 文 献

1 Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol, 1965, 8: 357—366
2 The AToL Project Home Page: atol.sdsc.edu
3 Driskell A C, Ané C, Burleigh J G, et al. Prospect for building the tree of life from large sequence databases. Science, 2004, 306: 1172—1174 [DOI]
4 Woese C R, Fox G E, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA, 1977, 74: 5088—5090
5 Bergey's Manual Trust. Bergey's Manual of Systematic Bacteriology, 2nd ed., vol 1-5, New York, Springer-Verlag, 2001—2008
6 Asai T, Zaporojets D, Squires C, et al. An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc Natl Acad Sci USA, 1999, 96: 1971—1976 [DOI]
7 Teichmann S A, Mitchison G. Is there a phylogenetic signal in prokaryote proteins? J Mol Evol, 1999, 49: 98—107 [DOI]
8 Snel B, Huynen M A, Dutilh B E. Genome trees and the nature of genome evolution. Annu Rev Microbiol, 2005, 59: 191—209 [DOI]
9 Ciccarelli F D, Doerks T, von Mering C, et al. Toward automatic reconstruction of a highly resolved tree of life. Science, 2006, 311: 1283—1287 [DOI]
10 Qi J, Wang B, Hao B L. Whole genome prokaryote phylogeny without sequence alignment: a *K*-string composition approach. J Mol Evol, 2004, 58: 1—11 [DOI]
11 Hao B L, Qi J. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J Bioinformatics Computat Biol, 2004, 2: 1—19 [DOI]
12 Qi J, Luo H, Hao B L. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucl Acids Res, 2004, 32 Web

Server Issue: W45—W47.
13 The NCBI ftp-site: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
14 本文补充材料可以从以下网址之一下载 http://www.itp.ac.cn/hao/Suppl440.pdf 或 http://tlife.fudan.edu.cn/Suppl440.pdf.
15 Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.5c., distributed by the author at: http://evolution.genetics. washington.edu/phylip.html
16 Michel H. The future of the molecular biosciences: consequences of the massive parallel approach. Sceince and Technology Development: A Retrospective View over the Past Century and a Prospective Look into the Future. In: Lu Y X, ed. Shanghai Education Press, 2000. p 70
17 Shi X L, Xie H M, Zhang S Y, et al. Decomposition and reconstruction of protein sequences: the problem of uniqueness and factorizable language. J Korean Phys Soc, 2007, 50: 118—123
18 Konstantinidis K T, Tiedje J M. Towards a genome-based taxonomy for prokaryotes. J Bacteriol, 2005, 187: 6258—6264 [DOI]
19 Garrity G M, Bell J A, Lilburn T G. Taxonomic Outline of the Prokayotes. Bergey's Manual of Systematic Bacteriology, 2nd Ed, New York: Spinger-Verlag, Rel. 5.0, 2004, DOI: 10.1007/bergeys-outline200405.
20 The NCBI Taxonomy Browser: http://www.ncbi.nlm.nih.gov/Taxonomy/
21 The taxonomic list at EBI: http://www.ebi.ac.uk/genomes/ bacteria.html
22 Margulis L, Schwartz K V. Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth, 3rd ed. W. H. Freeman, 1998.
23 Woese C R, Kandler O, Wheelis M L.Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA, 1990, 87: 4576—4579 [DOI]
24 Nakabachi A, Yamashita A, Toh H, et al. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science, 2006, 314: 267 [DOI]
25 Johnson Z I, Zinser E R, Coe A, et al. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science, 2006, 311: 1737—1740 [DOI]
26 Coleman M C, Sullivan M B, Martiny A C, et al. Genome islands and the ecology and evolution of *Prochlorococcus*. Science, 2006, 311: 1768—1770 [DOI]
27 Woese C R, Olsen G J, Ibba M, et al. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Revs, 2000, 64: 202—236 [DOI]
28 Pringshein E G. The relationship between bacteria and *Myxophyceae*. Bacterialogical Revs, 1949, 13: 47
29 职晓阳, 蔡 曼, 杨玲玲, 等. 建立放线菌门的证据. 微生物学通报, 2006, 33: 181—183
30 Garnier T, Eiglmeier K, Camus J C, et al. The complete genome of *Mycobacterium bovis*. Proc Natl Acad Sci USA, 2003, 100: 7877—7882 [DOI]
31 Murray R G E. The higher taxa, or, a place for everything…? In Bergey's Manual of Systematic Bacteriology, 1st ed., vol. 4. Baltimore: Williams & Wilkins, 1989. 2329—2332
32 Doolittle W F, Papke R T. Genomics and the bacterial species problem. Genome Res, 2006, 7: 116
33 Goffeau A. Life with 482 genes. Sceince, 1995, 270: 445—446
34 Burggraf S, Stetter K O, Pouviere P, et al. *Methanopyrus kandleri*: an archeal methanogen unrelated to all other known methanogens. Sys Appl Microbiol, 1991, 14: 346—381
35 Huynen M, Snel B, Bork P. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. Science, 1999, 286: 1443a [DOI]