中国科学: 生命科学 2017 年 第 47 卷 第 2 期: 159 ~ 170
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述 中国科学院 科学与技术前沿论坛 后基因组时代的物理生物学

基因组时代的计算微生物学

左光宏[1], 郝柏林[1,2*]

1. 复旦大学物理系, 理论生命科学研究中心, 上海 200433;
2. 中国科学院理论物理研究所, 北京 100190

* 联系人, E-mail: hao@mail.itp.ac.cn

收稿日期: 2016-11-06; 接受日期: 2016-11-22; 网络版发表日期: 2017-01-22
国家重大科学研究计划(批准号: 2007CB814800, 2013CB834100)资助

摘要 微生物基因组和RNA序列构成生物学数据的重要部分. 从基因组出发而且不用序列联配的CVTree和基于16S rRNA序列联配的LVTree, 是两套原始数据和计算过程相互独立的构建原核生物亲缘树和分类系统的途径. 这两套途径的自动化, 使亲缘关系和分类系统成为大数据分析的副产品, 可以帮助后继乏人的分类学摆脱困境. 特别是基于基因组的CVTree, 既提供了大范围研究的工具, 又在种以下具有16S rRNA序列分析所不能企及的高分辨力, 可以提出和解决一批新问题, 开辟若干新方向. 本文是相关研究工作的扼要综述.

关键词 原核生物, 亲缘树, 分类系统, 不用序列联配的基因组比较, CVTree, LVTree

生物是物, 生物有理; 生物有形, 生物有数. “夫圣人者, 原天地之美, 达万物之理”[1]; 物之至美者, 莫过生物. 形和数乃数学的对象. 用数学研究和表达物理, 就是理论物理. 物理学和生物学的相互促进由来已久. 在英文和某些其他西方语言里, 医师(physician)和物理学者(physicist)源于同一个字根. 物理手段决定生物研究的深度. 1859年, 达尔文的《物种起源》全书没有提到细菌或微生物, 但确实几次说到显微镜. 微生物学是显微镜下的生物学. 计算机也是一种物理手段, 微生物学正在被计算机发展所推进.

1943年, 理论物理学家薛定谔在都柏林发表“什么是生命”的著名演讲时, 听众席里坐着日后成为中国两弹一星功勋之一的彭桓武. 1978年, 彭桓武在创立中国科学院理论物理研究所时把理论生物列入研究方向. 1993年, 中国物理学会委托清华大学组织为期3天的“物理学与生物学”讨论会; 1995年为祝贺彭桓武先生80华诞, 专门举行了题为理论物理与生命科学的香山会议[2].

虽然本次科学技术前沿论坛的标题中写有“后基因组时代”, 本文主旨仍是强调基因组时代正方兴未艾. 微生物基因组数据的大量涌现和积累, 为从整体和细节两个方面推进微生物学研究提供了前所未有的机遇. 这首先是从基因组数据出发, 阐明细菌的亲缘关系和建立客观的分类系统. 一向以“湿”实验为基础的微生物学, 也必然要借助信息技术来实现综合. 计算微生物学, 已经成为重要的研究方向. 本文将结合本研究组磨剑10余年所锤炼的CVTree网络服务器, 介绍已经取得的成果和今后的发展前景, 同时与基于

引用格式: 左光宏, 郝柏林. 基因组时代的计算微生物学. 中国科学: 生命科学, 2017, 47: 159–170
Zuo G H, Hao B L. Computational microbiology in genomic era. Sci Sin Vitae, 2017, 47: 159–170, doi: 10.1360/N052016-00312

16S rRNA序列分析的结果略做比较.

1 原核生物分类学的危机

原核生物是古菌(Archaea)和细菌(Bacteria, 过去曾称为真细菌Eubacteria)的总称, 本文中有时也统称为细菌. 它们是地球上最成功的物种, 同真核生物一起组成地球上的三大"生命超界"(domains of life)[3,4]. 有人估计地球上存活的原核细胞总数达到10^{30}的量级[5], 即每克地球物质, 不论其所处环境是否适合生命存在, 平均携带着100个细菌细胞. 虽然对于如何定义细菌的"种", 存在长期争论, 地球上细菌的"种"数不少于10^7[6]. 然而, 目前具有合法发表的名称和一定谱系信息的细菌还没有超过14000种, 即只有10^4的量级(这是根据文献[7]的数字, 加上每年发表约800个新种所做估计). 基于如此稀疏采样的微生物学的成功, 乃是由于世界的统一性来自其物质性. 基于同理, 模式生物的研究揭示出生物的普遍特性.

20世纪微生物分类学的主要成就有三. (ⅰ) 建立了为细菌命名的国际规则[8]("规则"的最新版本正在修订出版中[9]). (ⅱ) 在伯杰基金会支持下持续多年出版以《伯杰手册》为代表的细菌鉴定[10,11]和分类[12]手册. (ⅲ) 根据沃斯(Carl Woese, 1928~2012)与合作者们的建议[13,14], 把核糖体小亚基中的16S rRNA序列作为比较细菌菌株的分子标记. 因此, 21世纪初细菌名字的总数虽然比100年前少, 但是其系统性和可参照性更强.

进入21世纪以来, 《伯杰细菌鉴定手册》仍然停留在1994年的第9版[11], 却以12年功夫出全了《伯杰细菌系统学手册》的第2版[15]. 随后, 全电子版的《伯杰古菌和细菌系统手册》简称BMSAB[16]在2015年上线. 另外一批作者编辑出版的多卷本《原核生物》, 在2013~2014年出齐了第Ⅳ版[17], 全书11卷中的6卷是分类系统. 如果说, 伯杰手册系列的分类描述以属为单位, 则《原核生物》各卷的描述基本上按科划分, 在入选时间上则比伯杰手册更近一些. 虽然如此, 亲缘树与分类系统的比较要跟上新种的发表, 就必须参考细菌分类方面的主要期刊, 首先是《国际微生物系统与演化杂志》(*International Journal of Systematic and Evolutionary Microbiology*, 简称IJSEM). 相对最为及时的原核生物种和种以上分类单元的名录是LPSN[18]. 这是由Jean Euzeby在1997年创建, 现在由其他人继续维护的网站. 它反映着IJSEM和其他主要微生物分类学出版物的最新状态, 其信息距初次发表一般只有几个月的延迟时间.

由于新一代测序技术的发展, 使得测序成本不断降低. 越来越多的生物学研究从基因组测序开始. 有关团队在提取了有用信息后, 没有兴趣、人力和物力去完成新菌株的鉴定、命名和发表. 作为20世纪重大成果的国际规则, 逐步成为妨碍广大非分类专业的微生物学工作者发表新细菌的"行会戒律". 例如, 提交可培养菌株和DNA杂交实验结果作为发表新种的前提, 就不再适应当前形势, 因为可培养的细菌绝对不超过菌种总数的百分之一, 而DNA杂交也并不是多数实验室可以保证质量的常规操作. 于是许多测序结果被冠以相当任意的名称, 作为"永久草图"(permanent draft)存档. 例如, 在本研究组已经下载的7万多个基因组中, 就有6000个以上在名称中包含bacterium一词, 而不是国际规则所要求的"林奈双名". 新种和新株的发现与日俱增, 专业的分类工作者逐年减少. 微生物分类学面临危机!

2 出路在于基因组测序

1995年发表了最早测序的两个细菌基因组, 它们是流感嗜血菌(*Haemophilus influenzae*)[19]和生殖道支原体(*Mycoplasma genitalium*)[20]. 到2008年已经测序的基因组数目接近1000个, 而且开始出现越来越多的"永久草图". 此后被测序的基因组, 特别是永久草图的数目急速上升. 2013年11月NCBI宣布改变细菌菌株基因组的发布方式, 实际上停止了发布以NC_标示的细菌基因组序列. 从2008年到数据尚不完全的2016年10月中旬, 永久草图的累计数目达到43121, 远超过同期产生的完全基因组数目(7982)(根据GOLD数据库[21]所提供的统计数字).

《伯杰手册》现在的总负责人维特曼(Williams B. Whitman)早在2011年就针对此种危机发出警告[22], 提出把DNA序列也作为描述新菌种的基础. 他最近提出了这样做的正式建议[23]. 事实上, 这一思路由来已久. 早在1987年, 一个负责调和各种细菌分类方法的专门委员会就在报告[24]中指出: "(人们)普遍同意DNA序列应是定义亲缘关系的参照标准, 而亲缘关系

应当定义分类系统. 此外, 命名系统(nomenclature)应当与基因信息一致并反映后者." 早在人类基因组计划甚嚣尘上的1998年, 沃斯就为微生物基因组学发表了"宣言"[25]; 同一时期, 他在另一篇文章[26]中说: "基因组测序的时代已经到来, 基因组学将是未来微生物学的中心. 此刻看起来人类基因组好像是焦点所在和基因组测序的首要目的. 但是, 请不要被蒙蔽. 长远的真正回报来自微生物基因组学."

虽然绝大多数微生物基因组测序对象的选择出于致病、耐药、能源、环境、生态等"功利"考虑, 只有GEBA(Genomic Encyclopedia of Bacteria and Archaea)[27,28]等少数计划针对亲缘关系和分类系统的研究. 然而, 由于已完成基因组数测序的微生物数量巨大, 它们的分类代表性已经相当宽广, 足以构建反映大部分已知门类的细菌和古菌亲缘树骨架. 现在已经到了实现沃斯富有远见设想的时候.

3 不用序列联配的基因组比较

原核生物的基因组在核苷酸数目和基因内容方面的差别很大. 不考虑那些高度退化的细菌内共生菌株. 生殖道支原体的基因组小到只有58万碱基对和480多个基因[20], 而目前已经测序的最大的基因组来自纤维堆囊菌(*Sorangium cellulosum*), 它有1480万碱基对、编码1万多个基因[29]. 怎样比对尺寸如此悬殊的DNA和蛋白质序列呢? 原核生物基因组的比较只能采用非序列联配(alignment-free)的方法.

本研究组不用序列联配来实现基因组比较的办法, 就是把氨基酸频度计数推广到K肽计数. 取一个基因组所编码的全部蛋白质序列, 规定一个不大的正整数K($K\geqslant 3$). 用宽度为K的滑动窗口移过每个蛋白质序列, 每次移动1个字母. 从一条长度为L个字母的蛋白质, 得到$(L-K+1)$个K肽. 考察从全部蛋白质得到的K肽, 把每种K肽重复出现的次数记录下来. 由于蛋白质共有20种氨基酸, $K=1$时K肽的种类为20种; 而$K=3$时, K肽的种类为$20^3=8000$种. 把所有可能的K肽按其中氨基酸字母的字典顺序排列, 在每个位置上填入相应K肽的出现次数, 得到一个初始的组分矢量(composition vector, CV). 每个细菌用相应的CV代表; 从原点出来的两个CV的距离, 代表两个物种的距离. 从距离矩阵出发用标准办法, 如邻接法构造亲缘树. 可能包括本研究组在内的许多研究者都尝试过这种简单方法, 却因为结果不好而止步不前.

对简单方法失败的反思, 使本研究组注意到木村资生的中性演化理论[30]. 现在已经被普遍接受的木村理论认为基因组中保留着大量"不好不坏"的中性突变的结果. 这些突变对上述K肽计数的贡献, 与物种分化过程没有直接关系, 应当作为某种背景减除掉. 还是根据木村理论, 突变在分子水平上随机发生, 而自然选择才决定演化的方向. 因此, 中性突变造成的背景可以用某种统计模型来减除. 对同一个基因组编码的全部蛋白质序列, 统计出K肽, $(K-1)$肽和$(K-2)$肽的数目, 然后依据某种随机模型来从$(K-1)$肽和$(K-2)$肽的数目预测特定K肽的数目, 并把结果同实际统计出来的数目比较. 如果实际计数的结果与预测的数目相同, 那这个计数结果就没有包含新的生物信息, 因为$(K-1)$肽和$(K-2)$肽的数目可能包含生物信息, 但是统计预测公式并不带来新的信息. 预测和实际计数的差别才是更有意义的量. 对初始CV的每个分量进行减除, 用减除后的差别代替原来的分量, 得到新的"重正化"后的CV.

本研究组使用$(K-2)$阶的马尔可夫预测来求得特定K肽的数目. 这个预测公式可以用两种方法推导出来, 或是借助概率论中联合概率与条件概率的关系[31~37], 或是使用最大熵原理[38]. 由于相应数学公式已经多次在文献[31~37]中描述过, 仅在此指出, 预测公式的选取不是唯一的, 目前也不是靠第一原理, 而是由实际结果取舍. 例如, 有人给出过只利用$(K-1)$肽数目的看起来更简单一些的式子[39], 但是所给出的亲缘树上, 有古菌(*Pyrobaculum aerophilum*)混入了细菌超界, 破坏了生命三大超界的划分.

用经过"重正化"的CV做物种的代表矢量. 把一个物种的CV投影到另外一个物种的CV上, 得到二者的关联C. 归一化以后的C取从+1(完全关联)到−1(完全反关联)的数值. 用C定义"距离"或非相似性$d=(1-C)/2$. 这是可以从0变到1的归一化了的"距离". 距离二字放在引号内, 因为它不是保证全部三角形不等式都能成立的普通的几何距离, 而是一种"准度规". 不去讨论这些目前尚未懂透的数学细节, 有兴趣的读者可以参看文献[34~37].

使用$(K-2)$阶的马尔可夫预测, 决定了最小的K值是3. 使用更大的K值将更突出物种特异性, 但是过大

的*K*值会导致接近星形的树(star tree), 反而不能正确反映物种间的亲缘关系. 可以比较严格地论证[34~37], 最佳*K*值的选取同涉及的蛋白质所包含的氨基酸总数有关, 对于古菌和细菌, 最佳*K*值是5和6; 对于病毒, 最佳*K*值是4和5; 对于真菌, 最佳*K*值是6和7. 这些不敏感的"对数"估值, 同本研究组十几年来的计算经验是一致的.

4 CVTree网络服务器

上一节里描述的算法, 思想简单、步骤明了. 但是具体实现起来, 涉及处理维数极高的矢量和矩阵. 例如, *K*=6时单个CV的维数是6400万, 而计算1万个细菌的亲缘树要用到10000×10000的矩阵. 为了帮助微生物学工作者直接使用这个方便工具并发掘其潜在威力, 专门设计和开发了名为CVTree的网络服务器, 为国内外用户提供免费服务. CVTree 1.0于2004年上线[40], 现在已经停止服务. CVTree 2.0于2009年上线[41], 目前仍在运行. 然而, 强烈建议读者使用2015年开始对外服务的功能更为强大的CVTree3[42]. 虽然这些网络服务器都带有在线使用说明书和可以单独打印的说明文件, 本研究组还是以最新的CVTree3为例, 扼要介绍其特点和功能.

使用者只需要在自己的浏览器里敲进地址 http://tlife.fudan.edu.cn/cvtree3/, 不必登录就进入用户界面. 可以(但是不必须)输入自己的电子邮件, 以便离线等待结果通知. CVTree3服务器目前在一个64核的集群计算机上运行. 由于采用了用空间换时间即保留此前各个作业中间运行结果的策略, 一个3000多基因组的作业大致可在20 min内完成. 一个涉及1.5万个基因组的作业, 要用96 h以上才能得到结果. 尽管目前已经收集了7万多个原核生物基因组, 但是不能把这些都内置到服务器中, 这是要避免某位漫不经心的用户"全选", 从而导致系统崩溃. 本研究组鼓励目标明确的同行, 与我们合作使用尚未公开的服务器从事大规模研究. 概括地说, CVTree3服务器有以下功能:

4.1 输入数据集合

CVTree3服务器内置有3000多个基因组, 供用户挑选. 这些基因组会不定期地更新. 用户可以多次上传自己的基因组数据, 每次可以上传100 M以内的压缩或未压缩文件. 内置和上传的每个基因组都应当带有如下的谱系信息:

<D>Bacteria<K>Bacteria<P>Proteobacteria<C>Gammaproteobacteria<O>Enterobacteriales<F>Enterobacteriaceae<G>Escherichia<S>Escherichia_coli<T>Escherichia_coli_str_K12_substr_MG1655

这里<D>, <K>, <P>, <C>, <O>, <F>, <G>, <S>和<T>分别代表超界(Domain)、界(Kingdom)、门(Phylum)、纲(Class)、目(Order)、科(Family)、属(Genus)、种(Species)和株(sTrain). 对于原核生物本来不必区分界和超界, 这里分别保留了<D>和<K>, 是为了以后把CVTree推广到真核生物. 如果某一个分类层次没有确定名称, 就用特定的指示字Unclassified标出. 例如, <F>Unclassified表示谱系中的"科"没有确定. 内置基因组的初始谱系信息来自NCBI的分类网页[43]. 虽然NCBI在每一条项目后面都声明它不是分类学的参考, 但它确是比较完全和动态的信息, 因为每一条分子信息的提交者都会指出相应的谱系信息, 尽管有些可能不符合一般公认的分类系统. 本研究组在形成内置的谱系文件时, 做过一些必要的调整和补充.

4.2 亲缘树的缩并和展开, 自动报告统计结果

用户可以选择一个或多个*K*值, 在一次运行中完成全部计算, 并由系统按分类级别报告统计结果. 这里的一个核心概念是单源性或单系性(monophyly). 1866年, Ernst Haeckel首次引入的这个概念, 涉及认定宏观动植物的祖先和后代关系. 对于无性繁殖为主的细菌, 采取一种实用观点, 即把考虑范围限制在CVTree的输入数据集合和与之相应的谱系信息. 这个集合包含着属于多种分类单元的成员. 如果一个分类单元, 例如一个"属", 包含了根据谱系信息属于它的所有成员, 而不包含其他外来成员, 则这个"属"是单源的. 在此意义下, 梭形菌属(*Clostridium*)就不是单源的, 因为迄今为止的所有分类系统, 例如, 文献[15~17]都认为梭形菌分成许多个互相不交叠的集团, 而本研究组的输入数据集合中的梭形菌也有可能来自不同的集团. 以上是对分类系统而言. 对于亲缘树上的分枝, 同样要判断是否单源. 如果长在一个树枝上的所有叶子代表着输入数据集合中一个单源分类单元里的全部成员, 那么这个枝是单源的. 这时可以把整个单源枝换成一个叶子, 并标上相应分类单元的名字和其中成员

的数目. 单源枝的此种缩并(collapsing)可以大为减少一棵大树的枝叶数目, 有利于一目了然地看清树的整体结构. 缩并后的枝叶还可以展开成原来的样子.

实际工作中, 往往对若干个*K*值一举计算出全部亲缘树. 然后考察一个分类单元, 例如一个“属”, 在不同*K*下是否单源. 一般说来, 随着*K*从3变到更大, 更多的树枝会趋向单源. 本研究组把这个过程称为“收敛”. 如果要对*K*=3~9手工考察每个门、纲、目、科和属是否收敛, 其工作量相当之大. CVTree服务器自动完成收敛考察和统计, 并以表格形式报告结果. 这个表格按照从<D>~<S>的分类阶梯, 顺序给出每个分类单元所包含的基因组数目, 以及在每个*K*值下相应树枝是否单源. 如果一个分类单元, 例如某个属, 在输入数据集合中只包含一个基因组, 则根据定义它只能是单源的; 这些平庸的单元可以在报告表中掩蔽掉不予显示.

4.3 询问特定基因组所处的分类位置

迅速找到一棵大树上感兴趣物种的办法, 就是在会话式显示树枝的页面上敲进该物种或分类单元的名称, 以代替原有的Search Query字样. 例如, 敲进属名*Corynebacterium*(棒杆菌属), 可以从CVTree的大树上切出图1. 图1只显示了*Corynebacterium*属的分枝情形. 它的单源性因*Turicella*的插入而被破坏. 本研究组以后还会继续讨论这个例子.

4.4 谱系修正和重新缩并

考察一棵大树时, 可能会发现某些谱系信息不合理, 或者有可能补充缺失的信息, 即把Unclassified改成确定的分类单元名字. 用户可以准备一份谱系修正文件(lineage modification file), 其中每一行给出一条修正建议, 形式为:

现有谱系<空格>新谱系 #注释

只要不因为二义性而引起非所预期的副作用, 可以不写出全部谱系信息, 而只写出需要修正的部分. 例如, 图1提示, 只要把*Turicella*属并入*Corynebacterium*属, 就可以使后者成为单源支. 相应的谱系修正建议是:

<G>Turicella <G>Corynebacterium # by inspecting CVTree

如果所作修正还涉及<O>或<F>的改变, 就不能写得这样简单了. 特别是为了使修正只针对特殊的菌株, 而没有其他副作用, 有时要明确写出<T>一级的信息.

提交谱系修正之后, 系统重新进行缩并和统计, 随即产生新的树图和收敛报告. 上面的修正导致单元属<G>Corynebacterium{122}. 谱系修正文件同用户上传的基因组数据一起, 在作业最后一次运行以后, 还在用户工作空间里保存一周, 然后由系统自动删除.

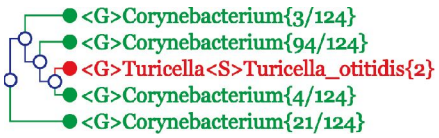

图 1 在CVTree显示页面上询问*Corynebacterium*(棒杆菌属)后显示的该属局部(网络版彩图)

*Turicella*的插入破坏了此属的单源性, 见本文后面的讨论

4.5 附加说明: 16S rRNA树的交互式显示

本文主旨在于讨论基于全基因组的亲缘关系和分类系统. 考虑到现有的原核生物分类系统主要基于16S rRNA序列分析, 从CVTree得到的结果可以与之比较. 2008年, 《应用与系统微生物学》(*Systematic & Applied Microbiology*)杂志编辑部和LPSN网页[18], 与一批欧洲微生物学工作者合作, 开始在线发表基于上万条经过人工审读的16S rRNA序列的亲缘树(all-species living tree, LVTree[44,45]), 其最近的版本是2015年9月的第123版[46]. 这棵树是超过百页的文件, 不易搜寻和查阅. 把原来为CVTree设计的缩并、展开、查询、修改谱系和再缩并, 以及报告各个分类单元是否为单源枝等各种功能, 全部移植到为LVTree编写的一个交互式显示程序LVTree Viewer上[47]. 不过使用这个显示程序只能查看最近的已经构建好的LVTree版本, 不能上传用户自己的序列, 也不能考察没有合法名称的菌株, 这是LVTree的设计原则决定的. 用户可以同时使用CVTree和LVTree这两套工具研究同一个分类问题.

5 原核生物分类系统的整体性研究

传统的细菌分类学着重单个菌种或类群的具体研究. 现在有了数百万16S rRNA序列和近十万基因组数据, 已经可以提出和解决一批整体性的问题. 下面列举的若干项目, 每一项都可以成为有丰富内容的研究方向. 限于篇幅, 只能以例示意, 点到为止.

5.1 细菌的高层次分类

直到不久以前, 细菌命名的国际规则[8,9]只承认到“纲”一级的分类单元. 最近刚刚提出在规则中承认“门”的建议[48]. Cavalier-Smith制造了“巨分类”(mega-classification)一词[49], 特指目、纲、门等高阶层次的分类. 现在可以同时参照CVTree和16S rRNA序列分析的结果, 看到原核生物巨分类成功和欠缺的大致轮廓. 本研究组不久前研究了古菌的巨分类[50], 在介绍CVTree3的论文[42]中部分地讨论了细菌巨分类的状况, 因此在这里只对细菌分类在门的水平上加以评述. NCBI[43]和BMSAB[16]等处都列举了29~30个确定的细菌门, 它们基本上限于可培养的原核生物. 但是对于尚没有合法命名和基本描述的“候补门”(candidate division), 则数目的估计差异很大, 从20个到上百个门, 主要来自对各种生态环境的16S rRNA普查[51,52]. 最近有作者提出了在16S rRNA序列的基础上可培养和未被培养菌株的统一分类建议[53]. 同基于基因组的CVTree结果比较, 才能判断统一分类的客观正确性. 从宏基因组数据提取单个菌株基因组的方法, 以及单细胞测序等技术的进步, 使得越来越多的来自未定门的菌株被测序. 目前已经有数百个这样的基因组, 但它们的质量良莠不齐. 本研究组建议把此种基因组两次送到CVTree服务器去构树, 靠第一次的结果剔除质量过差或重复的数据, 再考察第二次构树的结果. 事实上在NCBI前几年发布的基因组中, 已经有少数来自未定门. 图2显示出3个这样的基因组, 它们各自都处于独立门的水平. 然而, 后两个基因组可能属于同一个门, 甚至就是名称不同的同一个基因组. 再者, WWE3和TM7这两个未定门果真靠得如此近吗? 这些问题都要靠包含更多未定门基因组的树来分析判断. 这是今后的研究课题.

5.2 单菌种属的考察

许多新属在初次命名时只包含一个种. 随着研究发展, 属中种的数目逐渐增加. 然而, 有些属自命名以来只包含单一菌种, 特称之为单菌种属(monospecific genus, 简称MSG). 把LPSN[18]所列举的MSG数目按其发表年代画成曲线, 得到图3所示的情形. 这是“大数据”时代看问题的新视角, 传统的微生物学工作者不会想到这么做.

单种属如果插入某个其他属并且破坏后者的单源性, 那就有可能反映分类问题. 图3中最左面的点是1803年首次描述的*Beggiatoa alba*, 它的正式名称是1845年定的. 直到2015年的BMSAB[16], 它仍然被列为该属里唯一的具有合法名称的种. 然而, 在已经测序的基因组里有5个与*Beggiatoa*有关. 把这些基因组上传到CVTree以后, 计算出如下的分枝情况:

图4中有一个BMSAB[16]和LPSN[18]中都没有列举的种*Beggiatoa leptomitiformis*. 它早在1998年就被命名和发表[54], 在2015年公布了其基因组的测序结果[55]. 可能因为最初的描述[54]是用俄文发表的, 它一直没有被收入国际上的主流细菌分类文献. 无论如何, 它终结了*Beggiatoa*的单种属地位. 此外, *Thioploca*是一个1907年发表的单种属, 目前只有这一个测过序的基因组. 其他3个名字中带有_sp_的基因组, 也对应没有合法种名的菌株. 本研究组对单种属的兴趣在于它们可以作为引子, 帮助挑选出一些值得研究的问题和可能的分类修正.

5.3 LVTree上的非单源性与CVTree分枝的单源性

本文在4.2节里已经解释了单源性概念. 非单源关系包括多源性(polyphyly, 也称多系)和并源性(paraphyly, 也称并系). 对于最初来自动物分类学的这些概念的定义, 存在着长期争论. 对于原核生物, 不理会这些争论, 而把多源和并源统称之为非单源性. 在16S rRNA树上出现的大量非单源分枝, 曾经困惑过《伯杰手册》的一些前主编们, 以至于他们试图用主成分分析等手段来代替通用的构树方法[56,57]. 事实上, 这是由于靠16S rRNA序列分析不能明确界定许多属

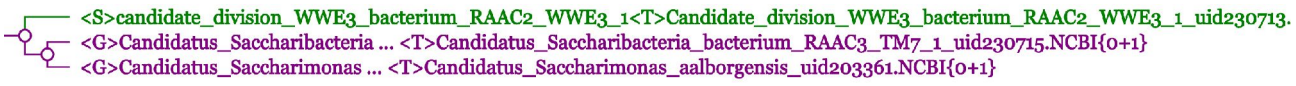

图 2　在包含10442个细菌基因组的CVTree上有3个未定门的基因组形成一枝. 它们可以暂时并为一个门(如TM7), 也可以分属2个或3个门(网络版彩图)

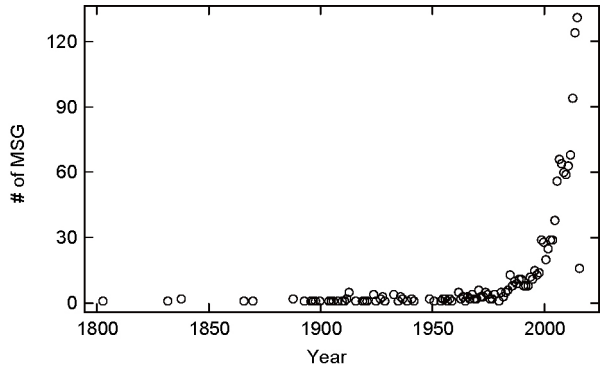

图 3　MSG数目随发表年代的变化曲线

此图根据LPSN[18]直到2016年10月底发表的数据绘制. 曲线最左端的点是1803年发表的*Beggiatoae alba*. 200多年来这个属只有这个唯一的有合法名称的种, 目前已经测序了一些有关菌株的范围. 这样, 也就更容易理解为何在《原核生物》第Ⅳ版[17]中, 把科作为划分章节的基本单元. 在基于全基因组的CVTree上, 情况有很大改善. 只举一个实例: 柄杆菌科(Caulobacteraceae)包含4个属, 它在LVTree上表现为一个单源枝, 但是在属的水平上, 只有一个*Asticcacaulis*属是单源的, 其他3个属均包含并源关系(图5).

图6是从CVTree树上剪下来的柄杆菌科分枝. 只要把不具有合法名字的Caulobacteraceae_ bacterium_ PMMR1归入*Phenylobacterium*属, 或者把它定义为一个新属, 则所有各个属都是单源的. 比较LVTree和CVTree, 这样的情形还有很多处. 总的说来, CVTree的分枝更符合现有分类系统中属的划分.

图 4　在*Beggiatoa*附近的分枝情况(网络版彩图)

这是从包含321个古菌和10441个细菌基因组的CVTree剪下的局部

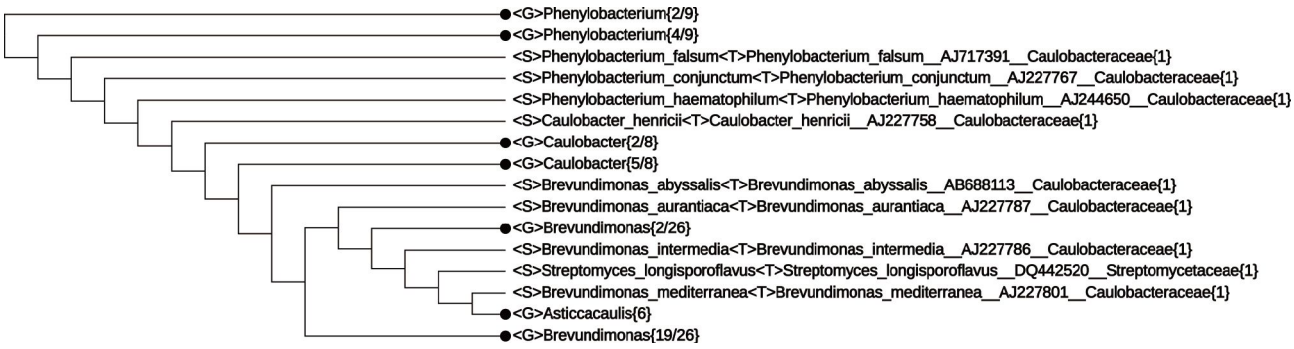

图 5　从LVTree上剪下来的柄杆菌科分枝

*Asticacaulis*属是单源的, 其他3个属处于并源关系. 在文献[47]中已经指出, *Streptomyces longisporoflavus*是一个标错的*Brevundimons*菌株. 即使纠正了这个错误, *Brevundimonas*属的单源性还是被*Asticcacaulis*破坏, 而这两个属又破坏了剩下两个属的单源性

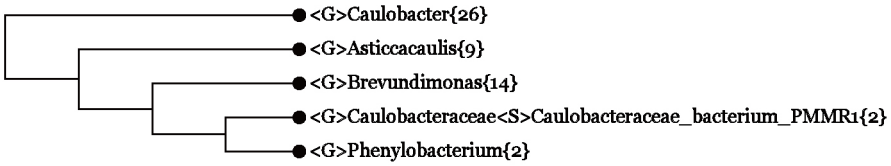

图 6　从CVTree上剪下来的柄杆菌科分枝

只要把不具有合法名称的Caulobacteraceae_bacterium_PMMR1归入*Phenylobacterium*属, 或为它建立一个新属, 所有各个属就都是单源的

6 CVTree在种和亚种以下的高分辨力

16S rRNA序列分析对种以下的菌株没有分辨能力, 这已是公认的事实[44,58], 而这正是全基因组方法可以发挥作用的领域. 上面5.3节所叙述的CVTree上的单源枝多于LVTree, 其实也是16S rRNA分析方法分辨力不够的表现. 事实上还可以举出CVTree高分辨力的更多表现.

6.1 细菌种以下的菌株分化关系

以沙门氏菌为例, 演示一下CVTree方法的分辨力. 沙门氏菌在人和动物中导致斑疹伤寒和急性肠炎两大类疾病. 1881年就学会培养沙门氏菌, 但是还不会同其他肠道菌区分. 1896年靠血清反应确定了导致斑疹伤寒的菌株. 1946年建立了靠血清型区分致病菌株的Kauffman-White系统. 人们发现多于2500种血清型, 曾经一度把血清型作为沙门氏菌的种. 但是, 1980年代DNA杂交实验又说明所有血清型属于同一沙门氏菌. 种以下菌株的命名和分类处于混乱状态, 直到2005年才基本上统一了认识, 原来沙门氏属下面只有两个种: *S. bongori*和*S. enterica*, 而后者包含6个亚种. 亚种与血清型关系如何呢? 现在已经测序了超过4500个沙门氏属的细菌基因组, 可以在CVTree上考察分枝顺序与血清型的关系. 2005年发表过一个新种*S. subterranea*, 从CVTree看它根本不属于此属, 应另行讨论. 本研究组构造了包含1424个沙门氏菌株的万株基因组大树. 图7是从这棵树上剪下来的沙门氏菌属的分枝情形. 在此图中, *S. enterica*的6个亚种全集中在从下往上数的第2行里, 即Salmonella_enterica{13/1421}之内. 图中的其他大分枝说明, 沙门氏菌中可能还可以区分出更多的种或亚种. 本研究组愿意同对此问题有兴趣的学者合作, 开展更大规模的沙门氏菌属内部关系的研究.

另一个例子是金黄葡萄球菌(*Staphylococcus aureus*). 除了与人类和平共处的菌株, 还有引起皮肤及其他疾病的菌株, 特别是手术后引起交叉感染的抗药菌株, 已经在医院中造成日益严重的问题. 现在已经完成测序的金黄葡萄球菌基因组总数超过了6800个. 用CVTree构建这些菌株的分枝顺序, 只是计算时间问题. 更重要的是把对这些菌株的实验研究结果和临床记录, 与菌株的分化过程联系起来. 期待与有关研究团队合作.

6.2 细菌的种群遗传学研究

比起真核生物, 对原核生物种群遗传学的研究极

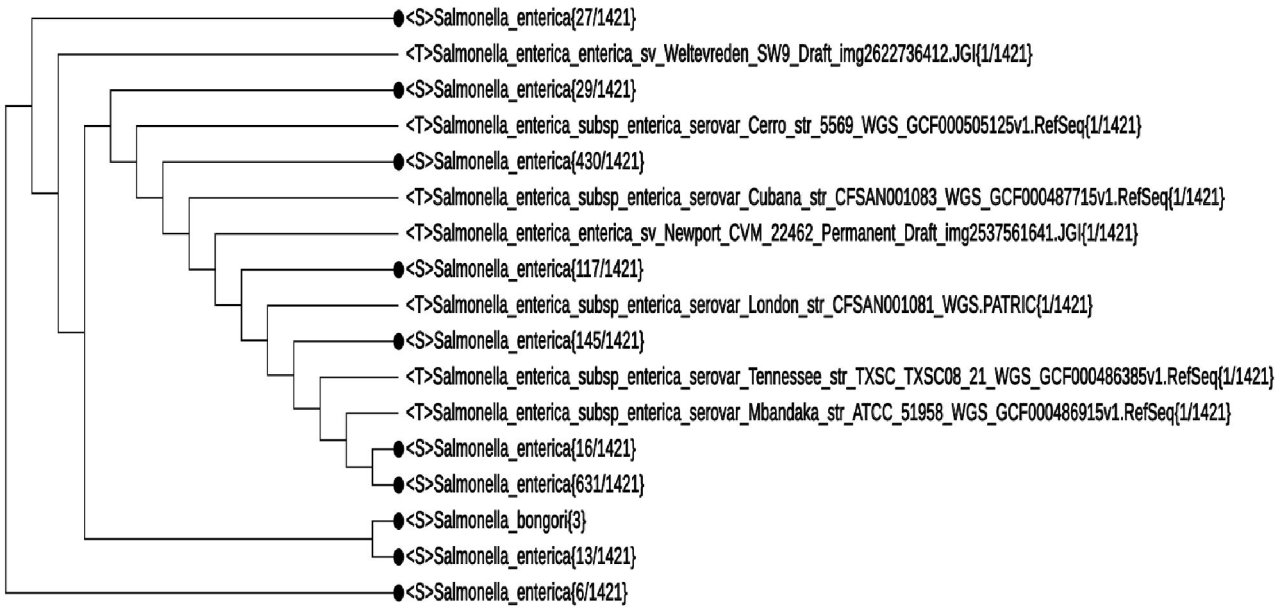

图 7　从万株基因组的CVTree上剪下来的*Salmonella*属分枝

这是由3个*S. bongori*和1421个*S. enterica*基因组构成的单源枝. 代表后者6个亚种的基因组全在从下往上数的第2行, 即Salmonella_enterica{13/1421}里

其不足. 基本上只对人体内的无害的大肠杆菌种群有一些结果[59], 这涉及亲缘组(phylogroup)A, B1和B2. CVTree上的*Escherichia coli*分枝不仅清楚对应这些无害亲缘组, 而且同致病亲缘组D, E等的划分也保持一致[42]. 但是, 大肠杆菌的血清型研究导致了数百个类型的更细致的划分, 它们与CVTree分枝的关联并不清楚. 同样是血清型, 现有的19个化脓性链球菌(*Streptococcus pyogenes*)基因组在CVTree上的分枝与血清型划分完全一致, 将来基因组数目曾多以后此种一致性能否保持, 仍有待观察. 肺炎链球菌(*S. pneumoniae*)的血清型和它们在CVTree上的分枝只是有关联, 但并不一致. 结核分支杆菌(*Mycobacterium tuberculosis*)的情形也是如此. 在临床微生物学中历史悠久的血清型检验, 或许将来会被基因组测序后的分枝顺序所取代.

6.3 细菌的生物地理

宏观动植物对地理环境的适应, 曾经对达尔文提出演化理论起过启发作用. 细菌是否也有地理分布和随环境变异的情形? 2003年发表在《科学》(*Science*)上的一篇论文[60]指出, 幽门螺旋菌(*Helicobacter pylori*)的变异反映了其人类宿主的迁徙. 现在至少有550多条幽门螺旋菌基因组已经测序, 对此问题可以进行更深入的研究. 一个更直接的, 不受中间宿主影响的实例, 来自冰岛硫叶菌("*Sulfolobus islandicus*", 这个名字尚未合法发表, 因而按"规则"要求打上引号). 现在此菌已经有多个采自欧亚大陆和北美大陆不同热温泉的菌株被测序. 使用DNA电子杂交和CVTree等多种分子方法的研究表明, 这些菌株还没有分化成不同的种, 尚属于不同的地理变种(geovars)[61].

6.4 用DNA杂交不能区分的菌株

自20世纪80年代以后, DNA-DNA杂交成为区分不同细菌种的标准方法. 然而, 有些临床上可以明确区分的菌株, 用DNA杂交却无法分辨. 例如, 耶尔森氏菌属的*Yersinia pseudotuberculosis*和*Yersinia pestis*两个种就不能用DNA区分, 于是有人建议把它们并成同一个种. 然而, 这个建议被细菌分类的法律委员会否定, 理由是为了避免在医生中引起危害公众健康的误解. 在CVTree上这两个种的菌株明确分开, 足以解除法律委员会的顾虑[62]. 又如大肠杆菌和几种志贺氏菌(*Shigella*)的关系, 许多鉴定方法都把志贺氏菌放进大肠杆菌的分枝内部, 许多人认为只是为了历史和临床原因, 才为它们保留了不同的名字. 不过, 在CVTree上却没有问题: 志贺氏菌和大肠杆菌一样, 都是埃希氏菌属(*Escherichia*)下的平等成员. 这一结论[63]虽然尚未被普遍接受, 今后的发展将继续检验其正确性.

6.5 对特定菌株的电子筛选

对同一种细菌的大量变异菌株进行筛选, 以便寻求在致病性、抗药性、代谢产物等方面具有特异表现的菌株, 乃是一种成本高昂、费时耗力的过程. 然而, 在积累了相当数量的实验结果以后, 可以把新菌株的基因组同已经研究过的菌株基因组混合构建CVTree, 并且把以往的筛选结果标注到各个分枝上, 这将能够帮助判断效果较好的变异方向, 提高筛选效率. 随着测序成本降低, 这类电子筛选方法将受到更多重视.

本文所论及的许多问题, 都还不是系统的研究结果, 而是需要进一步发展的研究方向. 希望这篇综述能引起广大微生物工作者的注意.

参考文献

1 庄子, 《知北游》
2 郝柏林, 刘寄星, 主编. 理论物理与生命科学. 上海: 上海科学技术出版社, 1997
3 Woese C R, Fox G E. Phylogenetic structure of the prokaryotic domains: the primary kingdoms. Proc Natl Acad Sci USA, 1977, 74: 5088–5090
4 Woese C R, Kandler O, Wheelis M L. Towards a natural system of organisms: proposal for the domains of Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA, 1990, 87: 4576–4579
5 Whitman W B, Coleman D C, Wiebe W J. Prokaryotes: the unseen majority. Proc Natl Acad Sci USA, 1998, 95: 6578–6583
6 Microbiology by numbers. Nat Rev Microbiol, 2011, 9: 628
7 Chun J, Rainey F A. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol, 2014, 64: 316–324

8 Lapage S D, Sneath P H A, Lessel E F, et al. International Code of Nomenclature of Bacteria (1990 Revision). Washington: American Society for Microbiology, 1992
9 Parker C T, Tindall B J, Garrity G M. International code of nomenclature of prokaryotes. Int J Syst Evol Microbiol, 2015, 1: 465–481
10 Breed R S, Murray E G D, Smith N R, et al. Bergey's Manual of Determinative Bacteriology. 7th ed. Baltimore: The Williams & Wilkins Co., 1957
11 Buchanan R E, Gibbons N E. Bergey's Manual of Determinative Bacteriology. 8th ed. Baltimore: The Williams & Wilkins Co., 1994
12 Bergey's Manual of Systematic Bacteriology. Baltimore: The Williams & Wilkins Co., 1984–1989. 1–4
13 Fox G E, Pechman K R, Woese C R. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. Int J Syst Bacteriol, 1977, 27: 44–57
14 Fox G E, Magrum L J, Balch W E, et al. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci USA, 1977, 74: 4537–4541
15 Bergey's Manual Trust. Bergey's Manual of Systematic Bacteriology. 2nd ed. New York Dordrecht Heidelberg London: Springer, 2001–2012. 1–5
16 Bergey's Manual Trust. Bergey's Manual of Systematics of Archaea and Bacteria. Hoboken: John Wiley & Sons, 2015
17 Rosenberg E, Editor-in-Chief. The Prokaryotes. 4th ed. Heidelberg New York Dordrecht London: Springer, 2013–2014. 1–11
18 Parte A C. LPSN—list of prokaryotic names with standing in nomenclature. Nucleic Acids Res, 2014, 42: D613–D616
19 Fleischmann R D, Adams M D, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus* influenzae Rd. Science, 1995, 269: 496–512
20 Fraser C M, Gocayne J D, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science, 1995, 270: 397–404
21 Genomes OnLine Database. [2016-10-31]. https://gold.jgi.doe.gov/statistics/
22 Whitman W B. Intent of the nomenclatural Code and recommendations about naming new species based on genomic sequences. Bull BISMiS, 2011, 2: 135–139
23 Ngo H T, Yin C S. *Luteimonas terrae* sp. nov., isolated from rhizosphere soil of *Radix ophiopogonis*. Int J Syst Evol Microbiol, 2016, 66: 1920–1925
24 Wayne L G, Brenner D J, Colwell R R, et al. Report of the Ad Hoc Committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol, 1987, 77: 463–464
25 Woese C R. A manifesto for microbial genomes. Curr Biol, 1998, 8: R781–R783
26 Woese C R. The quest for Darwin's grail. ASM News, 1999, 65: 260–263
27 Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature, 2009, 462: 1056–1060
28 Kyrpides N C, Hugenholtz P, Eisen J A, et al. Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. PLoS Biol, 2014, 12: e1001920
29 Han K, Li Z F, Peng R, et al. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. Sci Rep, 2013, 3: 2101
30 Kimura K. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press, 1983
31 Hao B, Qi J, Wang B. Prokaryotic phylogeny based on complete genomes without sequence alignment. Mod Phys Lett B, 2003, 17: 91–94
32 Qi J, Wang B, Hao B. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Biol, 2004, 58: 1–11
33 Gao L, Qi J, Hao B. Simple Markov subtraction essentially improves prokaryote phylogeny. AAPPS Bull, 2006, 16: 3–7
34 李强. 关于K串组成的一个试探性的进化模型以及序列的唯一重建问题. 博士学位论文. 上海: 复旦大学, 2009
35 Zuo G, Li Q, Hao B. On K-peptide length in composition vector phylogeny of prokaryotes. Comput Biol Chem, 2014, 53: 166–173
36 李强, 左光宏, 郝柏林. 从完全基因组出发建立原核生物亲缘关系和分类系统时遇到的数学问题. 中国科学: 物理学 力学 天文学, 2014, 44: 1301–1310
37 郝柏林. 来自基因组的一些数学. 上海: 上海科技教育出版社, 2015
38 Hu R, Wang B. Statistically significant strings are related to regulatory elements in the promoter region of *Saccharomyces cerevisiae*. Physica A, 2001, 290: 464–474
39 Yu Z G, Zhou L Q, Anh V V, et al. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment. J Mol Evol, 2005, 60: 538–545
40 Qi J, Luo H, Hao B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res, 2004, 32: W45–W47
41 Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res, 2009, 37: W174–W178

42 Zuo G, Hao B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. Genomics Proteomics Bioinformatics, 2015, 13: 321–331

43 The NCBI Taxonomy database. [2016-10-31]. https://www.ncbi.nlm.nih.gov/taxonomy

44 Yarza P, Richter M, Peplies J, et al. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol, 2008, 31: 241–250

45 Yarza P, Spröer C, Swiderski J, et al. Sequencing orphan species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst Appl Microbiol, 2013, 36: 69–73

46 All-Species Living Tree project. [2016-10-31]. http://www.arb-silva.de/projects/living-tree/

47 Zuo G, Zhi X, Xu Z, et al. LVTree viewer: an interactive display for the all-species living tree incorporating automatic comparison with prokaryotic systematics. Genomics Proteomics Bioinformatics, 2016, 14: 94–102

48 Garrity G M, Whitman W B, Schink B, et al. Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol, 2015, 65: 4284–4287

49 Cavalier-Smith T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol, 2002, 52: 7–76

50 Zuo G, Xu Z, Hao B. Phylogeny and taxonomy of archaea: a comparison of the whole-genome-based CVTree approach with 16S rRNA sequence analysis. Life, 2015, 5: 949–968

51 Youssef N H, Couger M B, McCully A L, et al. Assessing the global phylum level diversity within the bacterial domain: a review. J Adv Res, 2015, 6: 269–282

52 Schloss P D, Handelsman J. Status of the microbial census. Microbiol Mol Biol Rev, 2004, 68: 686–691

53 Yarza P, Yilmaz P, Pruesse E, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Micro, 2014, 12: 635–645

54 Grabovich M Y, Churikova V V, Dubinina G A, et al. Mixotrophic and lithoheterotrophic of the freshwater filamentous sulfur bacterium *Beggiatoa leptomitiformis* D-402. Mikrobiologia, 1998, 67: 383–388

55 Fomenkov A, Vincze T, Grabovich M Y, et al. Complete genome sequence of the freshwater colorless sulfur bacterium *Beggiatoa leptomitiformis* Neotype Strain D-402^T. Genome Announc, 2015, 3: e01436-15

56 Garrity GM, Lilburn TG. Mapping taxonomic space: an overview of the road map to the second edition of Bergey's Manual of Systematic Bacteriology. WFCC Newsl, 2002, 35: 5–15

57 Garrity G M, Lilburn T G. Self-organizing and self-correcting classifications of biological data. Bioinformatics, 2005, 21: 2309–2314

58 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

59 Tenaillon O, Skurnik D, Picard B, et al. The population genetics of commensal *Escherichia coli*. Nat Rev Micro, 2010, 8: 207–217

60 Falush D, Wirth T, Linz B, et al. Traces of human migrations in *Helicobacter pylori* populations. Science, 2003, 299: 1582–1585

61 Zuo G, Hao B, Staley J T. Geographic divergence of "*Sulfolobus islandicus*" strains assessed by genomic analyses including electronic DNA hybridization confirms they are geovars. Antonie van Leeuwenhoek, 2014, 105: 431–435

62 Hao B. CVTrees support the Bergey's systematics and provide high resolution at species levels and below. Bull BISMiS, 2011, 2: 189–196

63 Zuo G, Xu Z, Hao B. Shigella strains are not clones of *Escherichia coli* but sister species in the genus *Escherichia*. Genomics Proteomics Bioinformatics, 2013, 11: 61–65

Computational microbiology in genomic era

ZUO GuangHong[1] & HAO BaiLin[1,2]

1 Department of Physics and T-Life Research Center, Fudan University, Shanghai 200433, China;
2 Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China

Microbial genomes and RNA sequences comprise an important part of biological big data. CVTree is a whole-genome based and alignment-free method whereas LVTree is based on alignment of 16S rRNA sequences. They provide two independent ways to construct phylogenetic trees and to extract taxonomic information for prokaryotes. The automation of these two approaches make the study of prokaryotic phylogeny and classification by-product of big data analysis and come as a rescue to the declining discipline of taxonomy. Especially, the whole-genome based CVTree not only provides a tool for large-scale study but also possesses high resolution power at the species level and below, a distinctive feature beyond the reach of 16S rRNA sequence analysis. These methods taken together may open new directions in microbiological research. This paper is a brief review of our recent work.

prokaryotic genomes, phylogeny, taxonomy, alignment-free comparison, CVTree, LVTree

doi: 10.1360/N052016-00312