

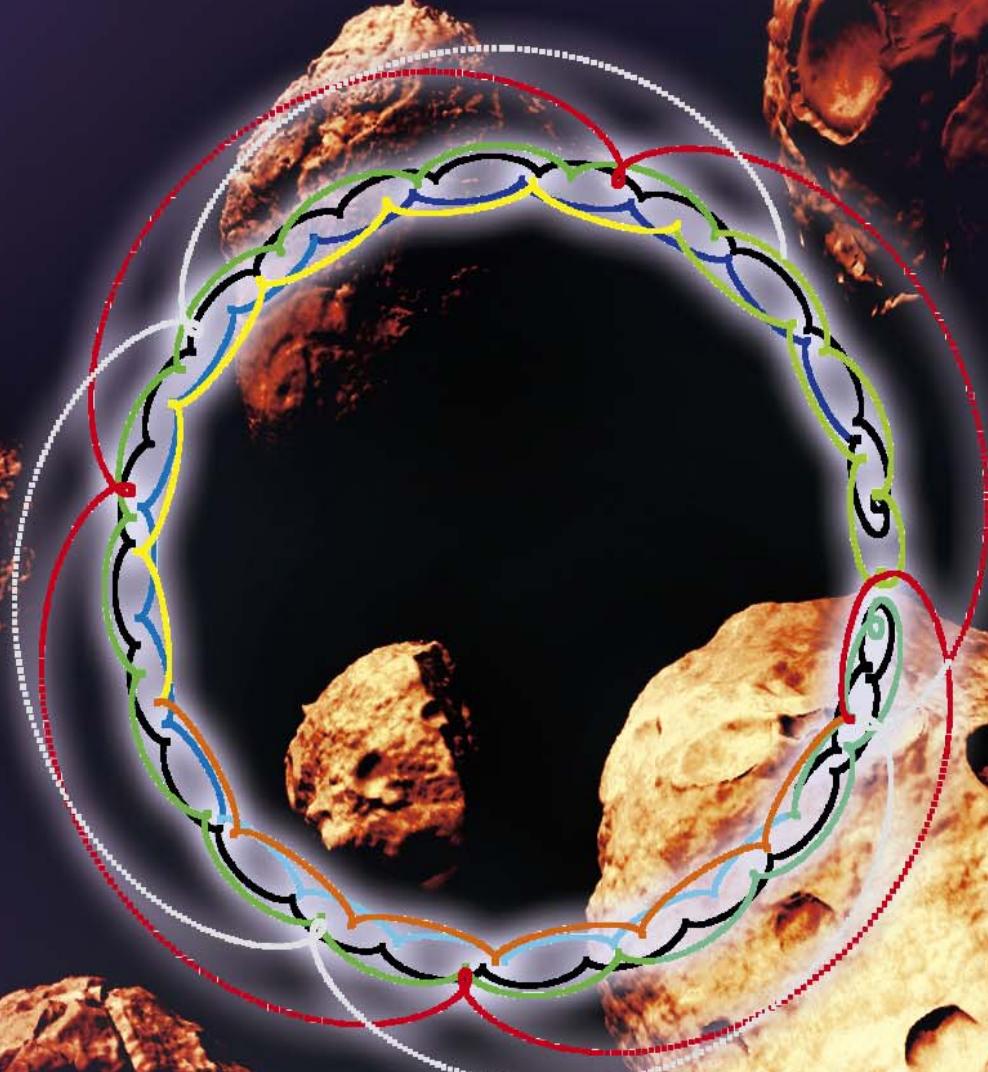
SCIENTIA SINICA Physica, Mechanica & Astronomica

# 中国科学

物理学  
力学 天文学

第44卷 第12期 2014年12月 1251-1368 ISSN 1674-7275 CN 11-5848/N

非线性科学专辑



中国科学院  
国家自然科学基金委员会

# 《中国科学》《科学通报》

荣誉总主编: 周光召      总主编: 朱作言

## 《中国科学: 物理学 力学 天文学》编辑委员会

主 编: 王鼎盛 中国科学院物理研究所

张 杰 上海交通大学

常务副主编: 龚旗煌 北京大学物理学院

副 主 编:

龙桂鲁 清华大学物理系

李树深 中国科学院半导体研究所

张肇西 中国科学院理论物理研究所

季向东 上海交通大学物理与天文系

金晓峰 复旦大学物理系

胡更开 北京理工大学宇航学院

洪友士 中国科学院力学研究所

徐仁新 北京大学物理学院

符 松 清华大学航天航空学院

景益鹏 上海交通大学物理与天文系

廖新浩 中国科学院上海天文台

编 委:

### 物理 I: 凝聚态物理、原子分子物理、光物理和声学等

王力军 清华大学物理系

王 炜 南京大学物理学院

吕正红 加拿大多伦多大学材料科学与工程学院

刘 明 中国科学院微电子研究所

杨金龙 中国科学技术大学化学物理系

汪卫华 中国科学院物理研究所

张仁和 中国科学院声学研究所

张 海 澜 中国科学院声学研究所

张淑仪 南京大学物理学院声学研究所

张 靖 山西大学光电研究所

陈难先 清华大学凝聚态物理中心

金奎娟 中国科学院物理研究所

施 靖 美国加利福尼亚大学河滨分校物理和天文系

闻海虎 南京大学物理学院

洪明辉 新加坡国立大学工学院电器和计算机工程系

袁建民 国防科学技术大学理学院物理系

蒋红兵 北京大学物理学院

谢心澄 北京大学物理学院

戴 宁 中国科学院上海技术物理研究所

### 物理 II: 理论物理、粒子物理、核物理等

王新年 美国劳伦斯伯克利实验室

左 维 中国科学院近代物理研究所

叶沿林 北京大学物理学院

任中洲 南京大学物理学院

庄鹏飞 清华大学物理系

许 怒 美国劳伦斯伯克利实验室

孙昌璞 中国工程物理研究院北京计算科学研究中心

李 翳 中国科学院理论物理研究所

陈晓松 中国科学院理论物理研究所

罗民兴 浙江大学物理系 浙江近代物理中心

周善贵 中国科学院理论物理研究所

赵 鸿 厦门大学物理系

胡红波 中国科学院高能物理研究所

柳卫平 中国原子能科学研究院

费少明 首都师范大学数学科学学院

姚为民 美国劳伦斯伯克利实验室

贺贤土 北京应用物理与计算数学研究所

唐孝威 浙江大学物理系

黄超光 中国科学院高能物理研究所

盛政明 上海交通大学物理与天文系

## 力学

龙 勉	中国科学院力学研究所	仲 政	同济大学航空航天与力学学院
李少凡	美国加州大学伯克利分校城市与环境工程系	李俊峰	清华大学航天航空学院
李家春	中国科学院力学研究所	杨基明	中国科学技术大学近代力学系
吴锤结	大连理工大学航空航天学院	邱志平	北京航空航天大学固体力学研究所
余振苏	北京大学工学院	罗 宏	美国北卡罗莱纳州立大学机械和航空航天工程系
金栋平	南京航空航天大学机械结构力学及控制国家重点实验室	郑晓静	西安电子科技大学
赵亚溥	中国科学院力学研究所	胡 晖	美国爱荷华州立大学航空航天工程系
茹重庆	加拿大阿尔伯特大学机械工程系	廖世俊	上海交通大学船舶海洋与建筑工程学院

## 天文学

毛淑德	中国科学院国家天文台	朱 紫	南京大学天文与空间科学学院
汲培文	国家自然科学基金委员会	严 镛 璞	美国密苏里大学哥伦比亚分校物理与天文系
李立新	北京大学科维理天文与天体物理研究所	李爱根	美国密苏里大学哥伦比亚分校物理与天文系
肖 龙	中国地质大学(武汉)地球科学学院	邹振隆	中国科学院国家天文台
张双南	中国科学院高能物理研究所	张思炯	中国科学院南京天文光学技术研究所
陈鹏飞	南京大学天文与空间科学学院	周又元	中国科学院国家天文台
钱永忠	美国明尼苏达大学物理与天文系	高 煜	中国科学院紫金山天文台
葛 健	美国佛罗里达大学天文系	韩占文	中国科学院云南天文台

## 《中国科学：物理学 力学 天文学》编辑部

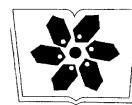
地 址:	北京东黄城根北街 16 号	《中国科学》杂志社, 100717
网 址:	www.scichina.com	phys.scichina.com
电 话:	(010) 64015835 (编辑部)	电子信箱: physics@scichina.org (编辑部)
	(010) 64019709 (发行部)	sales@scichina.org (发行部)
	(010) 64008316 (广告部)	ads@scichina.org (广告部)
传 真:	(010) 64016350	
主 任:	黄延红	
责任编辑:	王 维 朱全娥 侯修洲 郭媛媛	
封面设计:	胡 煜	



链接到网站



国家自然科学基金资助



中国科学院科学出版基金资助

第 44 卷 第 12 期 2014 年 12 月

## 目 次

### 非线性科学专辑

#### 前言

- “非线性科学”专辑·前言 ..... 1251  
王炜, 孙义燧

#### 评述

- 海王星特洛伊小天体动力学 ..... 1252  
周礼勇, 管璞, 孙义燧

- 准脆性固体的灾变破坏及其物理前兆 ..... 1262  
郝圣旺, 白以龙, 夏蒙梦, 柯孚久

- 大气、海洋动力学中一些非线性偏微分方程的研究 ..... 1275  
郭柏灵, 黄代文, 黄春研

- 弱 KAM 理论和 Hamilton-Jacobi 方程 ..... 1286  
李霞, 严军

- 三维反应扩散系统中的螺旋波和失稳控制 ..... 1291  
王宏利, 欧阳颀

- 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 ..... 1301  
李强, 左光宏, 郝柏林

---

**封面说明** 非线性科学研究方兴未艾, 引力作用下的天体运动是最广为人知的非线性运动例子之一。即使是最简单的天体力学模型, 也可能表现出复杂的非线性特征。圆型限制性三体问题中, 行星绕太阳的公转轨道被简化为圆形, 而太阳系中的小天体则被简化为在它们的引力场中运动的零质量粒子。封面在为数众多的小天体背景上显示了一条小天体由束缚于行星轨道的位置(黑色至棕色)经历多个共振并最终逃逸(红色和灰色)的复杂运动过程, 详见周礼勇等人文(P1252).

---

p53 信号网络的非线性动力学研究.....	1311
张小鹏, 刘锋, 王炜	
由基因调控网络数据分析揭示振荡斑图的功能结构.....	1319
张朝阳, 黄旭辉, 郑志刚, 胡岗	
基于复杂网络的信号检测与传递.....	1334
刘宗华	

## 论文

相对论强激光在等离子体中传输的成丝不稳定性和斑图动力学 .....	1344
黄太武, 周沧涛, 贺贤土	
强激光场氦原子非序列双电离过程中光子动量分配.....	1356
陶建飞, 刘杰	
椭圆偏振场中原子隧穿电离的残存问题研究 .....	1363
黄凯云, 傅立斌, 刘杰	



## 非线性科学专辑 · 评述

# 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

李强<sup>①②††</sup>, 左光宏<sup>①††</sup>, 郝柏林<sup>①\*</sup>

① 复旦大学物理系和理论生命科学研究中心, 上海 200433;

② 科学院和马普学会计算生物学伙伴研究所, 上海 200031

\* 联系人, E-mail: hao@mail.itp.ac.cn

†† 同等贡献

收稿日期: 2014-04-23; 接受日期: 2014-07-14

国家重点基础研究发展计划(编号: 2007CB814800, 2013CB834100)、上海市基础研究计划(编号: B111)和复旦大学物理系应用表面物理国家重点实验室资助项目

**摘要** 我们课题组近 10 年所发展的组分矢量(CVTree)方法, 已经成为从完全基因组出发而不使用序列联配来构建细菌亲缘关系的有效手段。在整个生物分类学日益后继乏人的背景下, 组分矢量(CVTree)方法伴随着基因组时代的发展, 成为人人可以方便使用的微生物分类的定义性工具。本综述不讨论 CVTree 的生物学结果, 而着重从非线性科学的角度, 介绍我们在研究过程中所提出和解决的数学问题。这将涉及组合学、图论、形式语言学等离散数学的篇章。我们特别希望, 本文所列举的一些尚未解决的数学问题能引发新的研究工作, 使 CVTree 方法的理论基础更为坚实。

**关键词** 无序列比对, 全基因组亲缘关系, 原核生物, 组合学, 欧拉圈, 可因式化语言

**PACS:** 02.50.Ga, 02.10.Ox, 02.70.Rr

**doi:** 10.1360/SSPMA2014-00124

## 1 引言

原核生物(Prokaryote)是古菌(Archaea)和细菌(Bacteria, 过去曾称为真细菌Eubacteria)的总称。这些单细胞微生物同真核生物一起并居三个生命超界中的两个, 在生物演化过程和地球生态系统中起着重要作用。然而, 由于形态特征有限, 原核生物的亲缘关系和分类系统的研究曾长期落后于多细胞动植物。特别由于缺少一切原核生物所共有的蛋白质序列, 分子水平上的演化研究更曾停滞不前。直到 1970

年代后期, Woese 及合作者<sup>[1]</sup>采用核糖体小亚基的 16S rRNA 序列作比较分析, 这一领域开始迅速发展。

历时 12 年才出版齐全的《伯杰系统细菌学手册》第二版<sup>[2]</sup>已经尽可能基于 16S rRNA 序列分析。这一事实隐含的原则性问题是: 现代原核生物的亲缘关系和分类系统愈益“合二为一”, 对它们的正确性必须有客观检验, 这种检验在数据和方法两方面都应当独立于传统的 16S rRNA 序列分析。

我们组所发展的组分矢量(简称 CVTree)方

**引用格式:** 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  
Li Q, Zuo G H, Hao B L. Some mathematical problems inspired by the study of whole-genome-based phylogeny and taxonomy of prokaryote (in Chinese). Sci Sin-Phys Mech Astron, 2014, 44: 1301–1310, doi: 10.1360/SSPMA2014-00124

法<sup>[3-8]</sup>, 不仅提供了对 16S rRNA 序列分析的独立检验, 而且在分辨能力上远远超过后者<sup>[9-11]</sup>. CVTree 方法已经成功地应用于构建病毒<sup>[12,13]</sup>、叶绿体<sup>[14]</sup>、细菌和古菌<sup>[6,15,16]</sup>, 以及真菌<sup>[17]</sup>的亲缘关系.

作为一种全新的方法, CVTree 的基础还有待分析与奠定. 近十年来, 我们已经提出和解决了一批问题<sup>[18-22]</sup>. 本文在综述这些结果的同时, 还要指出一些尚未研究的课题.

## 2 组分矢量方法概述

组分矢量方法有三个区别于其他构树途径的主要特点: 它基于全基因组数据; 它不使用序列联配; 它的正确性靠直接同分类系统比较来检验.

基于全基因组, 就避免了选取序列片段、特别是同源蛋白所带来的主观任意性. 基因横向传递不再是构树的严重障碍, 而仅仅是基因组演化的一种机制. 虽然测序对象的选择主要取决于功利考虑, 但测序技术日益提高、成本不断降低, 已经使得有全基因组做代表的原核生物的分类覆盖相当广泛. 基于全基因组的亲缘关系和分类系统研究已经现实可行.

原核生物基因组极其多样. 生殖道支原体 (*Mycoplasma genitalium*) 的基因组只有 48 万碱基对和不到 500 种蛋白质; 目前已经测序的最大基因组来自纤维堆囊菌 (*Sorangium cellulosum*), 它有 1300 多万碱基对、编码 9380 个基因. 如何“联配”尺寸如此悬殊的序列呢?

为了不用序列联配来比较基因组, 我们取每个基因组所编码的全部蛋白质产物, 并确定一个小整数  $K$ . 用宽度为  $K$  的滑动窗口, 扫过每条由  $L_i$  个氨基酸组成的蛋白质序列, 得到  $\sum_{i=1}^M (L_i - K + 1)$  个  $K$ -肽的集合, 其中  $i = 1, 2, \dots, M$  是蛋白质序列的编号. 把一个基因组所编码的全部蛋白质中的氨基酸总数记为  $L = \sum_{i=1}^M L_i$ . 对于固定的  $K$ , 最多有  $20^K$  种  $K$ -肽. 把它们按氨基酸字母的字典顺序排列, 填入相应的  $K$ -肽计数, 构造出一个具有  $20^K$  个分量的组分矢量 (Composition Vector, CV). 需要说明的是, 使用蛋白质序列作为我们研究的对象, 是经过大量的探索之后才确定下来的. 最初我们尝试使用全基因组 DNA

序列, 以及编码部分即 CDS 做, 虽然也能得到一些有意义的结果, 但是总体来说并不是特别理想.

设物种  $A$  和  $B$  的组分矢量分别是  $A(a_1, a_2, \dots, a_{20^K})$  和  $B(b_1, b_2, \dots, b_{20^K})$ . 它们之间的关联  $C(A, B)$  由标量积决定:

$$C(A, B) = \frac{\sum_{i=1}^{20^K} a_i b_i}{\sqrt{\sum_{i=1}^{20^K} a_i^2} \sqrt{\sum_{j=1}^{20^K} b_j^2}}. \quad (1)$$

$C(A, B)$  是已经归一的关联, 它的变化范围是  $[-1, 1]$ . 进一步定义物种  $A$  和  $B$  之间的关联“距离”或“非相似性”(Dissimilarity):

$$D(A, B) = \frac{1 - C(A, B)}{2}, \quad (2)$$

$D(A, B)$  的变化范围是  $[0, 1]$ . “距离”二字加上引号的原因将在后面解释.

直接使用由公式 (2) 得到的距离矩阵来构树, 结果并不好. 其原因在于, 由简单计数构造的组分矢量中, 包含着中性突变所导致的与物种分化关系不大的贡献. 为了突出物种特异性, 必须设法减除掉中性突变造成的背景. 根据木村资生的中性演化理论<sup>[23]</sup>, 突变在分子水平上随机发生, 因此可以采用某种统计模型来扣除背景. 我们采用  $K - 2$  阶的马可夫预测<sup>[3,6,19]</sup>, 从实际数出来的  $K - 2$  肽和  $K - 1$  肽的数目来预测每一种  $K$  肽的数目. 然后取预测和实际计数的差值作为新的组分矢量的分量. 我们使用的公式是

$$f^0(\alpha_1 \alpha_2 \cdots \alpha_K) = C \frac{f(\alpha_1 \alpha_2 \cdots \alpha_{K-1}) f(\alpha_2 \alpha_3 \cdots \alpha_K)}{f(\alpha_2 \alpha_3 \cdots \alpha_{K-1})}, \quad (3)$$

其中  $\alpha_i$  是 20 个氨基酸字母之一. 公式 (3) 右端用到两种  $K - 1$  肽和一种  $K - 2$  肽的实际出现频度, 而左面是预测出来的  $K$  肽的频度  $f^0$ ; 常数  $C$  来自概率和频度之间的变换:

$$C = \frac{\sum_i (L_i - K + 1) \sum_j (L_j - K + 3)}{\sum_i (L_i - K + 2)^2}, \quad (4)$$

当蛋白质的长度  $L_i \gg K$  时,  $C$  很接近 1. 除了导致好的结果, 公式 (3) 的优点是可以用两种独立的办法推导出来: 或是借助联合概率与条件概率的关系加上马可夫假定<sup>[6,19]</sup>, 或是利用最大熵原理<sup>[24]</sup>. 我们取

$K$  肽的实际计数和预测值之差  $(f - f^0)/f^0$ , 作为组分矢量的新分量. 前面公式(1)和(2)里面, 就是使用这样“重正化”以后的组分矢量来代表物种 A 和 B.

### 3 蛋白质序列分解为 $K$ -肽及重构问题

一条由  $L$  个氨基酸组成的蛋白质序列, 使用宽度为  $K$  的滑动窗口, 总可以分解成  $L - K + 1$  个  $K$ -肽. 现在提出一个逆问题: 给定这一批  $K$ -肽, 要求重新构成长度为  $L$  的氨基酸序列, 每条  $K$ -肽必须而且只许用一次, 全部用完. 这个逆问题是有解的, 因为至少可以回到作为出发点的那条蛋白质序列. 问题: 重构是否唯一? 显然,  $K$  足够大时重构唯一. 不唯一时有多少重构序列? 把  $K$  换成  $K+1$  时, 重构数目会减少, 在何种  $K$  值下, 重构成为唯一的?

谢惠民指出<sup>[27]</sup>, 上述重构数目可由图论中的有向欧拉圈数目决定. 我们以一条具体的蛋白质为例, 加以说明. 下面是一种冬季比目鱼的抗冻蛋白质前体 (PDB 蛋白质结构数据库序列号 ANPA\_PSEAM), 它只包含 82 个氨基酸:

MALSLFTVGQLIFLFDWTMRITEASPDPAAKAAPA  
AAAAAPAAAAPDTASDAAAAAALTAANAKAAAELTAA  
NAAAAAAATARG

给定  $K = 5$ , 把第 1 个 5 肽看做由前 4 个字母代表的状态向后 4 个字母代表的状态的跃迁:

$MALSL : MALS \rightarrow ALSL,$

沿蛋白质序列右移一个字母, 把第 2 个 5 肽看做由前 4 个字母代表的状态向后 4 个字母代表的状态的跃迁:

$ALSLF : ALSL \rightarrow LSLF,$

前一个跃迁的末态自然就是后一个跃迁的初态. 如法炮制, 直到最后一个 5 肽代表的跃迁:

$ATARG : ATAR \rightarrow TARG.$

把每个状态画成一个图的顶点; 如果遇到重复出现的状态, 就只画一次顶点, 而把相应的跃迁箭头指回

来; 最后用一条辅助跃迁, 从最后一个末态指向第 1 个初态. 这样一来, 一条蛋白质导致一条封闭的路经, 同时就定义了一个欧拉图. 那条路经就是图上的一个封闭的欧拉圈. 问题归结为, 同一个欧拉图上还有没有其他的欧拉圈? 一共有多少个不同的欧拉圈?

其实在上述做图过程中, 不必保留所有的顶点. 例如, 一串进出各一次的顶点可以只留一个, 而不影响欧拉圈的数目. 这条 82 个氨基酸的蛋白质, 最终对应如图 1 所示的只有 9 个顶点的欧拉图:

欧拉圈的数目是图论中解决得比较好的问题. 图 1 的结构由两个矩阵描述: 一个对角的度矩阵

$$\mathbf{M} = \text{diag}(d_1, d_2, \dots, d_9);$$

其中  $d_i$  是第  $i$  个顶点的度 (对于欧拉图: 入度 = 出度 = 度). 另一个连接矩阵  $\mathbf{A} = \{a_{ij}\}$ , 其元素等于从顶点  $i$  连到顶点  $j$  的弧线数目. 两者之差  $\mathbf{C} = \mathbf{M} - \mathbf{A}$  称为克希荷夫矩阵 (也叫做拉普拉斯矩阵). 具体到图 1 有:

$$\mathbf{C} =$$

$$\left\{ \begin{array}{cccccccccc} 2, & -1, & 0, & 0, & 0, & 0, & -1, & 0, & 0, \\ 0, & 2, & -2, & 0, & 0, & 0, & 0, & 0, & 0, \\ 0, & 0, & 2, & -2, & 0, & 0, & 0, & 0, & 0, \\ 0, & 0, & 0, & 2, & -2, & 0, & 0, & 0, & 0, \\ -1, & 0, & 0, & 0, & 4, & -2, & -1, & 0, & 0, \\ 0, & -1, & 0, & 0, & -1, & 2, & 0, & 0, & 0, \\ 0, & 0, & 0, & 0, & 0, & 0, & 2, & -2, & 0, \\ 0, & 0, & 0, & 0, & 0, & 0, & 0, & 2, & -2, \\ -1, & 0, & 0, & 0, & -1, & 0, & 0, & 0, & 2, \end{array} \right\}.$$

克希荷夫矩阵的特点是它所有的代数余子式  $\Delta$  都相同. Mathematica 软件里有专门计算代数余子式的函数, 结果是  $\Delta = 192$ . 如果图中没有从一个顶点回到自身的弧线, 即  $a_{ii} = 0, \forall i$ , 任何两个顶点之间也没有并行的两条或更多弧线, 即  $a_{ij} = 0, 1, \forall i \neq j$ , 则相应的图称为简单图. 图论里有一个著名的 BEST 定理<sup>[25]</sup>, 缩写来自四位数学家的姓氏, 他们是 Bruijn, Aardenne-Ehrenfest, Smith 和 Tutte. 对于简单欧拉图, 给定  $K$  时欧拉圈的数目是

$$R(K) = \Delta \prod_i (d_i - 1)! \quad (5)$$

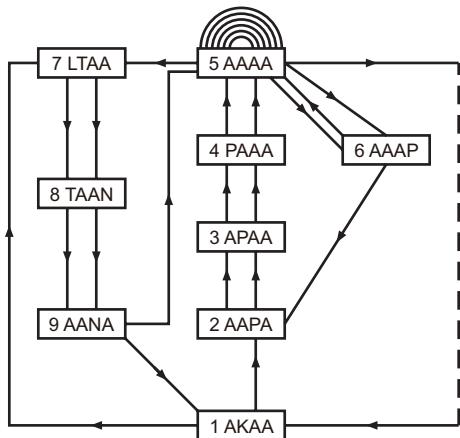


图 1 对应蛋白质序列 ANPA\_PSEAM 的欧拉图, 虚线来自把末态连回到初态的辅助线

**Figure 1** Eulerian cycle for protein sequence ANPA\_PSEAM. The dash line is an auxiliary line which links the end to the beginning of Eulerian figure.

图 1 显然不是简单欧拉图。然而, 只要在每一条并行的弧线上添加一个辅助顶点, 它就成为简单图; 仍然可以使用公式 (2) 计算圈数, 只是矩阵变得更大。谢惠民借助矩阵的初等变换证明, 可以不增大矩阵, 只是把公式 (5) 推广成

$$R = \frac{\Delta \prod_i (d_i - 1)!}{\prod_{ij} a_{ij}!}, \quad (6)$$

上式中允许某些  $a_{ii} \neq 0$ , 以及一些  $a_{ij} > 1$ 。上式分母中的  $a_{ij}!$  就是为了抵消并行弧线所导致的重复计数。我们后来得知, 公式 (6) 早曾在文献 [26] 中给出, 不过那里考虑的是指定了起点的欧拉圈, 因此该起点的度  $d$  要出现在分子中。

对于蛋白质列 ANPA\_PSEAM 所导致的图 1, 欧拉圈的数目是  $R(5) = 1512, R(6) = 60, R(7) = 2, R(8) = 1$ , 即  $K = 8$  时该蛋白质序列具有唯一的重构。

#### 4 欧拉圈数目与可因式化语言

2000 年, 郝柏林和谢惠民讨论蛋白质序列分解成  $K$  肽后重构的唯一性, 谢惠民发现了它与欧拉圈数目关系, 随后在 Santa Barbara 参加“统计物理与生物信息”活动期间, 郝同张淑誉一起做了一批实例, 手工绘制了不少欧拉图, 算了一批拉普拉斯矩阵。由于该项研究并不在我们研究的主线上, 前述研究结果

曾经只在一份束之高阁的电子预印本里<sup>[27]</sup>有所表述。直到一篇内容完全与生物学无关的文章<sup>[28]</sup>引用了预印本<sup>[27]</sup>, 才再次引起我们注意。文献[28]宣称存在着有限状态自动机, 它可以识别在特定的  $K$  值下, 一个符号序列是否具有唯一重构。不过, 文章[28]只包含存在性证明, 并没有实际构造出相应的自动机。关于语言和自动机的基本概念, 可参看<sup>[29]</sup>。

首先给定由  $m$  个符号组成的字母集  $\Sigma = \{q_1, a_2, \dots, a_m\}$ 。 $\Sigma$  中任意多个符号的各种各样的串, 组成一个大集合  $\Sigma^*$ 。 $\Sigma^*$  的任何子集  $L \subset \Sigma$  称为一个语言。关键在于如何定义这个子集。

可因式化语言  $L$  的定义是: 如果一个符号串  $x \in L$ , 则  $x$  的任何子串都属于  $L$ 。我们在 1990 年代发展单峰映射的符号动力学时, 就知道符号动力学中允许字组成可因式化语言。后来在研究细菌基因组中的缺失短串时, 又借助可因式化语言解决了冗余缺失串的计数问题<sup>[30,31]</sup>。

现在取  $\Sigma^*$  上一切在指定  $K$  下具有唯一重构的符号串的集合来构成  $L$ 。根据定义,  $L$  就是一个可因式化语言, 因为一个具有唯一重构的符号串的任意子串必然具有唯一的重构。考虑到来自基因组的符号序列的有限性, 我们可以只研究有限的语言  $L$ 。这样的  $L$  必然属于正规语言。这不仅使得文献[28]的存在性定理成为显然事实, 而且给出了构造相应有限状态自动机的途径<sup>[32]</sup>。

一个有限状态自动机由 5 个元素构成:

$$\{Q, \Sigma, \delta, q_0, F\}, \quad (7)$$

这里  $Q = \{q\}$  是状态的集合,  $q_0 \in Q$  是初始状态。每个状态有 3 个成分, 即  $Q = P \times N \times C = \{(p; n; c)\}$ 。 $p$  标记最近读进来的符号  $a_p \in \Sigma$ 。初始状态  $q_0$  还没有读进任何符号, 我们形式上引进一个不属于  $\Sigma$  的符号  $a_0$ , 于是可以写  $P = \Sigma \cup 0$ 。 $n$  是  $m+1$  个符号的表  $\{n_0, n_1, n_2, \dots, n_m\}$ , 用来更新在  $p$  之后读进来的下一个符号。对于初始状态  $n = \{\epsilon, \epsilon, \epsilon, \dots, \epsilon\} \equiv \epsilon^{m+1}$ ,  $\epsilon$  表示“空”或不存在。于是可以写  $N = (\Sigma \cup \epsilon)^{m+1}$ 。 $c$  是  $m$  个拨动开关的表  $\{c_1, c_2, \dots, c_m\}$ 。拨动开关的两个状态可以记作 *WHITE* 和 *BLACK*。最初  $c = \text{WHITE}^m$ , 初始状态  $q_0 = (0; \epsilon^{m+1}; \text{WHITE}^m)$ 。读进一个禁止字后,  $c$  成为 *BLACK* $^m$ 。只要  $c$  没有成为全黑即 *BLACK* $^m$ ,

$q$  就是一个可以接受的状态. 一旦  $c$  成为全黑, 它就一直保持全黑, 自动机判断所读入的串不具有唯一重构.  $F \in Q$  是接受语言  $L$  的状态, 即  $F = (p; n; c \neq BLACK^m)$ . 关键是从  $Q \times \Sigma$  到  $Q$  的转移函数  $\delta(q, a)$ . 我们不去详细描述, 而只在图 2 中给出实现  $\delta$  的程序.

这个看起来似乎很简单的自动机, 是一类特别的确定性有限状态自动机. 它在任一时刻的状态唯一, 状态转移图的结点数目也是固定的. 但是对于  $K$  串这样的字母表, 结点数目就太大了, 以至于在实际程序中不能真正模拟, 只可动态增加.

从自动机理论 [29] 知道, 就功能而言, 非确定性自动机同确定性自动机是等价的. 从一个非确定性自动机出发, 可以通过“子集合构造”得到确定性自动机; 确定性自动机一般并不唯一, 但存在着一个结点数目最少的最小确定性自动机. 最近, 文献 [33] 证明, 欧拉圈问题对应的最小确定性自动机具有稍多于指数、量级为  $\exp[O(m \log m)]$  的状态数目. 这与我们构造的自动机的状态数是一致的, 最多差指数上的常数因子. 文献 [33] 的结果和证明都是基于李强提供的论据. 不过, 看来这里还存在着远非平庸的有待解决的问题.

现在我们至少可以用三种办法写出程序, 来处理给定蛋白质序列的重构唯一性问题:

(1) 直接实现重构的程序, 返回一条条重构出的序列. 这个程序给出的信息最多, 即所有能实现的重构序列. 然而, 还必须给定一个重构数目的上限, 例

```

1 procedure δ((p,n,c),a)
2   if (np ≠ ε) & (np ≠ a) then
3     i ← p
4     repeat
5       ci ← BLACK
6       i ← ni
7       until i = p
8     endif
9     if ca = BLACK then
10      c ← BLACKm
11    endif
12    p ← np ← a
13  endp rocedure

```

图 2 实现转移函数  $\delta(q, a)$  的程序 [32]

Figure 2 An algorithm of transition function  $\delta(q, a)$  [32].

如当重构超过 1 万种时, 就给出信息并停止计算, 否则程序可能无休止地运行下去.

(2) 实现推广的 BEST 公式 (6) 的程序, 它只给出特定  $K$  值下重构序列的数目, 而不输出任何序列. 实现这个程序时, 必须用一些技巧来归并那些不影响欧拉圈数目的顶点. 这些技巧的文字描述, 反而比用算法语言写出的程序更为繁琐, 故此处从略.

(3) 一个判断给定序列是否具有唯一重构的有限状态自动机. 它给出的信息最少: “是” 或 “否”. 在 “否”的情形下, 可以把  $K$  换成  $K+1$ , 直到得出 “是”, 即使得重构数为 1 的  $K$  值. 事实上, 它包含着导致重构不唯一的具体  $K$  串及其位置的信息, 只是没有输出.

用这些工具武装起来之后, 可以针对实际的蛋白质数据库, 检查在何种  $K$  值下序列具有唯一的重构. 计算表明 [34], 对于并不大的  $K$  值, 自然界里的多数蛋白质具有唯一的重构, 从不唯一到唯一的转变, 发生在  $K = 5, 6, 7$  的狭窄区间内. 这一事实是对我们使用  $K$  肽集合来代替蛋白质一级序列的支持. 另一方面, 自然界里存在着一些蛋白质, 它们具有极大的重构数. 文献 [20] 列举了一批这样的蛋白质. 它们多是实现机械或抗冻功能的纤维蛋白, 而不是做为酶的珠蛋白. 例如, 丁香假单胞菌 (*Pseudomonas Syringae*) 在超冷下的冰成核蛋白 ICEN\_PSESY 只有 1200 个氨基酸, 它的  $R(11) = 1.55675 \times 10^{27}$ , 当  $K = 46$  时它才具有唯一的重构.

还可以提出一个问题: 给定  $K$  值和多条蛋白质序列, 其中每一条都具有唯一的重构; 对于来自所有这些蛋白质的  $K$  肽集合, 一起来进行重构, 那还能有多少蛋白质保持正确的唯一重构? 看来, 只能对于每个蛋白质集合给出具体的计算结果, 而没有一般性的答案.

## 5 K-肽数目的统计预测

为了减除中性突变背景所使用的统计预测公式 (3), 并不是唯一的选择. 原则上可以设想其他的统计方法. 例如, 文献 [35] 直接写出而不是推导出一个从  $K-1$  肽概率预测  $K$  肽概率的线性关系:

$$\begin{aligned} p^0(\alpha_1 \alpha_2 \cdots \alpha_K) &= \frac{1}{2}(p(\alpha_1 \alpha_2 \cdots \alpha_{K-1})p(\alpha_K) \\ &\quad + p(\alpha_1)p(\alpha_2 \alpha_3 \cdots \alpha_K)). \end{aligned}$$

然而, 在它所导致的亲缘树上, 一个古菌 (*Pyrobaculum Aerophilum*) 混入了细菌枝, 从而破坏了三个生命超界的划分.

我们不知道是否存在最佳预测的数学判据. 传统的统计再抽样检查, 包括自举法 (Bootstrap) 和刀切法 (Jackknife), 充其量只能说明所构树对于输入数据集的稳定性和自洽性, 而并不能证明它的客观正确性. 为了满足囿于传统再抽样方法的要求, 我们发展了针对 CVTree 方法的刀切和自举算法 [22], 说明组分矢量方法确实可以通过各种检验, 特别是基因组数据集合从 2002 年初的不足 70 个, 增加到 2013 年 10 月底接近 2700 个, CVTree 的质量不断改进, 更可以说是通过了规模不断扩大的反刀切法检验.

然而, 我们要强调指出, 判断亲缘树质量的根本方法, 是在从门到种的各个层次上同细菌分类系统的公认结果做直接比较. 基因组序列来自对样品的测序, 可以视为实验数据; CVTree 是一套理论分析框架; 分类系统乃是几代微生物学家近 200 年的研究成果, 他们涵盖了从形态特征到 16S rRNA 序列的各种观察和数据. 亲缘关系和分类系统的直接比较, 在 20 世纪末还是做不到的事情, 因为那时人们还在质疑细菌蛋白质序列中究竟是否含有亲缘信息 [36], 而当时所构造的基于基因组的树不能分辨门以下的分类单元 [37]. 亲缘树和分类系统目前达到的高度一致, 同时是对两者客观正确性的证明, 是进行微生物学研究的基础性事实.

## 6 最佳 $K$ -值的选择

在 CVTree 方法中, 肽段长度  $K$  看起来像、而实际上不是参数, 因为我们从来不调整它的数值, 对所有的基因组一视同仁地使用同一个  $K$  值. 比较不同  $K$  值下亲缘树的“收敛”情况, 更提供了考察构树质量的新角度. 我们多年来的实际构树经验表明, 对于病毒最佳的  $K$  值范围是 5 [13], 对于细菌是  $K = 5$  和 6 [6], 对于真菌是  $K = 7$  [17]. 许多从方法论角度考

察 CVTree 的作者曾建议使用更长的  $K$  来构树, 他们提出完全组分矢量 (CCV [38])、改进的完全组分矢量 (ICCV [39]) 等方法, 也有人引用截断的后缀树来处理更宽的  $K$  值范围 [40], 但是都没有得出比 CVTree 更好的结果.

定性地说, 比较长的  $K$  值可以突出肽段的物种特异性. 但是如果选取的都是太长的肽段, 最终得到的将是一棵星形树 (star tree): 每个物种自成一体一类, 互相无关. 要反映不同物种之间的相互关系, 就必须恰当计入一些较短的肽段. 从我们使用的  $K - 2$  阶马可夫预测公式 (3) 出发, 可以把以上定性考虑定量化. 式中最长的  $K$  应反映物种特异性, 因此它们的数目不能太多, 而应当少于在一个同等长度的随机氨基酸序列中出现此种肽段的概率.

假定各种氨基酸的出现概率都是  $1/20$ , 那么一种特定  $K$  肽的出现概率是  $1/20^K$ . 在总长度为  $L$  的随机氨基酸序列中, 此种  $K$  肽的总数估计是  $L/20^K$  个. 我们所关心的反映物种特异性的某种  $K$  肽的数目应当“大于”随机情况, 或者说在随机序列中它的出现应当是小概率事件, 即

$$\frac{L}{20^K} \ll 1. \quad (8)$$

另一方面,  $K - 2$  肽的数目不能太少, 它们才能把不同的物种联系起来. 换言之, 我们至少应当有

$$\frac{L}{20^{K-2}} \gg 1. \quad (9)$$

对上面两个不等式的左右都取对数 (采用以 10 为底的对数比较简便), 解出  $K$  并且放到一起, 有

$$\frac{\log L}{1 + \log 2} < K < 2 + \frac{\log L}{1 + \log 2}.$$

对于典型的病毒、细菌和真菌基因组所编码的蛋白质产物总氨基酸数目, 可以分别取  $L = 10^5$ ,  $L = 10^6$  和  $L = 10^7$ , 于是对于病毒得到

$$3.8 < K < 5.8, \quad K = 4, 5;$$

对于细菌和古菌有

$$4.6 < K < 6.6, \quad K = 5, 6;$$

对于真菌有

$$5.4 < K < 7.4, \quad K = 6, 7.$$

对数估值的适用范围比较宽松。这里给出的最佳  $K$  值范围，也与我们多年的计算经验一致。

没有必要采用更大的  $K$  值，还有一个简单直观的原因。对于一个给定的基因组，不同  $K$  肽的总数在  $K$  值比较小时的增长速度低于  $20^K$ ，而在  $K$  值变大以后，受限于一条下降的直线  $L = M(K-1)$  其中  $M$  是蛋白质数目， $L = \sum_{i=1}^M L_i$  是各条蛋白质所包含的氨基酸总数，这在前面介绍组分矢量方法的第2节里都已经引入。图3是一批古菌和细菌的  $K$  肽总数随  $K$  的变化曲线，可见更大的  $K$  值已经把不再有实质贡献。

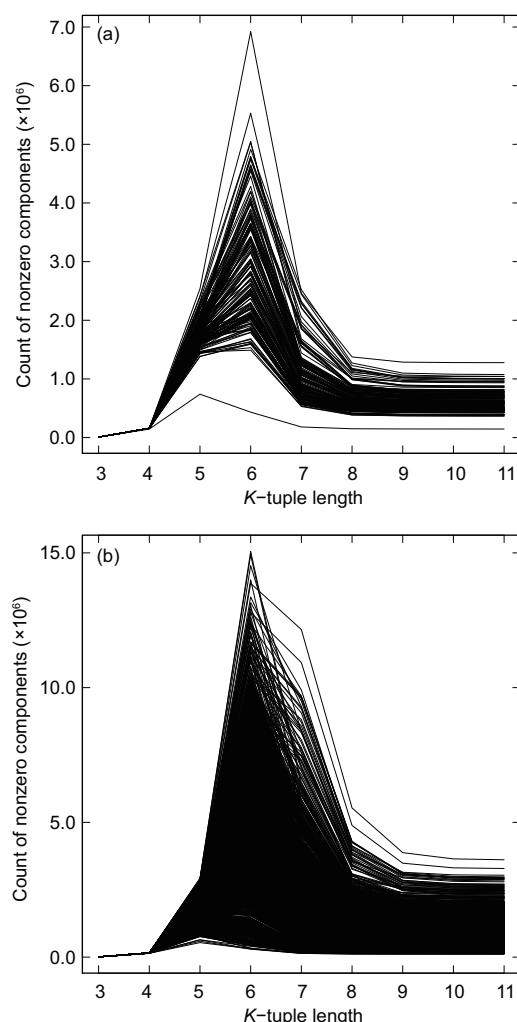


图3 一批  $K$  肽总数随  $K$  的变化曲线  
(a) 古菌; (b) 细菌

**Figure 3** Number of  $K$ -strings as function of  $K$  for whole genomes of archaea (a) and bacteria (b).

## 7 三角形不等式和准度规

组分矢量之间的“距离”公式(2)可以稍加变换<sup>[41]</sup>，写成：

$$D(A, B) = \frac{1}{2} \left( 1 - \frac{\mathbf{A}^T \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} \right) = \frac{1}{4} \left\| \frac{\mathbf{A}}{\|\mathbf{A}\|} - \frac{\mathbf{B}}{\|\mathbf{B}\|} \right\|^2, \quad (10)$$

其中  $\|\mathbf{A}\|$  是代表物种  $A$  的、以  $a_i$  为分量的组分矢量  $\mathbf{A}$  的模。可见， $D(A, B)$  是一个欧几里德距离的平方。欧几里德距离满足包括三角形不等式在内的的距离三公理，但是它的平方就不一定满足三角形不等式。事实上，由公式(2)或(10)所定义的距离矩阵不能保证所有的三角形不等式都成立<sup>[21]</sup>。

以2012年3月初基于  $N = 1570$  个基因组构建的CVTree为例。这时三角形的总数是  $N(N-1)(N-2)/6 = 643750240$  个。不同  $K$  值下，不等式不成立的三角形数目见下面的表1。

应当指出，不等式被破坏的三角形数目极其稀少，而且所涉及的基因组与亲缘树上位置有问题的物种没有任何关联。表1再次表明， $K = 5$  和  $6$  给出最好的结果。

只满足比三个“距离公理”更弱条件的度规在有些文献中称为准度规<sup>[42]</sup>，它仍然强于“非相似性”。我们使用的关联“距离”(公式(2))是一种准度规。

从我们的计算结果还可以提出一个问题：欧几里德距离的平方所定义的准度规，虽不能保证全部三角形不等式成立。但是，在我们的具体情况下，被破坏的不等式只占极小的比例。有没有数学判据，把这一类准度规挑选出来？这是一个可以提给古希腊数学家的问题，我们不知道是否已经有答案。

## 8 一些没有解决的问题

我们在前文的叙述中已经提到一批没有解决的问题。CVTree作为一套操作方法，其结果达到与细菌

表1 不同  $K$  值下被破坏的三角形不等式数目

**Table 1** Numbers of missing triangle inequality of different  $K$  values

$K$	3	4	5	6	7
绝对数目	12 501	415	0	0	3
相对比例 (%)	$1.87 \times 10^{-3}$	$6.44 \times 10^{-5}$	0	0	$4.6 \times 10^{-7}$

分类系统高度一致,但是它还面临许多奠基性问题,其中有一些可能只具有学术意义,如下所述。

(1) CVTree 方法里没有严格定义的统计量 (statistic). 特定的  $K$  串出现在一条长度为  $L$  的随机序列中的概率及其分布性质,是离 CVTree 方法甚远的过于简单的统计量。对于特定的  $K$  串同时出现在两条随机序列中的概率及其分布, Waterman 及合作者引入的  $D_2$  统计量,向前进了一步。但是要同 CVTree 联系起来,就必须考虑互相重叠的  $K$  串计数和随后的减除手续;前者破坏关于序列随机性的独立同分布假设,后者更是非线性操作。关于非联配方法的一篇最新综述<sup>[43]</sup> 虽然把我们使用的(2)式称为非相似性测度,但是并没有关于其统计性质的分析,因此它还构不成一种统计量。

(2) 枝长的标定与遗传距离的关系。这个问题的提法与所研究的对象的分类广度有关。传统的亲缘树研究通常涉及较窄的分类范围,例如一个亚门(脊椎动物)、一个纲(哺乳动物)、一个目(灵长目)等等,这时基因序列突变速率恒定是一个合理的假设。

**致谢** 陈国义、王彬、戚继、罗红、史晓黎、高雷、卫海滨、孙健冬、汪浩、徐昭、虞洪杰、周婵、夏立等博士的贡献,以及谢惠民教授在数学问题上所给予的指导表示感谢;郑伟谋教授对 CVTree 方法不完善之处的持续批评也一直是我们研究的动力。

## 参考文献

- 1 Fox G E, Pechman K R, Woese C R. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. *Int J Syst Bacteriol*, 1977, 27: 44–57
- 2 Bergey's Manual Trust. *The Bergey's Manual of Systematic Bacteriology*. 2nd ed. New York: Springer-Verlag, 2001–2012
- 3 Hao B L, Qi J, Wang B. Prokaryotic phylogeny based on complete genomes without sequence alignment. *Mod Phys Lett*, 2003, B17: 91–94
- 4 Hao B L, Qi J. Vertical heredity vs. horizontal gene transfer: a challenge to bacterial classification. *J Syst Sci Complexity*, 2003, 16: 307–314
- 5 Hao B L, Qi J. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. In: *The Proceedings of CSB2003*. IEEE, 2003. 375–384
- 6 Qi J, Wang B, Hao B L. Whole proteome prokaryote phylogeny without sequence alignment: A  $K$ -string composition approach. *J Mol Evol*, 2004, 58: 1–11
- 7 Qi J, Luo H, Hao B L. CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucl Acids Res*, 2004, 32: W45–W47
- 8 Xu Z, Hao B L. CVTree update: A newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucl Acids Res*, 2009, 37: W174–W178
- 9 Hao B L. CVTrees support the Bergey's Systematics and provide high resolution at species level and below. *Bull BISMIS*, 2011, 2(Part 2): 189–196
- 10 Zuo G H, Xu Z, Hao B L. *Shigella* Species are not strains of *Escherichia coli* but sister members in the genus *Escherichia*. *Genomics, Proteomics Bioinf*, 2013, 11: 61–65
- 11 Zuo G H, Hao B L, Staley J T. Geographic divergence of “*Sulfolobus islandicus*” strains assessed by genomic analyses including electronic DNA hybridization confirms they are geovars. *Antonie van Leeuwenhoek J Microbiol*, 2014, 105: 431–435
- 12 Gao L, Qi J, Wei H B, et al. Molecular phylogeny of Coronaviruses including human SARS-CoV. *Chin Sci Bull*, 2003, 48: 1170–1174

CVTree 所研究的首先是涉及一切有基因组数据的原核生物,目前已经涵盖 30–40 个门。很难设想,生活在地球上不同环境里的微生物的基因序列会以近似恒定的速率发生突变。因此,亲缘树的拓扑和枝长两个特性,拓扑才是更重要的。而这正是 CVTree 的长处。尽管如此,我们还是给出过一个关联“距离”与遗传距离的变换关系<sup>[44]</sup>,不过它不能保证树的拓扑结构完全不变。

(3) 如果一棵亲缘树上的枝长与物种分化时间一致,则枝长具有相加性。这时必有两个物种到第三个物种的距离相等,且大于或等于它们之间的距离。这是距离公理中对三角形不等式的更强的要求,即三角形必须是等腰或等边的。这种距离称为超度规(关于超度规的基本概念,可以参看文献[45])。一般说来,从序列集合得到的距离矩阵并不满足超度规要求。原则上有很多办法使原来的距离矩阵超度规化,在从后者构建的树上,根的位置应当自然得出。然而,目前并没有一种判据来挑选出最具生物学意义的超度规化手续。

- 13 Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol Biol*, 2007, 7: 41
- 14 Chu K H, Qi J, Yu Z G, et al. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol Biol Evol*, 2004, 28: 70–76
- 15 Gao L, Qi J, Sun J D, et al. Prokaryote phylogeny meets taxonomy: Comparasion of composition vector trees with systematic bacteriology. *Sci China Ser C-Life Sci*, 2007, 50: 587–599
- 16 Sun J D, Xu Z, Hao B L. Whole-genome based Archaea phylogeny and taxonomy — a composition vector approach. *Chin Sci Bull*, 2010, 55: 2323–2328
- 17 Wang H, Xu Z, Gao L, et al. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol*, 2009, 9: 195
- 18 Wei H B, Qi J, Hao B L. Prokaryote phylogeny based on ribosomal proteins and aminoacyl tRNA synthetases by using the compositional distance approach. *Sci China Ser C-Life Sci*, 2004, 47: 313–321
- 19 Gao L, Qi J, Hao B L. Simple Markov subtraction essentially improves prokaryote phylogeny. *AAPPS Bull*, 2006, 16: 3–7
- 20 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
- 21 Li Q, Xu Z, Hao B L. Composition vector approach to whole-genome-based prokaryotic phylogeny: Success and foundations. *J Biotechn*, 2010, 149: 115–119
- 22 Zuo G H, Xu Z, Yu H J, et al. Jackknife and bootstrap tests of the Composition vector trees. *Genomics, Proteomics Bioinf*, 2010, 8: 262–267
- 23 Kimura M. Theory of Neutral Evolution. Cambridge: Cambridge University Press, 1983
- 24 Hu R, Wang B. Statistically significant straings are related to regulatory elements in the promoter region of *Saccharomyces cerevisiae*. *Physica*, 2001, A290: 464–474
- 25 Fleischner H. Eulerian Graphs and Related Topics, Part 1. New York: Elsevier, 1991, 2: IX80
- 26 Hutchinson J P. On words with prescribed overlapping subsequences. *Utilitas Math*, 1975, 7: 241–250
- 27 Hao B L, Xie H M, Zhang S Y. Compositional representation of protein sequences and the number of Eulerian loops. arxiv: physics/0103028
- 28 Kontorovich L. Uniquely decodable n-gram embeddings. *Theor Comput Sci*, 2004, 329: 271–284
- 29 Hopcroft J, Ullman J. Introduction to Automata Theory, Langauges and Computation. Boston: Addison-Wesley, 1979
- 30 Hao B L. Fractals from genomes — exact solutions of a biology-inspired problem. *Physica*, 2000, A282: 225–246
- 31 Hao B L, Xie H M, Yu Z G, et al. Factorizable language: From dynamics to bacterial compleate genomes. *Physica*, 2000, A288: 10–20
- 32 Li Q, Xie H M. Finite automata for testing composition-based reconstructibility of sequences. *J Comput Syst Sci*, 2008, 74: 870–874
- 33 Kontorovich A, Trachtenberg A. Deciding unique decodability of bigram counts via finite automata. *J Comput Syst Sci*, 2014, 80(2): 450–456
- 34 Xia L, Zhou C. Phase transition in sequence unique reconstruction. *J Syst Sci Complexity*, 2007, 20: 18–29
- 35 Yu Z G, Zhou L Q, Ahn V V, et al. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from compleate genomes without sequence alignment. *J Mol Evol*, 2005, 60: 538–545
- 36 Teichmann A A, Mitchison G. Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol*, 1999, 49: 98–107
- 37 Hyunam M, Snel B, Bork P. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science*, 1999, 286: 1443a
- 38 Wu X M, Wan X F, Wu G, et al. Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method. *Int J Bioinf Res Appl*, 2006, 2: 219–248
- 39 Lu G Q, Zhang S P, Fang X. An improved string composition method for sequeunce comparison. *BMC Bioinf*, 2008, 9(Suppl 6): SI5
- 40 Apostolico A, Denas O, Dress A. Efficient tools for comparative substring analysis. *J Biotechn*, 2010, 149: 120–126
- 41 Chan R H, Wang R W, Yeung H M. Composition vector method for phylogenetics — a review. In: The Proceedings of The Ninth International Symposium on Operations Research and Its Applications (ISORA'10). Chengdu, 2010. 13–20
- 42 Heinonen J. Lectures on Analysis on Metric Spaces, Univeritext. New York: Springer, 2001, doi: 10.1007/978-1-4613-0131-8
- 43 Song K, Ren J, Reinert G, et al. New developments of alignment-free squence comparison: measures, statistics, and next-generation sequencing. *Briefing Bioinf*, 2014, 30(20): 2949–2955
- 44 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
- 45 Rammal R, Toulouse G, Virasoro M A. Ultrametricity for physicists. *Rev Mod Phys*, 1986, 58: 765–788

# Some mathematical problems inspired by the study of whole-genome-based phylogeny and taxonomy of Prokaryote

LI Qiang<sup>1,2††</sup>, ZUO GuangHong<sup>1††</sup> & HAO BaiLin<sup>1\*</sup>

<sup>1</sup> *T-Life Research Center, Department of Physics, Fudan University, Shanghai 200433, China*

<sup>2</sup> *CAS-MPG Partner Institute for Computational Biology, Shanghai 200031, China*

Our composition vector tree (CVTree) method is an effective and efficient phylogenetic method for prokaryote. It is whole-genome-based, alignment and easy-to-use. And benefit from the development of the gene technology, it is competent for one of the definite tools in the study of prokaryotic taxonomy. In this paper, instead of showing the biological result, we focused on the mathematical problems that were issued in the study of CVTree. To resolve these problems, some discrete mathematics knowledge, including combinatorics, graph theory and formal language, were engaged in our study. We also posed some unsolved fundamental problems for the CVTree method. We hope these problems will inspire some new researches in the theoretic field.

**alignment-free, whole-genome-based phylogeny, prokaryote, combinatorics, Eulerian cycles, factorizable language**

**PACS:** 02.50.Ga, 02.10.Ox, 02.70.Rr

**doi:** 10.1360/SSPMA2014-00124

## 征 稿 简 则

**简介:**《中国科学:物理学 力学 天文学》(中文版)和 *SCIENCE CHINA Physics, Mechanics & Astronomy*(英文版)是中国科学院和国家自然科学基金委员会共同主办、《中国科学》杂志社出版的学术刊物,主要报道物理学、力学和天文学基础研究与应用研究等方面具有创新性和高水平的最新研究成果,月刊。

**收录情况:**《中国科学:物理学 力学 天文学》与其英文版 *SCIENCE CHINA Physics, Mechanics & Astronomy* 是两个完全独立的刊物,前者被《中国科学引文数据库》、《中国期刊全文数据库》、《中国科技论文与引文数据库》和《中国数字化期刊群》等收录,并进入《中文核心期刊要目总览》;后者被 SCI, EI, Astrophysics Data System, Current Contents, Google Scholar, Index to Scientific Reviews, INSPEC, Mathematical Reviews, MathSciNet 等收录。

**栏目:**《中国科学:物理学 力学 天文学》设有以下 4 个栏目:

**评述:**综述所研究领域的代表性成果和研究进展,评论研究现状,提出今后研究方向的建议。要求作者在该领域从事过系统的研究工作,或者所做工作与该领域的研究紧密相关(10000 字左右,附 600 字左右的摘要)。

**论文:**报道物理学、力学和天文学各领域具有创新性、高水平和重要科学意义的最新科研成果(8000 字左右,附 300 字左右的摘要)。

**快报:**简明扼要地及时报道物理学、力学和天文学各领域具有创新性和新颖性的科研成果(4000 字左右,附 300 字左右的摘要)。

**评论:**评介过去或近期在本刊或国内外重要刊物上发表的重要研究成果(1500 字左右,附 200 字左右的摘要)。

**投稿:**请使用在线投稿的方式,访问本刊网站 [www.scichina.com](http://www.scichina.com) 或 [phys.scichina.com](http://phys.scichina.com),点击“作者投稿系统”,进入“学术期刊管理系统”,首次投稿时需注册一个“作者账户”。注册完成之后,按照提示与引导进行投稿。如果不能在线投稿,请与编辑联系,另行约定投稿方式。本刊编委可推荐经其本人审阅后的稿件,如同意具名推荐该稿件(首页注明“xxx推荐”或“Recommended by xxx”),经核实时送主编终审。编委本人作为作者参与的论文并同意具名负责的稿件(首页注明“xxx供稿”或“Contributed by xxx”)同样办法处理。专业主编将依据稿件的具体情况尽快予以答复。

**审稿:**稿件将由编委会组织同行专家进行评审,并做出录用与否的决定。评审过程大约需要 60~90 天。评审结束后,无论录用与否,编辑部将及时向作者转达评审意见。作者若在 90 天内没有收到编辑部有关稿件的取舍意见,请及时与编辑部联系。作者在通知编辑部后,可以改投他刊。本刊不受理“一稿多投”之稿件。

**文章署名:**通讯作者应保证稿件内容经全体作者认可并同意署名。投稿后,署名的改变要有全体原作者签名同意的书面材料。

**录用:**稿件被录用后,全体作者必须签署“著作权转让声明书”,将该论文(各种语言版本)所享有的复制权、发行权、信息网络传播权、翻译权、汇编权在全世界范围内转让给《中国科学:物理学 力学 天文学》的出版单位《中国科学》杂志社。全体著作权人授权《中国科学》杂志社根据实际需要独家代理申请上述作品的各种语言版本(包含各种介质)的版权登记事项。著作权转让声明书可以从本刊网站上下载。

**中国科学 物理学 力学 天文学**  
SCIENTIA SINICA Physica, Mechanica & Astronomica

第 44 卷 第 12 期 2014 年 12 月出版

版权所有,未经许可,不得转载

主 管	中 国 科 学 院	出 版	《中国科学》杂 志 社
编 辑	中 国 科 学 院	印 刷 装 订	北京中科印刷有限公司
	《中国科学》编辑委员会	总 发 行 处	北京报刊发行局
	北京(100717)东黄城根北街 16 号	订 购 处	全 国 各 邮 电 局
主 编	王鼎盛 张杰		《中国科学》杂志社发行部

刊号: ISSN 1674-7275 代号: 国 外 BM40G  
CN 11-5848/N 国内邮发 80-211 每期定价: 138.00 元 全年定价: 1656.00 元

广告经营许可证: 京东工商广字第 0429 号

# 超精细无液氦低温光学恒温器

Montana Instruments 长期潜心研制全自动无液氦低温光学测量平台，它集中了一系列的新专利技术，突破了传统恒温器所面临的振动大、噪声大、温度不稳等问题，它采用无液氦液氮制冷，变温范围在 3 K-350 K，样品震动位移不超过 5 nm，温度波动长时间保持在 10 mK 以内，升降温以及抽真空由计算机全程控制，同时产品的设计以及全自动化软件使用户更容易在此平台上实现各种低温光学实验。



## CRYOSTATION

应用领域：拉曼、红外、穆斯堡尔谱、荧光、AFM、MFM、SEM、共聚焦、量子点、信息存储.....



超大样品腔 (19 cm 直径, 4.3 K)



磁场选件 (3 K, 1 T)

## 系统参数

- 系统温区：3-350 K 变频制冷机提供制冷，无需液氦。
- 自 动 化：系统全[自动](#)控制（降温，升温，抽真空，洗气，监测）
- 降 温 时 间：108 min (300 K降至3 K)
- 温 度 稳 定 性：<10 mK peak to peak, <2 mK RMS )
- 震 动 稳 定 性：<5 nm peak to peak; <1 nm RMS )
- 底 座 稳 定 性：采用特种合金材料，在全温区无热胀冷缩引起的位移
- 制 冷 功 率：100 mW (在4.2 K 同时打开5个窗口)
- 样 品 空 间：高40 mm, 直径53 mm
- 测 量 引 线：28根内置引线
- 光 学 测 量：宽的入射角，数值孔径可达0.87
- 引 入 磁 场：1.5 T-9 T (选件)
- 共 聚 焦 距 离：2 mm
- 低 温 显 微 成 像：可选AFM/MFM/CFM等显微镜(选件)

美国 Quantum Design 中国子公司  
北京 : 010-85120277/78/79/80  
上海 : 021-52280980  
广州 : 020-89202739  
网址 : [www.qd-china.com](http://www.qd-china.com)

ISSN 1674-7275

