

文章编号: 0529-6579 (1999) 01-0124-127

外类群对构建基因树的影响^{*}

钟 扬¹, 唐先华¹, 施苏华², 黄柳林², 谈凤笑²

(¹中国科学院武汉植物研究所, 武汉 430074; ²中山大学生命科学学院)

摘 要: 以金缕梅科枫香亚科和睡莲目的分子数据为例, 初步探讨了不同的单一外类群、复合外类群和随机外类群对构建基因树的影响. 基因树比较的结果表明: 与内类群亲缘关系较近的单一外类群能反映较多信息, 而复合外类群与其中 1 个或多个单一外类群的结果相同, 应用随机外类群则可获得一个粗略的分支关系.

关键词: 分子系统学, 基因树, 外类群, 复合外类群, 随机外类群

分类号: Q 19 Q 751 **文献标识码:** A

分子系统学研究的主要目标是根据分子生物学数据重建基因树, 进而推断出物种树^[1-4]. 长期以来, 已经发展了大量的系统树构建方法, 主要有简约法、相容法、距离法和最大似然法等^[5-9]. 在上述方法中, 外类群选择往往作为 1 个首要的步骤. 尤其在基于形态学数据的系统发育研究中, 外类群比较法 (outgroup comparison method) 已被广泛应用于性状的极化^[6, 7]. 然而, 在实际工作中还存在一些问题, 如复合外类群是否一定优于单一外类群, 复合外类群的规模应当多大, 究竟哪个单一外类群最为合适, 有关这些问题的研究仅在部分类群中取得了成功的经验, 尚缺乏普遍的规律.

目前, 关于分子系统发育分析中外类群问题的专门讨论还比较少. 本文以金缕梅科枫香亚科和广义睡莲目若干种的分子数据为例, 比较了由单一外类群、复合外类群以及随机外类群 (random outgroup) 所构建的基因树, 初步探讨了不同外类群对基因树的影响, 以期引起广泛的研究和深入的讨论.

1 材料与方法

1.1 研究类群与分子数据

数据集 1: 金缕梅科 (Hamamelidaceae) 枫香亚科 (Liquidambaroideae) 5 种的核糖体 DNA ITS 区序列 (699 碱基对, 待发表). 单一外类群分别为金缕梅 (*Hamamelis mollis* Oliver)、壳菜果 (*Mytilaria laosensis* Lec.) 和华榛 (*Corylus chinesis* Franch.).

数据集 2: 金缕梅科枫香亚科 6 种的叶绿体 DNA matK 基因序列 (1 010 碱基对, 待发表). 单一外类群分别为金缕梅、壳菜果和毛桃木莲 (*Magnolia moto* Dandy).

数据集 3: 广义睡莲目 (*Nymphaeales*) 9 种的叶绿体 DNA rbcL 基因序列 (1 183 碱基对)^[8]. 外类群为金鱼藻 (*Ceratophyllum demersum* L.).

* 收稿日期: 1998-10-15 第一作者简介: 钟扬, 男, 34 岁, 研究员

1.2 随机外类群与复合外类群 统计学意义上, 与内类群亲缘关系最远的应为随机外类群, 即由计算机生成的与内类群相同长度的随机序列。本项工作中应用的随机序列编码, 只从 A, G, C, T 4 个碱基中随机抽取, wG/wC 约为 50%。

复合外类群既可由多个单一外类群, 也可由单一或复合外类群与随机外类群组合产生。对不能直接处理复合外类群的程序, 可按 IUPAC (IUB) 核酸编码表将复合外类群转换成单一序列。

1.3 基因树构建方法 应用 PHYLIP 3.5c 软件包^[9] 中的 SEQBOOT 程序生成自展 (bootstrap) 数据集, 随机种子数 (random seed) 为 1, 自展重复次数为 1 000; 分别采用 DNA-PARS 和 DNAPENNY 程序构建最简约的系统树 (有根树); 采用 DNADIST 程序 (Kimura 2-parameter 距离) 和 NEIGHBOR 程序 (neighbor-joining 法和 UPGMA 法) 构建无根树; 采用 CONSENSE 程序获得多数一致树 (majority-rule consensus tree)。

2 结果与分析

图 1 示出了金缕梅科枫香亚科 5 种 (其中阿丁枫取 2 个样本) 的 6 个基因树。由于内类群数据集和构树方法及程序完全一致, 因而各基因树间的差异可以反映不同外类群的作用。图 1a, b, c 中分别应用了 3 个单一外类群。仅从自展数据支持率看, 图 1a, b 优于图 1c。若取阈值 50%, 即采用 50% 多数一致树, 图 1a 和 b 中的树不变, 而图 1c 中树的信息将明显减少。例如, 半枫荷 (*Semiliquidambar cathayensis* Chang, *S. c.*) 将与阿丁枫 (*Altingia chinensis* Oliver ex Hance, *A. c.*, *A. c.* *) 合并。从外类群与内类群的亲缘关系看, 华榛 (*C. c.*) 属壳斗科 (Fagaceae), 远于金缕梅 (*H. m.*) 和壳菜果 (*M. l.*)。这一结果表明, 不同基因树所反映的信息量大小可能和外类群与内类群亲缘关系的远近相关。应用 3 个单一外类群的复合外类群 (Com) 构建的树 (图 1d) 与图 1a, b, c 的拓扑结构完全一

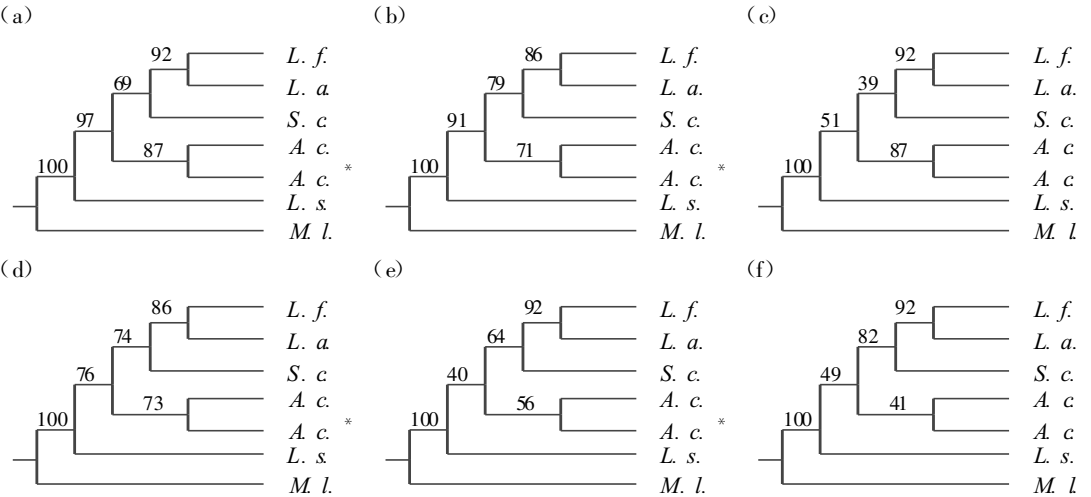


图 1 金缕梅科枫香亚科 5 种基因树的比较, 数字为自展数据支持率/%

Fig. 1 Comparason of 5 gene trees from Liquidambariaceae

L. f. = *Liquidambar formosana*; *L. a.* = *L. acahycina*; *L. s.* = *L. styraciflua*; *S. c.* = *Semiliquidambar cathayensis*; *A. c.* = *Altingia chinensis*; *M. l.* = *Mytilaria laosensis*; *H. m.* = *Hamamelis mollis*; *C. c.* = *Corylus chinensis*; d Com 由 *M. l.*, *H. m.*, *C. c.* 复合; e 随机外类群 Ran; f 复合外类群 Ran * 由 Com 和 Ran 复合

致，并没有增加新的信息，而支持率则与图 1a, b 接近. 图 1e, f 则表明应用随机外类群（或随机外类群与复合外类群进一步复合的外类群），可以获得与图 1a, b, c 大致相同的结果. 为了进一步说明复合外类群和随机外类群对构建基因树的影响，我们应用 3 个数据集、4 种构树方法和不同外类群的组合获得了大量的基因树. 对这些基因树的比较结果详见表 1. 结果表明：复合外类群与其中一个或多个外类群构建的基因树相同，而通过随机外类群也可获得一个粗略的分支关系. 例如，对数据集 2，随机外类群的结果与一个单一外类群的结果相同（50%多数一致树）. 对其他数据集甚至可以获得更为一致的结果. 值得注意的是，对方法组合 IV 而言，同一数据集不同外类群构建的基因树完全一致. 事实上，UPGMA 法是数量分类中常常应用的一种聚类分析方法，本身并不需要采用外类群. 其结果是可以用作一个初级分类的^[10].

表 1 复合外类群、随机外类群与单一外类群构建的基因树比较

Tah 1 Comparison between gene trees based on complex or random outgroups and single outgroups

种名		类群	方法组合 ¹⁾					
			I	II	III	IV		
数据集 1	内类群	<i>L. formosana</i>	复合外类群 ²⁾	全同 ⁴⁾	金缕梅、壳 菜果 ⁵⁾	金缕梅、壳菜果	全同	
		<i>L. acalycina</i>						
		<i>A. chinensis</i>	复合外类群 ³⁾	全同	金缕梅、壳 菜果	全同	全同	
		<i>L. styraciflua</i>						
		<i>S. cathayensis</i>						
	外类群	<i>M. laosensis</i>						
		<i>H. mollis</i>						
		<i>C. chinensis</i>						
	数据集 2	内类群	<i>L. formosana</i>	复合外类群	壳 菜 果	金缕梅(50%)	壳菜果(50%)	全同
			<i>L. acalycina</i>					
<i>A. chinensis</i>			随机外类群	金缕梅(50%) ⁶⁾	壳菜果(50%)	壳菜果(50%)	全同	
<i>A. hainanensis</i>								
<i>L. styraciflua</i>								
外类群		<i>S. cathayensis</i>						
		<i>M. laosensis</i>						
		<i>H. mollis</i>						
数据集 3		内类群	<i>M. moto</i>					
			<i>N. nucifera</i>					
	<i>N. lutea</i>							
	<i>B. schreberi</i>							
	<i>N. odorata</i>							
	<i>N. luteum</i>							
	<i>E. ferox</i>							
	<i>V. crizama</i>							
	<i>B. longifolia</i>							
	<i>C. caroliniana</i>							

外类群 *C. demesum*

1) 所有单一外类群的复合; 2) 单一外类群或复合外类群与随机外类群的复合; 3) 方法组合 I 使用 Seqboot→Dnapars→Consense 程序; II 使用 Seqboot→Dnapenny→Consense 程序; III 使用 Seqboot→Dnadist→Neighbor (Neighbor-Joining 法)→Consense 程序; IV 使用 Seqboot→Dnadist→Neighbor (UPGMA 法)→Consense 程序; 4) 与所有单一外类群的构建的分支树结构一致; 5) 与表中所列的一致; 6) 50%多数一致树相同

3 讨 论

本文仅以金缕梅科枫香亚科和睡莲目若干种的分子数据为例, 初步探讨了不同的单一外类群、复合外类群和随机外类群对构建基因树的影响. 其结果是否具有普遍意义, 还有待深入研究. 进一步研究的问题包括: ① 其他有机体类群(如动物等)及不同数据集. 表 1 中数据集 2 的结果变化最大, 是否与数据集本身有关? 如序列排列的质量. ② 其他方法组合和其他计算机程序. 我们应用 PAUP3. 1. 1^[11] 对部分数据集进行了验算, 结果相似.

参考文献:

- [1] LI W H. Molecular evolution. Sunderland: Sinauer Associates, 1997
- [2] DOYLE J J. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst Biol*, 1992, 17 (1): 144~163
- [3] DOYLE J J. Trees within trees: genes and species, molecules and morphology. *Syst Biol*, 1997, 46(3): 537~553
- [4] MADDISON W P. Gene trees in species trees. *Syst Biol*, 1997, 46(3): 523~536
- [5] FELSENSTEIN J. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet*, 1988, 22: 521~565
- [6] 钟扬, 李伟, 黄德世. 分支分类的理论与方法. 北京: 科学出版社, 1994
- [7] WATROUS L E, WHEELER Q D. The out-group comparison method of character analysis. *Syst Zool*, 1981, 30 (1): 1~11
- [8] LES D H, GARVIN D K, WIMPEE C F. Molecular evolutionary history of ancient aquatic angiosperms. *Proc Natl Acad Sci(USA)*, 1991, 88: 10119~10123
- [9] FELSENSTEIN J. PHYLIP (Phylogeny Inference Package) version 3.5c. Seattle: University of Washington, 1997
- [10] 钟扬, 陈家宽, 黄德世. 数量分类的方法与程序. 武汉: 武汉大学出版社, 1990
- [11] SWOFFORD D L. PAUP: phylogenetic analysis using parsimony, Version 3.1.1. Champaign: Illinois Natural History Survey, 1993

Effect of Outgroups on Construction of Gene Trees

ZHONG Yang^{*}, TANG Xianhua, SHI Suhua, HUANG Yelin

Abstract: Taking the molecular data from the subfamily Liquidambaroideae (Hamamelidaceae) and the order Nymphaeales *sensu lato* as examples, we constructed and compared the gene trees based on different combinations of data sets, tree-building methods, and outgroups. The gene tree using a single outgroup which has closer relationship with the ingroup can reflect more phylogenetic information. The same tree might be obtained using a single outgroup or more outgroups and a complex outgroup which includes this single outgroup(s). A random outgroup is defined as a random code sequence generated by using a computer program. Random outgroups can be used to infer sketchy cladistic relationships among particular taxa.

Keywords: molecular systematics, gene tree, outgroup, complex outgroup, random outgroup

^{*} Wuhan Institute of Botany, Academia Sinica, Wuhan 430074, China.