植物分类学报 46 (6): 795–807 (2008)
Journal of Systematics and Evolution (formerly Acta Phytotaxonomica Sinica)
doi: 10.3724/SP.J.1002.2008.08081
http://www.plantsystematics.com

基因树冲突与系统发育基因组学研究

邹新慧 葛 颂*
(系统与进化植物学国家重点实验室, 中国科学院植物研究所 北京 100093)

Conflicting gene trees and phylogenomics

Xin-Hui ZOU Song GE*
(*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences*, Beijing 100093, China)

Abstract With more and more sequence data available, it has been a widespread practice to apply multiple genes to reconstruct phylogenies at different hierarchical levels. The phenomenon of conflicting gene trees has accordingly become a remarkable and difficult problem. It is increasingly understood that the difference between gene tree and species tree and the causes behind should be fully appreciated in molecular phylogenetic studies. In this paper, we have explored the major causes resulting in conflicting gene trees, including stochastic errors, systematic errors and biological factors. We also introduced a newly developed discipline, phylogenomics, and demonstrated its power and great potential in resolving difficult phylogenetic problems using our recent phylogenomic study of *Oryza* as an example. Furthermore, we discussed some strategies and approaches in elucidating conflicting gene trees and provided some suggestions and recommendations for molecular phylogenetic studies using multiple genes.

Key words conflicting gene trees, gene tree, molecular phylogenetics, phylogenomics, species tree.

摘要 随着越来越多的基因序列被运用于系统发育重建中, 随之产生的基因树冲突已成为分子系统发育研究中日益突出的问题。因此, 在分子系统发育研究中, 应正确理解基因树和物种树之间的差异, 充分注意和分析基因树冲突的原因, 正确解释分子系统发育的结果。本文通过一些典型实例分析了在多基因系统发育研究中引发基因树冲突的三类主要原因: 随机误差、系统误差和生物学因素。在此基础上, 对近年来兴起的系统发育基因组学进行了介绍, 并以稻属*Oryza*研究为例, 阐述了系统发育基因组学方法在解决基因树冲突以及系统发育研究中的优势和应用价值, 并进一步探讨了解决基因树冲突的策略和方法, 以期为分子系统发育研究提供一些启示和帮助。

关键词 基因树冲突; 基因树; 分子系统发育学; 系统发育基因组学; 物种树

地球上的一切生命形式都有一个共同的起源, 无论动物、植物、真菌、原生生物, 还是原核生物, 它们都藉由一部共同的进化历史而有着或近或远的关联。重建所有生物的进化历史并以一种树状结构即系统发育树(phylogenetic tree)的形式来表示生物类群之间的进化关系, 一直是系统发育学研究的核心问题, 也是进化生物学研究的重要内容之一(Li, 1997; Futuyma, 1998; Nei & Kumar, 2000)。建立可靠的系统发育关系不仅是生物分类和命名的基础, 也是阐明类群起源和扩散、探讨性状演化以及揭示物种形成机制的前提(Futuyma, 1998; Soltis & Soltis, 2000)。

早在19世纪中叶, 德国生物学家海克尔就发表了他著名的“系统树”, 将整个生命世界形象地描绘成一棵大树(Haeckel, 1866)。在随后的100多年里, 生物学家们主要通过表型特征如形态、解剖、生理等性状进行物种间进化关系的研究, 为生命之树构筑了基本框架, 但这棵大树仍存在大量的缺失环节(Futuyma, 1998)。20世纪60年代以后, 随着分子生物学技术的迅猛发展, 分子数据开始被广泛运用于系统发育研究, 如蛋白质电泳、DNA-DNA杂交、免疫学等(Crawford, 2000)。自20世纪80年代以后, 随着PCR (polymerase chain reaction)技术的出现以及DNA测序技术的不断完善, 利用大量分子序列资料进行系统发育重建成为可能。分子数据, 尤其是DNA序列, 其数据的丰富性、在所有生物体中的可比性, 以及数据分析的规范性等特点使它成为进化

2008-06-10 收稿, 2008-08-02 收修改稿。
* 通讯作者(Author for correspondence. E-mail: gesong@ibcas.ac.cn; Tel.: 86-10-62836097; Fax: 86-10-62590843)。

生物学研究的重要手段, 而建立在数学和统计学基础上的系统发育树的构建理论和方法也由此获得了迅速发展, 形成了分子系统发育学(molecular phylogenetics)这一新的研究领域, 即利用生物大分子的信息来推断生物进化历史, 或者说重建生物类群的系统发育关系(Li, 1997; Nei & Kumar, 2000)。

由于不同的DNA片段在进化速率上存在较大差异, 因此可以在几乎所有分类学水平上推断生物类群之间的进化关系, 如用于研究近缘物种人与猿之间的亲缘关系, 以及古老的进化事件如叶绿体和线粒体的起源等问题(Li, 1997; Nei & Kumar, 2000)。随着分子序列证据的广泛应用, 人们对现存陆生植物的各大类群之间、具体科属内部的物种之间的进化关系乃至栽培作物的起源等问题的认识也发生了根本性的变化(Crawford, 2000; Soltis & Soltis, 2000)。Chase等(1993)分析了来自种子植物代表性类群的499条叶绿体*rbc*L序列, 首次基于分子证据全面探讨了被子植物的系统发育关系, 是植物分子系统发育重建研究的典范, 为后来进一步完善被子植物系统发育框架(APG II, 2003)奠定了重要的基础。随着DNA序列测定技术的发展及其成本的大幅下降, 出于对单基因片段信息量有限以及基因树不等同于物种树的担忧, 很多研究已不再满足于用一个基因的序列来重建类群的系统发育关系。Qiu等(1999)利用来自叶绿体、线粒体和核基因组的5个基因片段以及Soltis等(1999)利用来自叶绿体和核基因组的3个基因片段对被子植物系统发育关系的研究, 是系统发育重建进入多基因序列分析的代表性研究之一。如今, 这种采用多基因序列进行系统发育重建的方法逐渐被广泛采用, 并已成为当前分子系统发育研究的一种基本方法(Wendel & Doyle, 1998; Crawford, 2000; Delsuc et al., 2005; Philippe et al., 2005a)。

随着越来越多的基因片段被用于系统发育研究, 基因树冲突(conflicting gene trees)现象也不断出现, 即, 对相同的类群, 用不同基因片段建树会得到不同的分支式样或系统发育关系, 并且, 这一现象已逐渐成为分子系统发育重建中难以避免的棘手问题之一。本文首先介绍了分子系统学研究中普遍发生的基因树冲突现象以及引起基因树冲突的主要原因; 然后以笔者对稻属*Oryza* L.的系统发育基因组学研究为例, 介绍了系统发育基因组学这一新的研究方向及其在解决基因树冲突中的重要作用, 以期为目前分子系统学中的多基因序列建树研究提供一些启示和帮助。

1 基因树冲突——分子系统发育研究中日益突出的问题

在分子系统发育研究中, 利用某一基因片段提供的信息所构建的系统发育树被称为基因树(gene tree), 而反映物种之间真实进化关系的系统发育树被称为物种树(species tree)。虽然基因树不能等同于物种树, 但基因树的分支式样能够反映物种的进化历史(de Queiroz et al., 1995; Wendel & Doyle, 1998)。当用多基因来建树以避免单基因建树信息量不足所带来的误差时, 人们发现, 不同基因片段可能会展现或多或少不同的分支式样甚至严重的分歧, 即基因树之间发生冲突。

随着分子证据的不断积累, 这种基因树之间出现不一致的案例也越来越多, 似乎已成为一种普遍现象(de Queiroz et al., 1995; Wendel & Doyle, 1998; Rokas & Carroll, 2006)。以植物为例, 基因树之间的冲突广泛存在于不同的分类学水平上, 无论是在属内种间(Wendel et al., 1995; Rieseberg et al., 1996; Doyle et al., 1999; Ge et al., 1999; Cronn et al., 2002; Mason-Gamer, 2008), 科内属间(Olmstead & Sweere, 1994; Soltis & Kuzoff, 1995; Kellogg et al., 1996; Seelanan et al., 1997; Guo & Ge, 2005; Koch et al., 2007), 还是目以上或被子植物主要谱系间(Goremykin et al., 2004; Soltis et al., 2004; Stefanovic et al., 2004; Wortley et al., 2005), 甚至种子植物和其他陆生植物类群之间的各个分类学水平上均发现了基因树冲突现象(Won & Renner, 2003; Bergthorsson et al., 2004; Qiu et al., 2006)。因此, 在分子系统发育研究中, 应理解基因树和物种树之间的差异, 充分注意和分析基因树冲突的原因, 正确解释分子系统发育研究的结果。

2 基因树冲突的主要原因

不论基因树冲突的程度如何, 对于一个特定的类群, 其进化历史是唯一的, 即物种树是唯一的。当所构建的多个基因树之间出现冲突时, 实质上是

体现了部分(或者全部)的基因树没能正确反映物种进化关系。因此, 探讨基因树冲突的原因实际上是探讨基因树不能正确反映物种树的原因。迄今, 已有多篇文献对系统发育冲突的原因及其检测方法进行了总结和分析(de Queiroz et al., 1995; Johnson & Soltis, 1998; Wendel & Doyle, 1998; Philippe et al., 2005a; Jeffroy et al., 2006)。下面我们结合一些实例着重对导致基因树冲突的几个主要原因进行介绍。基因树冲突的原因大致可以归纳为以下三大方面。

(1)随机误差(stochastic error), 又可称为取样误差(sampling error)。在分子系统发育研究中, 用基因树正确推断物种树的前提是基因片段代表了它所来自的基因组, 但如何选择合适的DNA片段进行系统发育关系重建一直是个颇具争议的问题(de Queiroz et al., 1995; Wendel & Doyle, 1998)。当某个基因片段较短或信息含量较低时, 进化过程中的“噪音”(如回复突变、趋同突变、平行突变等)可能会由于取样误差而掩盖其真实的系统发育信息, 从而导致基因树的分支式样与物种树不同(Cummings et al., 1995; Wendel & Doyle, 1998)。分子系统学研究早期出现的一些基因树冲突往往是由于所用片段信息量不足而产生的, 即所谓的软性冲突(soft incongruence)(Seelanan et al., 1997)。例如, Olmstead和Sweere (1994)在利用3套分子证据对茄科Solanaceae系统发育关系进行分析时发现, Solanoideae亚科系统位置的不确定完全是由于信息量不足而引起的, 并非基因树之间有冲突。目前普遍认为, 通过增加数据量和应用更合适的分析方法能够消除随机误差对系统发育重建的影响(Cummings et al., 1995; de Queiroz et al., 1995; Wortley et al., 2005)。

(2)系统误差(systematic error)。在系统发育重建过程中, 建树所用的方法是否适合分子数据实际的进化模式, 是决定所得系统树正确与否的重要因素之一。然而, 分子进化的复杂性和多样性常常使我们现有的方法并不能很好地拟合基因真实的进化模式, 于是所建系统树就不能反映真实的进化关系(Delsuc et al., 2005; Brinkmann & Philippe, 2008)。更为重要的是, 与随机误差不同, 系统误差不会因为数据量的增加而减小, 在某些情况下随着信息量增加, 系统误差会不断加大, 导致某些分支式样得到强烈的统计支持但却未能正确地反映物种树(Huson & Bryant, 2006; Jeffroy et al., 2006)。

比较常见的系统误差主要有三种, 第一种是核苷酸成分差异所导致的系统误差(又被称为“compositional signal”)。目前我们常用的方法及模型都假设同一基因在不同生物类群中进化时都遵循相同的碱基替代模式, 因而对同一基因而言, 不同的类群拥有相似的或恒定的(stationary)碱基频率(Kumar & Gadagkar, 2001)。当这一假设不成立时, 如类群间因突变偏性(biased substitution)或选择等因素使碱基频率显著不同时, 具有相似碱基频率的类群就会被错误地聚为一支, 尽管这种相似是独立获得的, 而并非由于具有最近共同祖先(Foster & Hickey, 1999; Collins et al., 2005)。比如, Phillips等对Rokas等(2003)研究酵母属*Saccharomyces* Meyen ex Reess的物种进化关系时所用的106个基因进行分析时发现, 若换用最小进化距离法进行系统发育估计, 则会得到与Rokas的结果不同但支持率达100%的系统树, 后来发现这是由于数据中较强的碱基成分差异造成的(Phillips et al., 2004)。

第二种系统误差来自类群之间进化速率的差异(又被称为“rate signal”)。由于世代长短不一, 居群大小相异, DNA复制时的保真度不同、DNA修复效率有高有低等因素, 不同生物类群的进化速率是不同的(Andreasen & Baldwin, 2001)。考虑到核苷酸只有四种状态, 当类群间进化速率高度不一致时, 进化较快的类群可能因为多重突变在某些碱基位点随机地获得相同的碱基, 当这种非同源相似(homoplasy)足以掩盖序列中真实的历史信息时, 就会造成长枝吸引(long branch attraction)(Felsenstein, 1978)。例如, 两侧对称动物(bilaterian animal)的进化关系一直存在几种矛盾的结论, Philippe等利用现有的数据资源, 通过增加物种取样和去除快速进化的基因位点等分析发现, 前人用多基因分析所得的结论(monophyly of Coelomata)实际上是由于取样太少而造成长枝吸引的结果(Philippe et al., 2005b)。再例如植物中引起广泛争议的被子植物基部类群*Amborella trichopoda* Baill.的系统位置问题, 虽然以*Amborella* Baill.、*Austrobaileya* C. White和*Nymphaea* L.为核心的ANITA群构成被子植物基部类群这一观点早已形成共识, 但对其中究竟哪一个类群是最基部的被子植物仍存在争议。尽管主流观

点认为*Amborella*处在被子植物的最基部(Qiu et al., 1999; Soltis & Soltis, 2000; Soltis et al., 2004), 但Goremykin 等 (2003, 2004) 分别测出 *Amborella trichopoda*和*Nymphaea alba* L.的叶绿体全基因组序列并与其他13个物种叶绿体全基因组进行比较分析后提出: 单子叶植物是所有其他被子植物的姊妹群。后来, Soltis等(2004)和Stefanovic等(2004)对Goremykin的数据重新进行了分析, 发现在增加一个关键类群以后, 或者选择合适的进化模型进行分析, *Amborella*仍然是被子植物的最基部类群, 他们认为系统发育分析中模型选择不当是导致Goremykin等得出单子叶植物为被子植物基部类群的原因, 因为单子叶植物叶绿体基因的进化速率明显加快, 从而被吸引到被子植物的最基部(Stefanovic et al., 2004)。

第三种系统误差为特定碱基位点的速率差异(又被称为“heterotachous signal”或“heterotachy”, 来自希腊语, 意指“different speed”, Philippe et al., 2005c)。目前常用的替代模型已经考虑到了基因内部的不同碱基位点因功能约束不同而具有相异的进化速率(Yang, 1994; Gu et al., 1995), 但几乎所有的方法都假设功能约束不随时间改变, 反映在进化速率上, 即特定碱基位点的替代速率随着时间推移在类群之间保持恒定。然而, 在实际的分子进化过程中, 基因组某一特定位点的进化速率在类群之间是不同的, 且不同位点的速率变异可能是独立的(Philippe & Lopez, 2001; Lopez et al., 2002; Jeffroy et al., 2006)。例如, Huelsenbeck用考虑这种特定碱基位点速率差异的covarion模型对10个蛋白质编码基因和1个核糖体基因进行分析后发现, 对其中9个基因而言, covarion模型被证实明显比不考虑位点速率差异的模型能更好地拟合现实数据(Huelsenbeck, 2002)。

(3)生物学因素(biological factor)。用基于共同祖先的同源性状来建树一直是系统学研究的重要前提。表现在DNA序列上, 只有当该基因片段是反映物种分歧事件的直系同源基因(orthologs)时, 才能得到物种树的正确估计。然而进化过程中杂交(hybridization)和渐渗(introgression)、基因水平转移(horizontal gene transfer)、基因重复后的拷贝丢失(hidden paralogs)以及谱系分选(lineage sorting)等过程常常让研究者很难判断基因片段是否为直系同源, 若用违背这一前提条件的基因构建基因树时, 它所反映的是该基因片段自身的进化式样而并非物种的分歧事件。

在植物中, 杂交和渗入远比动物常见, 其本质上的网状(reticulate)进化关系很容易以基因树冲突的形式表现出来, 这种基因树的冲突尤其容易发生在单亲遗传的叶绿体或线粒体片段与双亲遗传的核基因片段之间(Soltis & Kuzoff, 1995; Doyle et al., 1999; Ge et al., 1999; Guo & Ge, 2005)。早在上世纪90年代中期, 就已经发现了100多例由于杂交和叶绿体基因渗入而造成基因树冲突的实例(Rieseberg et al., 1996)。基因水平转移则是一种基因跨物种的非生殖传播方式, 其效果与杂交和渗入有某种相似之处, 但它可以发生在亲缘关系很远的物种之间。早期曾发现在细菌的进化过程中存在大量的基因水平转移, 近10多年来在真核生物包括植物中也存在大量的基因水平转移(Wendel & Doyle, 1998)。例如, Bergthorsson等(2004)测序了被子植物基部类群*Amborella*的线粒体基因组, 并与其他陆生植物的线粒体基因进行了比较分析。结果发现, 在*Amborella*已知的31个线粒体编码蛋白基因中, 有20个基因的至少1个拷贝来自其他的陆生植物(其中有6个来自苔藓植物)。可以预期, 基于这些经历了基因水平转移的片段进行系统发育重建, 其表现出的分支式样(基因树)不可能正确地反映物种之间的进化关系(物种树)。

也许在引起基因树冲突的诸多生物学因素中, 基因重复后的拷贝丢失和谱系分选是系统发育重建中不可避免但又很难可靠鉴别的因素。近年来, 比较基因组学分析表明, 拟南芥(Simillion et al., 2002)、水稻(Paterson et al., 2004; Wang et al., 2005)、脊椎动物(Gu et al., 2002)、酵母(Wolfe & Shields, 1997)等生物中均发生了古多倍化事件, 使研究者认识到古多倍体(paleo-polyploid)在生物界中的普遍存在, 因此基因重复后又丢失拷贝造成直系同源基因和旁系同源基因(paralogs)混淆这一因素不可忽视。如图1所示, 一个祖先种经历二次物种形成事件后形成了A、B、C三个物种, 如果祖先种中某个基因发生重复形成了2个拷贝, 当不发生拷贝丢失时, 所有3个后代物种中都存在2个拷贝, 用这2个拷贝分别建树都能得到正确的物种树(图1, 上)。可是, 如果祖先种中的2个基因拷贝在后代物种中发

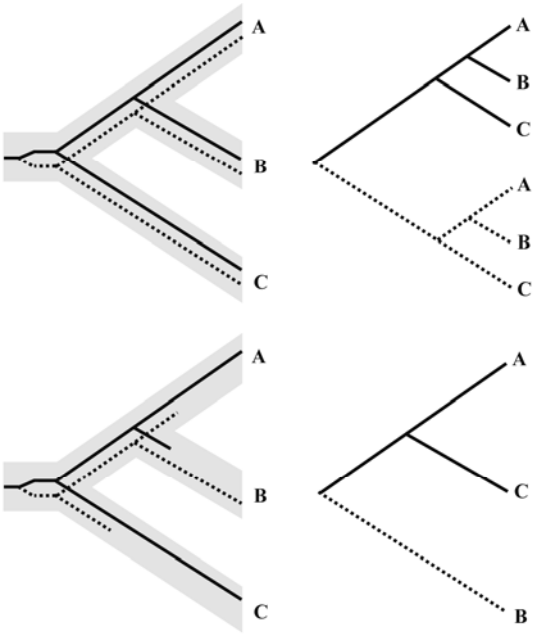

图1 物种形成过程中基因重复(或谱系分选)对系统发育重建的影响 左图显示的是物种树背景下(灰色)因基因重复而产生的两个基因拷贝(或因祖先居群中多态性而存在的两个等位基因)的进化历史, 右图显示的是两种不同情况下的基因树, 2个基因拷贝或等位基因分别用实线和虚线表示。字母A、B和C分别代表三个不同的物种。图的上半部分显示的是不发生拷贝丢失或谱系分选的情况, 图的下半部分显示的是发生拷贝丢失或谱系分选的情况。

Fig. 1. The effect on phylogenetic reconstruction by hidden paralogs (or Lineage Sorting). Left is the history of two gene copies arising from duplications (or two alleles arising from polymorphisms in the ancestral population) drawn in the context of a species tree (gray bars), and right is the gene tree drawn from the left. Two gene copies (or two alleles) are drawn in solid lines and broken lines, respectively. Letters A, B and C denote three different species. The upper half of the figure shows the case when there is no loss of gene copies or lineage sorting, and the lower half of the figure shows the case when loss of gene copies or lineage sorting does happen.

生差异性丢失, 利用该基因构建的基因树就和物种树不一致(图1, 下), 也即构建系统发育关系的基因是旁系同源片段(paralogs)。谱系分选的后果与基因重复后又丢失拷贝相似, 只需将图1中的2个拷贝(实线和虚线)理解为2个不同的等位基因而已。但值得注意的是, 谱系分选的发生与否与祖先种有效群体大小以及两次物种形成事件的间隔时间密切相关(Nei, 1987; Pamilo & Nei, 1988), 它常常发生于物种形成事件间隔短且祖先群体大的近缘类群中, 如人、黑猩猩和大猩猩(Chen & Li, 2001), 果蝇及其近缘种(Pollard et al., 2006), 稻属及其近缘种(Zou et al., 2008)的进化过程中。

除了上述主要因素之外, 还有一些其他的因素也可能导致基因树冲突, 需要在实际研究工作中加以充分注意, 比如技术上的原因如测序错误, 以及基因和基因组进化方面的原因如位点之间的协同进化、基因内部的重组、RNA编辑(RNA editing)等(Wendel & Doyle, 1998; Philippe et al., 2005a)。

3 系统发育基因组学——基因组时代的产物

面对普遍存在的基因树冲突以及产生基因树冲突的种种复杂原因, 一个随之而来的问题是, 随着分子信息的不断增多, 基因树冲突的情况可能会越来越普遍, 那么分子系统发育重建是否仍然有效? 自从1995年世界上第一例对流感嗜血杆菌 *Haemophilus influenzae* Rd.的全基因组序列报道以来(Fleischmann et al., 1995), 全基因组测序计划在全球蓬勃开展, 至2008年1月, 已经完成了866个物种的全基因组序列, 另外还有2654个物种的全基因组测序工作正在进行中(Genomes On Line Database)。随着基因组计划的实施, 开展系统发育研究的分子证据迅速积累, 一个由系统发育生物学和基因组学交叉形成的新学科——系统发育基因组学(Phylogenomics)应运而生(Eisen, 1998; Delsuc et al., 2005; Philippe et al., 2005a), 它不仅为我们从基因组层面上进行系统发育重建提供了契机, 同时也为我们正确理解和处理基因树冲突提供了有力的手段。

从系统发育生物学的角度看, 基因组学的丰富数据既包括了大量序列信息, 同时还蕴藏着有关重复基因、DNA片段缺失/插入、转座子丢失/插入等信息, 为系统发育研究提供了丰富的资料, 使得利用大规模基因组水平的数据进行系统发育分析成为可能(Delsuc et al., 2005)。系统发育基因组学是利用基因组水平的海量数据信息进行系统发育分析的新兴学科, 它是后基因组时代的产物, 也是未来进化生物学研究的重要趋势之一。随着数据量的不断增加, 分子系统发育重建中的随机误差问题将逐渐消失, 引起基因树冲突的各种生物学因素也能被逐一分析, 同时大量数据还具有“缓冲”作用, 将生物学因素的影响降至最低, 最终更可靠地推断出真实的类群间进化关系。系统发育基因组学尽管在植物类群中的应用才刚刚开始(Qiu et al., 2006; Zou

et al., 2008), 但近10年来已逐步成功运用到各大门类生物类群的系统发育重建工作中, 如原核生物(Daubin et al., 2002; Comas et al., 2007)、原生生物(Bapteste et al., 2002)、真菌(Rokas et al., 2003)、动物(Takezaki et al., 2004; Rokas et al., 2005; Savard et al., 2006)以及模式生物如果蝇及人类与其近缘物种的进化关系研究中(Chen & Li, 2001; Patterson et al., 2006; Pollard et al., 2006), 解决了一些长期困惑生物学家的问题。

用于系统发育研究的基因组数据主要包括三类: 一级序列(primary sequences), 基因组特征(whole-genome features)和基因组稀有变异(Rare genomic changes)(Delsuc et al., 2005)。其中, 用于建树的基因组特征包括基因种类(gene content或gene repertoire)、基因顺序(gene order)、低聚核苷酸串(DNA strings)等。基因组稀有变异则包括插入缺失(indel)、逆转座子插入(retroposons integration)、基因分裂与融合(gene fusion and fission)、内含子得失(gain or loss of introns)等(Rokas & Holland, 2000)。相对核苷酸序列的4种可能状态, 基因组特征和基因组稀有变异这两类性状具有数量相当可观的状态空间, 发生回复突变及趋同的机率极低, 因而对非同源相似有高度的免疫性, 同时又绕开了具争议的对位排列步骤, 具有核苷酸序列或氨基酸序列所不能比拟的优势, 因此在解决系统发育的一些关键和疑难问题时具有重要价值。比如Venkatesh等利用氨基酸序列的插入缺失和内含子得失这些基因组稀有变异性状, 解决了有颌脊椎动物的系统发育关系这个争论了近一个世纪的难题(Venkatesh et al., 2001)。又如, 水稻*Oryza sativa* L.及其6个近缘野生种具有共同的A基因组类型, 由于这些物种形成历史短且存在一定的种间基因流, 其系统发育关系尤其是A基因组基部类群一直存在争议, 已有证据确定的基部类群既有*O. meridionalis* Ng也有*O. longistaminata* Chev. & Roehr. (Zhu & Ge, 2005)。Zhu和Ge (2005)采用快速进化的核基因内含子序列重建了稻属A基因组的系统发育关系。在确定A基因组基部类群时, 他们发现了3个微型反向重复转座元件(MITE)具有重要的系统发育信息, 是很好的基因组稀有变异性状。如图2所示, 稻属A基因组的基部类群有3种可能(图2: A–C), Zhu和Ge (2005)发现3个MITE均插入到除*O. meridionalis*以外所有类群相关基因的内含子中, 这一结果强烈支持*O. meridionalis*为A基因组的基部类群(图2: A), 因为若图2A为真实物种树, 则3个MITE的存在只需要3次进化事件就可以完成; 可是, 如果承认其他任何一种进化关系(图2: B, C)为真实物种树, 3个MITE的存在就必须通过至少6次进化事件才能完成, 所以图2A的进化关系是最简约的。

鉴于对基因组特征和基因组稀有变异两类性状的利用起步较晚, 相关算法和模型还不完善, 且现有基因组数据仅覆盖生物类群的很小一部分, 因

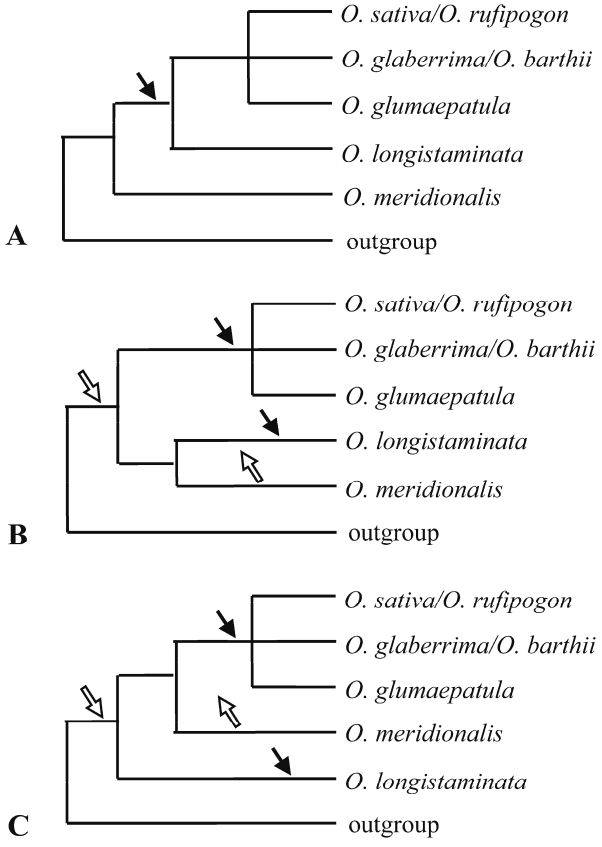

图2 通过MITE插入性状确定稻属A基因组的基部类群 A、B、C分别表示3种可能的系统发育关系, 向下箭头表示MITE的获得, 向上箭头表示MITE的丢失。**A.** 每个MITE的存在只需一次插入就可以解释。**B, C.** 每个MITE的存在均需两次插入(实心箭头)或一次插入和一次丢失(空心箭头)才能解释(引自Zhu & Ge, 2005)。

Fig. 2. The basal lineage of A-genome in *Oryza* was resolved by MITEs (miniature inverted-repeat transposable elements). Three phylogenetic hypotheses regarding the basal lineage of the A-genome species were shown in Fig. 2A, 2B and 2C. Downward and upward arrows stand for the insertion and excision of MITEs, respectively. In Fig. 2A, the most parsimonious explanation of the occurrence for a MITE requires only one insertion event. In Fig. 2B and 2C, at least two evolutionary events are required for accounting for the occurrence of a MITE, i.e., two independent insertions (closed arrows), or alternatively one insertion and one deletion (open arrows) (from Zhu & Ge, 2005).

此在系统发育研究中的应用还有很大的发展空间(Boore, 2006)。相对而言, 一级序列在系统发育研究中的利用在方法上要成熟得多。分子系统学研究长期积累下来的对分子序列进化规律的认识以及利用单基因重建类群进化历史的原理和方法都可以运用到系统发育基因组学研究中。目前应用较多的多基因建树策略主要包括直接合并分析(total evidence, or supermatrix approach)(Kluge, 1989)和独立分析(separate analysis, or supertree approach)(Miyamoto & Fitch, 1995)。直接合并分析是先将来自不同基因的序列首尾拼接成一个统一的数据矩阵再进行系统发育估计的方法, 最初由Kluge (1989)提出, 基于哲学家Rudolf Carnap提出的总体证据原则, 即对任何理论或假说的判断都必须基于全部证据, 认为合并数据能得到最大化的系统发育信息, 尤其是能揭示数据中隐藏的系统发育信息(Rieppel, 2005)。独立分析方法则考虑到数据之间不同程度的异质性, 先对单个基因分别建树, 然后再对所得到的一系列系统树进行整合, 最后以一棵系统树来代表最终的系统发育关系。系统树整合时可以采用简约矩阵法(matrix representation with parsimony, MRP)、严格一致法(strict consensus)、半严格一致法(semi-strict consensus)、平均一致法(average consensus procedure)等, 整合而成的系统树含有源数据集当中所有的类群(Bininda-Emonds et al., 2002; Delsuc et al., 2005; Philippe et al., 2005a)。比如, Salamin等对现有55个已发表的涉及禾本科Poaceae的系统树进行整合, 用简约矩阵法建立了一棵含395属、迄今类群覆盖度最高的禾本科系统树(Salamin et al., 2002)。

直接合并分析策略充分利用了各个数据集的所有信息, 独立分析策略的直接操作对象是系统树, 利用的仅是从各数据中总结出的树形信息, 因而忽略了很多系统发育信息。然而, 直接合并分析需要所有基因具有相同的一组类群, 独立分析策略则仅要求不同数据之间含有部分重叠的类群, 并且对各种类型的数据都适用(如形态性状、分子标记、分子序列等), 因此有利于对各类数据进行信息整合, 它常被用于将不同性质和不同取样的数据整合以构建分类学水平较高的大系统树乃至生命之树(Tree of Life)(Bininda-Emonds et al., 2002)。

4 基因树冲突与系统发育基因组学研究——以稻属研究为例

目前, 用系统发育基因组学方法来探讨类群的进化关系在植物界中还非常少。作为重要的模式植物, 水稻两个亚种的全基因组测序工作均已完成, 这为利用系统发育基因组学手段解决稻属的系统发育关系以及探讨基因树冲突的机制提供了绝佳的机会和基础。

从上世纪初开始, 作为世界上最重要的作物之一, 水稻一直受到生物学家的关注, 水稻所在稻属的系统发育关系研究也在许多国家展开(Nayar, 1973; Second, 1985; Wang et al., 1992), 但直到上世纪末才出现第一篇采用多基因片段对整个稻属进行系统发育重建的报道(Ge et al., 1999)。在这项研究中, Ge等(1999)采用2个单拷贝核基因(*Adh1*和*Adh2*)和1个叶绿体基因(*mat*K)片段构建了稻属全部23个物种的系统发育关系, 并据此定义了稻属第10个基因组(HK); 与此同时, 提出了稻属全部10个基因组的进化关系, 并揭示了稻属多倍体的起源方式及其亲本来源(图3: A)。然而, 该研究也引出了一个问题: *Adh2*和*mat*K基因树强烈支持A和B基因组为姊妹群, 但*Adh1*基因树却支持A和C基因组为姊妹群, 也即在A、B、C基因组之间关系上出现基因树冲突, 在图3A中用多歧分支表示。此外, 稻属的基部类群归属也没能得到明确分辨(Ge et al., 1999)。值得一提的是, A、B、C基因组之间3种可能的系统发育关系均得到过去不同研究证据的支持, 稻属基部类群的归属也涉及了G、F和其他多倍体基因组类群(Zou et al., 2008)。

为此, 我们以分别来自稻属全部6个二倍体基因组的代表物种为对象, 以稻属近缘属假稻属*Leersia* Sol. ex Sw.为外类群, 基于水稻全基因组序列设计并扩增了遍布12条染色体的142个核基因片段。通过对142个基因的直接合并分析, 不管是采用最大似然法(ML)、最大简约法(MP)还是贝叶斯方法(BI), 我们均得到了一棵有完全分辨、树形相同且所有分支均得到显著统计支持的系统树(图3: B)。在该树上, A和B基因组为姊妹群, G基因组为稻属的基部类群。随后, 我们进一步用各种检测方法评估了合并数据的系统误差, 结果发现, 合并数据的系统

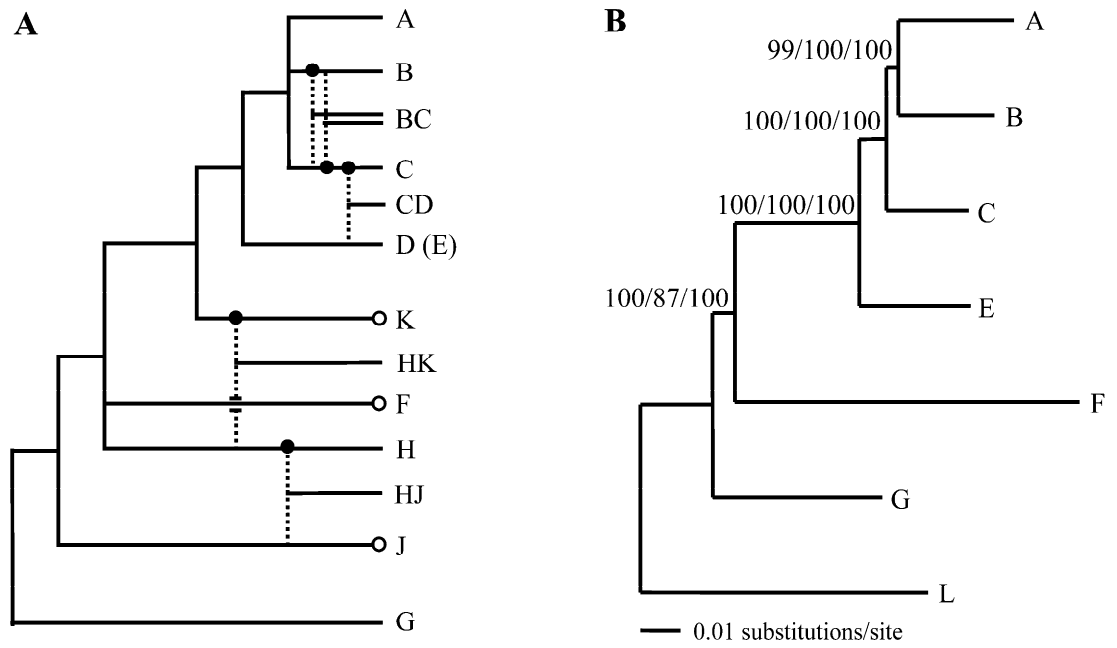

图3 稻属各基因组之间的系统发育关系 **A.** 基于*Adh1*、*Adh2*和*mat*K所得到的稻属基因组间进化关系示意图, 虚线显示异源多倍体的起源, 实心圆代表异源多倍体的母本供体, 空心圆代表迄今还未发现(或已灭绝)的基因组类型(引自Ge et al., 1999)。**B.** 基于142个基因位点合并序列用最大似然法(ML)建立的系统树, 最大简约法(MP)和贝叶斯方法(BI)也得到了相同的树形。大写字母A、B、C、E、F和G分别代表稻属中所有的二倍体基因组类型, L代表外类群。各分支上的数字分别为ML、MP方法的自展检验支持率和BI方法的后验概率(引自Zou et al., 2008)

Fig. 3. The phylogenetic relationship of rice genomes. **A,** Evolutionary relationships of the rice genomes inferred from *Adh1*, *Adh2*, and *mat*K gene phylogenies. Dashed lines indicate origins of allotetraploids. Solid circles indicate the maternal parents of the tetraploids and open circles indicate the unidentified diploid genomes (from Ge et al., 1999). **B,** The maximum likelihood (ML) tree inferred from the concatenated sequences of 142 genes. The same topology was obtained from maximum parsimony (MP) and Bayesian inference (BI). Capital letters A, B, C, E, F, and G represent all recognized diploid genome types of *Oryza*, and L represents the outgroup. Numbers above branches indicate bootstrap supports of ML and MP, and posterior probability of BI, respectively (from Zou et al., 2008).

发育重建并未受到核苷酸成分差异、进化速率差异以及特定碱基位点速率差异等系统误差的影响。因此, 该系统树(Zou et al., 2008)应该反映了类群真实的进化关系。

在上述解决稻属主要谱系系统发育关系的过程中, 一个有趣的现象是, 当我们对142个基因位点分别进行独立建树时, 得到40多种分支式样相互冲突的系统树, 这为深入探讨基因树冲突的原因提供了很好的机会。通过一致性网络方法(consensus networks)(Huson & Bryant, 2006)对单基因树进行分析后发现, 53%的基因支持(AB)C关系, 支持(AC)B和(BC)A关系的基因分别占21%和26%; 同理, 支持G为稻属基部类群的基因占45%, 支持F基因组为基部类群的基因占30%, 而支持FG为单系共同构成稻属基部类群的基因占25% (图4: A)。简言之, 绝大多数的单基因树支持物种树。

那么, 基因树冲突的原因是什么? Zou等(2008)进行的一系列统计分析表明, 随机误差、系统误差以及生物学因素中的基因水平转移、杂交/渐渗、基因重复后的拷贝丢失等均无法解释上述基因树冲突, 而谱系分选才是基因树冲突的主要原因。这一解释得到了基因在染色体上分布式样的支持。如图4B所示, 当我们将支持不同分支式样的基因标注在染色体上时, 可以发现支持每种分支的基因随机分布在12条染色体上(卡方检验, P=0.233–0.823), 即单基因树的分支式样与其在染色体上的位置没有明显相关性, 这是区别杂交/渗入和谱系分选的重要标准之一。更为重要的是, 理论研究表明, 当连续物种形成事件间隔很短(如辐射进化), 谱系分选的后果将十分严重, 尤其是在祖先有效群体很大时(Pamilo & Nei, 1988)。为此, 我们采用基于溯祖理论(Coalescence theory)的方法, 根据ABC三个基因组间物种形成的间隔及有效群体大小, 得出如果在ABC基因组进化过程中发生谱系分选, 那么基因树

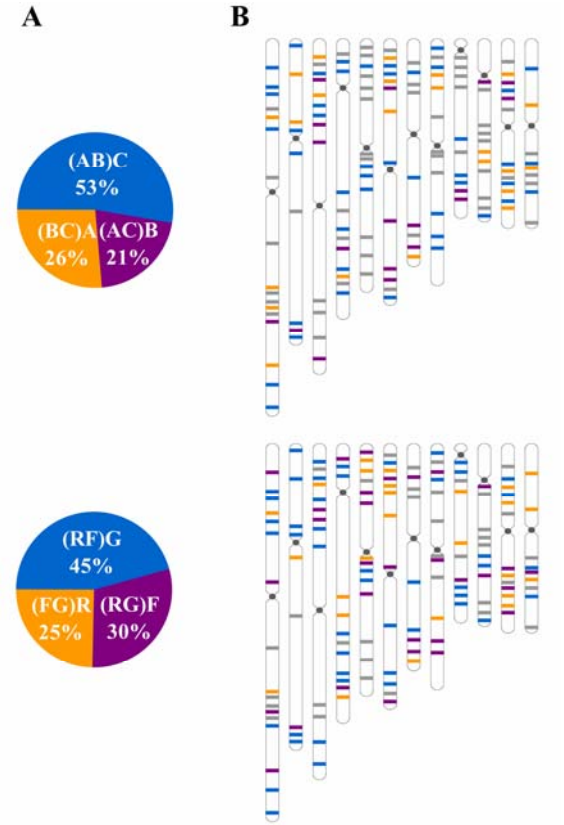

图4　基因树冲突及其特点　**A.** 饼图显示的是支持ABC基因组间3种不同关系的基因百分比(上)和支持三种不同稻属基部类群基因的百分比(下), R表示A、B、C和E基因组的集合。**B.** 支持各种分支的142个基因(颜色对应于饼图)在水稻12条染色体上分布(引自Zou et al., 2008)。

Fig. 4. Genome-wide incongruence among gene trees and its characteristics. **A,** Pie graphs indicate the proportions of gene trees that support alternative relationships between A, B, and C genomes (above) and the basal lineages in *Oryza* (below). R represents the rest of the genome types, including A-, B-, C-, and E-genomes. **B,** Illustration of the relative physical locations of the 142 sampled genes (in the corresponding colors) on the 12 rice chromosomes (from Zou et al., 2008).

反映物种树的概率将不大于0.62，与我们实际得到的结果0.53相吻合(Zou et al., 2008)。因此，根据142个基因的系统发育基因组研究，我们认为在稻属的进化过程中发生了两次物种快速分化事件，由于谱系分选作用造成在利用现有物种基因序列来重建这些分化事件时基因树不能正确反映物种树，呈现出基因组水平的基因树冲突现象。

类似稻属快速物种形成而导致系统发育树中某些分支难以分辨或出现基因树冲突的情况在许多动植物类群中也有报道(Takezaki et al., 2004; Rokas et al., 2005; Rokas & Carroll, 2006)，要解决这类快速分化的物种进化关系往往需要大量的基因位点或信息。例如，对上述稻属142个基因数据的进一步随机抽样分析表明(图5)，若用以95%概率获得正确系统发育估计做衡量标准，以基因为单位取样时，至少需要120个基因才能正确分辨A、B和C基因组之间的关系；而要正确分辨稻属基部类群，ML法需要80个以上的基因，而MP法需要120个以上的基因。若以碱基位点为单位取样，则两种方法用40 kb就足以正确分辨A、B和C基因组之间的关系，这相当于抽取46个基因；而要正确分辨稻属基部类群，ML法需要40 kb，MP法则需要80 kb(相当于92个基因) (图5)。

5　基因组时代的系统发育研究——机遇与挑战

面对日益突出的基因树冲突现象，系统发育基因组学的兴起给我们带来了难得的机会，同时也给系统发育重建研究提出了新的挑战。多基因系统发育重建的大量实例以及我们对稻属系统发育基因组学的研究，提供了如下几点重要的启示。首先，系统树冲突本身提供了重要的生物学信息，通过一定的分析(包括增加基因片段、增加取样类群或采用不同的分析方法等)可以揭示一些重要的生物学现象和过程，如杂交/渐渗和适应性辐射或快速物种形成等等。需要强调的是，尽管本文的重点在讨论基因树(或分子证据)之间的冲突及其原因，不可忽视的是，系统发育冲突可以发生在不同类型的证据

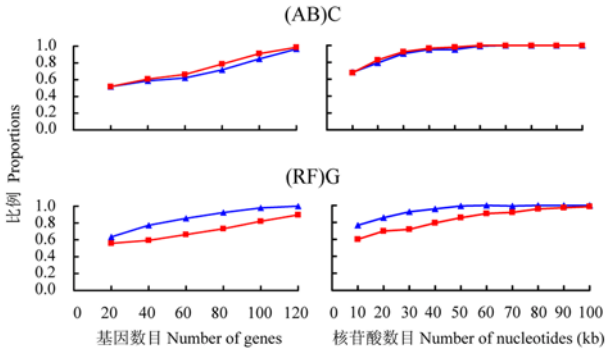

图5　对142个基因进行不同规模随机抽样分析所得到的正确分支的比例　ML法和MP法结果分别用三角和方块表示，基因组类型用与图4相同的大写字母来表示。

Fig. 5. The proportions of correct clades based on resampling of 142 gene sequences at various scales. Results of ML and MP analyses are indicated by triangles and diamonds, respectively. Genome types are represented with the same capital letters as in Fig. 4.

之间, 如形态学证据和分子证据之间(de Queiroz et al., 1995; Futuyma, 1998; Wendel & Doyle, 1998), 其发生原理和解决办法与对待基因树冲突并没有原则上的不同。

其次, 通过采用基因组水平的多基因序列数据, 一些系统发育难题有可能迎刃而解。如在稻属研究中(Zou et al., 2008), 系统发育基因组学的方法克服了少数基因片段所存在的信息量不足以及谱系分选对构建系统树所带来的“噪音”, 在存在广泛基因树冲突的前提下获得了对物种树的正确估计, 充分说明系统发育基因组学在解决类群进化关系问题中的巨大潜力和广阔的应用前景, 而这一趋势已被来自大量动物类群的研究所充分证实(Chen & Li, 2001; Philippe et al., 2005b; Rokas et al., 2005; Savard et al., 2006; Patterson et al., 2006; Pollard et al., 2006)。

第三, 考虑到迄今已完成全基因组测序的数百个物种仅仅是地球生物多样性中的沧海一粟, 系统发育基因组学方法在短期内还不能广泛应用到大量动植物类群中, 多基因系统发育研究方法仍然是现实有效的手段, 因为通过少数基因片段的研究可以分辨出系统发育树上大多数分支, 同时揭示出基因树冲突之所在, 这些反映基因树冲突的分支本身就是系统发育分析的重要结果。如前述稻属的研究中, 早期3个基因片段所构建的系统发育框架以及基因树冲突现象并没有因为142个基因的系统发育基因组学研究而改变, 只是表现基因树冲突的两个分支得以彻底解决。生物进化过程中的诸多因素等都会影响系统发育基因组学方法解决系统发育关系的能力, 比如对古老的快速分化事件而言, 由于现存系统发育信息的缺乏以及高度非同源相似的存在, 可能在现有技术条件下用再多的基因序列也无法解决, 这时候, 基因组特征和基因组稀有变异等在解决进化关系问题上可能更有价值(Delsuc et al., 2005; Rokas & Carroll, 2006; Whitfield & Lockhart, 2007)。

第四, 在系统发育基因组学研究时, 选取基因片段的数目及每个片段的长度也是值得认真考虑的问题。前述稻属的系统发育基因组学研究表明, 以碱基位点为单位的随机取样比以基因为单位的随机取样更容易获得正确的系统树(图5), 这可能是由于同一基因内的碱基位点连锁较强的缘故(Cummings et al., 1995)。因此, 在运用系统发育基因组学方法时, 在数据量一定的情况下, 随机选用多个独立的短片段比选用少数的长片段可能更为有效。

最后, 应该引起重视的是, 生物类群的快速分化发生在生物进化历史的各个时期和各种生物类群中。从5亿多年前寒武纪大爆发中动物各大类群的辐射进化(Rokas et al., 2005), 到白垩纪被子植物的起源和分化(Soltis et al., 2004; Lockhart & Penny, 2005), 再到近期果蝇及其近缘种(Pollard et al., 2006)以及人、黑猩猩与大猩猩的分化(Chen & Li, 2001; Enard & Paabo, 2004; Patterson et al., 2006), 物种快速分化事件几乎遍布生命之树的各个部分。与类群快速分化相伴随的谱系分选过程使得物种基因组处在一种系统发育信息相互矛盾的镶嵌状态(mosaics)(Enard & Paabo, 2004; Pollard et al., 2006), 用不同的片段很可能得到不同的基因树。因此, 在面对可能是由快速分化而形成的生物类群时, 我们应该保持高度警惕, 不能轻易相信基于单基因树所建立的系统发育关系, 不论这一基因树是否完全分辨或者得到多高的统计支持。

系统发育重建已经进入了基因组时代, 随着大量的基因或基因组信息不断应用到系统发育重建的研究中, 基因树的冲突可能会成为一种普遍现象。尽管基因树冲突为系统发育重建带来了困难, 但我们也应该意识到, 当排除随机误差和系统误差的影响之后, 基因树之间冲突的存在为生物进化机制研究提供了宝贵线索, 为我们进一步解读生命的本质问题打开了一扇大门, 这或许比得到一棵物种树更有意义。

致谢 国家自然科学基金(30430030, 30121003)资助。

参考文献

Andreasen K, Baldwin BG. 2001. Unequal evolutionary rates between annual and perennial lineages of checker mallows (*Sidalcea*, Malvaceae): evidence from 18S-26S rDNA internal and external transcribed spacers. Molecular Biology and Evolution 18: 936–944.

APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of Linnean Society 141: 399–436.

Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Müller M,

Philippe H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proceedings of the National Academy of Sciences USA 99: 1414–1419.

Bergthorsson U, Richardson AO, Young GJ, Goertzen LR, Palmer JD. 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proceedings of the National Academy of Sciences USA 101: 17747–17752.

Bininda-Emonds ORP, Gittleman JL, Steel MA. 2002. THE (SUPER) TREE OF LIFE: Procedures, Problems, and Prospects. Annual Review of Ecology and Systematics 33: 265–289.

Boore JL. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology & Evolution 21: 439–446.

Brinkmann H, Philippe H. 2008. Animal phylogeny and large-scale sequencing: progress and pitfalls. Journal of Systematics and Evolution 46: 274–286.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim KJ, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang QY, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA. 1993. Phylogenetics of seed plants: an analysis of nucleotide-sequences from the plastid gene *rbc*L. Annals of the Missouri Botanical Garden 80: 528–580.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. American Journal of Human Genetics 68: 444–456.

Collins TM, Fedrigo O, Naylor GJ. 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. Systematic Biology 54: 493–500.

Comas I, Moya A, Gonzalez-Candelas F. 2007. From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma—Proteobacteria as a test case. Systematic Biology 56: 1–16.

Crawford DJ. 2000. Plant macromolecular systematics in the past 50 years: one view. Taxon 49: 479–501.

Cronn RC, Small RL, Haselkorn T, Wendel JF. 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany 89: 707–725.

Cummings MP, Otto SP, Wakeley J. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Molecular Biology and Evolution 12: 814–822.

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Research 12: 1080–1090.

de Queiroz A, Donoghue MJ, Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. Annual Review of Ecology and Systematics 26: 657–681.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361–375.

Doyle JJ, Doyle JL, Brown AH. 1999. Incongruence in the diploid B-genome species complex of *Glycine* (Leguminosae) revisited: histone H3-D alleles versus chloroplast haplotypes. Molecular Biology and Evolution 16: 354–362.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research 8: 163–167.

Enard W, Paabo S. 2004. Comparative primate genomics. Annual Review of Genomics and Human Genetics 5: 351–378.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27: 401–410.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Manna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. Journal of Molecular Evolution 48: 284–290.

Futuyma DJ. 1998. Evolutionary biology. Sunderland, MA: Sinauer Associates.

Ge S, Sang T, Lu BR, Hong DY. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proceedings of the National Academy of Sciences USA 96: 14400–14405.

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Molecular Biology and Evolution 20: 1499–1505.

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Molecular Biology and Evolution 21: 1445–1454.

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Molecular Biology and Evolution 12: 546–557.

Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nature Genetics 31: 205–209.

Guo YL, Ge S. 2005. Molecular phylogeny of Oryzeae (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. American Journal of Botany 92: 1548–1558.

Haeckel E. 1866. Generelle Morphologie der Organismen. Vol. 2: Allgemeine Entwickelungsgeschichte der Organismen. Berlin: Reimer.

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Molecular Biology and Evolution 19:

698–707.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23: 254–267.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics 22: 225–231.
Johnson LA, Soltis DE. 1998. Assessing congruence: empirical examples from molecular data. In: Soltis DE, Soltis PS, Doyle JJ eds. Molecular systematics of plants II: DNA sequencing. Boston: Kluwer. 297–348.
Kellogg EA, Appels R, MasonGamer RJ. 1996. When genes tell different stories: The diploid genera of Triticeae (Gramineae). Systematic Botany 21: 321–347.
Kluge AG. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Systematic Zoology 38: 7–25.
Koch MA, Dobes C, Kiefer C, Schmickl R, Klimes L, Lysak MA. 2007. Supernetwork identifies multiple events of plastid *trn*F (GAA) pseudogene evolution in the Brassicaceae. Molecular Biology and Evolution 24: 63–73.
Kumar S, Gadagkar SR. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics 158: 1321–1327.
Li WH. 1997. Molecular evolution. Sunderland, MA: Sinauer Associates.
Lockhart PJ, Penny D. 2005. The place of *Amborella* within the radiation of angiosperms. Trends in Plant Science 10: 201–202.
Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Molecular Biology and Evolution 19: 1–7.
Mason-Gamer RJ. 2008. Allohexaploidy, introgression, and the complex phylogenetic history of *Elymus repens* (Poaceae). Molecular Phylogenetics and Evolution 47: 598–611.
Miyamoto MM, Fitch WM. 1995. Testing species phylogenies and phylogenetic methods with congruence. Systematic Biology 44: 64–76.
Nayar NM. 1973. Origin and cytogenetics of rice. Advances in Genetics 17: 153–292.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford: Oxford University Press.
Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics: an empirical approach using 3 molecular data sets in the Solanaceae. Systematic Biology 43: 467–481.
Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5: 568–583.
Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences USA 101: 9903–9908.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441: 1103–1108.
Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005a. Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36: 541–562.
Philippe H, Lartillot N, Brinkmann H. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular Biology and Evolution 22: 1246–1253.
Philippe H, Lopez P. 2001. On the conservation of protein sequences in evolution. Trends in Biochemical Sciences 26: 414–416.
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evolution Biology 5: 50.
Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21: 1455–1458.
Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genetics 2: e173.
Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404–407.
Qiu YL, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L, Rest J, Estabrook GF, Hendry TA, Taylor DW, Testa CM, Ambros M, Crandall-Stotler B, Duff RJ, Stech M, Frey W, Quandt D, Davis CC. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proceedings of the National Academy of Sciences USA 103: 15511–15516.
Rieppel O. 2005. The philosophy of total evidence and its relevance for phylogenetic inference. Papeis Avulsos de Zoologia 45: 77–89.
Rieseberg LH, Whitton J, Linder CR. 1996. Molecular marker incongruence in plant hybrid zones and phylogenetic trees. Acta Botanica Neerlandica 45: 243–262.
Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biology 4: e352.
Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology & Evolution 15: 454–459.
Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310: 1933–1938.
Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.
Salamin N, Hodkinson TR, Savolainen V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). Systematic Biology 51: 136–150.
Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Research 16: 1334–1338.
Second G. 1985. A new insight into the genome differentiation in *Oryza* L. through isozymic studies. In: Sharma AK, Sharma A eds. Advances in chromosome and cell genetics. New Delhi: Oxford and IBH. 45–78.

Seelanan T, Schnabel A, Wendel JF. 1997. Congruence and consensus in the cotton tribe (Malvaceae). Systematic Botany 22: 259–290.

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences USA 99: 13627–13632.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, Soltis PS. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends in Plant Science 9: 477–483.

Soltis DE, Kuzoff RK. 1995. Discordance between nuclear and chloroplast phylogenies in the *Heuchera* Group (Saxifragaceae). Evolution 49: 727–742.

Soltis ED, Soltis PS. 2000. Contributions of plant molecular systematics to studies of molecular evolution. Plant Molecular Biology 42: 45–75.

Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402: 402–404.

Stefanovic S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evolutionary Biology 4: 35.

Takezaki N, Figueroa F, Zaleska-Rutczynska Z, Takahata N, Klein J. 2004. The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of forty-four nuclear genes. Molecular Biology and Evolution 21: 1512–1524.

Venkatesh B, Erdmann MV, Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proceedings of the National Academy of Sciences USA 98: 11382–11387.

Wang X, Shi X, Hao B, Ge S, Luo J. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytologist 165: 937–946.

Wang ZY, Second G, Tanksley SD. 1992. Polymorphism and phylogenetic-relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. Theoretical and Applied Genetics 83: 565–581.

Wendel JF, Doyle JJ. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. In: Soltis DE, Soltis PS, Doyle JJ eds. Molecular systematics of plants II: DNA sequencing. Boston: Kluwer. 265–296.

Wendel JF, Schnabel A, Seelanan T. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proceedings of the National Academy of Sciences USA 92: 280–284.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends in Ecology & Evolution 22: 258–265.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.

Won H, Renner SS. 2003. Horizontal gene transfer from flowering plants to *Gnetum*. Proceedings of the National Academy of Sciences USA 100: 10824–10829.

Wortley AH, Rudall PJ, Harris DJ, Scotland RW. 2005. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. Systematic Biology 54: 697–709.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39: 306–314.

Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. New Phytologist 167: 249–265.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biology 9: R49.