植物学报 *Chinese Bulletin of Botany* 2023, **58** (2): 261–273, www.chinbullbotany.com
doi: 10.11983/CBB22224

·特邀专家方法·

基于分子数据的系统发生树构建

彭焕文[1, 2, 3], 王伟[1, 2, 3*]

[1]中国科学院植物研究所, 系统与进化植物学国家重点实验室, 北京 100093; [2]国家植物园, 北京 100093
[3]中国科学院大学, 北京 100049

摘要 系统发生学是研究生物类群间进化关系的学科。随着测序技术、分析方法和计算能力的改进, 分子数据被广泛应用, 促进了系统发生学的快速发展。系统发生树已成为生态学和比较生物学等研究领域的有力工具。然而, 许多研究在进行系统发生树构建时更侧重各种软件的使用, 一些基本原则或注意事项有时会被弱化甚至忽视。该文详细介绍了基于分子数据进行系统发生树构建的工作流程和基本方法, 包括类群取样、分子标记选择、序列比对、分区及模型选择、序列联合分析以及拓扑结构检验等关键步骤。此外, 该文还为系统发生树构建常用的3种方法(最大简约法、最大似然法和贝叶斯法)提供了相应的软件操作流程和运行命令, 以期为相关研究提供参考。

关键词 系统发生树构建, 分子系统学, 联合分析, 核苷酸替换模型

彭焕文, 王伟 (2023). 基于分子数据的系统发生树构建. 植物学报 **58**, 261–273.

1 系统发生构建概述

系统发生构建(phylogenetic reconstruction)指利用各种性状推断生物类群间的进化关系, 这种进化关系通常以系统发生树(phylogenetic tree)的形式来展示。传统方法根据广义形态性状(外部形态、解剖、胚胎、孢粉和植物化学等)构建类群之间的亲缘关系。但形态性状易受环境饰变影响, 即同塑性(homoplasy)较高, 且可获得的性状较少。Nandi等(1998)利用251个广义形态性状以及叶绿体*rbcL*基因序列构建了被子植物科水平的系统发生树, 这是迄今为止包含形态性状数目最多的植物系统发生分析。相对于形态性状, 分子数据具有数量多、可遗传以及易确定同源性等优点, 能在亲缘关系较远的类群之间进行直接比较(Nei, 1996; Whelan et al., 2001)。因此, 利用分子数据建立的系统发生关系能更真实地反映类群的进化历史。自Chase等(1993)利用*rbcL*基因构建当时规模最大的种子植物系统发生树以来, 分子系统学(molecular systematics)作为植物系统学的分支学科逐渐趋于成熟。在过去的30年中, 尤其是随着第二代和第三代测序技术的出现, 分子数据呈井喷式增长并被上传至公共数据库, 可服务于全球的科研人员, 极大地促进了分子系统学的发展, 也使植物生命之树重建研究取得显著成果(葛颂, 2022; Liu et al., 2022)。同时, 系统发生树被广泛应用于生态学和比较生物学研究, 并为公众所熟知(Benton and Ayala, 2003)。在2009年, Soltis等(2009)就宣称植物分子系统学, 甚至整个生物系统学的发展已进入黄金时代(golden era)。

总之, 系统发生树是所有生物学研究的基础, 不仅对进化至关重要, 也成为生态学及比较生物学等研究领域的有力工具(Soltis et al., 1999; 范凯等, 2021; 康凯程等, 2021)。然而, 在测序技术和建树方法快速发展的同时, 利用分子数据进行系统发生树构建的一些基本原则或注意事项(如何合理地进行类群取样和分子标记选择, 如何从海量数据中筛选出正确的序列, 如何选择合适的建树方法等)时常被弱化甚至忽视。本文对基于分子数据(主要是DNA序列)进行系统发生构建的流程和方法进行概括介绍, 以期为相关研究提供参考。

收稿日期: 2022-09-19; 接受日期: 2022-11-12
基金项目: 中国科学院战略性科技先导专项(B类) (No.XDB31030000)和国家自然科学基金(No.32170210, No.32011530072, No.31770233, No.31770231)
* 通讯作者。E-mail: wangwei1127@ibcas.ac.cn

2 系统发生树构建的科研设计和工作流程

利用分子数据可对生物类群或基因家族进行系统发生树构建，本文以对生物类群分析为例进行介绍。系统发生树构建主要包括取样、矩阵组装和系统树构建3部分(图1)，具体包括类群取样、分子标记选择、序列获得、序列比对、分区及模型选择、序列联合分析和拓扑结构检验等关键步骤。

2.1 取样

2.1.1 类群取样

合理地进行类群取样是正确构建系统发生树和分析物种间进化关系的前提。类群取样分为内类群(ingroup)取样和外类群(outgroup)取样。

2.1.1.1 内类群取样

内类群即研究的目标类群。内类群取样应具有代表性，即能代表所研究类群在分类、形态变异以及地理分布上的多样性。对于形态异质或地理广布的类群，取样的代表性至关重要。例如，防己科通常为大的木质藤本，但科内的宽筋藤属(*Tinospora*)却既有木本物种，又有草本物种，且广泛分布于旧世界的热带地区(包括热带亚洲、澳洲和非洲)。仅包括亚洲和澳洲木本物种的系统发生分析支持该属为一单系群(Wang et al., 2012)，但Wang等(2017)在类群取样涵盖了习性和地理分布的多样性之后，发现2个草本物种形成1个独立的支系，与其它木本物种关系较远，据此为这

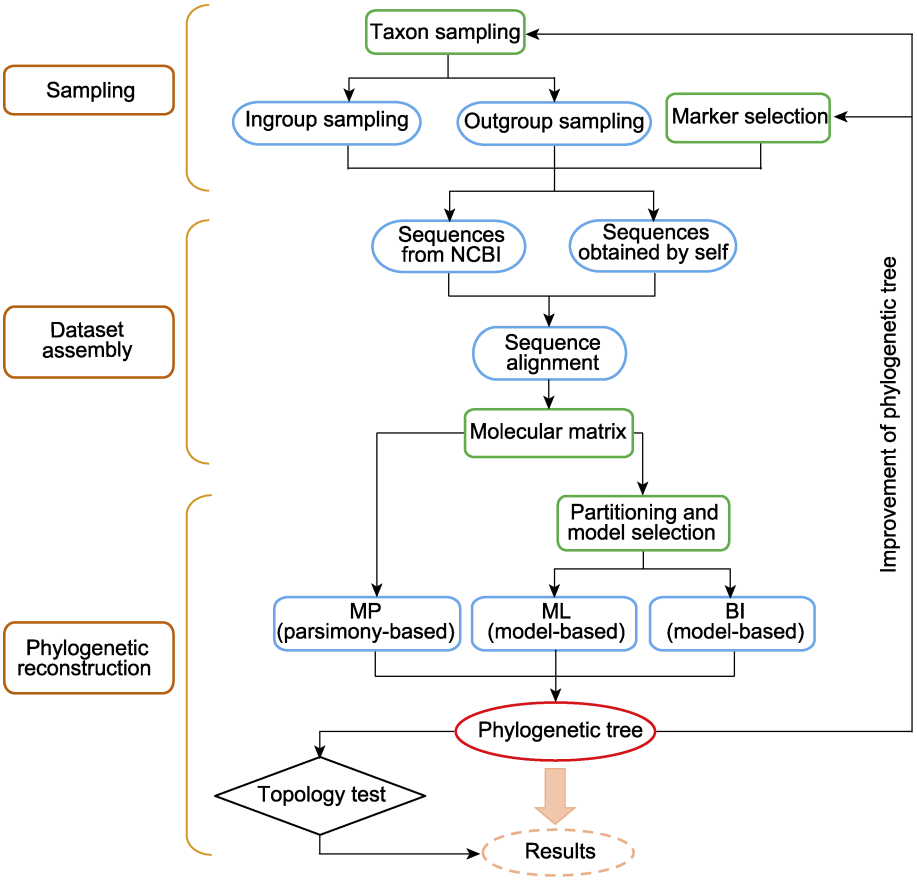

图1 基于分子数据构建系统发生树的流程
MP: 最大简约法; ML: 最大似然法; BI: 贝叶斯法

Figure 1 The pipeline of phylogenetic tree reconstruction based on molecular data
MP: Maximum parsimony; ML: Maximum likelihood; BI: Bayesian inference

2个草本物种建立了新属——青牛胆属(*Paratinospora*); 同时, 发现非洲的*Tinospora caffra*也是1个独立的支系, 结合形态证据将其归入重新恢复的琉萼藤属(*Hyalosepalum*)。随后, 通过进一步扩大类群取样, Lian等(2019)发现传统上置于宽筋藤属的物种隶属5个不同的支系, 即青牛胆属、琉萼藤属(包括原宽筋藤属非洲的*T. caffra*和*T. oblongifolia*)、*Fawcettia* (包括原宽筋藤属澳洲的*T. tinosporoides*)、*Chasmanthera* (包括原宽筋藤属非洲的*T. uviformis*)以及由其它宽筋藤属物种组成的狭义宽筋藤属。

2.1.1.2 外类群取样

外类群是研究者为正确推断内类群的系统发生关系, 在其研究对象之外选取的与内类群有密切关系的分类单元。首先, 鉴于形态性状的同塑性较高, 外类群取样最好根据大尺度分子系统学的结果, 而非形态学结果。其次, 所选外类群与内类群之间的距离不应太远, 且注意避免选取具有长枝的类群, 以减少进化噪音和规避长枝吸引(long-branch attraction)。此外, 选取的外类群数量通常不少于2个(Lozano-Fernandez, 2022)。例如, 在研究对象是1个属的情况下, 建议在该属所在族的其它属中各选1个代表种作为外类群, 再选取邻近族的代表属用于在建树时置根。这样才能检测内类群是否为单系。

2.1.2 分子标记的选择

不同基因、基因的不同区域以及编码氨基酸不同位置的密码子等具有不同的功能, 因而其进化速率差别极大。根据进化速率的不同, 分子标记可分为进化速率慢的标记(slow marker)和进化速率快的标记(fast marker)。进化速率慢的标记含有的进化噪音较少, 非同源相似性水平较低, 且序列比对相对容易, 如基因编码区。进化速率快的标记受到的功能限制较小, 更接近中性突变, 且含有更多的信息位点, 但因不同物种序列长度变异较大导致序列比对困难, 如基因间隔区和内含子序列。在实际应用中, 如何根据研究分类阶元的高低选择合适的分子标记并无具体要求。Borsch等(2003)研究发现, 在重建基部被子植物进化关系方面, 使用进化速率快的*trnT-trnF*序列和使用多个进化速率慢的保守基因效果相似, 因此建议在进行系统发生重建时使用fast marker更经济高效。Jian等(2008)在对虎耳草目的系统发生研究中发现, 使用slow marker构建的系统发生树能很好地解决深层水平(deep level)的系统发生关系, 但浅层水平(shallow level)或近期分化支系之间的关系则未能得到解决; 相反, 使用fast marker构建的系统发生树则能很好地解决相对近期分化支系之间的关系, 但未能解决深层次的关系; 通过将2类分子标记进行联合分析可很好地解决所有水平的系统发生关系。因此, 在实际系统发生构建中, 我们建议根据所研究类群的特点将2类分子标记联合使用, 即快慢结合。

2.2 矩阵组装

2.2.1 分子序列获得

分子序列获得的方式有2种: 自测序和从公共数据库下载。针对自测序数据, 需先根据基于正向和反向引物测序获得的序列图谱, 掐头去尾, 将两端杂乱的区域删除(图2), 再拼接成完整序列。在人工校对序列时需仔细, 首先检查2个或2个以上引物测得序列重叠的区域是否存在兼并位点(图2), 确定兼并位点产生的原因; 然后将拼接好的序列在NCBI数据库(https://www.ncbi.nlm.nih.gov/)中进行BLAST搜索, 确认所获得序列是否为目的物种的目标片段; 最后将获得的新序列与已获得(之前自测序或从公共数据库下载)的序列进行比对(alignment), 逐条检查新获得序列的变异位点, 并结合测序峰图确认变异位点的真实性, 对于蛋白编码序列还需检查是否能正确翻译成氨基酸。上述操作均可在Geneious软件(Kearse et al., 2012)中完成。此外, 对于分布广泛的类群, 实地采样存在一定困难, 因此研究者通常需要从腊叶标本的DNA中获取所需序列, 以保证取样具有代表性。然而, 标本由于保存方式或年代久远等问题, DNA降解严重且含量较低, 极易被其它材料污染。包含污染的序列将会误导系统发生构建, 从而影响后续各种分析, 如分化时间估算和生物地理推断。对此, Wang (2018)针对如何合理正确地利用腊叶标本提出了实验步骤和分析流程, 希望能作为研究人员在植物系统学研究中使用腊叶标本的参考指南, 有助于尽早发现和避免来自其它植物或生物的潜在序列污染。

对于从公共数据库中下载的数据, 使用时同样不可掉以轻心。公共数据库虽然存储了大量的序列, 为系统发生树构建时密集的类群取样提供了可能, 但其

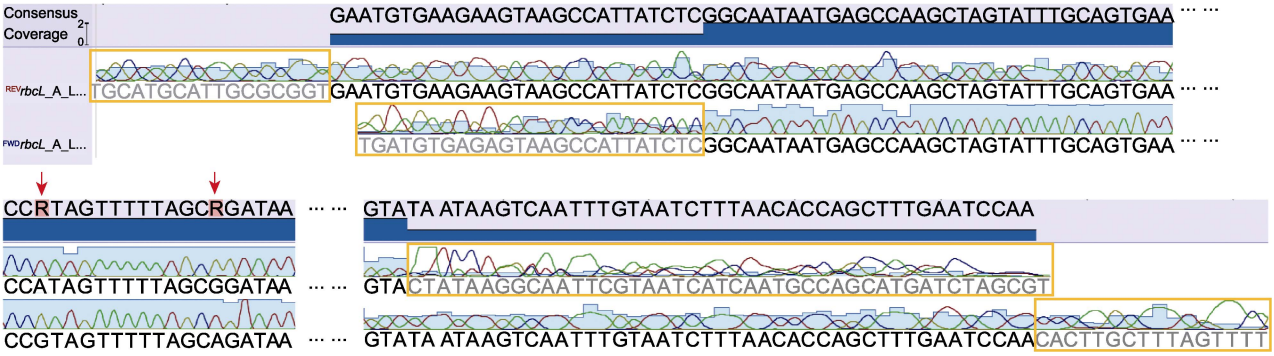

图2 Sanger测序获得的双向引物DNA序列色谱图

橙色方框示序列两端不明确碱基，红色箭头示2个色谱图中相互冲突的碱基。

Figure 2 DNA sequence chromatograms obtained through Sanger sequencing using both forward and reverse primers

Orange boxes show the ambiguous bases in both ends of the sequences, and red arrows show the conflicting base calls between two chromatograms.

包含许多污染或鉴定错误的序列(Wang et al., 2014b; 向小果和王伟, 2015)。因此，研究者在使用时需对下载的序列进行初步分析，删除错误序列。例如, Wang等(2014b)在对毛茛族进行系统发生分析时发现，*Kumlienia hystricula*的*psbJ-petA*序列(GenBank accession No.GU258008)实为毛茛属(*Ranunculus*)的序列，*Ranunculus orthorhynchus*的*psbJ-petA*序列(GenBank accession No.HQ338267)实为极地毛茛属(*Arcteranthis*)的序列。除明显的错误序列外，还可能存在隐藏的问题序列。一般默认序列的种内差异小于种间差异，因此在构建系统发生树时，针对同一物种，可使用来自不同样品的序列联合建树。但由于物种鉴定或实验污染等原因，来源于不同样品名义上属于同一物种的序列可能掺杂着其近缘种的序列，这是通过BLAST搜索难以避免的问题。下文将详述对这种序列的处理方法。

2.2.2 分子序列比对

序列比对常用的软件有Clustal X (Thompson et al., 1997)、MAFFT (Katoh and Standley, 2013)和Muscle (Edgar, 2004)等。同时, BioEdit (Hall, 1999)、MEGA (Tamura et al., 2021)和Geneious等软件中不仅内嵌了上述序列比对程序，还可在自动比对后对矩阵进行手动调整。对于蛋白质编码基因，在调整矩阵时必须以三联体密码为1个单位考虑序列的插入(insertion)或缺失(deletion) (图3A)。研究者可先找到起始密码子，根据三联体密码利用BioEdit等软件进行手动校正，也可以先将核苷酸序列翻译成氨基酸序列后进行比对和手动校正，再依此生成相应的核苷酸序列。对于非编码序列，凭经验依据同源性最高准则调整自动比对好的矩阵，在系统发生分析前删除难排的区域(difficult-to-align region)或多聚碱基(polybase)区域(图3B, C)，尤其是稍微改变位置就能对系统树的拓扑结构产生很大影响的区域。删除人为判定的难排区域可在PAUP软件(Swofford, 2002)中根据如下命令完成。

```
begin assumption;
charset ambiguous=1–24 105–126; #定义需删除区域
end;
begin paup;
include all;
exclude ambiguous; #删除指定区域
export file=filename.nex format=nexus interleaved=no; #生成删除指定区域的nexus格式文件
end;
```

在数据量庞大的情况下，可借助Gblocks (Castresana, 2000)或trimAl (Capella-Gutiérrez et al., 2009)等软件自动删除难排区域，提取保守序列建树。本文简要介绍Gblocks软件的基本操作方法(软件操作均以Windows系统为例)。

(1) 双击Gblocks.exe，打开软件界面，进入主菜单;

(2) 键入o，输入要打开的文件路径，如E:\Gblocks\test\test.fasta (软件支持NBRF/PIR和FASTA格式文

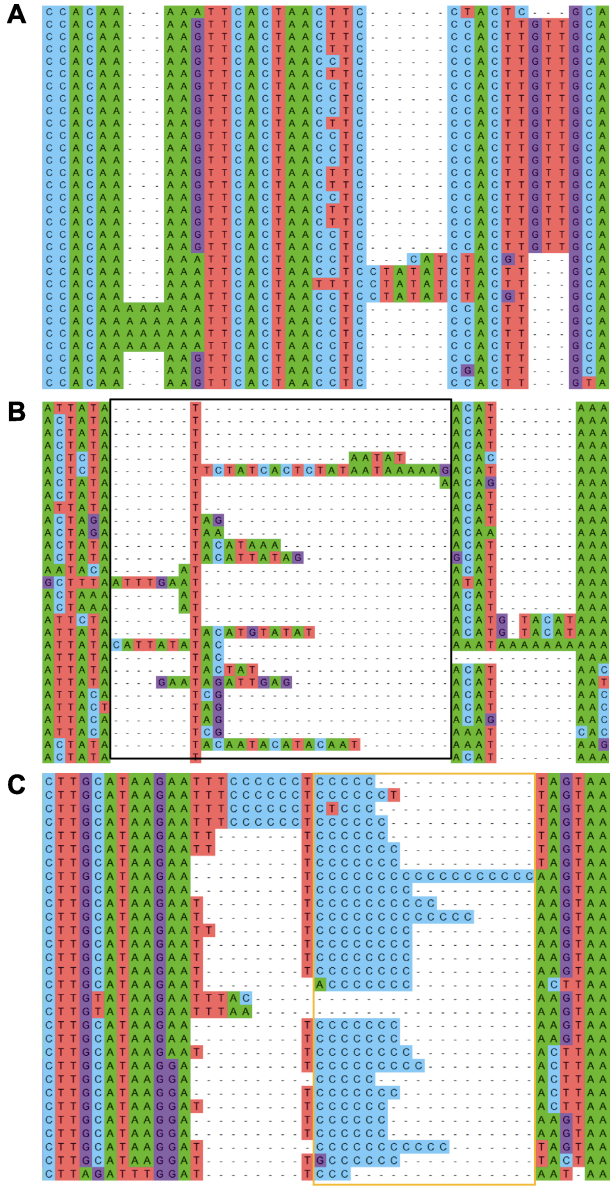

图3 序列比对示例

(A) 编码蛋白的序列比对; **(B)** 含有难排区域(黑色方框)的非编码序列比对; **(C)** 含有多聚碱基区域(橙色方框)的非编码序列比对

Figure 3 Examples for sequence alignment

(A) Alignment for protein-coding sequences; **(B)** Alignment for non-coding sequences with a difficult-to-align region (indicated by black box); **(C)** Alignment for non-coding sequences with a poly-base region (indicated by orange box)

件);

(3) 键入t, 选择序列类型: Protein, DNA和Codons;

(4) 键入b, 进入参数设置菜单, 前4个参数一般使用默认数值, 调整第5个参数可选择存在缺失的位点的保留比例: None、With Half和All表示保留的比例分别为0%、小于50%和100%;

(5) 键入g, 获得提取的保守序列文件。

详见网址: http://molevol.cmima.csic.es/castresana/Gblocks.html

2.3 系统发生树构建

2.3.1 分区策略及模型选择

如前所述, DNA序列不同区域的进化速率可能存在很大差异, 研究者可通过建立序列进化模型来估测核苷酸替换速率, 提高系统发生推断的可靠性。在利用基于模型的统计学方法构建系统发生树时, 首先要对比对好的DNA序列进行分区, 再赋予每个分区(partition)各自的序列进化模型。如果所用DNA片段较少, 分区通常人为性较大, 可根据来源(核DNA vs质体DNA)、区域(基因vs基因间隔区)、功能(蛋白编码基因vs核糖体基因)、是否表达(内含子vs外显子)和密码子位置(第1密码子vs第2密码子vs第3密码子)等原则进行划分, 也可根据软件确定最佳分区策略。如果所用基因较多, 进行系统发育基因组学分析则必须借助软件。本文简要介绍PartitionFinder软件(Lanfear et al., 2017)的操作方法。该软件可以确定核苷酸、氨基酸及形态数据集的最佳分区策略, 比较预先设定分区的优劣, 并确定每个分区的最优进化模型。

(1) 运行环境设置。该软件在Python 2.7.10及以上、3.0以下版本环境中运行, 需安装numpy等多个依赖包。为简便起见, 笔者建议先安装Anaconda (https://www.anaconda.com), 在其中的CMD.exe Prompt中安装Python 2.7并运行PartitionFinder软件。Python 2.7安装及环境激活命令如下。

```
conda create -n python27 python=2.7 anaconda
activate python27
```

(2) 文件准备。需准备联合矩阵文件test.phy和配置文件partition_finder.cfg (软件包中自带示例文件)。

partition_finder.cfg文件中需更改的参数如下。

- alignment=test.phy; #联合矩阵文件名称
- models=GTR, GTR+G, GTR+I+G; #备选进化模型, 可列举, 可根据将使用的建树软件确定备选模型(如models=mrbayes), 或检验所有模型(models=all)等

- model_selection=aicc; #最优模型选择的依据准则，有AIC、AICc和BIC三种可选项
- [data_blocks] #根据序列特征(如基因、密码子和间隔区)预先设定的子分区，如

codon1=1-1000\3; #某基因第1密码子
codon2=2-1000\3; #某基因第2密码子
codon3=3-1000\3; #某基因第3密码子
intron=1001–2000; #某基因内含子

- search=greedy; #软件搜寻最佳策略的方式，有all、greedy和rcluster等多种选项。all通常用于小数据集，greedy用于~10 loci的数据集，rcluster用于~100 loci的数据集

(3) 软件运行。在CMD.exe Prompt中进入工作目录，运行如下命令。

python <PartitionFinder.py> <partition_finder.cfg>

<PartitionFinder.py>: PartitionFinder.py的完整存在路径;

<partition_finder.cfg>: partition_finder.cfg的完整存在路径。

(4) 结果查看。运行结束后，打开analysis\best_scheme.txt文件，可查看最佳分区策略及最优模型。同时，该文件提供常用建树软件中进行分区和模型设置的代码。

详见网址: http://www.robertlanfear.com/partitionfinder/tutorial/

2.3.2 系统发生树构建常用方法

根据所使用的数据类型，系统发生树构建方法可分为基于距离和基于性状两大类。

基于距离的方法是先将序列矩阵转化为遗传距离的数据矩阵，然后再进行系统树构建。常用方法包括最小进化法(minimum evolution)、最小二乘法(least squares)、邻接法(neighbor-joining, NJ)和聚类法(UPGMA)等。当前，生物类群系统发生研究中较少使用基于距离的建树方法。但邻接法由于运算速度快，能分析超大数据集，因而可用于序列筛选时构建临时树、分析蛋白序列并判断序列功能以及土壤微生物的门类鉴定等。

基于性状的方法可直接使用比对好的序列矩阵，例如，用DNA序列或氨基酸序列进行系统树的推断。常用方法包括最大简约法(maximum parsimony, MP)、最大似然法(maximum likelihood, ML)和贝叶斯法(Bayesian inference, BI)等。MP依据简约性原则建树，即最接近真实系统关系的树是所要求的性状状态改变数最少的树，利用该方法不仅能分析核苷酸、氨基酸序列，也可分析形态矩阵、对分子矩阵中插入和缺失编码的性状矩阵以及联合的形态和分子数据。MP受序列比对影响较大，且对进化速率异质性较敏感，易产生长枝吸引。ML和BI均基于统计学方法，引入序列进化模型，用于以高统计置信度估计系统发生关系。它们可在一定程度上消除长枝吸引对系统树构建的影响，但对模型依赖性较大，对计算能力要求也较高(Yang and Rannala, 2012)。这3种方法是当前构建系统发生树最常用的方法。由于不同分析方法对数据集的敏感性不同(Smith, 2013; Lu et al., 2018)，我们建议在实际工作中至少选用MP、ML和BI中的2种方法进行系统发生树构建。

2.3.3 系统发生树构建常用软件操作流程

根据在系统发生树构建中应用的广泛程度，本文简要介绍系统发生树构建常用软件的操作流程。

(1) PAUP

PAUP (Swofford, 2002)是利用最大简约法建树的常用软件。使用该软件建树需先将比对好的序列转换成NEXUS文件，然后将执行代码粘贴在序列之后并保存，最后打开PAUP，选择File–Open (选中NEXUS文件)–Execute即可运行软件。基本执行代码如下(详见网址: http://paup.phylosolutions.com/)。

```
begin paup;
log file=MPhs.log;
set autoclose=yes warntree=no warnreset=no increase=auto;
outgroup Ginkgo_biloba; #外类群名称必须与序列矩阵中的名称保持一致；可以有多个，中间用空格隔开。
hsearch addseq=random nreps=1000;
savetrees file=MPhs.tre format=altnexus brlens=yes root=yes;
showtrees;
pscores /ci=yes ri=yes rc=yes hi=yes scorefile=pscore.txt;
contree /majrule=yes showtree=yes treefile=MPhscon.tre;
bootstrap nreps=1000 keepall=yes treefile=
```

MPhsbt.tre brlens=yes/addseq=random nreps=10;

gettrees file=MPhsbt.tre storebrlens=yes storetreewts=yes warntree=no;

contree /majrule=yes usetreewts=yes showtree=yes treefile=MPhsbtcon.tre;

log stop;

end;

计算过程保存在MPhs.log文件中; 最大简约树保存在MPhs.tre文件中; 简约性指数保存在pscore.txt文件中; 同等简约树的严格一致树和50%多数一致树保存在MPhscon.tre文件中; 1 000次自举法检验的简约树保存在MPhsbt.tre文件中, 其50%多数一致树保存在MPhsbtcon.tre文件中。

此外, 通过最大简约法建树还可使用TNT软件(http://www.lillo.org.ar/phylogeny/tnt/)。

(2) RAxML

RAxML (Stamatakis, 2014)是利用最大似然法建树的常用软件。该软件可对序列进行分区, 但所有分区只能指定同一个序列进化模型, 一般选择最复杂的GTR+Γ+I模型。基本操作方法如下(详见网址: https://cme.h-its.org/exelixis//web/software/raxml/)。

解压软件包后, 将可执行文件(.exe)、分区指定文件(.txt)和序列文件(PHYLIP格式)置于同一工作目录下, 通过Windows控制台命令窗口进入该工作目录, 执行下列命令。

raxml -f a -s test.phy -m GTRGAMMAI -x 12345 -# 1000 -n RM1.trees -o Ginkgo_biloba -q partition.txt

raxml是可执行文件名称, -s指定序列文件, -m指定序列进化模型, -n指定输出文件后缀, -o指定用于置根的序列, -q指定分区文件。

分区文件格式如下:

DNA, Subset1=1–1524

DNA, Subset2=1525–2080, 3301–3778

DNA, Subset3=2081–3300, 3779–4182

用FigTree软件(http://tree.bio.ed.ac.uk/software/figtree/)打开生成的RAxML_bipartitions.RM1.trees文件, 可查看系统发生树及每个节点的自展值(bootstrap, BS)。

在利用最大似然法建树时, 若数据量较大(如基因组数据), 还可使用IQ-Tree (http://www.iqtree.org/)或RAxML-ng (https://github.com/amkozlov/raxml-ng)软件, 以提高运算速度。

(3) MrBayes

MrBayes (Ronquist et al., 2012)是利用贝叶斯法建树的常用软件。该软件不仅能对序列进行分区, 还可单独为每个分区指定不同序列进化模型。与PAUP类似, MrBayes要求用NEXUS序列文件, 并将执行代码粘贴在序列最后, 双击mrbayes_x64.exe打开软件, 输入exe <test.nex>运行软件(<test.nex>为test.nex的存在路径), 基本执行代码如下(详见网址: https://nbisweden.github.io/MrBayes/index.html)。

begin mrbayes;

log start filename=test.log;

outgroup Ginkgo_biloba; #指定用于置根的类群, 只能有1个

set autoclose=yes nowarn=yes;

charset Subset1=1–1524; #指定分区

charset Subset2=1525–2080 3301–3778;

charset Subset3=2081–3300 3779–4182;

partition region=3: Subset1, Subset2, Subset3;

set partition=region;

lset applyto=(1) nst=6 rates=invgamma; #指定Subset1的序列模型(可根据jModelTest或PartitionFinder输出文件进行设置)

lset applyto=(2) nst=6 rates=invgamma; #指定Subset2的序列模型

lset applyto=(3) nst=6 rates=gamma; #指定Subset3的序列模型

prset applyto=(all) ratepr=variable;

unlink statefreq=(all) revmat=(all) shape=(all) pinvar=(all) tratio=(all);

mcmc ngen=1000000 samplefreq=1000 printfreq=1000 nchains=4 savebrlens=yes; #设置运行代数和取样频率

sump filename=test.nex burnin=250; #忽略前250棵(25%)抽样树并总结替代模型参数

sumt filename=test.nex burnin=250 conformat=simple; #忽略前250棵(25%)抽样树并总结后验概率树

end;

利用贝叶斯法构建系统发生树时, 需用达到稳态之后计算得到的抽样树来总结后验概率树。判断分析是否达到稳态的方法是在Tracer软件(Rambaut et al., 2018)下打开log文件, 检验各项参数的有效样本

大小(effective sample size, ESS)是否大于200。分析完成后，可用FigTree软件打开生成的test.nex.con.tre文件，查看系统发生树及每个节点的后验概率(posterior probability, PP)。

利用贝叶斯法建树还可使用PhyloBayes (http://www.atgc-montpellier.fr/phylobayes/)和RevBayes (https://revbayes.github.io/)软件。

(4) MEGA

与上述命令行界面的软件不同，MEGA采用图形用户界面，对用户更友好。该软件可以利用ML、MP、NJ和UPGMA等多种方法构建系统发生树(详见网址: https://megasoftware.net/)。软件安装完成后，点击菜单栏中的Phylogeny按钮，在下拉选项中选择建树方法，在弹出的窗口中选择比对好的序列文件(FASTA格式)，再根据提示设置运行参数，点击Compute即可运行。目前，MAGA较少用于生物类群的系统发生树构建。

2.3.4 序列联合分析

测序技术的快速发展和分子数据井喷式增长为系统发生分析带来了极大便利，分子系统学也从最初的利用同工酶标记(allozymes)、随机扩增多态性DNA标记(RAPD)、选择性扩增限制性片段长度多态性(AFLP)和简单重复序列标记(SSR)等，发展到当前可极为方便地利用来自细胞核、线粒体和叶绿体3套基因组的DNA序列进行系统发生分析。但如何准确分析多个分子标记的数据仍存在一些误区。众所周知，植物的3套基因组拥有不同的进化历史，其中核基因组可通过双亲遗传，而线粒体和叶绿体基因组则是单亲遗传。利用来自不同遗传体系的数据建立的系统发生树可能存在明显冲突。Wang等(2014a)提出处理这一问题的详细指导方案。现以植物中最常用的叶绿体基因组和核基因组为例介绍序列联合分析方法。

2.3.4.1 鉴别显著冲突

目前鉴别显著冲突最常用的2种方法包括不相合长度差异检验(incongruence length difference test, ILD test)和基于树的比较(tree-based comparisons)。

ILD test根据计算得出的*P*值判断2个数据集之间是否存在显著冲突，该分析在PAUP软件中进行，代码如下。

```
begin paup;
log file=test.log;
set autoclose=yes warntree=no warnreset=no increase=auto;
charset cpDNA=1–5891;
charset nrDNA=5892–7010;
charpartition genes=1: cpDNA, 2: nrDNA;
showcharparts;
hompart partition=genes nreps=100/addseq=random nreps=10;
log stop;
end;
```

然而，ILD test仅对2个树的拓扑结构进行整体比较，并不考虑引起冲突的节点支持率的高低，而且有时难以发现仅在局部产生冲突的节点(Wang et al., 2007)。此外，当研究类群较多时，ILD test的运算时间较长。因此，我们建议使用基于树的比较方法，该方法将不同数据集构建的系统树可视化后比较拓扑结构之间是否存在冲突以及冲突是否显著。根据Wang等(2014a)的建议，BS≥70和PP≥0.95的节点存在的冲突为显著冲突。

2.3.4.2 为冲突提供可能的合理解释

引起冲突的可能原因包括两大类: 人为原因和生物学原因。人为原因包括序列错误、长枝吸引和进化饱和(evolutionary saturation); 生物学原因包括使用旁系同源基因(paralogous gene)、不完全谱系分选(incomplete lineage sorting)和杂交(hybridization) (Wang et al., 2014a)。

(1) 序列错误。序列错误可能是由类群鉴定错误或实验过程中出现污染导致，在自测序和从公共数据库下载的序列中均可能出现。由于叶绿体基因组为环形结构，很少发生重组，在利用多个叶绿体DNA片段进行系统学研究时大多不做单基因系统发生分析。但对于本文2.2.1节提到的近缘种之间可能存在的序列污染问题，就需要通过对单基因树进行比较，若采用同一物种某一个体测得的基因片段得出的系统位置与其它个体测得的片段发生严重冲突，则需要仔细排查，删除问题序列。例如，Wang等(2014b)在对毛茛族的研究中，通过系统树比较、BLAST搜索和序列比对检查，发现前人使用的10条叶绿体*psbJ-petA*序列错误是造成毛茛族叶绿体*matK*和*psbJ-petA*系统树

冲突的原因, 这对该族的分化时间和祖先分布区推断造成很大误导。该研究对如何利用公共数据库的数据以及如何避免使用错误序列等进行了详细分析。与Sanger测序相比, 二代和三代测序产生的庞大数据量使污染序列更难以识别。基于基因树间共祖距离(patristic distance)的双峰分布检测, Owen等(2022)开发出识别及排除可能被交叉污染的基因的流程, 为基因组数据污染问题提供了解决方案。

(2) 长枝吸引。在系统树中, 进化速率明显高于其它类群的支系积累了更多的变异位点(自衍征), 导致在构建系统树(尤其是用简约法建树)时可能会产生较大的进化噪音, 得到错误的拓扑结构。根据系统发生树的分枝长度来判断是否有长枝出现(有长枝不代表一定存在长枝吸引)。如果有很长的分枝出现, 先把具长枝的类群去掉, 再重新进行分析, 若其它类群的关系发生变化, 说明存在长枝吸引(Fan and Xiang, 2003)。增加具长枝类群的物种取样和增加性状可在一定程度上克服长枝吸引的影响。如果叶绿体和核基因系统树的冲突是长枝吸引造成, 那么2套基因组数据可以进行联合分析。

(3) 进化饱和。进化饱和通常出现在高度分化(即演化历史比较古老)的类群中, 或者早期分化的支系以及进化速率较快的DNA片段中。它所引起的冲突并非由不同的进化历史造成, 因此这种冲突能直接进行联合分析。

(4) 旁系同源。旁系同源通常出现在多拷贝的核基因位点上。ITS是应用最普遍的核DNA片段。判断ITS是否存在旁系同源的方法是PCR产物电泳是否有多条带, 以及测序产物是否出现套峰等。如果出现这两种情况, 需要进行克隆实验验证。

(5) 杂交和不完全谱系分选。杂交和不完全谱系分选产生的系统树其拓扑结构几乎相同。区分二者的常用方法有3种: ① 通过比较是否存在形态性状的过渡状态以及物种的分布区、物候与生境等是否重叠来判断发生杂交的可能性(Xiang et al., 2005); ② 比较进化事件的最小数目(van der Niet and Linder, 2008); ③ 最小遗传距离法(Joly et al., 2009)。此外, 根据溯祖理论(coalescent theory), 祖先多态性在大约5 *Ne* (有效群体大小)世代内合并, 即多态性消失, 如果假设的*Ne*比自然界实际观察到的大很多, 即可将不完全谱系分选排除(Pelser et al., 2010)。

2.3.4.3 处理冲突并进行联合分析

在人为原因导致的冲突中, 错误序列必须删除, 长枝吸引和进化饱和引起的冲突则可以直接进行联合分析。在生物学原因导致的冲突中, 旁系同源基因和存在不完全谱系分选的基因也必须删除; 对于杂交造成的冲突, 比较2个系统树即可确定杂交支系的2个亲本, 那么杂交支系的进化历史就可得到澄清。对杂交支系可能的处理方式有3种: (1) 常规做法是移除引起冲突的类群再进行联合分析, 但该方式不能在联合树上展示杂交支系; (2) 先将引起冲突的类群删除并联合建树, 再根据2套数据独立进行分析获得系统树, 将杂交支系插入到联合树上(a-posterior insertion) (e.g., van der Niet and Linder, 2008); (3) 将冲突的支系作为2个支系进行计算: 一支仅有叶绿体DNA数据, 其核DNA按照缺失处理; 另一支仅有核DNA数据, 叶绿体DNA按照缺失处理(Pelser et al., 2010)。

2.3.5 拓扑结构检验

拓扑结构检验是针对同一个数据集, 测试2棵或更多棵系统树的拓扑结构是否存在显著差异。在构建系统发生树时, 采用不同的建树方法可能会得出略有不同的拓扑结构, 那么这些备选拓扑结构之间是否存在优劣? 我们可以通过统计学方法进行检验。常用的非参数检验方法有近似无偏(approximately unbiased)检验(Shimodaira, 2002)、KH (Kishino-Hasegawa)检验(Kishino and Hasegawa, 1989)和SH (Shimodaira-Hasegawa)检验(Shimodaira and Hasegawa, 1999)。这些检验可通过Tree-Puzzle (Schmidt et al., 2002)和Consel (Shimodaira and Hasegawa, 2001)软件进行。此外, 还可使用参数自举似然比检验——SOWH (Swofford-Olsen-Waddell-Hillis)检验(Goldman et al., 2000)以及利用贝叶斯因子(Kass and Raftery, 1995)来判断不同拓扑结构的差异显著性。拓扑结构检验也可针对传统分类单元, 是确定其是否为单系群的有效方法。

在获得系统树之后, 还应判断其能否解决所关心的科学问题。若不能, 则应重新调整取样策略, 以优化系统树(图1)。

3 总结与展望

系统发生树是进化生物学研究的基础, 合理地利用海

量数据进行系统发生树构建是推断生物进化历史的关键。本文从类群取样、分子标记选择、序列筛选，以及利用不同来源的数据集进行多DNA片段联合分析和系统树构建常用方法等方面，为从事系统发生研究和以系统发生树为工具进行生态学和比较生物学等方面研究的科研人员提供了指导和建议。目前，在科技飞速发展的大环境下，生命之树构建已取得了许多重要成就，被子植物、裸子植物和蕨类植物等大类群在目和科水平上的系统发生关系得到了较好解决。同时，科学家也一直在建树方法、运算能力和科学传播等方面不断努力(王伟和刘阳, 2020)。首先，基于急剧扩增的数据量，尤其是随着基因组数据的广泛应用，研究者对开发新的建树软件和新的序列进化模型的需求日益迫切。例如，Xi等(2012)针对金虎尾目的研究提出了基于贝叶斯混合模型的后验数据分区方法；Goremykin等(2013)在探讨被子植物最早分化类群的研究中提出了新的模型(CAT+GTR+Γ+covext model)以考虑碱基组成的异质性；基于溯祖理论的方法为解决核基因组系统发生分析中不完全谱系分选的干扰提出了方案(Liu et al., 2015; Mirarab et al., 2021)。其次，海量数据为构建超矩阵(super matrix)和超大系统发生树(super tree)提供了机会。例如，Folk等(2019)通过构建包括虎耳草目72%物种的全球尺度的超大系统发生树，揭示了生态机会对促进该类群快速多样化的重要作用；Sun等(2020)通过组装近2万种蔷薇类物种水平的超矩阵进行系统发生分析，揭示了中新世中期以来全球的持续降温导致热带和非热带地区极不平衡的物种多样化动态。但利用超矩阵建树对计算机的运算能力要求极高，如何兼顾运算速度和建树的准确性是当前亟待解决的问题。最后，科学研究的主要目的之一是科学传播，即通过科普的方式向社会大众传播科学知识，促进公众对科学的理解、支持和参与，在获得系统树之后，如何向公众展示也是系统发生学未来的努力方向。

参考文献

范凯，叶方婷，毛志君，潘鑫峰，李兆伟，林文雄 (2021). 被子植物小热激蛋白家族的比较基因组学分析. 植物学报 **56**, 245–261.

葛颂 (2022). 中国植物系统和进化生物学研究进展. 生物多样性 **30**, 22385.

康凯程，牛西强，黄先忠，胡能兵，隋益虎，张开京，艾昊 (2021). 辣椒R2R3-MYB转录因子家族的全基因组鉴定与比较进化分析. 植物学报 **56**, 315–329.

王伟，刘阳 (2020). 植物生命之树重建的现状、问题和对策建议. 生物多样性 **28**, 176–188.

向小果，王伟 (2015). 植物DNA条形码在系统发育研究中的应用. 生物多样性 **23**, 281–282.

Benton MJ, Ayala FJ (2003). Dating the tree of life. *Science* **300**, 1698–1700.

Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W (2003). Noncoding plastid *trnT-trnF* sequences reveal a well resolved phylogeny of basal angiosperms. *J Evol Biol* **16**, 558–576.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.

Castresana J (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552.

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 Jr GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Ann Missouri Bot Gard* **80**, 528–580.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.

Fan CZ, Xiang QY (2003). Phylogenetic analyses of Cornales based on 26S rRNA and combined 26S rDNA-*matK*-*rbcL* sequence data. *Am J Bot* **90**, 1357–1372.

Folk RA, Stubbs RL, Mort ME, Cellinese N, Allen JM, Soltis PS, Soltis DE, Guralnick RP (2019). Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. *Proc Natl Acad Sci USA* **116**, 10874–10882.

Goldman N, Anderson JP, Rodrigo AG (2000). Likelihood-based tests of topologies in phylogenetics. *Syst Biol* **49**, 652–670.

Goremykin VV, Nikiforova SV, Biggs PJ, Zhong BJ, De-

lange P, Martin W, Woetzel S, Atherton RA, Mclenachan PA, Lockhart PJ (2013). The evolutionary root of flowering plants. *Syst Biol* **62**, 50–61.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**, 95–98.

Jian SG, Soltis PS, Gitzendanner MA, Moore MJ, Li RQ, Hendry TA, Qiu YL, Dhingra A, Bell CD, Soltis DE (2008). Resolving an ancient, rapid radiation in Saxifragales. *Syst Biol* **57**, 38–57.

Joly S, McLenachan PA, Lockhart PJ (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am Nat* **174**, E54–E70.

Kass RE, Raftery AE (1995). Bayes factors. *J Am Stat Ass* **90**, 773–795.

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649.

Kishino H, Hasegawa M (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29**, 170–179.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B (2017). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol* **34**, 772–773.

Lian L, Ortiz RDC, Jabbour F, Chen ZD, Wang W (2019). Re-delimitation of *Tinospora* (Menispermaceae): implications for character evolution and historical biogeography. *Taxon* **68**, 905–917.

Liu GQ, Lian L, Wang W (2022). The molecular phylogeny of land plants: progress and future prospects. *Diversity* **14**, 782.

Liu L, Wu SY, Yu LL (2015). Coalescent methods for estimating species trees from phylogenomic data. *J Syst Evol* **53**, 380–390.

Lozano-Fernandez J (2022). A practical guide to design and assess a phylogenomic study. *Genome Biol Evol* **14**, evac129.

Lu LM, Cox JC, Mathews S, Wang W, Wen J, Chen ZD (2018). Optimal data partitioning, multispecies coalescent and Bayesian concordance analyses resolve early divergences of the grape family (Vitaceae). *Cladistics* **34**, 57–77.

Mirarab S, Nakhleh L, Warnow T (2021). Multispecies coalescent: theory and applications in phylogenetics. *Annu Rev Ecol Evol Syst* **52**, 247–268.

Nandi OI, Chase MW, Endress PK (1998). A combined cladistic analysis of angiosperms using *rbcL* and non-molecular data sets. *Ann Missouri Bot Gard* **85**, 137–214.

Nei M (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet* **30**, 371–403.

Owen CL, Marshall DC, Wade EJ, Meister R, Goemans G, Kunte K, Moulds M, Hill K, Villet M, Pham TH, Kortyna M, Lemmon EM, Lemmon AR, Simon C (2022). Detecting and removing sample contamination in phylogenomic data: an example and its implications for Cicadidae phylogeny (Insecta: Hemiptera). *Syst Biol* **71**, 1504–1523.

Pelser PB, Kennedy AH, Tepe EJ, Shidler JB, Nordenstam B, Kadereit JW, Watson LE (2010). Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. *Am J Bot* **97**, 856–873.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol* **67**, 901–904.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**, 539–542.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.

Shimodaira H (2002). An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492–508.

Shimodaira H, Hasegawa M (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**, 1114.

Shimodaira H, Hasegawa M (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247.

Smith MR (2013). Likelihood and parsimony diverge at high taxonomic levels. *Cladistics* **29**, 463.

Soltis DE, Moore MJ, Burleigh G, Soltis PS (2009). Molecular markers and concepts of plant evolutionary relationships: progress, promise, and future prospects. *Crit*

Rev Plant Sci **28**, 1–15.

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

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Sun M, Folk RA, Gitzendanner MA, Soltis PS, Chen ZD, Soltis DE, Guralnick RP (2020). Recent accelerated diversification in rosids occurred outside the tropics. *Nat Commun* **11**, 3333.

Swofford DL (2002). PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.

Tamura K, Stecher G, Kumar S (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol* **38**, 3022–3027.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.

van der Niet T, Linder HP (2008). Dealing with incongruence in the quest for the species tree: a case study from the orchid genus *Satyrium*. *Mol Phylogenet Evol* **47**, 154–174.

Wang W (2018). A primer to the use of herbarium specimens in plant phylogenetics. *Bot Lett* **165**, 404–408.

Wang W, Del Ortiz RC, Jacques FMB, Chung SW, Liu Y, Xiang XG, Chen ZD (2017). New insights into the phylogeny of Burasaieae (Menispermaceae) with the recognition of a new genus and emphasis on the southern Taiwanese and mainland Chinese disjunction. *Mol Phylogenet Evol* **109**, 11–20.

Wang W, Del Ortiz RC, Jacques FMB, Xiang XG, Li HL, Lin L, Li RQ, Liu Y, Soltis PS, Soltis DE, Chen ZD (2012). Menispermaceae and the diversification of tropical rainforests near the Cretaceous-Paleogene boundary. *New Phytol* **195**, 470–478.

Wang W, Li HL, Chen ZD (2014a). Analysis of plastid and nuclear DNA data in plant phylogenetics—evaluation and improvement. *Sci China Life Sci* **57**, 280–286.

Wang W, Li HL, Xiang XG, Chen ZD (2014b). Revisiting the phylogeny of Ranunculeae: implications for divergence time estimation and historical biogeography. *J Syst Evol* **52**, 551–565.

Wang W, Wang HC, Chen ZD (2007). Phylogeny and morphological evolution of tribe Menispermeae (Menispermaceae) inferred from chloroplast and nuclear sequences. *Perspect Plant Ecol Evol Syst* **8**, 141–154.

Whelan S, Liò P, Goldman N (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* **17**, 262–272.

Xi ZX, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, Davis CC (2012). Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci USA* **109**, 17519–17524.

Xiang QY, Manchester SR, Thomas DT, Zhang WH, Fan CZ (2005). Phylogeny, biogeography, and molecular dating of cornelian cherries (*Cornus*, Cornaceae): tracking Tertiary plant migration. *Evolution* **59**, 1685–1700.

Yang ZH, Rannala B (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303–314.

Phylogenetic Tree Reconstruction Based on Molecular Data

Huanwen Peng[1, 2, 3], Wei Wang[1, 2, 3*]

[1]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China;* [2]*China National Botanical Garden, Beijing 100093, China*
[3]*University of Chinese Academy of Sciences, Beijing 100049, China*

Abstract Phylogenetics is a discipline reconstructing evolutionary relationships of organisms. With improvements in sequencing technique, analytic methods, and computation power, the molecular data have been used widely and have promoted greatly the rapid development of molecular phylogenetics. The phylogenetic tree has become a powerful tool in many areas of biology, such as ecology and comparative biology. Currently, phylogenetic studies mainly focus on phylogenetic tree reconstructions by using various software, however, some fundamental principles or matters that should be paid attention when performing phylogenetic analyses are sometimes weakened or even ignored. Here, we present the workflow and methods in details for phylogenetic tree reconstruction based on molecular data, including taxon sampling, molecular marker selection, sequence alignment, partitioning and model selection, combined analysis of multiple markers, and topological test. Currently, the widely used methods of phylogenetic reconstructions are maximum parsimony, maximum likelihood, and Bayesian inference. We thereby provide the detailed operating flows and corresponding commands for these three methods, respectively. We expect this paper will provide a reference for relevant researches.

Key words phylogenetic tree reconstruction, molecular systematics, combined analysis, nucleotide substitution model

Peng HW, Wang W (2023). Phylogenetic tree reconstruction based on molecular data. *Chin Bull Bot* **58**, 261–273.

---

* Author for correspondence. E-mail: wangwei1127@ibcas.ac.cn

(责任编辑: 白羽红)