

• 药材与资源 •

基于 ITS 条形码及机器学习的黄檀属物种分子鉴别研究

邱家荣¹, 刘巧珍¹, 代江鹏¹, 谭智杰¹, 林月霞¹, 高晓霞^{2*}, 朱 爽^{1*}

1. 广东药科大学生命科学与生物制药学院, 广东 广州 510006

2. 广东药科大学药学院, 广东 广州 510006

摘要: 目的 提高黄檀属的物种鉴别成功率, 并将机器学习方法与传统的基于距离/系统发育树的方法进行比较, 筛选最优的 ITS 条形码分析方法。方法 所使用的黄檀属物种 ITS 序列来自实验获得的 3 条以及从 NCBI 下载的 399 条共 96 个物种。以条形码 ITS 作为分子标记, 对比距离法、系统发育树法及机器学习方法在黄檀属物种的鉴别成功率。结果 在基于机器学习方法的分析中, 黄檀属物种的平均鉴别成功率为 39.59%, 其中 BLOG 能识别出 42 个黄檀属物种, 其正确序列分类占比为 95.75%。另外, SMO、Naïve Bayes、JRip、J48 能够识别出 34 个物种, 分别获得了 79.10%、58.71%、72.64%、76.37% 的正确序列分类占比。基于系统发育树法与距离法的分析分别获得 28.13% 和 36.46% 的鉴别成功率。结论 基于机器学习的黄檀属 ITS 条形码基原识别比距离法/系统发育树法拥有更高的鉴别成功率和社会经济效率。建议优先利用基于 ITS 条形码的机器学习方法对黄檀属物种进行基原识别。

关键词: ITS; 黄檀属; 机器学习; DNA 条形码; 基原识别

中图分类号: R286.12 文献标志码: A 文章编号: 0253-2670(2024)11-3825-10

DOI: 10.7501/j.issn.0253-2670.2024.11.024

Molecular identification of *Dalbergia* species based on ITS barcode with machine learning approachesKUANG Jiarong¹, LIU Qiaozhen¹, DAI Jiangpeng¹, TAN Zhijie¹, LIN Yuexia¹, GAO Xiaoxia², ZHU Shuang¹

1. School of Life Sciences and Biopharmaceutics, Guangdong Pharmaceutical University, Guangzhou 510006, China

2. School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou 510006, China

Abstract: Objective To improve the identification success rate of *Dalbergia* and screen out the best ITS analysis methods, compare the machine learning methods with the traditional distance-based and phylogenetic tree-based methods to screen the optimal ITS barcode analysis method. **Methods** A total of 402 ITS sequences of *Dalbergia* species used in this study were collected by experiments (three ITS sequences) and downloaded from NCBI (399 ITS sequences) for a total of 96 species. The barcode ITS was used as a molecular marker to compare the success rate of distance method, phylogenetic tree method and machine learning method in the identification of *Dalbergia* species. **Results** In the analysis based on machine learning methods, the average identification success rate of *Dalbergia* species was 39.59%, of which 42 *Dalbergia* species could be recognized by BLOG, and the percentage of their correct sequence classification was 95.75%. In addition, SMO, Naïve Bayes, JRip and J48 can identify 34 species with the correct sequence distribution rate of 79.10%, 58.71%, 72.64% and 76.37%, respectively. The distance-based and phylogenetic tree-based methods obtained the species identification success rate of 36.46% and 28.13%, respectively. **Conclusion** ITS barcoding identification of *Dalbergia* based on machine learning approaches has higher identification success rate and socio-economic than traditional methods. It is recommended to prioritize the use of machine learning approaches to identify *Dalbergia* based on ITS barcode.

Key words: ITS; *Dalbergia*; machine learning; DNA barcoding; species identification

收稿日期: 2023-12-06

基金项目: 广东省基础与应用基础研究基金自然科学基金面上项目 (2022A1515011268)

作者简介: 邱家荣, 男, 研究方向为中药分析与鉴定。Tel: 18575771352 E-mail: 2830783107@qq.com

*通信作者: 朱 爽, 硕士生导师, 副教授, 研究方向为中药分析与鉴定。Tel: (020)39352021 E-mail: 15683727@qq.com

高晓霞, 硕士生导师, 教授, 研究方向为药品检验与分析。Tel: (020)39352136 E-mail: gaouxia91@163.com

黄檀属 *Dalbergia* L. f. 植物属于豆科 (Fabaceae) 的蝶形花亚科 (Papilionoideae), 为攀援状灌木、乔木植物或木质植物, 在南美洲、非洲和亚洲的热带、亚热带地区广泛分布。黄檀属具有许多较高价值的物种, 可以作为植物中药材和优质建材^[1-2]。许多黄檀属物种的提取物被证实具有镇痛、抗炎抗菌、抗麻醉、抗毒性、抗癌、抗骨质疏松、抗白血病和解热活性等功效, 被广泛用于治疗咳嗽、腹泻、癫痫、溃疡等症状^[1,3-8]。

由于以前人们主要以形态学来识别区分黄檀属, 使得许多与其相似的属外物种被归类于黄檀属, 导致属内分类十分混乱。根据《濒危野生动植物种国际贸易公约》(CITES) 的国际贸易法规, 数种黄檀属物种被列入保护名单^[9-10]。由于黄檀树木的生长期较长, 而市场对其的需求量极大, 资源严重短缺, 且消费者普遍不具备专业的鉴定知识, 导致不法分子利用形态相似的、价格低廉的其他黄檀木进行替代, 损害了生产者和消费者的权益, 严重影响了市场的健康发展。因此需要不断完善黄檀属物种的快速、准确的物种鉴定方法。

DNA 条形码技术指的是利用一段或几段具有一定变异的、长度较短、两端高度保守的 DNA 片段进行物种基原识别的一项分子鉴定技术^[11]。该技术利用了 DNA 条形码片段的变异稳定性 (种内变异小, 种间变异大) 来实现对物种的基原识别。ITS 作为 DNA 条形码技术中植物基原识别的主要条形码之一, 自 2010 年以来一直蓬勃发展, 已被证实具有较好的物种基原识别能力^[12-16]。目前, ITS 已被用于区分 722 种以上的植物, 并被推荐用于验证中草药的真实性^[17]。如芸香、黄芪、冬虫夏草、沉香以及黄檀一些中药材正品以及各自的混伪品能够利用 ITS 进行快速有效鉴别^[1]。

随着计算机科学的迅速发展, 人工智能与 DNA 条形码技术实现了同步接轨, 比如在 DNA 条形码的数据集分析中引入机器学习方法 (machine learning algorithm, MLA)^[18]。机器学习主要是设计和分析一些让计算机可以自动“学习”的算法, 这些算法是一类从数据中获得规律, 并利用这些规律对未知数据进行预测。BLOG (barcoding with logic) 和 WEKA (Naïve Bayes、SMO、JRip 和 J48) 是应用于 DNA 条形码领域的机器学习方法^[19]。已有研究者将机器学习方法引入 8 种黄檀属的 DNA 条形码分析中, 结果显示, SMO 识别成功率最高, 优于

距离法和邻接法 (neighbor-joining method, NJ) 树; BLOG、J48 和 JRip 等分类器也获得了较高的识别成功率, 极大地提高了 DNA 条形码的物种识别效率^[20-21]。目前基于 DNA 条形码的黄檀属物种的研究所覆盖的物种数量较少, 只涉及有限的热门物种; 或所覆盖的物种数量较高但并无提供明确的鉴别成功率^[1]。此外, DNA 条形码数据集分析方法是 DNA 条形码技术最重要的环节之一, 对于相同的数据集, 不同的 DNA 条形码分析方法会产生不同的物种鉴别效率。

本研究以 ITS 条形码作为分子标记, 分别用距离法、系统发育树和机器学习法 (BLOG、Naïve Bayes、SMO、JRip 和 J48)^[22]对黄檀属物种的 402 条 ITS 序列进行分析和比较。本研究通过覆盖大范围黄檀属物种数目 (大数据集) 的前提下, 比较分析基于机器学习、距离/系统发育树法在 ITS 条形码的黄檀属基原识别成功率及优劣, 筛选出最优的 ITS 条形码分析方法, 为保护黄檀属濒危物种, 维护市场秩序, 保障临床用药安全等提供科学依据和技术支撑。

1 材料和筛选

本研究从 NCBI 收集的黄檀属物种 ITS 条形码序列, 通过其已有信息反复检查并进行手动初步筛选整理, 其中初步筛选条件包括: (1) 序列长度区间保持为 500~800 bp, 并用 MEGA v 7.0^[23]进行对比剪切, 因为不同研究者上传的序列所采用的引物不同可能导致起点和终点有所差异; (2) 查找序列对应的文献并检查该序列是否有效可用; (3) 排除未进行黄檀属下级类别确认的序列; (4) 收集部分仅有极少数序列的物种。截至 2023 年 1 月 20 日, 结合本课题组实验得到的 3 条降香黄檀的 ITS 序列 (登录号为 KY489987~KY489989), 本研究共获得 402 条序列涵盖 96 个黄檀属物种, 序列登录号列于表 1。

2 方法

2.1 基于距离法的黄檀属物种 ITS 序列分析

采用 MEGA v 7.0 软件对黄檀属 ITS 序列进行 Clustal 比对和分析, 计算保守位点 (Conserved sites)、变异位点 (variable sites) 和简约信息位点 (parsimony informative sites); 采用 TaxonDNA^[24]软件基于 Kimura-2-Parameter (K2P) 双参数算法模型计算 ITS 序列以得到黄檀属的物种数、种内距离 (intraspecific distances) 和种间距离 (interspecific distances), 并将结果绘制成散点图进行“Barcoding gap”分析; 采用最佳匹配 (best match, BM) 和最

表 1 从 NCBI 获取的黄檀属植物序列
Table 1 Sequences of *Dalbergia* obtained from NCBI

物种拉丁名	登录号
<i>D. miscolobium</i>	EF451059~EF451070、AB828666、MH465109、DQ787405
大叶黄檀 <i>D. foliolosa</i>	EF451067、AF189022
<i>D. miscolobium</i> haplotype	JQ582850~JQ582876
<i>D. congestiflora</i>	AF068140
印度黄檀 <i>D. sissoo</i>	AF189023、EF451079、KM276184~KM276204、KX057860、KM521407~KM521409、AB828659、MH465104~MH465105
阔叶黄檀 <i>D. latifolia</i>	MG920556~MG920558、JX856441~JX856443、KM276112~KM276126、MH465107、KM521381、AB828646、MN099442
披针叶黄檀 <i>D. lanceolaria</i>	KM276105~KM276111、KM676496、KM521379、AB828645、JX856439、JX506484
披针叶黄檀亚种 <i>D. lanceolaria</i>	KM276151~KM276164
东非黑黄檀 <i>D. melanoxylon</i>	KM276127~KM276150、AB828650、MN076253、MN114132、MH465108
亚马孙黄檀 <i>D. spruceana</i>	MH465102~MH465103、AB828660
<i>D. ecastaphyllum</i>	AB828628、EF451071~EF451072、KM521373~KM521374
越南黄檀 <i>D. tonkinensis</i>	KY213796~KY213801、MN153525、FR854138~FR854142、MN121058、MN117711
<i>D. frutescens</i>	EF451078、AB828631~AB828632、MN076252
<i>D. decipularis</i>	EF451077、MH465106
<i>D. brasiliensis</i>	EF451076
巴西黑黄檀 <i>D. nigra</i>	EF451074~EF451075、MH465097~MH465098
<i>D. monetaria</i>	EF451073、KM521384~KM521385、AB828651
<i>D. villosa</i>	EF451068
<i>D. elegans</i>	EF451066
<i>D. cuiabensis</i>	EF451065
<i>D. acuta</i>	EF451064
<i>D. boehmii</i> subsp. <i>boehmii</i>	MN114131
南亚黄檀 <i>D. volubilis</i>	KM276212~KM276225
<i>D. tamarindifolia</i>	KM276205~KM276211
锈色黄檀 <i>D. rubiginosa</i>	KM276165~KM276183
<i>D. horrida</i>	KM276096~KM276104、AB828638
<i>D. cenatensis</i>	KM276089~KM276095、KM521361~KM521362、KM521359、AB828620~AB828621
<i>D. sissooides</i>	MK418451
降香黄檀 <i>D. odorifera</i>	KM521377、KM521388~KM521389
托叶黄檀 <i>D. stipulacea</i>	KM676489、KM521410~KM521413、AB828661、HG004827
绒叶黄檀 <i>D. velutina</i>	KM521417~KM521420
<i>D. trichocarpa</i>	KM521415~KM521416、AB828662
<i>D. subcymosa</i>	KM521414
毛叶黄檀 <i>D. sericea</i>	KM521406、AB828658
多裂黄檀 <i>D. rimosa</i>	KM521402~KM521404
多裂黄檀变种 <i>D. rimosa</i> var.	KM521376、AB828655
斜叶黄檀 <i>D. pinnata</i>	KM521401、KM521399、AB828654、KP092714~KP092715、HG004794
卵叶黄檀 <i>D. ovata</i>	KM521398
奥氏黄檀 <i>D. oliveri</i>	KM521392~KM521397、KM521390、KP719021、AB828652、FR854133~FR854137
<i>D. nigrescens</i>	KM521386、FR854128~FR854132
刀状黑黄檀 <i>D. cultrata</i>	KM521368~KM521370、AB828626、MW04435352
大金刚黄檀 <i>D. dyeriana</i>	KM521372、MH710796
<i>D. entadoides</i>	KM521375、HG313773
<i>D. hupeana</i>	KM521378、MN076254、MN496335

表 1 (续)

物种拉丁名	登录号
<i>D. hupeana</i>	MH711574、MH710919、KP338139
交趾黄檀 <i>D. cochinchinensis</i>	KM521363~KM521367、KP719022、AB828624~AB828625、FR854123~FR854127
<i>D. cana</i>	KM521357~KM521358
秧青 <i>D. assamica</i>	KM521354~KM521356、AB828610、FR854118~FR854122
<i>D. thorelii</i>	AB828668
<i>D. revoluta</i>	AB828667
<i>D. lateriflora</i>	AB828665
<i>D. inundata</i>	AB828664
赛州黄檀 <i>D. cearensis</i>	AB828663
<i>D. sakanensis</i>	AB828657
<i>D. rostrata</i>	AB828656
<i>D. parviflora</i>	AB828653
<i>D. melanocardium</i>	AB828649
<i>D. martinii</i>	AB828648
<i>D. maritima</i>	AB828647
<i>D. lactea</i>	AB828643
<i>D. junghuhnii</i>	AB828641
<i>D. kurzii</i>	AB828642
<i>D. humbertii</i>	AB828640
<i>D. hostilis</i>	AB828639
<i>D. havilii</i>	AB828637
藤黄檀 <i>D. hancei</i>	AB828636、HG313774
马达加斯加黄檀 <i>D. greveana</i>	AB828635
<i>D. godefroyi</i>	AB828634
<i>D. glomerata</i>	AB828633
<i>D. falcata</i>	AB828629
<i>D. floribunda</i>	AB828630
<i>D. dongnaiensis</i>	AB828627
<i>D. capuronii</i>	AB828623
<i>D. canescens</i>	AB828622
<i>D. bracteolata</i>	AB828619、MN264657、MN264656
<i>D. bojeri</i>	AB828618
<i>D. boehmii</i>	AB828617
<i>D. bintuluensis</i>	AB828616
<i>D. bignoniae</i>	AB828615
降香奇楠 <i>D. benthamii</i>	AB828614
<i>D. baronii</i>	AB828613
南岭黄檀 <i>D. balansae</i>	AB828612、KP092712~KP092713
<i>D. aurea</i>	AB828611
<i>D. armata</i>	AB828609
<i>D. arbutifolia</i>	AB828608
<i>D. afzeliana</i>	AB828607
<i>D. adamii</i>	AB828606
<i>D. acariiantha</i>	AB828605
<i>D. abrahamii</i>	AB828604
<i>D. abbreviata</i>	AB828603
<i>D. dialoides</i>	HG313775
钝叶黄檀 <i>D. obtusifolia</i>	MW044355、MW044354
滇黔黄檀 <i>D. yunnanensis</i>	MH808442~MH808443
<i>D. microphylla</i>	MN264658
<i>D. obovata</i>	MH465110
微凹黄檀 <i>D. retusa</i>	MH465099~MH465101
滇南黄檀 <i>D. kingiana</i>	HG004883
<i>D. hupeana</i>	KP338139

近距离匹配 (best close matc, BCM) 功能来计算 ITS 在黄檀属的鉴别成功率。

2.2 基于系统发育树的黄檀属物种 ITS 序列分析

分子系统发育分析是根据生物大分子序列差异来评估物种或分子间进化关系。本研究采用 NJ、最大似然法 (maximum likelihood, ML) 和贝叶斯推断法 (bayesian inference, BI) 对全部黄檀属 ITS 序列构建系统发育树来进行聚类分析。

采用 MEGA v 7.0 对黄檀属 402 条 ITS 序列进行序列比对, 随后构建最大似然树和最大简约树。构建最大似然树: 利用 MEGA 中的 Find Best DNA/Protein Models (ML) 功能对序列分析并模拟得出最适合的算法模型, 随后根据得出的模型, 完全删除 (complete deletion) 空位/缺失数据 (gaps/missing data) 并选择迭代 1 000 次开始构建发育树。构建最大简约树: 选择 Subtree-Pruning-Regrafting (SPR) 算法模型, 完全删除空位/缺失数据并迭代 1 000 次可作信度分析开始构建发育树。

采用 Phylosuite 软件的 Mrbayes 模块构建贝叶斯树, 由 ModelFinder 分析出最适合的建树模型, MCMC 参数中 Generations (代数) 参数取值为 5×10^6 , 其他参数默认, 开始构建贝叶斯树。

2.3 基于机器学习的黄檀属物种 ITS 序列分析

采用 BLOG 对所有黄檀属物种 ITS 序列进行学习分析并获得适用于黄檀属各物种识别的逻辑和决策规则。采用 WEKA 中的 4 个常用的分类器 Naïve Bayes、SMO、JRip 和 J48 对所有黄檀属 ITS 序列进行学习分析, 最后计算其鉴别成功率。

3 结果与分析

3.1 黄檀属物种 ITS 序列特征和距离计算

筛选获得的黄檀属 ITS 序列共计 402 条, 涵盖 96 个物种, 其中保守位点为 306 个, 占比 33.08%; 变异位点为 615 个, 占比 66.49%; 简约信息位点为 421 个, 占比 44.51%; 种内距离范围为 0~14.50%, 其平均种内距离为 3.04%; 种间距离范围为 0~42.50%, 其平均种间距离为 12.96%。

理想的 DNA 条形码种内距离应明显小于种间距离, 两者之间存在显著的差距, 即为 “Barcoding gap” [25]。TaxonDNA 的 “Extreme Pairwise” 功能确定每条序列的最小种间距离和最大种内距离, 将不确定最大种内距离的数据 (即 No matching conspecific sequence) 删除, 剩余全部数据绘制成散点图, 如图 1 所示。在散点图上添加一条 1:1 斜线

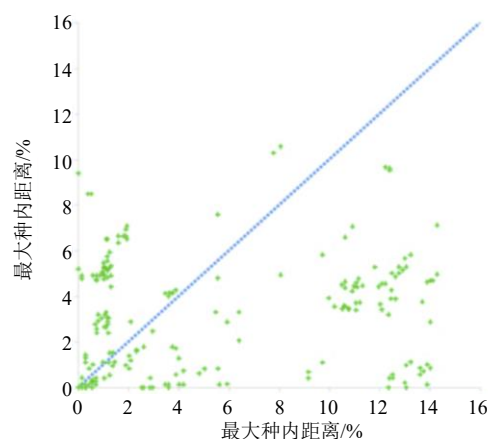


图 1 黄檀属 ITS 序列的 barcoding gap 分布

Fig. 1 Barcoding gap distribution of ITS sequences of *Dalbergia*

进行分隔, 根据 “Barcoding gap” 分析, 落在斜线上方的点呈现出 “Barcoding gap”, 占比为 43.25%。存在 “Barcoding gap” 的 2 条序列意味着其在 TaxonDNA 分析划界建立的阈值下不存在种内距离与种间距离的重叠, 则可划分为同一物种 [26], 反之则属于不同物种。本研究使用的序列集包含了部分单序列的物种和质量不好的序列, 这些序列则被划分到 “Barcoding gap” 中。

在基于 TaxonDNA 2 种算法 BM 和 BCM 的分析中, 基于 BM 算法的黄檀属物种 ITS 条形码正确分配率为 75.37%; 基于 BCM 算法的正确分配率为 72.88%。利用 BM 算法和 BCM 算法能够成功识别的物种有钝叶黄檀、刀状黑黄檀、*D. hupeana* Benth、越南黄檀、东非黑黄檀、阔叶黄檀、*D. frutescens* var. *tomentosa* Tndl、大金刚黄檀、印度黄檀、亚马孙黄檀、微凹黄檀、巴西黑黄檀、降香黄檀、交趾黄檀、奥氏黄檀、斜叶黄檀、南岭黄檀、披针叶黄檀、托叶黄檀、*D. trichocarpa* L.、多裂黄檀、*D. nigrescens* L.、多裂黄檀变种、赛州黄檀、*D. cana* L.、秧青、南亚黄檀、*D. tamarindifolia* L.、锈色黄檀、*D. horrida* L.、*D. miscolobium* haplotype L.、*D. entadoides* L.、*D. ecastaphyllum* L.、*D. frutescens* var. *frutescens* L., 共计 35 种。

3.2 基于系统发育树的黄檀属物种 ITS 分析

利用筛选得到的 402 条黄檀属物种 ITS 序列构建的 ML 树、NJ 树和 BI 树分别见图 2~4。本研究以 BI 树来进行主要分析, ML 树和 NJ 树作辅助分析, 无特殊说明的节点平均支持率按 BI/ML/NJ 的顺序标明。

在 BI 树中, 锈色黄檀、*D. cenatensis* L.、斜叶

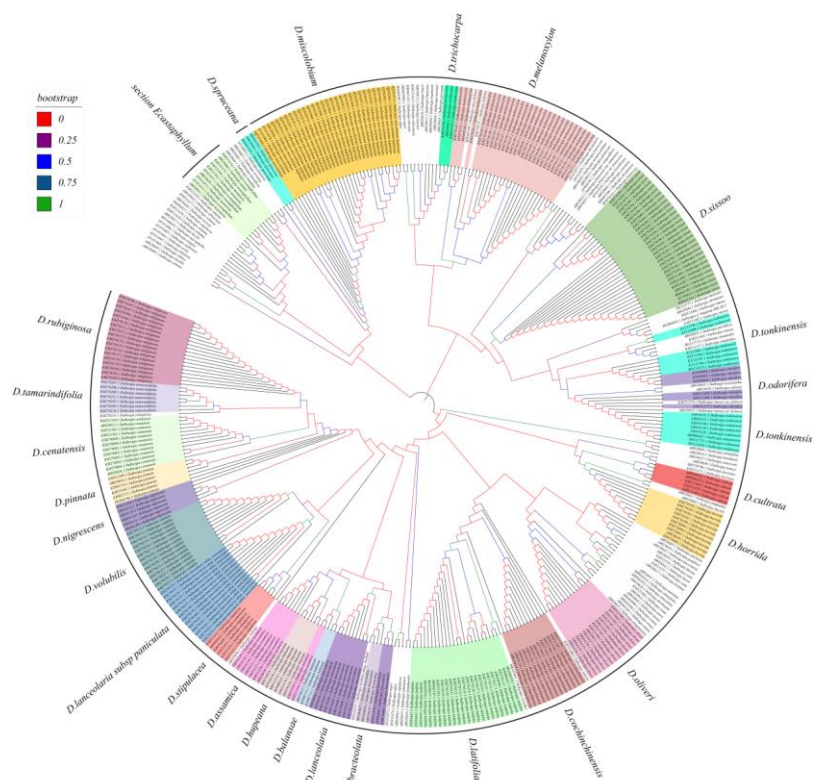


图 2 基于黄檀属 ITS 序列构建的 ML 树

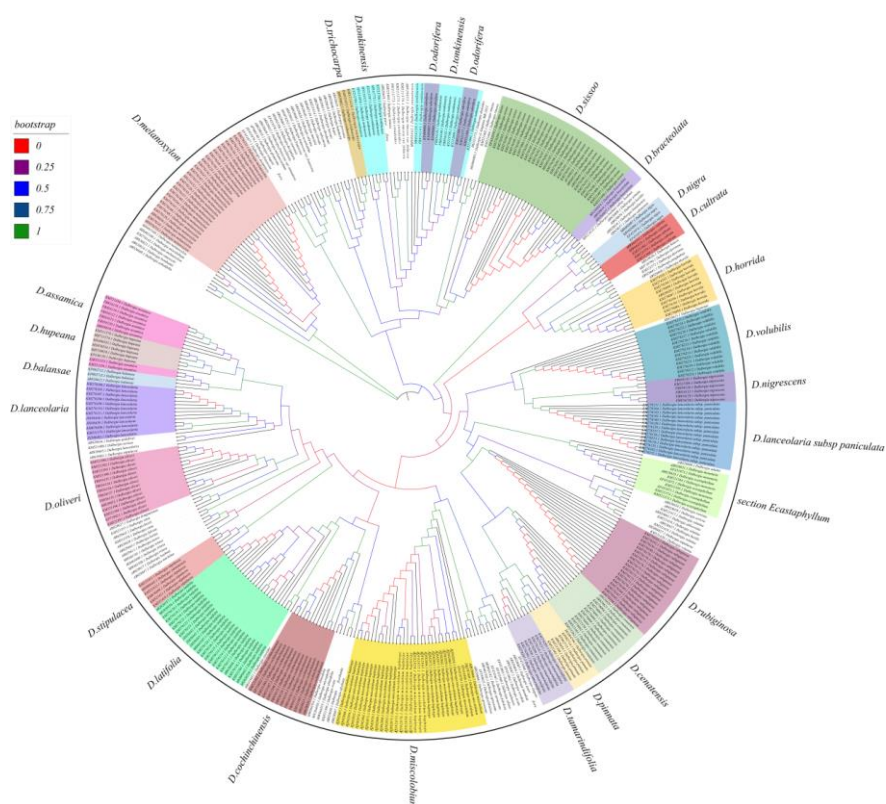
Fig. 2 ML tree based on barcode ITS of *Dalbergia*

图 3 基于黄檀属 ITS 序列构建的 NJ 树

Fig. 3 NJ tree based on barcode ITS of *Dalbergia*

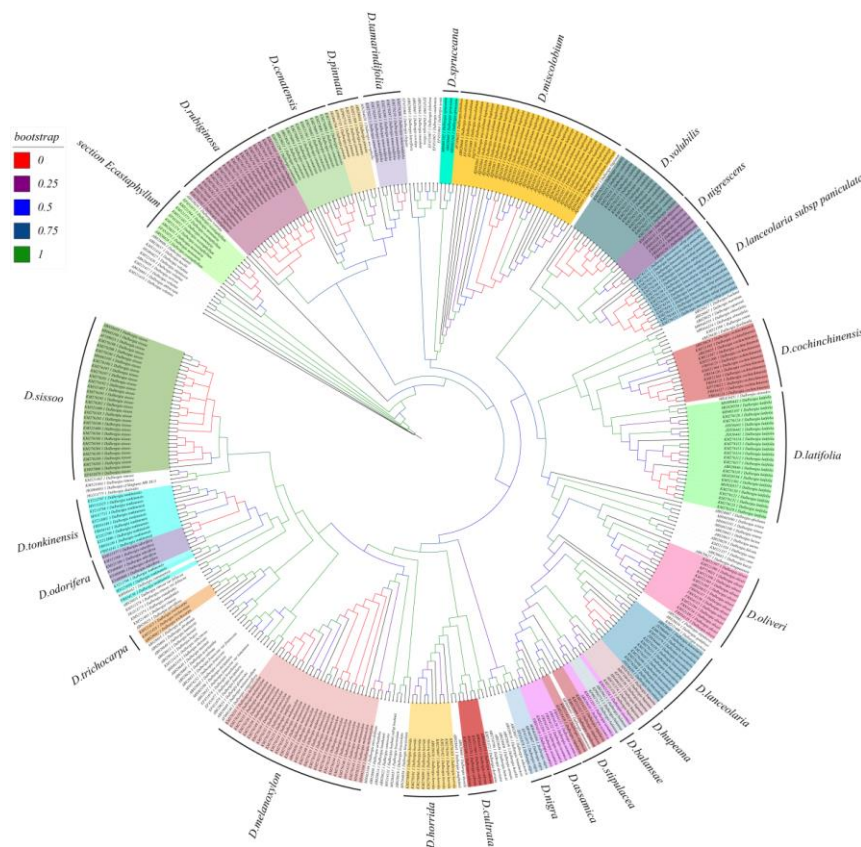


图 4 基于黄檀属 ITS 序列构建的贝叶斯树

Fig. 4 BI tree based on barcode ITS of *Dalbergia*

黄檀和 *D. tamarindifolia* 单独聚为一支，支持率为 56.8%；而在 ML 树和 NJ 树中，该 4 个物种同样单独聚为一支，支持率分别为 41% 和 53%。斜叶黄檀和 *D. tamarindifolia* 在 BI 树中聚为一支，而在 ML 树和 NJ 树中单独聚为一支，平均支持率都大于 90%。*D. adamii*、*D. monetaria* 和 *D. ecastaphyllum* 均聚为一支，支持率分别为 100%、64%、97.9%。*D. monetaria* 和 *D. ecastaphyllum* 为姐妹物种，均被纳入为 section *Ecastaphyllum*^[3]。*D. miscolobium* haplotype、*D. miscolobium*、*D. spruceana* 和 *D. foliolosa* 聚为一支，支持率分别为 100%、83%、97.9%。同样的，在 BI、ML 和 NJ 树中，南岭黄檀全部序列和秧青的部分序列均聚为一支，支持率为 100%、94%、97.7%；南岭黄檀的剩余序列单独聚为一支，但节点平均支持率较低，为 51.4%、55%、27.4%，可信度较低。南岭黄檀以高支持率验证了秧青和南岭黄檀为同义关系^[1]。披针叶黄檀亚种，*D. nigrescens*、南亚黄檀、交趾黄檀、奥氏黄檀、披针叶黄檀、巴西黑黄檀、刀状黑黄檀、*D. horrida*、印度黄檀均分别单独聚为一支，其支持率均大于 90%。

越南黄檀与降香黄檀聚为一支，说明越南黄檀和降香黄檀的亲缘关系较近。其中发现有一条阔叶黄檀序列在 3 棵树中都东非黑黄檀聚为一支，且支持率较高，分别为 90.1%、94%、100%，猜测该条序列可能发生某些位点的变异，导致分子系统发育树将该序列归到东非黑黄檀里。另外，许多只含少量序列 ($n \leq 3$) 的物种几乎都镶嵌或互相镶嵌在不同的分支里或者分支旁，且支持率参差不齐。综合上述分析，基于 BI/ML/NJ 算法的三棵系统发育树能够准确区分黄檀属的 96 个物种里的 27 个。

黄檀属 ITS 序列的变异位点占比达到 66.49%，因此具有较好的物种鉴别能力，在构建系统发育树时易于进行属内种间物种的区分^[1]。本研究结果进一步表明，含序列数多的物种大多数能够在系统发育树中聚类成分支且支持率较高，少数物种由于序列数极少，无法在得到明确的区分。其原因可能为人们主要根据形态学识别黄檀属，导致形态相似的近缘物种被命名并归类入黄檀属，随着 DNA 条形码技术的兴起，出现了“表里不一”的情况；由于冷门物种的治疗功效或利用价值并未完全发掘，人

们趋向于获取与研究热门物种的 DNA 条形码，导致物种 DNA 条形码序列数量较少，聚类受限。

3.3 基于机器学习的黄檀属物种 ITS 分析

基于 BLOG 算法，黄檀属序列随机分为了 142 条训练集序列和 259 条查询集序列，其训练集和查询集的信息统计见表 2。训练集中正确分类序列占比较低，为 36.36%，原因是只含一条序列的物种也被统计入内。而查询集中由于单条序列物种无法准确确认适用的逻辑规则，因此不能正确进行分类而被归入未分配序列，其正确分类序列占比达 95.75%。BLOG 能够准确识别黄檀属 96 种中 42 个物种，物种识别成功率为 43.75%，其中 *D. melanoxylon* 中共 22 条序列，12 条被归为正确分类序列（占 54.55%），10 条被归为未分类序列（占 45.55%）；*D. latifolia* 中共 20 条序列，19 条序列被归为正确分类序列（占 95%），1 条被归为未分类序列（占 5%）。其他 2 条或以上的序列的物种均为正确分类。

表 2 机器学习 BLOG 中训练集和查询集信息统计

Table 2 Statistics of trainfile and testfile information in machine learning BLOG

指标	训练集	查询集
序列数	143.00	259.00
正确分类序列数	52.00	248.00
正确分类序列占比/%	36.36	95.75
错误分类序列数	70.00	0
错误分类序列占比/%	48.95	0
未分类序列数	21.00	11.00
未分类序列占比/%	14.69	4.25

WEKA 的 4 个分类器中，基于 SMO、Naïve Bayes、JRip、J48 的分类器分别获得了 79.10%、58.71%、72.64%、76.37% 的正确分类序列占比。其中，SMO 分类器表现最好；而 Naïve Bayes 表现最差。He 等^[20]在对黄檀属进行鉴别分析以及冯婷婷等^[22]在对钩藤属和沉香属进行鉴别分析时也发现 Naïve Bayes 表现不佳，可能是因为 Naïve Bayes 算法是一类利用概率统计只是进行分类的算法，其按照贝叶斯定理先假设一个属性值对给类的影响独立于其他属性的值，即先验证概率后调整因子后计算后验概率^[23]，而由于各物种的序列数量不一致使得先验概率出现偏差而影响后验概率，使得分类准确率下降。

影响机器学习对物种鉴别成功率的主要原因是

数据集中各物种所含的序列数量。利用基于机器学习算法对 DNA 条形码进行学习分类的重要条件之一为拥有一套完整的参考数据集，即每个物种至少需要 4 个序列样本以方便建立参考数据库，每个物种的序列必须包括所有可能的核苷酸多态性（变异位点）^[18-19]。但如果没有使用足够序列数量的参考库则可能会出现拟合度不足或拟合过度的现象^[14]。本研究部分物种序列数量低于 4 个，使得正确分类序列占比降低以及产生了较大的相对误差。此外，在 5 种机器学习算法中，平均绝对误差以及均方差误差都较小，全部低于 0.1，即把这些序列数量较少物种的序列加入数据集中进行训练分析，结果显示能够鉴别物种的正确分类序列占比几乎没有受到这些物种序列的干扰。其原因可能是物种序列数量的增加，扩大了黄檀属物种序列的样本量，使得机器学习能够更好地学习分析这些序列的核苷酸多态性（变异位点）并用于训练集的分析，从而得到更加准确的逻辑和决策规则。

3.4 机器学习法与距离法/系统发育树法的比较分析

以条形码 ITS 作为分子标记，分别用机器学习法与距离法/系统发育树法对黄檀属 402 条 ITS 序列进行比较分析。研究结果表明基于机器学习的 BLOG 耗时低且鉴别成功率最高（43.75%），能够准确识别的黄檀属物种高达 42 种；WEKA 的 4 个分类器均能准确识别 34 种，鉴别成功率为 35.42%。而基于系统发育树的鉴别成功率最低（28.13%），仅能够准确识别 27 种；基于距离法能够成功准确识别 35 种，鉴别成功率为 36.46%。在机器学习法的 5 种分类器算法对黄檀属物种 ITS 序列的分析对比中，BLOG 算法的正确分类序列占比最高（95.75%），SMO 分类器算法位于第 2（79.10%），而 Naïve Bayes 算法最低（58.71%）。机器学习法与距离法/系统发育树法用于黄檀属物种大数据集比较分析的特征如下。

（1）距离法虽然耗时较低，能够适用于大型数据集，而由于不同物种的溯源祖深度存在差异，种间和种内距离常有重叠而无法区分物种，其结果往往不具有生物学意义^[27-28]。在效率层面，距离法在黄檀属物种 ITS 序列分析中，从数据分析到结果产出所用的时间较短，但其不能给出准确的生物学分类意义和高可读性的结果，因此还需其他分析方法参与物种鉴别。

(2) 系统发育树方法通过构建树枝状分支的系统发育树来分析物种间的亲缘关系^[22]。该方法的优点是能够比较直观的展示出每个物种的分支、距离以及可信度,即可读性较高。相较于距离法,基于系统发育树的 DNA 条形码分析,尤其是当物种种群具有复杂的进化史时,更加具有进化生物学意义^[1]。对于黄檀属内分类混乱的情况,基于系统发育树的方法能够明确直观地展示出各物种的进化情况。在效率层面,系统发育树由于需要进行一定次数的迭代计算分支的可信度,导致该方法的耗时较长,特别是处理大数量的序列集时,该方法所产生的结果可读性会有所降低,往往会出现消耗的时间长、可读性不占据优势的问题,最终导致效率低下。此外在结果产出后,后续人工审核步骤可能会由于人为因素而导致结果分析错误。

(3) 机器学习方法中 5 个分类器基于各自不同的算法,有效的弥补了距离法和系统发育树法的短板。在效率层面,机器学习算法在分析同样的数据集时所用的时间与距离法接近,同时能够获得比距离法和系统发育树法更高的鉴别成功率。BLOG 算法与 WEKA 中 RIPPER 算法类似,以分析识别序列中变异位点的逻辑规则,并将逻辑规则套用分析条形码序列以鉴别样本的种类^[19]。由 BLOG 算法产出的逻辑规则及 WEKA 的 4 个分类器产出的决策规则或矩阵都具有很高的可读性,其结果的清晰度也能减少后续人为审核步骤的主观性。

4 讨论

本研究以 ITS 条形码作为分子标记,分别用距离法、系统发育树和机器学习对黄檀属 402 条 ITS 序列进行分析和比较。研究结果表明机器学习法与距离法/系统发育树法都需要利用数据量来提高其准确性,但是由于算法的不同,导致产生的结果有差异。对于黄檀属 ITS 条形码数据集,利用基于机器学习方法进行黄檀基原识别能够获得更高的鉴别成功率,适合大范围物种数目,且具有更高的社会经济效率。此外 BLOG 能够分析并获得适用于黄檀属物种识别逻辑和决策规则,简单易懂,能够在黄檀属木材司法鉴定中作为重要的分类识别依据。

本研究表明机器学习方法可作为一种可靠、有效的黄檀属 DNA 条形码分析工具,是对传统 DNA 条形码分析方法的有力补充。机器学习方法能够弥补距离法和系统发育树方法的一些不足,更加契合市场、海关和医院等需求快速正确鉴别植物药材的

场所,为打击黄檀相似掺假,维持市场秩序以及用药安全提供技术支撑。综上所述,本研究建议优先利用机器学习对黄檀属 ITS 条形码进行高效准确的基原识别。

利益冲突 所有作者均声明不存在利益冲突

- [1] 李奇威. 基于 DNA 条形码的三种药用植物分子系统学研究 [D]. 广州: 广东药科大学, 2018.
- [2] Hassold S, Lowry P P 2nd, Bauert M R, *et al.* DNA barcoding of Malagasy rosewoods: Towards a molecular identification of CITES-listed *Dalbergia* species [J]. *PLoS One*, 2016, 11(6): e0157881.
- [3] Park K R, Yun H M, Quang T H, *et al.* 4-Methoxydalbergione suppresses growth and induces apoptosis in human osteosarcoma cells *in vitro* and *in vivo* xenograft model through down-regulation of the JAK2/STAT3 pathway [J]. *Oncotarget*, 2016, 7(6): 6960-6971.
- [4] Yun H M, Park K R, Quang T H, *et al.* 4-Parvifuran inhibits metastatic and invasive actions through the JAK2/STAT3 pathway in osteosarcoma cells [J]. *Arch Pharm Res*, 2017, 40(5): 601-609.
- [5] Kim E N, Kim Y G, Lee J H, *et al.* 6,7,4'-Trihydroxyflavone inhibits osteoclast formation and bone resorption *in vitro* and *in vivo* [J]. *Phytother Res*, 2019, 33(11): 2948-2959.
- [6] Kushwaha P, Khedgikar V, Ahmad N, *et al.* A neoflavonoid dalsissoal isolated from heartwood of *Dalbergia sissoo* Roxb. has bone forming effects in mice model for osteoporosis [J]. *Eur J Pharmacol*, 2016, 788: 65-74.
- [7] Saha S, Shilpi J A, Mondal H, *et al.* Ethnomedicinal, phytochemical, and pharmacological profile of the genus *Dalbergia* L. (Fabaceae) [J]. *Phytopharmacology*, 2013, 6: 56.
- [8] Rasamiravaka T, Jedrzejowski A, Kiendrebeogo M, *et al.* Endemic malagasy *Dalbergia* species inhibit quorum sensing in *Pseudomonas aeruginosa* PAO1 [J]. *Microbiology*, 2013, 159(Pt 5): 924-938.
- [9] Barrett M A, Brown J L, Morikawa M K, *et al.* Conservation. CITES designation for endangered rosewood in Madagascar [J]. *Science*, 2010, 328(5982): 1109-1110.
- [10] Hassold S, Lowry P P 2nd, Bauert M R, *et al.* DNA barcoding of Malagasy rosewoods: Towards a molecular identification of CITES-listed *Dalbergia* species [J]. *PLoS One*, 2016, 11(6): e0157881.
- [11] Jiao L C, Yin Y F, Cheng Y M, *et al.* DNA barcoding for identification of the endangered species *Aquilaria sinensis*:

- Comparison of data from heated or aged wood samples [J]. *Holzforschung*, 2014, 68(4): 487-494.
- [12] Ren B Q, Xiang X G, Chen Z D. Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers [J]. *Mol Ecol Resour*, 2010, 10(4): 594-605.
- [13] Al-Qurainy F, Khan S, Tarroum M, *et al.* Molecular authentication of the medicinal herb *Ruta graveolens* (Rutaceae) and an adulterant using nuclear and chloroplast DNA markers [J]. *Genet Mol Res*, 2011, 10(4): 2806-2816.
- [14] Lee S Y, Mohamed R, Faridah-Hanum I, *et al.* Utilization of the internal transcribed spacer (ITS) DNA sequence to trace the geographical sources of *Aquilaria malaccensis* Lam. populations [J]. *Plant Genet Resour*, 2018, 16(2): 103-111.
- [15] 张高曼, 赵立佳, 刘光富, 等. 基于 ITS 和 ITS2 序列的药用石斛 DNA 条形码鉴定研究 [J]. 中草药, 2024, 55(2): 575-587.
- [16] 励娜, 王丹, 陈一龙, 等. 基于 ITS2 DNA 序列的雷公藤属药材资源谱系地理学研究 [J]. 中草药, 2022, 53(17): 5476-5483.
- [17] Chen S L, Pang X H, Song J Y, *et al.* A renaissance in herbal medicine identification: From morphology to DNA [J]. *Biotechnol Adv*, 2014, 32(7): 1237-1244.
- [18] Weitschek E, Fiscon G, Felici G. Supervised DNA Barcodes species classification: Analysis, comparisons and results [J]. *BioData Min*, 2014, 7(1): 4.
- [19] Weitschek E, van Velzen R, Felici G, *et al.* BLOG 2.0: A software system for character-based species classification with DNA Barcode sequences. What it does, how to use it [J]. *Mol Ecol Resour*, 2013, 13(6): 1043-1046.
- [20] He T, Jiao L C, Yu M, *et al.* DNA barcoding authentication for the wood of eight endangered *Dalbergia* timber species using machine learning approaches [J]. *Holzforschung*, 2019, 73(3): 277-285.
- [21] 何拓. 基于机器学习的黄檀属与紫檀属木材识别方法研究 [D]. 北京: 中国林业科学研究院, 2019.
- [22] 冯婷婷. 基于机器学习的三种药用植物 DNA 条形码分子鉴别研究 [D]. 广州: 广东药科大学, 2021.
- [23] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets [J]. *Mol Biol Evol*, 2016, 33(7): 1870-1874.
- [24] Ross H A, Murugan S, Li W L S. Testing the reliability of genetic methods of species identification via simulation [J]. *Syst Biol*, 2008, 57(2): 216-230.
- [25] Zhu S, Li Q W, Chen S C, *et al.* Phylogenetic analysis of *Uncaria* species based on internal transcribed spacer (ITS) region and ITS2 secondary structure [J]. *Pharm Biol*, 2018, 56(1): 548-558.
- [26] Manzanilla V, Kool A, Nguyen N L, *et al.* Phylogenomics and barcoding of *Panax*: Toward the identification of ginseng species [J]. *BMC Evol Biol*, 2018, 18(1): 44.
- [27] Guo M Y, Xu Y Q, Ren L, *et al.* A systematic study on DNA barcoding of medicinally important genus *Epimedium* L. (Berberidaceae) [J]. *Genes*, 2018, 9(12): 637.
- [28] Meyer C P, Paulay G. DNA Barcoding: Error rates based on comprehensive sampling [J]. *PLoS Biology*, 2005, 3(12): e422.

[责任编辑 时圣明]