生物多样性 2008, **16** (5): 417–425 doi: 10.3724/SP.J.1003.2008.08215
Biodiversity Science http: //www.biodiversity-science.net

植物DNA条形码研究进展

宁淑萍[1, 3] 颜海飞[1] 郝 刚[2] 葛学军[1*]

1 (中国科学院华南植物园, 广州 510650)
2 (华南农业大学生命科学学院, 广州 510642)
3 (中国科学院研究生院, 北京 100049)

摘要: DNA条形码(DNA barcoding)已成为近5年来国际上生物多样性研究的热点, 即通过使用短的标准DNA片段, 对物种进行快速、准确的识别和鉴定。该技术在动物研究中已得到广泛的应用, 所采用的标准片段是线粒体COI基因中约650 bp长的一段。然而在植物中DNA条形码的研究进展相对缓慢, 目前尚处于对所提议的各片段比较和评价阶段, 还未获得一致的标准片段。由于植物中线粒体基因组进化速率较慢, 因此条形码片段主要在叶绿体基因组上进行选择, 被提议的编码基因片段主要有*rpoB*, *rpoC1*, *matK*, *rbcL*, UPA, 非编码区片段有*trnH-psbA*, *atpF-atpH*, *psbK-psbI*, 此外还有核基因ITS。已有的研究表明以上任何一个单片段都不足以区分所有植物物种, 因而不同的研究组相继提出了不同的片段组合方案, 目前被广泛讨论的组合主要有5种。本文综述了DNA条形码序列的优点、标准、工作流程、分析方法和存在的争议, 重点论述了植物条形码研究中被提议的各序列片段和组合的研究现状。

关键词: DNA条形码, 物种识别, 分析方法, 条形码评价

Current advances of DNA barcoding study in plants

Shuping Ning[1, 3], Haifei Yan[1], Gang Hao[2], Xuejun Ge[1*]

1 *South China Botanical Garden, Chinese Academy of Sciences, Guangzhou* 510650
2 *College of Life Sciences, South China Agricultural University, Guangzhou* 510642
3 *Graduate University of the Chinese Academy of Sciences, Beijing* 100049

Abstract: DNA barcoding has become one of hotspots of biodiversity research in the last five years. It is a method of rapid and accurate species identification and recognition using a short, standardized DNA region. DNA barcoding is now well established for animals, using a portion of the mitochondrial cytochrome *c* oxidase subunit 1 (COI or *cox1*) as the standard universal barcode. However, in plants, progress has been hampered by slow substitution rates in mitochondrial DNA. A number of different chloroplast regions have been proposed. There has been considerable debate, but little consensus regarding region choice for DNA barcoding land plants. Direct comparative assessment of different barcoding regions is now a priority to enable a standard barcoding solution to be agreed in plants. The proposed chloroplast barcoding regions mainly include five coding (*rpoB*, *rpoC1*, *matK*, *rbcL*, UPA) and three non-coding (*trnH-psbA*, *atpF-atpH*, *psbK-psbI*) regions. In addition, nrITS is also suggested as a potential plant barcode. Limited by the universality and resolvability of single barcoding region, five combinations of these regions are proposed. In this review, the advance of these barcoding regions, both their universality of primers and resolving power are reviewed. The advantages, standards, workflow and existent dispute of DNA barcoding are summarized.

Key words: DNA barcoding, species identification, analysis methods, barcoding evaluation

DNA条形码(DNA barcoding)是利用一个或少数几个DNA片段对地球上现有物种进行识别和鉴

收稿日期: 2008-08-23; 接受日期: 2008-09-18
基金项目: 国家科技基础条件平台工作重点项目: 植物标本标准化整理、整合及共享平台建设
* 通讯作者 Author for correspondence. E-mail: xjge@scbg.ac.cn

定(Kress *et al.*, 2005)的一项新技术, 是近年来进展最迅速的学科前沿之一。虽然十多年前研究者就已经采用小的基因片段对病毒、细菌、原生生物等缺乏足够形态特征的类群进行识别(Niesters *et al.*, 1993; Pace, 1997; Allander *et al.*, 2001; Hamels *et al.*, 2001), 在一些多细胞真核生物的研究中也有应用(Brown *et al.*, 1999; Doukakis *et al.*, 1999; Jackson *et al.*, 1999; Vincent *et al.*, 2000; Wells *et al.*, 2001; Wells & Sperling, 2001), 但20世纪末该技术并未扩展至整个生物界。真正将条形码技术引入生物界并提出"DNA条形码"概念的是加拿大University of Guelph教授、加拿大皇家学会会员Paul Hebert。Hebert等(2003a)选取线粒体细胞色素*c*氧化酶亚基I (cytochrome *c* oxidase subunit 1, COI)的一段序列在动物界不同分类水平上(门、目、种)进行分析, 发现无论在哪个分类水平上该基因都具有良好的识别能力, 从而提出建立以一段650 bp长的COI基因序列为基础的条形码识别方法。大量研究结果证明了COI条形码对动物物种的识别和鉴定切实可行(如: Hebert *et al.*, 2003b, 2004; Hajibabaei *et al.*, 2006; Yoo *et al.*, 2006; Yancy *et al.*, 2008)。截至2008年3月, 在DNA条形码数据库中已经收录了来自50,039种生物的363,584条序列, 其中来自13,761种生物的136,338条序列符合DNA条形码标准。这些物种中, 98%以上来自动物界(其中昆虫纲最多, 达65%以上) (Frézal & Leblois, 2008)。

生物条形码联盟(Consortium for the Barcode of Life, CBOL)在题为*Barcoding Life: Ten Reasons* 的小册子中清楚地阐述了DNA条形码的优点(http://phe.rockefeller.edu/barcode/)。概括起来有: (1)不受个体形态特征限制。采用一小块或一小片材料识别一个物种, 即使样本受损也不会影响识别结果。(2)不受个体发育阶段影响。有些物种在不同发育时期有明显差异, 不容易识别, 但其条形码不会发生变化。(3)对于分类学中难以区分的类群, 采用DNA条形码可以抛开形态相似的假象, 从基因水平上提供一种分类依据。(4)核苷酸序列组成的数据库可以被视为数字化的数据库, 提供明确的信息, 不仅弥补了形态描述的不足, 而且可以加快已知物种的识别速度, 同时便于新物种的发现, 将会使分类学科的发展更加快速和深入。(5)如果设想的条形码扫描仪可以实现, 将会减少对传统分类学人力和物力的需求, 会更有益于分类学家缺乏的国家, 尤其是发展中国家。

理想的DNA条形码应该符合以下几个标准: (1)在种间有明显的遗传变异和分化, 同时种内变异足够小; (2) 片段足够短, 便于一个反应完成测序工作, 而且便于DNA提取和PCR扩增, 尤其是对存在DNA降解的材料(如: 保存已久的腊叶标本、处理过的民间药材); (3) 存在保守区域, 便于设计通用引物。

DNA条形码在动物中采用线粒体基因, 而植物中至今尚未获得广泛认同的条形码标准片段, 当前工作重点依然是选择合适的片段并对其进行评价(Pennisi, 2007; Kress & Erickson, 2008)。本文综述了植物DNA条形码研究中被提议的序列片段和研究现状, 以及目前常用的数据分析方法, 以期能使国内植物学工作者对此有更多的了解和认识。

1 植物条形码研究中的候选片段

在植物中DNA条形码的研究进展相对缓慢, 主要有两方面原因: (1)植物线粒体基因组进化速率较慢, 遗传分化小, 因此动物中的标准片段COI不适用于植物; (2)系统学研究中常用的片段变异较小, 不适合用作条形码片段(Chase *et al.*, 2005; Kress *et al.*, 2005)。由于核基因组通常具有多拷贝的特性, 且物种内变异较大, 引物通用性差, 并且扩增时对模板DNA的质量要求高, 不适用于存在DNA降解的材料(Kress *et al.*, 2005), 因此, 植物中最可能的条形码还是从叶绿体基因组中选择(Chase *et al.*, 2005; Cowan *et al.*, 2006)。虽然叶绿体基因相对保守, 但仍然包含许多变异区域, 同时叶绿体基因组有其自身的优势: 单亲遗传避免了基因重组; 植物个体中均有大量的叶绿体, 即使DNA高度降解也容易扩增。

生物条形码联盟(CBOL)最初建议的植物条形码片段均为叶绿体片段: *matK*, *rpoC1*, *rpoB*, *accD*, *nhdJ*和*YCF5*。但因为后3个片段在一些主要的植物类群中有缺失, 如*YCF5*在苔藓类植物中缺失, *accD*在禾本科植物中缺失, 而*ndhJ*在松属植物中缺失, 在部分兰花中变短或功能丧失, 因此它们在第二阶段的更新中已被排除(http://www.kew.org/barcoding/)。此外, 一些研究者也建议了其他的片段, 例如: Kress等(2005)建议ITS和*trnH-psbA*两个片段,

Chase等(2005)和Newmaster等(2006)建议*rbcL*。更多的信息可以通过以下几个网站查询: www.barcoding.si. edu、www.barcodeoflife.org、http://phe.rockefeller. edu/BarcodeConference/、http://www.kew.org/barco-ding/。

由于越来越多的研究表明单靠某一个片段不太可能对所有的植物物种进行准确鉴定, 研究者又相继提出了不同的片段组合方案。片段组合的观点最早在Kress等(2005)的文中有所提及, 他们预测ITS + *trnH-psbA*将是被子植物中具有广泛应用价值的组合, 但并未做分析。2007年Chase等明确提出两套组合方案: *rpoC1* + *rpoB* + *matK* 和 *rpoC1*+ *matK* + *trnH-psbA*。同年Kress和Erickson (2007)又提出使用*rbcL* + *trnH-psbA*对陆生植物进行识别和鉴定。2007年9月在台北举行的第二届国际生物条形码大会上韩国植物学家Ki-Joong Kim等提出*matK* + *atpF-atpH* + *psbK-psbI* 和 *matK* + *atpF-atpH* + *trnH-psbA*两个组合(Pennisi, 2007; www.dnabarcodes-2007.org)。以上5种被提议的植物条形码组合方案构成了当前广泛讨论和评价的内容。在这些叶绿体片段中, *rpoB*, *rpoC*, *rbcL*、*matK*是编码区片段, *trnH-psbA*, *atpF-atpH*和*psbK-psbI*是非编码区片段。

2　各条形码片段和组合方案的应用现状

究竟哪些片段或组合在植物条形码研究中具有更好的应用前景？目前还没有一致意见。对片段的选择和评价应该考虑以下几个方面: (1)引物的通用性; (2)物种内部的变异程度; (3)区分物种的能力; (4)生物信息学的分析和应用(Kress & Erickson, 2008)。以此为标准, 以下就每个片段及组合分别进行介绍。

2.1　单片段情况

2.1.1　*matK*

相对于其他编码区片段, *matK*片段进化速率快, 但不同分支类群间很难进行扩增和测序, 因此其作为条形码最主要的争议是引物通用性差(Chase *et al.*, 2007; Hollingsworth, 2008), 不同类群往往需要采用不同的引物。使用生物条形码联盟植物工作组(Plant Working Group of the Consortium for the Barcode of Life, PWG-CBOL)建议的*matK*引物, Sass等(2007)未能在苏铁目中得到理想的扩增效果; Kress和Erickson (2007)对48属96个物种检测的扩增成功率仅为39.3%; Newmaster等(2008)在肉豆蔻科内毛楠属(*Compsoneura*)8个物种中的扩增同样失败。即使使用了10对该片段的不同引物, Fazekas等(2008)在32属92种251个个体中仅获得87.6%的扩增成功率。近些年, 该工作组和Ki-Joong Kim投入了相当多的工作开发该片段的通用引物, 但至今未取得理想的结果(Fazekas *et al.*, 2008)。这一片段已有的部分引物序列见表1。

与上述结果截然不同的是, Lahaye等(2008b)采用*matK*的390F/1326R引物(见表1)在所研究的1,667个植物材料中得到100%的扩增率, 并且单独使用*matK*或与*trnH-psbA*组合使用都可以正确识别90%以上物种。最近, Lahaye等(2008a)又再次强调*matK*可以作为单个片段应用于植物条形码, 而对于疑难类群再针对性地增加片段。但是, 该实验设计本身存在一些问题, 结论也有待商榷, 例如: 他们所分析的材料96%是兰科植物, 不能说明其他科的情况(Kress & Erickson, 2008); 有些物种缺乏重复个体, 而且没有包含同属内关系最近的姊妹种(Hollingsworth, 2008), 如果材料是姊妹种, 识别率可能会降低(Lahaye *et al.*, 2008b)。最近, Fazekas等(2008)采用Lahaye等介绍的390F/1326R引物获得的扩增成功率低于50%。我们使用该对引物在不同科、属间的扩增效果也不理想(待发表)。因此*matK*的引物通用性和鉴定效果还有待更多的实验数据检验, 开发广泛适用于植物各类群的通用引物是*matK*的一个工作重点。

2.1.2　*trnH-psbA*

*trnH-psbA*片段是进化速率最快的叶绿体间隔区之一, 片段两端存在75 bp的保守序列, 便于设计通用引物(Shaw *et al.*, 2005)。该片段引物(表1)通用性较好, 扩增成功率较高(Kress *et al.*, 2005; Kress & Erickson, 2007; Fazekas *et al.*, 2008; Lahaye *et al.*, 2008a, b; Newmaster *et al.*, 2008), 并且平均长度较短(大多数在450 bp左右), 有利于对降解材料的扩增(Shaw *et al.*, 2005)。但该片段中普遍存在插入/缺失事件, 甚至在近缘种间也存在(Aldrich *et al.*, 1988), 从而导致了不同植物间片段长度变异较大。在Kress等(2005)分析的53科80属99种植物中, *trnH-psbA*扩增长度为247–1,221 bp, 间隔区(排除了引物结合区和外显子侧翼区)为119–1,094 bp, 其中

表1 当前常用的部分条形码片段的引物序列
Table 1 Different primers for several plant barcoding regions

Gene	Primer	Direction	Sequence 5′→3′
matK	2.1	f	CCTATCCATCTGGAAATCTTAG
	2.1a	f	ATCCATCTGGAAATCTTAGTTC
	5	r	GTTCTAGCACAAGAAAGTCG
	3.2	r	CTTCCTCTGTAAAGAATTC
	2.1-Myristicaceae	f	CCTATCCATCTGGATATCTTGG
	5-Myristicaceae	r	GTTCTAGCACACGAAAATCG
	390F	f	CGATCTATTCATTCAATATTTC
	1326R	r	TCTAGCACACGAAAGTCGAAGT
	Angiosperms-KIM	F	ATCCATCTGGAAATCTTAGTTC
		r	GTTCTAGCACAAGAAAGTCG
	Plants-KIM	f	CRATCWATTCATTCAATATT
		r	CGTACAGTACTTTTGTGTTT
	matK_Kew	f	AATATCCAAATACCAAATCC
		r	ACCCAGTCCATCTGGAAATCTTGGTTC
trnH-psbA	psbA3	f	GTTATGCATGAACGTAATGCTC
	trnH-05	r	CGCGCATGGTGGATTCACAATCC
rbcL-a		f	ATGTCACCACAAACAGAGACTAAAGC
		r	GTYAAATCAAGTCCACCYCG
		f	ATGTCACCACAAACAGAGACTAAAGC
		r	CTTCTGCTACAAATAAGAATCGATCTC
atpF-atpH_KIM	atpF		ACTCGCACACACTCCCTTTCC
	atpH		GCTTTTATGGAAGCTTTAACAAT
psbK-psbI_KIM	psbK		TTAGCCTTTGTTTGGCAAG
	psbI		AGAGTTTGAGAGTAAGCAT

92%的物种扩增所得的片段长度为340–660 bp, 而且均具有独特的间隔区序列, 符合理想的条形码标准。Kress和Erickson (2007)采用9个候选片段(*rpoB*、*rpoC1*、*matK*、*trnH-psbA*、*rbcL*、ITS、*accD*、*nhdJ*和*YCF5*)比较分析了39目43科48属的96个种(包括海藻、苔藓和地钱类、蕨类、裸子植物和被子植物), Fazekas等(2008)采用另外9个相似的片段(*rpoB*、*rpoC1*、*matK*、*trnH-psbA*、*rbcL*、ITS1、UPA、*atpF-atpH* 和*psbK-psbI*)对32属92种251个植物个体进行了分析, 结果均表明所有片段中*trnH-psbA*在扩增成功率和物种识别率方面表现是最好的。

区分近缘或近期分化的物种是任何DNA条形码面临的一个挑战(Newmaster *et al.*, 2008)。肉豆蔻科是被子植物中一个古老的类群, 但又包含了一些近期分化的物种。在Newmaster等(2008)对该科内毛楠属开展的研究中, 只有*trnH-psbA*可以在每个种产生特异片段, 并能成功识别70%的种。Lahaye等(2008b)使用该片段对兰科植物进行鉴定, 识别率达90%以上。这些结果说明*trnH-psbA*对近缘种也有较好的识别。不过在某些类群中, 如伞形科独活属(*Heracleum*)和禾本科甜茅属(*Glyceria*), *trnH-psbA*却不能提供足够的变异以识别物种(Whipple *et al.*, 2007; Logacheva *et al.*, 2008)。

目前使用*trnH-psbA*最大的困难是非同属物种间的比对, 主要是由插入/缺失过多所引起 (Kress *et al*., 2005; Lahaye *et al.*, 2008b)。然而, 比对容易与否并不是条形码必需的条件, 一旦建立恰当的条形码数据分析方法, 插入/缺失还将会增加物种识别所需要的信息(Kress *et al.*, 2005)。在现阶段研究中, 比对可以先在同属种间进行, 之后再对所有物种进行比对, 并且尽量增加插入/缺失以保证同属种的同源性(Lahaye *et al.*, 2008a)。比较已有的多数研究结果, 我们认为*trnH-psbA*是非常有用的条形码片段之一, 即便不能单独使用, 也将可以成为组合方案中的一部分。

2.1.3 *rbcL*

由于在GenBank中有大量的*rbcL*序列数据, 并且其具有通用、易扩增、易比对的特点, *rbcL*被提议作为条形码片段。但是*rbcL*的变异主要存在于种以上水平, 物种水平上通常变异不够大(Kress & Erickson, 2007; Sass *et al.*, 2007; Fazekas *et al.*, 2008; Lahaye *et al.*, 2008b; Newmaster *et al.*, 2008)。通过使

用距离法对GenBank中大约10,300条长度大于1,000 bp的*rbcL*序列进行比较分析, Newmaster等(2006)发现尽管*rbcL*不能识别全部物种, 但可以区分不少同属植物。因此, 几位研究者都曾建议将*rbcL*与另外一个或多个片段组合使用。此外, *rbcL*的整体长度较长(至少1,300 bp), 需要使用4个引物并进行双向测序才能完成整个基因的测序(Kress *et al.*, 2005), 但理想的条形码要求片段长度较短, 因此有些研究仅选取其中一段进行扩增, 如*rbcL-a* (Kress & Erickson, 2007)。虽然该片段的引物通用性相对较好, 但也并不是对所有植物类群都适用。我们采用Kress和Erickson(2007)介绍的*rbcL-a*引物扩增后发现, 不同科、属间的扩增成功率存在很大差异(待发表)。

2.1.4 nrITS

核基因组的核糖体DNA ITS片段广泛分布于可进行光合作用的真核生物(除蕨类植物外)和真菌中, 是系统学研究中最常用的片段之一, 在GenBank中也积累了大量数据。Kress等(2005)最早将其视为植物条形码候选片段。组成该片段的不同部分(ITS1、ITS2和5.8S)序列变异差别较大, 5.8S最为保守, ITS1的识别效果好于ITS2 (Chase *et al.*, 2005)。Kress和Erickson(2007)的研究中, ITS1在成功扩增的材料中可以达到81.5%的正确识别率, 但该片段的扩增成功率仅为60.4%, 分析所有研究材料时正确识别率就下降为45.8%。因此扩增成功率是ITS作为条形码应用的一个限制因素。除此以外, 仍有下列原因导致ITS在一些类群中不适合作植物条形码: (1)其长度变异大, 多数物种扩增片段长度超过1,100 bp, 需要使用中间引物才能扩增获得整个基因; (2)存在长的poly-G、poly-C和poly-A, 导致测序和序列分析困难(Sass *et al.*, 2007); (3)核基因本身存在多拷贝的特性, 在种内序列变异较大, 进一步降低了该片段作为条形码的应用性(Kress & Erickson, 2007)。

除ITS外, 研究者们也考虑过系统学研究中表现较好的一些低拷贝核基因以及它们的内含子, 但由于缺乏通用引物最终被排除作为条形码片段的可能性(Kress *et al.*, 2005)。

2.1.5 其他片段

Taberlet等(2007)设计了扩增*trnL*内含子及部分P6 loop片段的引物。这两个片段具有引物保守、容易扩增等优点, 但最大的弱点是进化速率过慢, 在物种水平上识别率较低(Chase *et al.*, 2007; Taberlet *et al.*, 2007; Kress & Erickson, 2007)。Presting (2006)提议将UPA片段(Universal Plastid Amplicon)作为光合作用植物的条形码, 已有研究表明该片段在海藻中存在一定变异, 但在陆生植物内没有显著变异(Sass *et al.*, 2007; Fazekas *et al.*, 2008; Newmaster *et al.*, 2008)。另外, 在多数研究中*rpoB*和*rpoC1*也容易扩增, 虽然Chase等(2007)认为*rpoC1*片段的识别率较高, 他们提出的两个组合中均包含了该片段, 但多数实验数据显示*rpoB*和*rpoC1*序列相对保守, 变异较小。

以上各片段的更多引物序列可参见Fazekas等(2008)。

2.2 多片段组合情况

在植物中很难找到像动物中COI一样通用的单个片段, 即便找到“完美”的植物条形码, 靠单亲遗传的一个片段来区分杂交种或存在基因渗透的类群也会存在问题(Newmaster *et al.*, 2006)。多数研究结果也显示, 采用单片段的识别率很低, 不能达到条形码的要求(Kress & Erickson, 2007; Sass *et al.*, 2007; Newmaster *et al.*, 2008; Fazekas *et al.*, 2008), 因此筛选植物条形码不能仅关注单个片段, 必要时应增加片段。Chase等(2005)用交通灯法(traffic light approach)详细论述了植物中筛选DNA条形码的方法: 首先用单亲的叶绿体条形码进行初步分析, 能被准确鉴定的物种用绿灯(green light)表示; 若部分鉴定存在问题, 用黄灯(yellow light)表示, 由使用者根据需要决定是否需要进一步精确识别; 如果识别非常不精确则用红灯(red light)表示, 使用者需要进一步精确识别。Newmaster等(2006)以等级(tier)分类的观点支持这一方法: 首先找一个核心(core)片段作为第一级分类标准, 然后再根据不同类群选择不同片段作为第二级标准进一步分析。不同植物类群间进化速率差异较大, 理想的片段组合应该能够检测出多重水平的差异(Newmaster *et al.*, 2008)。到目前为止, 主要提出的片段组合有以下几种:

2.2.1 *rpoC1* + *rpoB* + *matK*或*rpoC1* + *matK*+*trnH-psbA*

这两套组合方案由Chase等在2007年提出, 是由相对保守的编码基因(*rpoC1*和*rpoB*)加进化相对较快的编码基因(*matK*)或非编码区(*trnH-psbA*)组成。*rpoC1*和*rpoB*引物通用性好, 扩增成功率高, 虽然进化较慢, 但也能区分相当数量的物种; *matK*序

列变异较大, 能够提供更多的识别。不过*matK*引物的通用性有待加强。另一个组合中保守的*rpoB*和长度一定的*matK*可以保证物种间进行广泛的比较, 加上高度变异的*trnH-psbA*就可以识别更多的物种。不过, 如前所述*trnH-psbA*的特性使得应用该组合时需要进一步发展序列分析策略。但在其他研究中, 这两个组合方案的识别效果并不理想(Sass *et al.*, 2007; Fazekas *et al.*, 2008), 其中*rpoC1* + *rpoB* + *matK*的识别率还要低于*rpoC1*+ *matK* + *trnH-psbA*。因此需要更多的实验数据来检测这两个组合的应用前景。

2.2.2 *rbcL* + *trnH-psbA*

该组合由Kress和Erickson于2007年提出, 他们选取了*rbcL*的一段*rbcL-a*。编码片段*rbcL*虽然变异较小但通用性好, 可以用作第一级分类的核心片段将一个未知样品锚定到科、属, 甚至是种, 不能被识别的样品则采用高变异的*trnH-psbA*进一步细分, 因此该组合尤其适用于被子植物中物种丰富的属。在Kress和Erickson(2007)所分析的48属96个物种中, *trnH-psbA*分别与*rpoC1*、*rpoB2*或*rbcL-a*的组合均拥有最高的通用性和物种识别能力(88%), 但*rbcL*被证明在陆生植物中很容易扩增, 而且能够在属和科水平上识别材料, 因此被认为是与*trnH-psbA*组合使用的最佳选择。Fazekas等(2008)认为条形码的识别能力与组合片段的数目相关, 虽然当采用4个片段时正确识别率达到最大值, 但*rbcL* + *trnH-psbA*是扩增和物种识别效果最好的两片段组合, 对物种的正确识别率(64%)与三片段组合的结果相近。

2.2.3 *matK* + *atpF-atpH* + *psbK-psbI*或 *matK* + *atpF-atpH* + *trnH-psbA*

Kim 等 2007 年 (Pennisi, 2007; www.dnabarcodes2007.org)在第二届国际生物条形码大会上提出这两种组合。Lahaye等(2008a)采用18科31种101个植物个体检测了新提出的这两个片段组合, 结果显示*psbK-psbI*和*atpF-atpH*的成功率较高，分别为98%和93.1%，当采用以上两种片段组合时，成功率均达到100%。在物种识别方面，采用UPGMA聚类法，*matK* + *atpF-atpH* + *psbK-psbI* 和 *matK* + *atpF-atpH* + *trnH-psbA*在物种单系性分辨率上分别为93.1%和89.3%。在Fazekas等(2008)对32属92种251个植物个体的分析中, *psbK-psbI*的扩增成功率仅有79%, 而*atpF-atpH*达到88%, 主要是在非种子植物(苔藓和蕨类植物)中扩增失败。在成功扩增的个体中, 单独使用*psbK-psbI*和*atpF-atpH*识别物种的正确率分别仅为44%和45%。*matK* + *atpF-*联单 *atpH* + *psbK* *-psbI*是正确识别率最高(69%)的三片段组合。虽然所检测的各片段效果相近, 没有出现特别理想的片段或组合, 但综合扩增成功率和物种识别率两方面表现, Fazekas等建议多片段联合时, 编码片段在*rbcL, rpoB, matK*中选择, 而非编码片段则在*trnH-psbA*和*atpF-atpH*中选择。

综上所述, 植物条形码需要采用多个片段组合。片段组合一定程度上可以降低种内变异带来的影响, 同时减少种内和种间变异的重叠(Newmaster *et al.*, 2006)。多片段组合应该由进化速率快慢不同的片段组成, 编码基因和非编码区组合是较好的选择。当前多数研究者倾向于*matK*和*trnH-psbA*这两个片段参与组合, 而第三个片段将可能是Kim等提出的*atpF-atpH*或*psbK-psbI*片段(Pennisi, 2007)。此外, 片段组合后分析应该分步进行, 编码基因受选择压力大, 变异通常较小但通用性好, 应该先用编码基因锚定到科或属, 再用变异更大的片段(编码或非编码的)区分到种。但是多数论文都是将不同片段的序列直接拼接进行分析, 序列演化速率的差异可能会影响到物种的正确识别率。因此，片段组合的分析方法有待进一步探讨。

3 DNA条形码的工作流程及分析方法

3.1 DNA条形码的工作流程

DNA条形码的工作流程与分子系统学研究的操作相似, 主要有以下步骤: (1)采集所需样品并提取DNA; (2)设计和合成通用引物; (3)进行PCR扩增, 筛选引物, 优化反应条件; (4)测序; (5)序列编辑、人工校正; (6)结果分析; (7)提交结果到相关数据库。

目前DNA条形码技术(工作流程)在植物中尚处于评估阶段, 但当该技术取得一致标准并完善后, 会建立相关的数据库来保存所得到的条形码序列, 届时将会对所提交的信息做出一定要求。目前只有动物的相关数据库Barcode of Life Database (BOLD) (http://www.boldsystems.org)。Ratnasingham和Hebert (2007)明确指出提交动物条形码序列应包含以下信息: (1)物种名称; (2)凭证标本信息(目录号和馆藏号); (3)采集号(采集人、采集日期和GPS定位地点); (4)标本鉴定人; (5) COI序列至少500 bp; (6) 用于

PCR扩增的引物; (7) 序列峰图。植物条形码数据库信息也可以参照此标准进行, 并要求配有照片, 以及采集地、形态特征等有关信息的文字描述。

3.2 DNA条形码的分析方法

在动物COI片段研究中, 所用到的分析方法比较简单。首先进行序列比对和人工校正, 剪去序列两端不可靠的碱基序列, 之后, 通过MEGA或PAUP计算种内和种间的Kimura-2-parameter distance (K2P)距离; 再根据距离计算结果建立Neighbour-joining tree (NJ树)。在数据较多的时候, 还可以进行多元尺度分析, 以图的形式更直观地反映物种水平的分辨效果(Hebert *et al.*, 2003a)。由于植物条形码研究还处于对片段的评价阶段, 分析方法与动物中已成熟的方法有所不同。主要分析内容和步骤如下:

(1) 序列比对和人工校正, 与动物条形码及分子系统学研究相同。

(2) 遗传距离计算: 种间距离通常采用pairwise uncorrected p-distance (Newmaster *et al.*, 2008)或Kimura-2-parameter distance (K2P) (Meyer & Paulay, 2005; Lahaye *et al.*, 2008a, b)模型计算。K2P是距离值很小时的最佳模型(Hebert *et al.*, 2003a), 也是生物条形码联盟(CBOL)推荐使用的距离计算模型(barcoding.si.edu/)。

种内距离通常采用3种参数表示(Meyer & Paulay, 2005; Lahaye *et al.*, 2008a, b): K2P 距离(K2P distance), 平均θ值和平均溯祖度(average coalescent depth)。其中平均θ值是指每个物种内不同个体间的平均K2P距离, 目的是消除不同物种因采样个体数不均引起的偏差; 平均溯祖度是指物种内所有个体间最大的K2P距离, 用以反映种内最大变异范围。K2P距离可以通过MEGA或PAUP计算, 在此基础上计算其余两个参数。究竟种内应该选用多少个体? Meyer和Paulay(2005)对此进行了分析, 通过比较选用2、5、10个个体时的种内遗传距离, 发现平均溯祖度随着采样数目的增加而增加, 由0.0049(n≥2), 到0.0057(n≥5), 再到0.0070(n≥10)。其他两个参数也有此特征, 因此应该尽可能地增加物种内的取样个体数。然而, 考虑到研究成本, 现在通常认为每个物种内不超过10个个体, 并最好包括5个不同居群。

当前植物条形码研究需要对各个片段的效果进行评估, 因此需要对不同片段在种内和种间的变异情况进行比较, 通常采用Wilcoxon Signed Rank Tests 进行检验。此项操作可以通过编写程序在PERL或R软件等统计分析软件中进行, 也可以通过网上的程序执行运算(http://faculty.vassar.edu/lowry/wilcoxon.html)。Newmaster等(2008)采用SPSS软件进行Kolmogorov-Smirnov检验, 其目的与Wilcoxon Signed Rank Tests相同。

(3)系统学分析: 条形码分析中通常采用标准的分子系统学方法(比如NJ、UPGMA、ML、MP、Bayes)建立多种系统树。然而, 建树的目的并不是利用条形码重建系统发育树, 而是为了检验每个物种的单系性, 即同一物种的不同个体能否紧密聚类到一起。不同的建树方法可能得到不同的效果, Lahaye等(2008b)对以上几种系统树进行了比较, 最终认为MP树和UPGMA树得到的物种正确识别率最高, 因此在他们最新的论文中只选用了这两种分析方法。但MP树所需要的运算时间长, 未必适合应用于大规模的数据计算。不同方法的运算时间差别很大, 而且适用的条件不同, 在使用时应根据需要进行选择。结果相差不大时应该选择最简单的树, 如NJ树, 这样才能达到条形码快速简便的效果。

(4) barcoding gap检验: 理想条形码检测到的同属内种间遗传变异应明显大于种内遗传变异, 并在两者之间存在显著差异, 形成一个明显的间隔区, 称作barcoding gap (Meyer & Paulay, 2005; Lahaye *et al.*, 2008a, b)。barcoding gap是评价DNA条形码理想与否的一个重要指标, 因此现阶段评价各片段时通常会进行barcoding gap检验。该检验实际上是用柱形图呈现种间、种内的遗传距离的分布频度, 采用Meier等(2006)开发的TaxonDNA软件结合一般的统计软件来完成。理想状况下, 柱形图上的种内变异集中在数值较小一侧, 而种间距离集中在数值较高一侧。

以上是当前植物条形码研究中最常用的分析方法, 文献中还有一些其他分析方法, 如: 相似法的BLAST、诊断法的DNA-BAR/DEGENBAR、多元尺度分析(multidimensional scaling)等, 但均未被广泛应用。

4 存在的争议和发展趋势

自提出DNA条形码概念以来, 大量生物学从业者持积极支持的态度, 但也有部分专家持怀疑和反

对态度。反对者往往强调如此短片段的DNA条形码不能提供物种水平上的可靠信息(Mallet & Willmott, 2003), 完全依靠遗传分化会导致错误的识别。另外, 对此方法的价值存在一些争议, 一些学者认为这个新技术会削弱或者取代、而非增强传统的以形态为基础的分类方法(Kress *et al.*, 2005)。无论是对动物还是植物的条形码研究最大的争议都是该方法能否适用于近缘和近期分化的物种, 这也是DNA条形码研究的难点。我们认为对于这样的类群可能应采用更多的分子标记(如SNP, SSR, AFLP等)来解决, 而不是局限于少量片段的序列。

由于各片段(尤其是多片段的组合选择)在不同类群中的应用效果不同, 目前尚未获得一致的植物条形码标准片段, 因此, 当前的研究热点仍然是选择和评价可能的条形码片段, 进行更大规模的分析和整体评价。另外, 植物条形码的分析方法也不够成熟, 需要生物信息学进一步发展开发出适合多片段和针对某些特殊片段(如*trnH-psbA*)的分析方法。正如林奈双名法为物种建立了一个形态特征的小型标签, 短的特征DNA序列构成了一个物种基因组的小型标签(Hollingsworth, 2008)。尽管还存在争议, 但DNA条形码已经在动物中得到广泛应用, 在植物中的研究也正在快速开展, 这将有助于非分类学专业工作者对有关材料进行快速、准确的鉴定。DNA条形码不能取代传统的分类学, 但作为数字信息DNA序列, 其准确性、丰富性以及独一无二的可重复性将使该技术成为分类学家有用的工具(Schindel & Miller, 2005)。随着生物技术的发展, 测序反应将会更快速、更便宜, 有利于构建更完整的公共序列数据库, 最终将使DNA条形码这一快捷、高效的技术越来越实用。

致谢: 中科院植物研究所马克平研究员对本项目的进行给予了很大帮助, 并对论文提出了中肯的意见, 特此致谢!

参考文献

Aldrich JBW, Cherney E, Merlin LC (1988) The role of insertions/deletions in the evolution of the intergenic region between *psbA* and *trnH* in the chloroplast genome. *Current Genetics*, **14**, 137–146.

Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proceedings of the National Academy of Sciences*, *USA*, **98**, 11609–11614.

Brown B, Emberson RM, Paterson AM (1999) Mitochondrial COI and II provide useful markers for *Weiseana* (Lepidoptera, Hepialidae) species identification. *Bulletin of Entomological Research*, **89**, 287–294.

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madrinan S, Petersen G, Seberg O, Jorgsensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L, Wilkinson M (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1889–1895.

Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon*, **55**, 611–616.

Doukakis P, Birstein VJ, Ruban GI, Desalle R (1999) Molecular genetic analysis among subspecies of two Eurasian sturgeon species, *Acipenser baerii* and *A. stellatus*. *Molecular Ecology*, **8**, S117–S127.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PloS One*, **3**, e2802.

Frézal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution*, **8**, 727–736.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences*, *USA*, **103**, 968–971.

Hamels S, Gala JL, Dufour S, Vannuffel P, Zammatteo N, Remacle J (2001) Consensus PCR and microarray for diagnosis of the genus *Staphylococcus*, species, and methicillin resistance. *Biotechniques*, **31**, 1364–1372.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **270**, 313–321.

Hebert PDN, Ratnasingham S, deWaard JR (2003b) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **270**, S96–S99.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *Plos Biology*, **2**, 1657–1663.

Hollingsworth PM (2008) DNA barcoding plants in biodiversity hot spots: Progress and outstanding questions. *Hered-*

ity, **101**, 1–2.

Jackson RB, Moore LA, Hoffmann WA, Pockman WT, Linder CR (1999) Ecosystem rooting depth determined with caves and DNA. *Proceedings of the National Academy of Sciences*, *USA*, **96**, 11387–11392.

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PloS One*, **2**, e508.

Kress WJ, Erickson DL (2008) DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences*, *USA*, **105**, 2761–2762.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences*, *USA*, **102**, 8369–8374.

Lahaye R, Savolainen V, Duthoit S, Maurin O, Bank Mvd (2008a) A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa). *Available from Nature Precedings*, <http://hdl.handle.net/10101/npre.12008.11896.10101>.

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, barraclough TG, Vincent S (2008b) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences*, *USA*, **105**, 2923–2928.

Logacheva MD, Valiejo-Roman CM, Pimenov MG (2008) ITS phylogeny of West Asian *Heracleum* species and related taxa of Umbelliferae-Tordylieae W.D.J. Koch, with notes on evolution of their *psbA-trnH* sequences. *Plant Systematics and Evolution*, **270**, 139–157.

Mallet J, Willmott K (2003) Taxonomy: renaissance or Tower of Babel? *Trends in Ecology & Evolution*, **18**, 57–59.

Meier RS, Kwong S, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Meyer CP, Paulay G (2005) DNA barcoding: Error rates based on comprehensive sampling. *PloS Biology*, **3**, 2229–2238.

Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany*, **84**, 335–341.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, **8**, 480–490.

Niesters HG, Goessens WH, Meis JF, Quint WG (1993) Rapid, polymerase chain reaction-based identification assays for Candida species. *Journal of Clinical Microbiology*, **31**, 904–910.

Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.

Pennisi E (2007) Wanted: A barcode for plants. *Science*, **318**, 190–191.

Presting GG (2006) Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. *Canadian Journal of Botany*, **84**, 1434–1443.

Ratnasingham S, Hebert PDN (2007) Bold: the barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Sass C, Little DP, Stevenson DW, Specht CD (2007) DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PloS One*, **2**, e1154.

Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature*, **435**, 17.

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu WS, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL (2005) The tortoise and the hare. II. Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, **92**, 142–166.

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, 8.

Vincent S, Vian JM, Carlotti MP (2000) Partial sequencing of the cytochrome oxydase b subunit gene I: A tool for the identification of European species of blow flies for postmortem interval estimation. *Journal of Forensic Sciences*, **45**, 820–823.

Wells JD, Pape T, Sperling FAH (2001) DNA-based identification and molecular systematics of forensically important sarcophagidae (Diptera). *Journal of Forensic Sciences*, **46**, 1098–1102.

Wells JD, Sperling FAH (2001) DNA-based identification of forensically important Chrysomyinae (Diptera: Calliphoridae). *Forensic Science International*, **120**, 110–115.

Whipple IG, Barkworth ME, Bushman BS (2007) Molecular insights into the taxonomy of *Glyceria* (Poaceae: Meliceae) in North America. *American Journal of Botany*, **94**, 551–557.

Yancy HF, Zemlak TS, Mason JA, Washington JD, Tenge BJ, Nguyen NLT, Barnett JD, Savary WE, Hill WE, Moore MM, Fry FS, Randolph SC, Rogers PL, Hebert PDN (2008) Potential use of DNA barcodes in regulatory science: Applications of the regulatory fish encyclopedia. *Journal of Food Protection*, **71**, 210–217.

Yoo HS, Eah JY, Kim JS, Kim YJ, Min MS, Paek WK, Lee H, Kim CB (2006) DNA barcoding Korean birds. *Molecules and Cells*, **22**, 323–327.

(责任编辑: 周玉荣)