应用昆虫学报 *Chinese Journal of Applied Entomology* 2013,50(1):283 - 287. DOI:10.7679/j.issn.2095 - 1353.2013.037

昆虫分子生物学方法

昆虫 DNA 条形码分析中的距离方法*

金 倩** 张爱兵***

(首都师范大学 生命科学学院 100048 北京)

摘 要 本文介绍了 DNA 条形码分析中常用的距离方法,包括简单的距离方法、基于阈值的距离方法和基于模糊成员关系的距离方法,并且以松毛虫属 7 个近缘种的COI 数据为例,运用基于阈值的距离方法进行演示分析。

关键词 DNA 条形码, 昆虫, 遗传距离

Distance-based DNA barcoding methods for insects

JIN Qian** ZHANG Ai-Bing***

(College of Life Sciences, Capital Normal University, Beijing 100048, China)

Abstract Three commonly used distance-based barcoding methods, minimum distance (MD), best close match (BCM), and a fuzzy-set theory-based method are introduced and discussed. We demonstrate an application of the BCM approach to an empirical dataset from seven Masson pine moth species.

Key words DNA barcoding, insects, genetic distance

DNA 条形码研究自加拿大科学家 Hebert 在 2003 年提出后(Hebert *et al.*,2003a,2003b)已经在不同生物类群中得到广泛的应用,基本上经历了从刚开始提出时的"广泛关注与争议"阶段,到"应用与争议并存"阶段,再到目前的"广泛应用与少量争议并存"阶段。昆虫作为后生动物中极其庞大的类群,在 DNA 条形码研究中占据着非常重要的分量,大量的 DNA 条形码研究都是以昆虫为研究对象展开的。

在 DNA 条形码研究中,如何分配携带未知 DNA 序列的物种到数据库中已知的物种是 DNA 条形码研究中基本和核心的问题之一。各种分析方法,比如基于进化树(NJ 等)、基于距离、基于人工智能等相继被领域内的学者提出(Zhang *et al.*, 2008, 2012a, 2012b)。基于进化树的分析方法由于对数据库中物种水平的单系性的严格要求,一定程度上限制了其应用,而距离方法具有原理简单、操作简便及没有单系性的要求等优点而具有很大的应用前景。DNA 条形码中的距离方法包括简单的距离方法(MD),基于阈值的距离方法(BCM),和基于模糊成员关系的距离方法(Zhang *et al.*, 2012a)。本文拟就其基本原理、目的、意义及实际应用等进行阐述。

简单距离方法的基本原理是通过计算未知问询序列和数据库中已知物种 DNA 序列的遗传距离,通常基于 K2P 进化模型进行计算,与未知序列遗传距离最小的序列所代表的物种即被认为是查询成功的物种。然而这种简单的距离方法往往会导致假阳性识别,尤其是当未知问询序列所代表的物种还没有被数据库收录时这种情况会时常发生。例如当一个新的 DNA 条形码项目启动时,研究人员无从知道一个新获得的 DNA 样本所代表的物种是否已在数据库中收录。基于此,BCM 方法和基于模糊成员关系的距离方法一定程度上克服了上述困难,减少了 DNA 条形码中假阳性识别的发生,它主要是通过对数据库中的样本进行统

* 资助项目:国家自然科学基金项目(31272340);北京市人才强教项目(PHR201107120)。

** E-mail:jinhongyu2001@163.com

*** 通讯作者,E-mail:zhangab2008@gmail.com

收稿日期:2012-12-21,接受日期:2012-12-28

计分析,给出一个识别相似性的统计阈值,即和问询序列遗传距离最小且达到一定阈值的数据库中的物种才被认为是识别成功的物种。

1 方法

1.1 昆虫类群样本收集及形态鉴定

研究人员首先要根据项目需求确定自己所研究的昆虫类群,比如,以鳞翅目枯叶蛾科松毛虫属昆虫为例,文献报道中国有该属昆虫 27 种,目前其常见和危害较为严重的 7 种松毛虫的 DNA 条形码库已初步建立(Dai *et al.*, 2012),然而为了全面掌握该类群的遗传变异,研究人员除需进一步补充该类群其它松毛虫的样本外,另一方面取样还要尽量覆盖每一种松毛虫不同的地理种群,因为已有研究表明,种群的遗传结构会对 DNA 条形码研究产生重要的影响。采集到的昆虫标本在构建 DNA 条形码库之初,需由相关类群的专家进行形态学鉴定,以保证后续鉴定的可靠性。所有这些标本在进行形态学鉴定的同时,还需取部分组织,如蛾类样本一侧的 3 只足(一般在昆虫冷冻后或新鲜标本状态时),放在无水乙醇中低温长期保存(-20C°),如图 1 所示。

1.2 构建昆虫类群本地 DNA 条形码数据库

取鉴定好的昆虫标本的组织,按常规方法进行昆虫总量 DNA 的提取,然后用通用COⅠ条形码引物(www. barcodinglife. org)进行 PCR 扩增,测序,如图 1 所示。所得COⅠ条形码 DNA 序列进行人工校对,并进行翻译测试,如果所得 DNA 序列能够根据无脊椎动物线粒体密码子表翻译成氨基酸/蛋白质,证明所得序列为目标序列,还可把所得序列和 GenBank 中的序列进行 BLAST 检索,以进一步确定所得序列为COⅠ基因。翻译测试可以在很多序列分析软件中完成,比如 Bioedit。每条序列在完成上述准备后,整理存放在一个. fasta 格式文件中,运用多序列比对算法进行序列比对(例如 ClustalX 软件),这一步非常关键,会影响到后续分析的准确性,比对后的序列理论上讲应该不会引入插入缺失(“indels”或“Gaps”),因为COⅠ作为蛋白质基因的保守性。如果在对齐后的序列中发现引入了“Gaps”,则需返回去重新检测相关序列的测序峰图,往往会发现这是由测序错误造成。此外,本地 DNA 条形码库还可以进行不断的更新。

1.3 应用距离方法和 DNA 条形码库识别未知样本

一旦本地 DNA 条形码数据库建成,则可用于未知样本的种类识别,这将为昆虫生态学研究提供很多便利,如图 1 所示,因为很多昆虫生态学工作者一方面由于分类学知识的匮乏或对分类学缺乏足够的兴趣,并不想把自身的大量研究精力消耗在所采集的昆虫标本的形态鉴定上,他们仅需鉴定出所研究昆虫标本的种类或身份就可满足其研究的需求,另一方面即使昆虫生态学工作者对分类学有足够的知识和兴趣,面对大量的昆虫生态学研究中的标本,也会使其疲于标本的识别与鉴定,而 DNA 条形码库为解决这个问题提供了一个不错的解决方案。

那么其要做的仅是测出所需研究昆虫样本的条形码基因,目前是COⅠ,与建立的 DNA 条形码库进行比较,未知标本的 DNA 序列还需和数据库中 DNA 序列进行一次比对,然后计算和每一条 DNA 序列的遗传距离,与其遗传距离最小的序列所对应的物种即为潜在的识别成功的物种(简单距离方法),一个统计学识别阈值可以被计算出来用于进一步的确认(BCM 和基于模糊成员关系的方法)。

2 应用分析

2.1 数据格式

现以松毛虫属 7 个近缘种的COⅠ数据为例(Dai *et al.*,2012),运用 BCM 方法进行条形码分析查询举例分析。输入数据为常用的. fasta/. fas 格式的文件。例如,

>COⅠ A111, Dendrolimus_punctatus

GGATTTGAGCTGGATAGTGGGAACTTCATTA-AGATTACTAA TTCGTGCTGAACTAGGTACTCCTG-GATCTTTAATTGGAGATGATCAAATTTATAATAC-TATTGTAACAGCTCATGCTTTCATTATAATTTTTT-TATAGTAATACCAATTATAATTGGGGGATTTGGT-AATTGATTAGTACCTTTAATATTAGGGGCCCCTG-ATATAGCATTCCCACGAATAAATAATATAAGATT-TTGATTATTACCACCCTCTCTTACCTTATTAATTT-CAAGAAGAATTGTAGAAAATGGAGCTGGAACTG-GATGAACTGTCTACCCCCCTTTATCATCTAATAT-

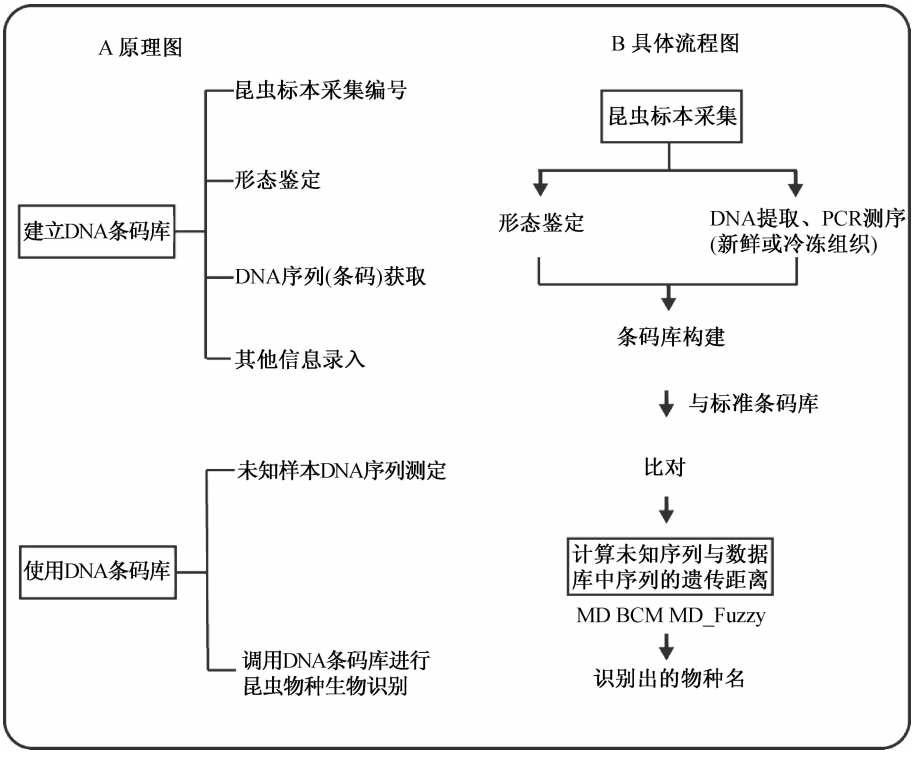

图 1　基于距离的 DNA 条形码方法原理图

Fig. 1　Distance-based DNA barcoding methods

TGCTCACGGGGGAAGATCCGTTGATTTAGCTATT-
TTCTCCCTACATTTAGCTGGAATTTCATCTATTTT-
AGGGGCAATTAATTTTATTACAACAATTATTAAT-
ATACGATTAAATAATATATCATTCGATCAAATAC-
CTTTATTTGTTTGAGCTGTAGGAATTACAGCATT-
TTTATTATTATTATCATTACCAGTTCTTGCTGGAG-
CAATTACTATATTATTAACTGATCGAAATTTAAAT-
ACATCATTTTTTGACCCTGCTGGAGGAGGGGATC-
CTATTTTATATCAACACTTATTTTGATTTTTTGGTC-
AC

2.2　运行程序进行初始设定和初步统计分析

运行程序 TaxonDNA/SpeciesIdentifier1. 7. 7-dev3(Meier *et al.*,2006)或从首都师范大学遗传多样性与进化课题组网站下载 http://life. cnu. edu. cn/shizishow. php? idh = 493),该程序用 Java 语言写成,在运行该程序前需安装 Java 运行环境。双击打开程序,如图 2(A)所示。在打开输入文件前,需进行初始设定,例如设定进化模型为 K2P,软件没有提供太多的进化模型可供选择,另一个选项是未校正的 P 距离,通常选用 K2P 模型即可。初步统计分析主要用于计算出适用于该数据集的遗传距离分布及阈值,用于后续计算。点击"Pairwise Summary"→点击"Calculate now",输出如程序右下角框体内所示(图 2:B)。

2.3　进行单个查询或是批量查询

进行单个序列查询,可点击"Actions and Views"框下的"Query against sequences",然后将待查询的序列粘贴到空白框内,点击"Query",所有匹配的序列计算后在"Query"下方的输出框中显示,点击每一个匹配上的序列,具体的匹配信息显示在右侧,如图 2(C)所示。

如果要进行批量查询,则需点击"Actions and Views"框下的"Best Match/Best Close Match",再选中"Compute from Pairwise Summary",然后阈值框中出现"3. 4800%"的阈值被选用,如图 2(D)所示,在这一步,也可指定一个主观的阈值,如 3%,但这往往会受到批评,然后点击"Recalculate",则计算结果会显示在下方的空白框中,成功的识别会被标注成"Successful match",并且显示是否在计算的阈值范围内,如图 2(D)所示。

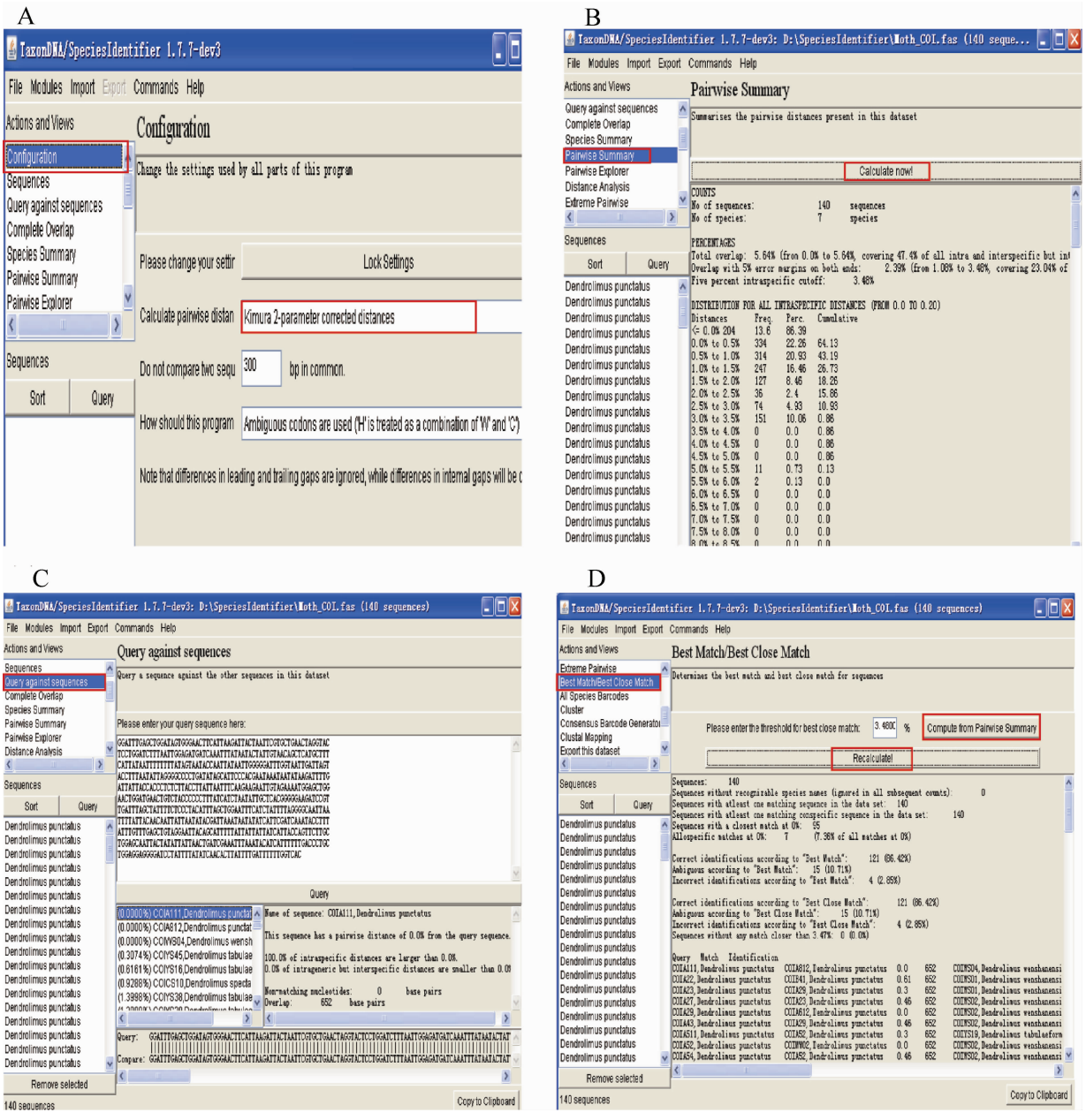

图 2　运用 BCM 方法进行物种识别的流程图

Fig. 2　The workflow of species identification with BCM

3　讨论

基于距离的方法具有概念和原理简单的优点,适用于大多数昆虫 DNA 条形码研究,然而为了保证距离方法推断的可靠性,对数据库中的每一物种尽可能进行较好的采样覆盖,即进行不同地理种群的采样,数据库中某些物种的缺失或是大量单一序列(singletons)的存在将会很大程度上削弱条形码识别的准确性与可靠性。另外在近缘种条形码识别中,单个条形码基因的识别往往缺乏鉴别力,在这种情况下,构建多基因的条形码识别系统将成为必要。

参考文献(References)

Dai QY, Gao Q, Wu CS, Chesters D, Zhu CD, Zhang AB, 2012. Phylogenetic reconstruction and DNA barcoding for closely related pine moth species (*Dendrolimus*) in China with multiple gene markers. *PLoS ONE*, 7(4): e32544.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR, 2003a. Biological identifications through DNA barcodes. *Proc. Biol. Sci.*, 270(1512):313 – 321.

Hebert PDN, Ratnasingham S, DeWaard JR, 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *P. Roy. Soc. B-Biol. Sci.*, 270(Suppl 1):S96 – S99.

Meier R, Shiyang K, VaidyaG, Ng PK, 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.*, 55(5):715 – 728.

Zhang AB, Muster C, Liang HB, Zhu CD, Crozier R, Wan P, Feng J, Ward RD, 2012a. A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Mol. Ecol.*, 21(8):1848 – 1863.

Zhang AB, Feng J, Ward RD, Wan P, Gao Q, Wu J, Zhao WZ, 2012b. A new method for species identification via protein-coding and non-coding dna barcodes by combining machine learning with bioinformatic methods. *PLoS ONE*, 7(2):e30986.

Zhang AB, Sikes DS, Muster C, Li SQ, 2008. Inferring species membership using DNA sequences with back-propagation neural networks. *Syst. Biol.*, 57(2):202 – 215.