生物多样性 BIODIVERSITY SCIENCE

2022, 30 (12): 22252, pages 1–14
doi: 10.17520/biods.2022252

•技术与方法• 土壤动物多样性: 物种与群落研究专辑

土壤动物的分子分类预测策略评估

徐聪[1], 张飞宇[1], 俞道远[2], 孙新[3], 张峰[1*]

1. 南京农业大学植物保护学院, 南京 210095; 2. 南京农业大学资源与环境科学学院, 南京 210095; 3. 中国科学院城市环境研究所, 福建厦门 130102

摘要: 土壤动物类群包含庞大的生物多样性, 由于传统的形态学鉴定技术很难满足该类群多样性调查和监测的巨大需求, 基于DNA等遗传物质的分子层面的鉴定技术(分子分类预测)逐渐登上舞台。然而, 分子分类预测能否在参考分子序列严重匮乏的土壤动物分类研究中实现有效鉴定、如何利用分子分类预测更为准确高效地获取土壤动物的分类信息, 是当下分子分类预测在土壤动物应用中的两大难题。为探究这两大难题, 本文基于宏条形码技术, 对5款常用的分子分类预测软件(VSEARCH、HS-BLASTN、EPA-NG、RAPPAS和APPLES; 前两款基于相似度算法, 其余基于系统发育位置算法)进行了准确性(科和属阶元)、运行速度和内存占用等性能的比较和评估。其中, 预测准确性的评估基于4类土壤动物(弹尾纲, 蜱螨亚纲, 环带纲和色矛纲)和3种分子标记(COI、16S和18S)展开。结果表明: EPA-NG在大部分场合下准确性最高, 尤其是在使用COI标记时, 准确性远高于其他工具。VSEARCH和HS-BLASTN准确性也较高, 基于16S和18S标记时, 它们的准确性和EPA-NG相当。此外, VSEARCH在所有软件中运行速度最快且内存占用最小, 这使得它在16S和18S的应用中比EPA-NG更具竞争力。RAPPAS和APPLES具有较低的假阳性, 但假阴性很高, 相对保守的算法使得它们无法将一些物种鉴定到低阶元。总体来说, 即使是在参考数据库缺少目标物种且小部分物种在分类上存在界定争议的前提下, 5款分子分类预测软件都能极为准确地将土壤动物预测至科级阶元, 因此分子分类预测在土壤动物应用中前景远大。COI标记在土壤动物科、属和种阶元上的覆盖度最广且能有效实现分子鉴定, 在目前最适合作为土壤动物尤其是土壤节肢动物的分子标记。在应用COI标记且参考数据库规模不大时, EPA-NG是分子分类预测的最佳选择; 而在应用16S、18S标记或参考数据库规模较大时, 更推荐使用VSEARCH。

关键词: 分子分类预测; 土壤动物; 生物信息学软件; 物种鉴定; 生物多样性

徐聪, 张飞宇, 俞道远, 孙新, 张峰 (2022) 土壤动物的分子分类预测策略评估. 生物多样性, 30, 22252. doi: 10.17520/biods.2022252.
Xu C, Zhang FY, Yu DY, Sun X, Zhang F (2022) Performance evaluation of molecular taxonomy assignment tools for soil invertebrates. Biodiversity Science, 30, 22252. doi: 10.17520/biods.2022252.

Performance evaluation of molecular taxonomy assignment tools for soil invertebrates

Cong Xu[1], Feiyu Zhang[1], Daoyuan Yu[2], Xin Sun[3], Feng Zhang[1*]

1 *College of Plant Protection, Nanjing Agricultural University, Nanjing 210095*
2 *College of Resources and Environmental Science, Nanjing Agricultural University, Nanjing 210095*
3 *Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, Fujian 130102*

ABSTRACT

Aims: Soil invertebrate communities are of extremely high diversity but still poorly studied in DNA-based diversity assessments. Since traditional morphological identifications have trouble in completing thousands of taxonomy assignments accurately with limited time, more and more biodiversity surveys turn to molecular taxonomy assignments. To promote biodiversity surveys on soil invertebrates, we made a comprehensive comparison for five popular taxonomy assignment tools (VSEARCH, HS-BLASTN, EPA-NG, RAPPAS and APPLES) targeting on different molecular markers (COI, 16S and 18S). Four soil invertebrate groups (Collembola, Acari, Clitellata and Chromadorea) were selected in the comparison representing three representative phyla of varied body-sizes.

收稿日期: 2022-05-09; 接受日期: 2022-08-18
基金项目: 国家科技基础资源调查专项(2018FY100303)和国家自然科学基金(31970434; 32270470)
* 通讯作者 Author for correspondence. E-mail: fzhang@njau.edu.cn

Methods: The databases of four soil invertebrate groups using three molecular markers were built with a filtering step. The commands of five taxonomy assignment tools were integrated into a script which would finally output the taxonomic information of query sequences. All of assignment accuracy, running speed and memory usage of five tools were estimated and compared.
Results: Our results indicated that EPA-NG performed best in accuracy for most cases, especially for COI. VSEARCH and HS-BLASTN remained high accuracy and showed similar accuracy performance when utilizing 16S and 18S markers. Moreover, shorter running time and lower memory usage made VSEARCH more popular applying in 16S and 18S than EPA-NG. RAPPAS and APPLES showed unstable performances in accuracy and were often too conservative to identify some species at generic or familial levels.
Conclusion: This study concluded that molecular taxonomy assignment could accomplish identifications of soil invertebrates in an accurate and efficient manner. COI marker is the most recommended marker applied in molecular taxonomy assignment for soil invertebrates because of its abundant repositories of reference sequences reflected in all of species, genus and family levels. When COI is utilized as marker, EPA-NG is the most recommended tool unless the reference database is too large. When 16S or 18S is utilized as marker, VSEARCH is most highly recommended.
Key words: taxonomy assignment; soil invertebrate; bioinformatics tool; identification; biodiversity

土壤生态系统高度复杂, 具有庞大的生物多样性(Decaëns, 2010)。以往, 专家和学者们对生物多样性的研究主要集中在水生和陆生(地上部分)生物; 近年来, 有关土壤生物的多样性研究才越来越受关注和重视(Bardgett & van der Putten, 2014; Phillips et al, 2020; Thakur et al, 2020)。土壤动物是土壤生态系统的重要组成部分, 在土壤的形成、改善土壤环境、促进土壤物质循环和能量流动等方面扮演着重要的角色(张志丹等, 2012; 战丽莉, 2013), 是陆地生态系统中种类和数量仅次于土壤微生物的类群(潘开文等, 2016)。不同土壤动物之间体型大小有很大的差异, 根据体长可分为3类: 大型(体长大于2 mm)、中型(体长在0.2 mm和2 mm之间)和小型(体长小于0.2 mm)土壤动物(Lavelle et al, 2006)。

传统的生物多样性调查和监测依赖于形态学方法进行分类和鉴定。随着生物调查和监测的规模日益增大, 形态学方法已经很难满足大批量的分类和鉴定需求。此外, 形态学鉴定需要丰富的类群分类知识, 尤其是在面对不完整或破损的样本时颇具困难(Jackson et al, 2014; Gueuning et al, 2019; van der Heyde et al, 2020)。

近年来兴起的高通量测序技术(high throughput sequencing, HTS)为多样性调查和监测提供了可靠、高效的鉴定(分类预测)新思路(Taberlet et al, 2012)。随着分子测序技术的发展, DNA宏条形码(metabarcoding)技术开始广泛应用于多样性调查和监测的分类预测环节(Bista et al, 2018; Arribas et al, 2021)。宏条形码技术通过对从环境样品(包括水、土壤、空气等)或生物混合样品中收集的混合DNA进行靶标标记(常称为“条形码”)扩增, 并利用高通量测序技术得到大量可操作分类单元(operational taxonomic units, OTUs), 最后将得到的OTUs与分子数据库中的参照序列进行比对以实现分子鉴定(Ji et al, 2013; Bohmann et al, 2014)。与形态学方法相对应, 像DNA宏条形码技术这样, 通过借助于测序技术得到生物DNA等遗传物质来实现其分子层面的分类鉴定的预测方法称为分子分类预测。

宏条形码技术最初应用于环境微生物, 环境中很多微生物不能直接分离培养, 只能通过混合样品来获取遗传信息(刘莉扬等, 2013)。近年来, 宏条形码技术开始扩展到动物研究领域, 例如鱼类(Stat et al, 2017; Sales et al, 2021)和两栖动物(Bálint et al, 2018)等。然而, 它在土壤动物中的应用还未得到普及, 该类群的条形码数据库也极度匮乏。尽管如此, 土壤生物多样性研究者们不断做出尝试以推进宏条形码技术在土壤动物中的应用。Oliverio等(2018)曾尝试使用宏条形码技术在目及科阶元对土壤动物进行分类预测, 结果表明分子分类预测技术能准确将土壤动物鉴定到高阶元; Arribas等(2021)将单倍型(amplicon sequence variants, ASVs)水平的宏条形码应用于土壤动物, 结果与基于阈值的OTU多样性趋势一致。然而, 基于宏条形码技术的分子分类预测对土壤动物在低阶元水平(科以下)的表现依旧缺乏定量研究。

准确而高效的分子分类预测对生物多样性调查和监测至关重要(Bazinet & Cummings, 2012), 分

子数据库的完备性和分子分类预测软件的选择决定了分类预测的准确性和效率。分子数据库的完备程度严重受制于分类工作者的进展等限制(时雷雷和傅声雷, 2014), 但分子分类预测工具和算法的迅猛发展使得分子分类在数据库受限的情况下仍有进一步提升的空间。近年来出现了各类分子分类预测软件, 主要基于两种算法。一种是传统的基于相似度的算法, 在覆盖率足够高的前提下, 通过将目标序列与数据库的序列进行比对, 得到与目标序列相似度最高的一条或多条参考序列的分类信息作为目标序列的分类预测参考。我们常用的BLAST (Altschul et al, 1990)就是基于这种算法实现分类预测的。另一种算法基于系统发育位置(phylogenetic placement, PP), 将目标序列放置到参考序列构建的系统发育树的各个分支上并计算目标序列在各个分支上的可能性, 最终计算出概率最高的分支从而预测出目标序列的分类信息(Berger & Stamatakis, 2011), 如EPA (RAxML) (Berger et al, 2011)等。

以往的比较和评估分子分类预测软件的研究主要基于相对完整的参考数据库(Brandt et al, 2021; Hleap et al, 2021; Mathon et al, 2021), 但不幸的是, 高度复杂多样的土壤动物的条形码数据库极度不完整, 这导致了在现实研究中将混合样品或环境样品中的土壤动物预测到物种水平十分困难。尽管如此, 我们依旧可以尝试在科级水平甚至属级水平进行分类预测。基于宏条形码技术的分子分类预测技术在动物应用中主要使用3种条形码: COI、16S和18S。其中, COI (线粒体细胞色素c氧化酶亚基I基因, cytochrome c oxidase I)的使用最为广泛, 因为动物的COI参考序列较其他条形码标记拥有更高的物种覆盖度(Hebert et al, 2003; Braukmann et al, 2019)。然而, Deagle等(2014)认为, COI的扩增引物保守性过低, 使用它来恢复目标类群可能会影响宏条形码分类预测技术的准确性。因此, 部分研究开始尝试用更为保守的16S (Elbrecht et al, 2016)和18S (Yang, 2013)作为分子标记。值得关注的是, 在Yang (2013)的研究中, 18S扩增出了比COI更为广泛的土壤动物类群。因此, 本研究同时采用了这3种条形码以表征不同程度的保守性和分子标记的差异。

本研究把多个分子分类预测软件的代码集成于一个可以统一执行的脚本, 可以高效快速地输出所有目标序列的各阶元预测信息。我们以准确性(采用*sensitivity*和*F1-score*两个指标; Gardner et al, 2019; Hleap et al, 2021)为主, 兼顾运行速度和内存资源占用, 对各个软件的性能进行了综合比较和评估。我们同时选取了4类代表性土壤动物: 弹尾纲(Collembola, 即跳虫)、蜱螨亚纲(Acari, 即蜱虫和螨虫)、环带纲(Clitellata, 即蚯蚓和蛭等)和色矛纲(Chromadorea, 即线虫)以表征不同类别、生活习惯和体型的土壤动物, 并分别构建了这些类群的COI、16S和18S数据库。软件准确性(科、属阶元水平)的比较和评估基于这些参考数据库展开。

1 材料与方法

1.1 数据库构建

本研究选取了弹尾纲、蜱螨亚纲、环带纲和色矛纲等4个类群作为土壤动物代表类群。这4个类群包含了节肢动物、环节动物和线虫动物, 囊括了不同的体型大小和生活习性, 且在土壤动物研究中比较热门, 具有代表性。

我们为4类土壤动物代表类群构建了各自的参考数据库, 包括COI (标准的658 bp片段)、16S (V4–V5片段, 约400 bp)和18S (V4片段, 约400 bp)的数据库。为了评估各个分子分类预测软件在更复杂的群落中的表现, 我们把4个类群的COI数据库组合并构建了一个额外的COI多类群数据库。由于分子分类预测的运行时间和计算资源损耗与数据库的大小显著相关, 因此在数据库中每个物种的序列只保留1条, 除非存在同种相似度差异较大(10%以上)的序列。同时, 我们会将每一条序列与同一类群中的其他序列进行比对, 相似度普遍较低的序列会被标记为可疑序列并在系统发育树上进一步观察。所有的序列均来自GenBank (NCBI)数据库和BOLDSystems (http://www.boldsystems.org; Ratnasingham & Hebert, 2007)数据库。我们通过自制脚本setup_filter.py来过滤高相似序列和可疑序列并在最终将每条参考序列和其分类信息绑定。这个脚本需要一个分类信息文件, NCBI的序列分类信息文件可通过taxonkit (Shen & Ren, 2021)获取, BOLDSystems的可通过treatBOLD.py获取。数据库构建所用到的脚本可在https://github.com/chachery/Perf_comp/的filter子文件夹中获取。

数据库中序列的分类信息从NCBI和BOLDSystems数据库中提取, 并经过目、科、属层级的系统发育树和相似度验证。尽管无法保证分类信息的完全正确, 但本研究通过多种办法尽量确保分类信息的准确性。分类信息的可能错误主要来源于3种情况: (1) NCBI和BOLDSystems间同一类群分类信息的不一致。这种情况较少, 本研究统一根据文献采用最新的分类信息。例如Lumbricidae, 我们依据Chelkha等(2020)将其归于Haplotaxida。(2)两条或多条序列一致(或高度相似)但分类信息不一致。这种情况往往是物种分类信息存在变动或者其中一条或多条序列分类信息有误导致。如果是物种分类信息存在变动, 则采用最新的分类信息。如果物种分类信息不存在变动(即其中一条或多条序列可能存在错误), 则根据系统发育树和相似度综合判断, 如依旧难以确认, 则根据序列在源数据库中的丰度来判断。序列分类信息不一致主要集中在物种水平, 而本研究的目标阶元是科和属, 如果仅是物种水平的分类信息错误极少影响本研究的预测准确性评估。(3)单条序列的分类信息错误。这种错误最难避免, 本研究主要通过两种手段来减少这种错误。其一, 将每条序列与其所属的属、科、目阶元的其他序列各进行一次系统发育树和相似度验证。其二, 每个物种只保留丰度最高的一条序列, 即这条序列是该物种中最可靠的序列。

1.2 分子分类预测软件

我们选取了5款主流的分子分类预测软件。这些软件具有较高的引用率, 拥有详细的使用教程和手册, 而且对用户免费开放。其中两款基于相似度算法: HS-BLASTN v0.0.5 + (Chen et al, 2015)(在算法逻辑与BLAST相同的基础上优化了运行速度)和VSEARCH v2.13.6 (Rognes et al, 2016); 另外3款基于系统发育位置PP算法: EPA-NG v0.3.7 (Barbera et al, 2019), RAPPAS v1.21 (Linard et al, 2019)和APPLES v2.0.0 (Balaban et al, 2020)。其中, 基于系统发育位置算法的系统发育树由FastTree2 v2.1.10 (Price et al, 2010)构建, 并在下游分析中借助于GAPPA v0.7.1实现预测结果的可视化(Czech et al, 2020)。

1.3 准确性的比较和评估

本研究所涉及的准确性的比较和评估基于现有的分类系统与系统发育研究而获得, 小部分物种中存在争议的科、属等阶元的确立综合了权威性杂志的分类划分、系统发育树判定和相似度判定。

我们把5款分子分类预测软件的运行代码嵌入了自制脚本“classify.sh”中, 并为该脚本设置了一些参数供使用者选择, 其中, 序列对齐参数借助于软件PAPARA v2.5 (Berger & Stamatakis, 2011)。脚本会在最终根据所选择的分类预测软件输出所有目标序列的各阶元预测信息以及可信值(如相似度)。我们基于1.1中得到的共计13个数据库, 对软件的准确性进行比较和评估。对于每个数据库, 我们进行100次重复, 每次从数据库中随机抽取100条序列(如果数据库序列数较少就适当减少)作为目标序列, 剩下的序列作为临时参考数据库, 然后用各个分类预测软件凭借“临时参考数据库”去预测“目标序列”的阶元信息(本研究以科和属为主), 并得到各个软件的*TP* (真阳性, 指同时存在于目标序列和预测结果中), *FP* (假阳性, 指在目标序列中并不存在, 却在预测结果中存在), *FN* (假阴性, 指在目标序列中存在, 却在预测结果中消失)。根据*TP*、*FP*和*FN*可计算得出*TPR* (true positive rate, 真阳率)和*PPV* (positive predictive value, 阳性预测值):

$$TPR = TP / (TP + FN) \quad (1)$$

$$PPV = TP / (TP + FP) \quad (2)$$

最终, 计算得到各个软件的两个准确性指标*sensitivity* (敏感度)和*F1-score* (F1分数) (Gardner et al, 2019; Hleap et al, 2021):

$$sensitivity = TPR \quad (3)$$

$$F1\text{-}score = 2 * TPR * PPV / (TPR + PPV) \quad (4)$$

其中, *sensitivity*指标能体现分类预测软件在混合样品或模拟社群中实现准确预测的概率, 而*F1-score*指标弥补了其不能凸显结果中错误预测分布情况的缺陷。

在软件的准确性比较和评估阶段, 我们用到了“doPick.sh”和“doCheck.sh”两个脚本。前一个脚本用于目标序列的抽取, 临时参考数据库的构建以及其他参考文件的同步; 后一个脚本调用了“classify.sh”, 用来实现所有预测软件的分类预测并同时得到它们的*TP*、*FP*和*FN*。软件准确性比较和评估所用到的脚本可在https://github.com/chachery/Perf_comp/

中获取, 相关数据和结果可在https://figshare.com/articles/dataset/data_and_results/19388075中获取。

1.4 运行速度和内存占用的比较

软件运行速度和内存占用的比较基于两种不同规模和复杂度的参考数据库展开。一个为序列数大于1,000的弹尾纲COI数据库, 另一个为序列数大于5,000的COI多类群数据库。我们用NCBI上下载的7,326条序列作为目标序列进行测试。在运行速度的比较中, 我们把运行时间折算成相对运行速度(以HS-BLASTN在线程数为1时的速度为单位速度), 分别计算出各个软件在线程数为1、4、8、16和32时的相对运行速度。在内存占用的比较中, 我们记录每个软件(单线程)进行预测时的RSS (resident set size)。所有软件测试都在乌班图18.04.1服务器(128 Gb内存, 16核/32线程)中运行。

2 结果

2.1 数据库构建

我们为4类土壤动物代表类群分别构建了COI、16S和18S的参考数据库(经过重复和错误序列的过滤, 最终每个种只保留一条最可靠的序列)。其中, COI数据库的生物多样性最为丰富, 4个类群(按弹尾纲、蜱螨亚纲、环带纲、色矛纲的顺序排列, 下同)分别筛选得到1,211、1,675、1,297和939条序列(即物种数), 并将它们混合得到COI多类群数据库(5,122条序列)。16S数据库的物种、属和科阶元的丰富度都比较低, 除了环带纲(物种、属和科阶元数分别为972、214和28, 表1)。18S数据库的属和科阶元丰富度比较高, 但物种数(分别为163、635、342和1,042, 表1)没有COI数据库丰富。

我们以COI为例, 将过滤后的精简版参考数据库与未过滤前的原始数据库进行运行时间上的比较。我们以100条COI序列作为目标序列, 使用单个线程的VSEARCH进行分类预测。结果表明, 弹尾纲和蜱螨亚纲的精简版参考数据库节省了80倍以上的时间, 环带纲和色矛纲节省了20倍左右的时间(附录2)。

2.2 准确性的比较和评估

2.2.1 COI数据库

对于COI数据库, 当采用*sensitivity*指标时, EPA-NG的表现最为出色, 且优势极为明显, 无论是科阶元(0.96 ± 0.02、0.86 ± 0.03、0.93 ± 0.03和0.88 ± 0.03; 按弹尾纲、蜱螨亚纲、环带纲、色矛纲的顺序排列, 下同)还是属阶元(0.74 ± 0.04、0.73 ± 0.04、0.70 ± 0.05和0.71 ± 0.04) (图1a–d, 附录4)。其次依次是HS-BLASTN (科: 0.84 ± 0.03、0.73 ± 0.05、0.86 ± 0.03和0.77 ± 0.04; 属: 0.64 ± 0.05、0.62 ± 0.05、0.58 ± 0.05和0.60 ± 0.05)和VSEARCH (科: 0.77 ± 0.04、0.71 ± 0.05、0.84 ± 0.04和0.77 ± 0.03; 属: 0.59 ± 0.04、0.60 ± 0.05、0.56 ± 0.06和0.57 ± 0.04), 而RAPPAS和APPLES的*sensitivity*指标较低(图1a–d, 附录4)。

当采用*F1-score*指标时, EPA-NG依旧保持着显著优势(科: 0.96 ± 0.02、0.87 ± 0.03、0.93 ± 0.02和0.89 ± 0.03; 属: 0.75 ± 0.04、0.74 ± 0.04、0.72 ± 0.05

表1 各个类群COI、16S和18S参考数据库中的物种、属和科阶元的数目
Table 1 The biodiversity showed in databases of different groups for COI, 16S and 18S

类群 Taxa	分子标记 Markers	物种数目 Species number	属数目 Genus number	科数目 Family number
弹尾纲 Collembola	COI	1,211	157	22
	16S	387	81	18
	18S	163	79	19
蜱螨亚纲 Acari	COI	1,675	460	190
	16S	456	76	28
	18S	635	459	220
环带纲 Clitellata	COI	1,297	255	39
	16S	972	214	28
	18S	342	203	42
色矛纲 Chromadorea	COI	939	249	86
	16S	170	64	28
	18S	1,042	430	135
4个类群合并 Merged	COI	5,122	1,121	337

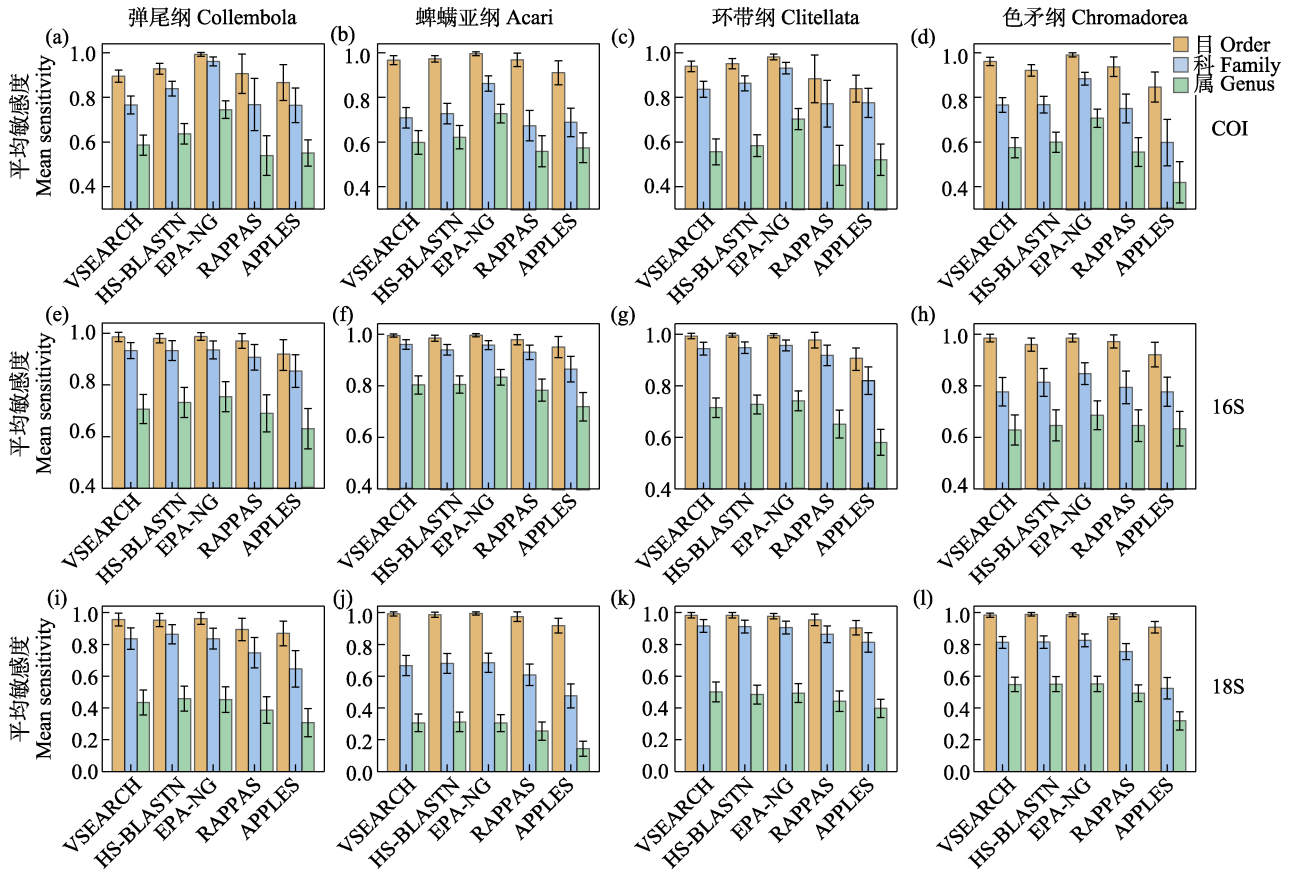

图1　5款分类预测软件在4个类群和3种分子标记应用中的平均敏感度

Fig. 1　Mean sensitivity of five taxonomic assignment tools among four groups and three markers

和0.73 ± 0.04, 图2a–d, 附录4)。APPLES (科: 0.85 ± 0.05、0.78 ± 0.05、0.86 ± 0.04和0.73 ± 0.09; 属: 0.67 ± 0.05、0.68 ± 0.06、0.64 ± 0.06和0.56 ± 0.09, 图2a–d, 附录4)在该指标下的表现超越了HS-BLASTN和VSEARCH, 尽管在预测色矛纲时的表现依旧不够亮眼。

2.2.2　16S数据库

对于16S数据库, 当采用*sensitivity*指标时, 尽管EPA-NG (科: 0.94 ± 0.03、0.96 ± 0.02、0.96 ± 0.02和0.85 ± 0.04; 属: 0.76 ± 0.06、0.83 ± 0.03、0.74 ± 0.04和0.69 ± 0.06, 图1e–h, 附录4)依旧位居首位, 但优势远不及应用COI数据库时。HS-BLASTN (0.93 ± 0.04、0.94 ± 0.02、0.95 ± 0.02和0.81 ± 0.05; 属: 0.73 ± 0.06、0.80 ± 0.03、0.73 ± 0.04和0.65 ± 0.06)和VSEARCH (0.93 ± 0.03、0.96 ± 0.02、0.94 ± 0.02和0.78 ± 0.06; 属: 0.71 ± 0.06、0.80 ± 0.04、0.72 ± 0.04和0.63 ± 0.06)紧随其后(图1e–h, 附录4)。另外, RAPPAS在应用16S时*sensitivity*指标比APPLES表现更佳(图1e–h)。

当采用*F1-score*指标时, 在科阶元上, VSEARCH (0.96 ± 0.02、0.98 ± 0.01、0.95 ± 0.02和0.85 ± 0.04), EPA-NG (0.94 ± 0.03、0.96 ± 0.02、0.96 ± 0.02和0.86 ± 0.04)和HS-BLASTN (0.95 ± 0.03、0.96 ± 0.02、0.95 ± 0.02和0.85 ± 0.05)的表现都很出色, 且十分接近; 在属阶元上, EPA-NG (0.77 ± 0.06、0.84 ± 0.03、0.76 ± 0.04和0.71 ± 0.05)以轻微的优势位居第一, 其次是VSEARCH (0.77 ± 0.05、0.84 ± 0.03、0.72 ± 0.04和0.70 ± 0.05)和HS-BLASTN (0.75 ± 0.05、0.83 ± 0.03、0.73 ± 0.04和0.68 ± 0.06) (图2e–h, 附录4)。

2.2.3　18S数据库

对于18S数据库, 当采用*sensitivity*指标时, EPA-NG (科: 0.84 ± 0.06、0.69 ± 0.06、0.91 ± 0.04和0.83 ± 0.04; 属: 0.45 ± 0.08、0.30 ± 0.05、0.49 ± 0.06和0.55 ± 0.05), HS-BLASTN (科: 0.86 ± 0.06、0.68 ± 0.06、0.91 ± 0.04和0.81 ± 0.04; 属: 0.46 ± 0.08、0.31 ± 0.06、0.48 ± 0.06和0.55 ± 0.05)和

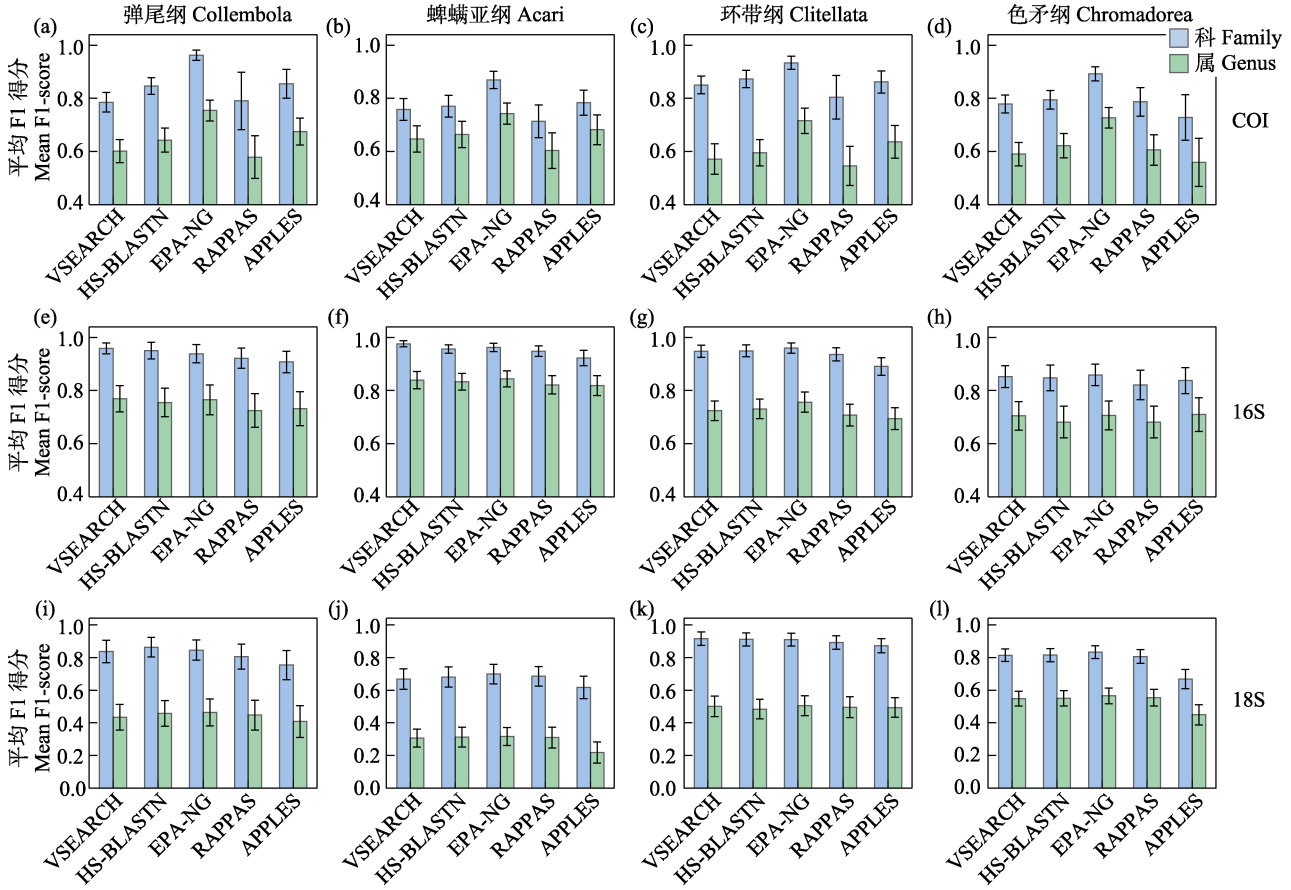

图2　5款分类预测软件在4个类群和3种条码应用中的平均F1分数

Fig. 2　Mean F1-score of five taxonomic assignment tools among four groups and three markers

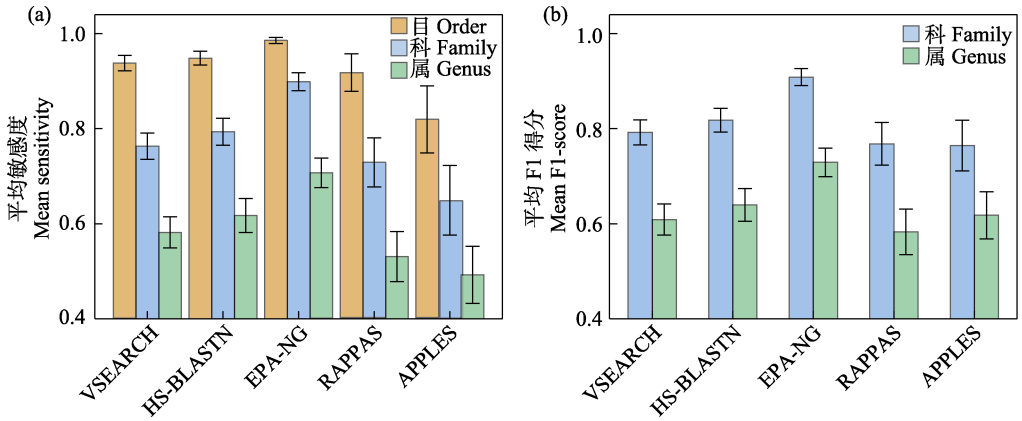

图3　5款分类预测软件在混合COI参考数据库应用中的平均敏感度(a)和平均F1分数(b)

Fig. 3　Mean sensitivity (a) and mean F1-score (b) of five taxonomic assignment tools with merged COI reference database

VSEARCH (科: 0.84 ± 0.07、0.67 ± 0.06、0.92 ± 0.04和0.81 ± 0.04; 属: 0.43 ± 0.08、0.31 ± 0.06、0.50 ± 0.06和0.55 ± 0.05)的表现十分接近, EPA-NG不再存在明显优势(图1i–l, 附录4)。同16S, RAPPAS在应用18S时在*sensitivity*指标上优于APPLES (图1i–l)。

当采用*F1-score*指标时, APPLES相对落后, 其他4款软件*F1-score*值都十分接近(图2i–l, 附录4)。

2.2.4　COI多类群数据库

对于4个类群的COI序列组成的额外数据库, 趋势和单个类群的COI数据库基本相同, 但

APPLES表现得更为乏力。当采用*sensitivity*指标时, 按排名顺序依次是EPA-NG (科: 0.90 ± 0.02; 属: 0.71 ± 0.03)、HS-BLASTN (科: 0.79 ± 0.03; 属: 0.62 ± 0.04)、VSEARCH (科: 0.76 ± 0.03; 属: 0.58 ± 0.03)、RAPPAS (科: 0.73 ± 0.05; 属: 0.53 ± 0.05)和APPLES (科: 0.65 ± 0.07; 属: 0.49 ± 0.06) (图3a, 附录4)。当采用*F1-score*指标时, 按排名顺序依次是EPA-NG (科: 0.91 ± 0.02; 属: 0.73 ± 0.03)、HS-BLASTN (科: 0.82 ± 0.03; 属: 0.64 ± 0.03)、VSEARCH (科: 0.79 ± 0.03; 属: 0.61 ± 0.03)、APPLES (科: 0.76 ± 0.05; 属: 0.62 ± 0.05)和RAPPAS (科: 0.77 ± 0.04; 属: 0.58 ± 0.05) (图3b, 附录4)。

2.3 内存占用和运行速度的比较

不同分子分类预测软件的计算资源(内存)占用和运行时间损耗存在较大差异。

我们把HS-BLASTN在单个线程时完成7,326条目标序列时的速度设为单位速度, 该过程损耗时间为91.6 s (137 s; 前后分别代表在应用大于1,000和大于5,000序列数的数据库时的时间损耗, 下同) (图4a, c, 附录3)。RAPPAS是唯一一款速度不随着线程数的增加而增加的软件, 无论使用多少线程, 它的运行时间都是172 s (400 s) (图4a, c, 附录3)。此外, RAPPAS也是唯一一款运行速度会随着鉴定次数缓慢加快的软件(附录1), 这可能运用了缓存或锻炼技术。VSEARCH是运行速度最快的软件, 它在单线程和24线程时运行时间分别为8.5 s和1 s (13.9 s和1.3 s) (图4a, c, 附录3)。其次是EPA-NG, 在单线程和24线程时运行时间分别为24 s和2.8 s (91.6 s和11.2 s) (图4a, c, 附录3)。然而, 在应用大于5,000序列数的数据库时, 线程数超过16的情况下, HS-BLASTN运行速度比EPA-NG更快(图4a, c)。APPLES是除RAPPAS之外运行速度最慢的, 在单线程和24线程时运行时间分别为719 s和50.9 s (3,000 s和208 s) (图4a, c, 附录3)。另外, 软件运行速度随线程数增加而加快的现象在线程数大于16时逐渐趋于不明显。

VSEARCH是占用内存最小的软件(11 Mb和21 Mb; 前后分别代表在单线程条件下应用大于1,000和大于5,000序列数的数据库时的内存占用大

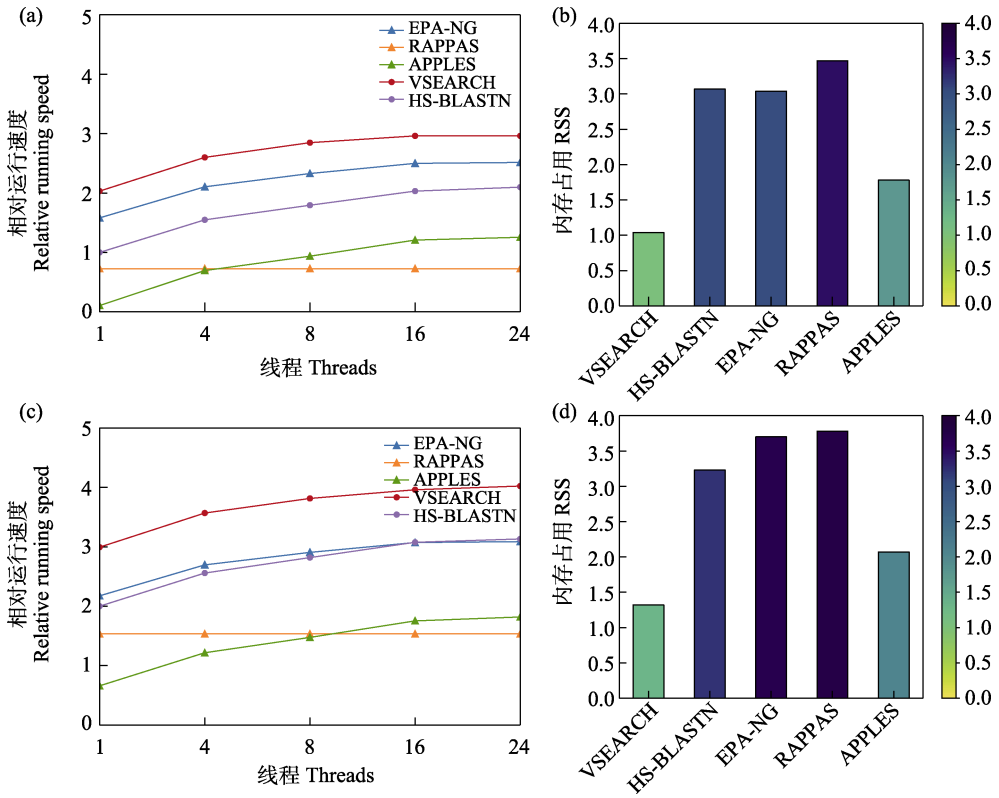

图4 5款分类预测软件在1,000量级(a, b)和5,000量级(c, d)参考数据库应用中的(相对)运行速度和内存占用

Fig. 4 Relative running speed and memory usage of five taxonomic assignment tools when applying reference databases with 1,000 sequences (a and b) and 5,000 sequences (c and d) respectively

小, 下同), 其次是APPLES (60 Mb和120 Mb), 尽管这两款软件的内存占用会随着线程数的增加而轻微增加(图4b, d, 附录3)。RAPPAS占用的内存最大(3 Gb和6 Gb), EPA-NG (1.1 Gb和5 Gb)和HS-BLASTN (1.2 Gb和1.7 Gb)的内存占用适中(图4b, d, 附录3)。

3 讨论

3.1 土壤动物的分子分类预测

土壤动物的构成复杂多样, 如何在分类学鉴定力不从心的境况中迅速、准确地探究多样性调查和监测所关注的物种丰富度和群落组成, 是土壤动物多样性研究一直以来亟需解决的一个巨大技术难题(傅声雷, 2018)。基于宏条形码技术的分子分类预测的出现和普及为土壤动物调查和监测中的快速鉴定带来了新的曙光(Kirse et al, 2021)。但是如何利用分子分类预测更为准确地获取土壤动物的分类信息, 依然是个大问题。本研究为解决上述问题, 采用3种分子标记对目前主流的5款分子分类预测软件在土壤动物应用中的分类预测效果进行了探究。

在本研究中, 我们模拟了在参考数据库缺少目标物种的场景中对土壤动物进行分子鉴定, 结果发现即使是在这种前提下, 5款分子分类预测软件都能将大部分土壤动物鉴定至科级阶元(附录4)。但由于参考数据库缺乏部分目标物种对应的属阶元, 且部分物种属阶元的确立因分类和鉴定工作的不完善(傅声雷, 2007)等原因导致至今仍存在争议, 成功恢复到属阶元的比例要一定程度低于恢复到科级阶元的比例。除本研究外, 郝金凤等(2017)、Hardulak等(2020)和Wang等(2018)分别将分子分类预测应用于蝗虫、甲虫和蚂蚁等并有效实现了这些类群的分子鉴定。总得来说, 分子分类预测能够准确高效地应用于参考数据库匮乏的土壤动物领域, 如果目标是参考数据库中缺乏的物种, 可以凭借近缘物种预测到属或者科阶元。

此外, 本研究所涉及的分子分类预测主要基于宏条形码技术。随着基因组学和测序技术的不断发展, 分子分类预测也开始往宏线粒体基因组、全基因组等方向发展。Vogler团队的研究表明不需要PCR扩增的宏线粒体基因组与宏条形码相比能提供更完整的序列、更精准的相对丰度(生物量)信息, 更高的物种检测率以及为生态和功能多样性的进化研究提供更精确的系统发育树指导, 但对应其成本也更高(Arribas et al, 2016; Gómez-Rodriguez et al, 2017)。

3.2 分子标记的选择

土壤动物的生物多样性极其丰富, 但其条形码数据库高度不完整, 缺乏大量物种(Decaëns, 2010)。总体来说, 对于物种水平, COI参考数据库丰富度最高; 而对于属和科水平, COI和18S参考数据库的丰富度都较16S更高。对于同一条形码, 各个类群的丰富度也存在差异。例如, 弹尾纲的18S数据库丰富度相对其他类群较低; 环带纲的16S数据库丰富度相对其他类群较高。

分子标记的选择很大程度上会影响基于宏条形码技术的调查和监测结果(Clarke et al, 2017)。在动物界, 凭借着拥有更多高质量的参考序列, 658 bp的COI片段被公认为标准DNA条形码(Hebert et al, 2003; Rodgers et al, 2017)。然而, Deagle等(2014)则认为COI片段过低的保守性会导致一些亲缘关系较近的物种间存在较低的相似性, 因此一些研究尝试使用更保守的16S或18S片段作为分子标记进行探究(Yang, 2013; Elbrecht et al, 2016)。在我们的研究中, 在应用16S时, 各个软件准确性层面的两个指标值都比较高, 这在一定程度上能反映16S在分类预测上的优势。然而, 在种、属和科3个水平上, 16S条形码的丰富度都十分匮乏(Braukmann et al, 2019)。

总的来说, 在土壤动物的分子分类预测中, 除少数特殊类群(如线虫等低等无脊椎动物, 它们的COI标记突变快、覆盖度不广且通用引物不容易扩增, 因此大部分研究使用18S作为分子标记; Ahmed et al, 2019)外, 优先推荐COI条形码作为分子标记。另外, Chesters和Zhu (2014)发现采用多种分子标记比仅用单个分子标记能区分更多的物种, 后续研究更证实了COI与其他分子标记的组合使用比仅使用COI时准确性更高(Chesters et al, 2015)。因此, 如果追求更高的准确性, 可以尝试COI与其他分子标记的组合。

3.3 准确性的比较和评估

不同软件间的准确性存在差异, 相同软件在不同宏条形码片段或不同类群的数据库中也有不同

的表现(Brandt et al, 2021)。在使用同一宏条形码片段时, 尽管5款所选的软件各自在不同类群中的表现存在差异, 但它们的排名趋势在不同类群中基本一致。各软件间的差异在不同条形码类型中有不同程度的体现。各软件间的准确性差异在COI条形码中最显著。在应用COI条形码时, EPA-NG在属和科水平上的准确性比其他软件有显著优势, 这种显著优势同时表现在*sensitivity*和*F1-score*两个指标上, 表明EPA-NG在应用COI条形码时同时具备较低的假阴性和假阳性。APPLES在*sensitivity*指标上不如HS-BLASTN和VSEARCH, 但在*F1-score*上表现比HS-BLASTN和VSEARCH更出色(除了色矛纲)。这体现了APPLES在应用COI条形码时拥有较低的假阳性(Balaban et al, 2020)和较高的假阴性, 这可能与它保守的算法有关, 不容易出现鉴定错误也不容易将序列鉴定到低阶元。另外, 在参考数据库加大的情况下, APPLES的准确性表现严重不稳定。在应用16S条形码时, 软件间的准确性差异远不及应用COI时。在使用*sensitivity*指标时, EPA-NG在属和科水平上的准确性表现都最出色, 尽管与其他软件的差距较应用COI时大大减小; 在使用*F1-score*指标时, EPA-NG、VSEARCH和HS-BLASTN都表现优异, 且差距极小。在应用18S条形码时, 各软件的准确性差异进一步缩小。EPA-NG、VSEARCH和HS-BLASTN的表现十分接近, 在使用*F1-score*指标时, RAPPAS的准确性也趋近于它们。

总的来说, EPA-NG的准确性表现在不同分子标记间最稳定, 始终保持着较高的准确性, 尤其是在使用COI条形码时, 表现出显著优势。同其他文章中软件比较和评估的结果一样(Brandt et al, 2021; Hleap et al, 2021), BLAST (本研究中使用HS-BLASTN代替, HS-BLASTN是快速版本的BLAST; Chen et al, 2015)在不同分子标记和类群中始终有较为稳定的出色表现(Bik et al, 2021; Hleap et al, 2021)。作为USEARCH的高质量开源替代, VSEARCH的准确性与BLAST十分接近, 尽管假阳性略高于BLAST (Edgar, 2010; Rognes et al, 2016)。APPLES的算法可能比较保守, 更适用于较高的分类水平(目及目以上) RAPPAS在准确性上表现并不突出, 其在免对齐上的优势在本研究中无法体现是非常可惜的。

3.4 其他性能的比较和评估

在运行速度上, 各个软件都能在数秒到数分钟内完成成千上万条序列的分类预测。尤其是VSEARCH、EPA-NG和HS-BLASTN, 在线程足够的情况下, 可以瞬间完成大批量鉴定。其中, VSEARCH的运行速度最快, 反映出其在算法上的优势(Rognes et al, 2016)。除RAPPAS之外, 其他软件的运行速度都随着线程的增加而加快, 但当线程数大于16时, 运行速度的增加并不显著。RAPPAS速度的优化可能通过缓存或者锻炼来实现, 它的运行速度会随着分类预测次数的增加而缓慢加快。随着参考数据库序列数增加, 分类预测的运行速度会随之减慢(Hleap et al, 2021), 此时基于相似度算法的软件比基于系统发育位置算法的软件运行速度损耗更小, 因此当参考数据库过大时, 基于系统发育位置算法的软件可能更加费时。

在内存占用上, VSEARCH的内存占用最小, 其次是APPLES, 它们的内存占用远小于其他3款软件, 尽管它们的内存占用会随着线程的增加而增加。EPA-NG随着参考数据库大小的增加内存占用也大大增加, 这可能会限制其在大型数据库应用中的发挥。除EPA-NG外, 其他软件在参考数据库大小增长至原来5倍左右的情况下, 内存占用增长在一倍以内, 这得益于算法在计算资源损耗上的不断优化。

总的来说, VSEARCH的运行速度最快, 内存占用最小。EPA-NG在参考数据库较小时运行时间和计算资源的损耗较少, 但当参考数据库较大时, 这些损耗会限制EPA-NG的发挥。HS-BLASTN在参考数据库较小时比EPA-NG损耗更多的运行时间和计算资源, 但当参考数据库较大时, 它比EPA-NG更节省运行时间和计算资源。APPLES内存占用较小, 但运行速度较慢。RAPPAS作为一款不需要对齐的基于系统发育位置算法的软件(Linard et al, 2019), 在不需要对齐的同时, 也付出了需要更多运行时间和内存占用的代价。

3.5 总结和展望

和大部分动物类群一样, 在4个土壤动物类群中, COI参考数据库的物种种类最繁多, 属和科水平的丰富度也很高, 这是很多动物领域的宏条形码研究选用COI作为分子标记的主要原因之一

表2　5款分类预测软件的推荐使用情况
Table 2　The recommendation on application of five taxonomic assignment tools

分类预测软件 Tools	推荐使用情况 Recommendation on application
EPA-NG	以COI作为分子标记且参考数据库不大的场合 COI is used as the marker and the reference database is small
VSEARCH	以16S或18S作为分子标记或者参考数据库较大的场合 16S/18S is used as the marker; the reference database includes thousands of sequences or more
HS-BLASTN	同VSEARCH, 但优先级不如VSEARCH Similar to VSEARCH
APPLES	仅预测较高阶元的场合 Predicting higher taxonomical hierarchy
RAPPAS	目标序列间长度差异较大的场合 When sequence lengths differ greatly

(Braukmann et al, 2019)。在应用COI条形码时, EPA-NG的准确性显著高于其他4款软件, 但随着参考数据库大小的增加, 运行时间损耗和内存占用增幅也比其他软件显著。因此, 在参考数据库不庞大的前提下(例如, 针对单个类群或类群较少时), EPA-NG是应用COI条形码时的最佳首选; 但当数据库足够庞大时(例如, 针对整个动物界或类群较多时), 运行时间和内存占用更节省的基于相似度算法的软件更具竞争优势(表2)。VSEARCH和HS-BLASTN准确性十分接近, 但是无论是在运行时间还是在内存占用上, VSEARCH的表现都更加出色(Rognes et al, 2016)。因此, 越来越多的应用宏条形码技术的研究报告使用VSEARCH来代替BLAST进行分子鉴定(Cavaliere et al, 2021; Lanzén et al, 2021; Torrell et al, 2021)。此外, VSEARCH还自带一些过滤、去噪和归一化等功能(Rognes et al, 2016), 所以本研究也同样推荐VSEARCH来取代BLAST。

单个条形码都有着它自身的局限性, 因此条形码间的组合使用以实现各自的优势互补或许会成为将来宏条形码技术的主流趋势(张宛宛等, 2017), 例如COI + 16S或COI + 18S等。现如今, 分子分类预测方法层出不尽, 除了主流的相似度算法和系统发育树放置算法, 其他算法也逐渐被开发出来, 例如基于机器学习的算法(Murali et al, 2018)等。将来的土壤动物分类预测可以不断去尝试这些新鲜的算法。另外, 相似度算法和系统发育位置算法以及其他算法都有自身的优劣, 将来或许可以尝试将两个或多个算法结合起来, 实现更精准更高效的分子分类预测, 例如EPA-NG和VSEARCH的组合等。

本文比较和评估了目前几款主流的分子分类预测软件在土壤动物应用中的性能, 虽然不能囊括所有的分类预测软件, 但希望能为今后基于宏条形码等技术的土壤动物调查和监测提供一些借鉴意义。

ORCID

俞道远 https://orcid.org/0000-0003-2984-0540
孙新 https://orcid.org/0000-0002-3988-7847
张峰 https://orcid.org/0000-0002-1371-266X

参考文献

Ahmed M, Back MA, Prior T, Karssen G, Lawson R, Adams I, Sapp M (2019) Metabarcoding of soil nematodes: The importance of taxonomic coverage and availability of reference sequences in choosing suitable marker(s). Metabarcoding and Metagenom, 3, e36408.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410.

Arribas P, Andújar C, Hopkins K, Shepherd M, Vogler AP (2016) Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. Methods in Ecology and Evolution, 7, 1071–1081.

Arribas P, Andújar C, Salces-Castellano A, Emerson BC, Vogler AP (2021) The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. Molecular Ecology, 30, 48–61.

Balaban M, Sarmashghi S, Mirarab S (2020) APPLES: Scalable distance-based phylogenetic placement with or without alignments. Systematic Biology, 69, 566–578.

Bálint M, Nowak C, Márton O, Pauls SU, Wittwer C, Aramayo JL, Schulze A, Chambert T, Cocchiararo B, Jansen M (2018) Accuracy, limitations and cost efficiency of eDNA-based community survey in tropical frogs. Molecular Ecology Resources, 18, 1415–1426.

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A (2019) EPA-ng: Massively parallel evolutionary placement of genetic sequences. Systematic Biology, 68, 365–369.

Bardgett RD, van der Putten WH (2014) Belowground biodiversity and ecosystem functioning. Nature, 515, 505–511.

Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs. BMC Bioinformatics, 13, 92.

Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Systematic Biology, 60, 291–302.

Berger SA, Stamatakis A (2011) Aligning short reads to reference alignments and trees. Bioinformatics, 27, 2068–2075.

Bik HM (2021) Just keep it simple? Benchmarking the accuracy of taxonomy assignment software in metabarcoding studies. Molecular Ecology Resources, 21, 2187–2189.

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu SL, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources, 18, 1020–1034.

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology and Evolution, 29, 358–367.

Brandt MI, Trouche B, Quintric L, Günther B, Wincker P, Poulain J, Arnaud-Haond S (2021) Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. Molecular Ecology Resources, 21, 1904–1921.

Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, de Waard JR, Sones JE, Zakharov EV, Hebert PDN (2019) Metabarcoding a diverse arthropod mock community. Molecular Ecology Resources, 19, 711–727.

Cavaliere M, Angeles IB, Montresor M, Bucci C, Brocani L, Balassi E, Margiotta F, Francescangeli F, Bouchet VMP, Pawlowski J, Frontalini F (2021) Assessing the ecological quality status of the highly polluted Bagnoli area (Tyrrhenian Sea, Italy) using foraminiferal eDNA metabarcoding. Science of the Total Environment, 790, 147871.

Chelkha M, Blanco-Pérez R, Bueno-Pallero FÁ, Amghar S, El Harti A, Campos-Herrera R (2020) Cutaneous excreta of the earthworm *Eisenia fetida* (Haplotaxida: Lumbricidae) might hinder the biological control performance of entomopathogenic nematodes. Soil Biology and Biochemistry, 141, 107691.

Chen Y, Ye WC, Zhang YD, Xu YS (2015) High speed BLASTN: An accelerated MegaBLAST search tool. Nucleic Acids Research, 43, 7762–7768.

Chesters D, Zheng WM, Zhu CD (2015) A DNA barcoding system integrating multigene sequence data. Methods in Ecology and Evolution, 6, 930–937.

Chesters D, Zhu CD (2014) A protocol for species delineation of public DNA databases, applied to the Insecta. Systematic Biology, 63, 712–725.

Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. Ecology and Evolution, 7, 873–883.

Czech L, Barbera P, Stamatakis A (2020) Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics, 36, 3263–3265.

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. Biology Letters, 10, 20140562.

Decaëns T (2010) Macroecological patterns in soil communities. Global Ecology and Biogeography, 19, 287–302.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460–2461.

Elbrecht V, Taberlet P, Dejean T, Valentini A, Usseglio-Polatera P, Beisel JN, Coissac E, Boyer F, Leese F (2016) Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. PeerJ, 4, e1966.

Fu SL (2007) A review and perspective on soil biodiversity research. Biodiversity Science, 15, 109–115. (in Chinese with English abstract) [傅声雷 (2007) 土壤生物多样性的研究概况与发展趋势. 生物多样性, 15, 109–115.]

Fu SL (2018) Strengthening the research on soil fauna diversity and their ecological functions using novel technology and field experimental facility. Biodiversity Science, 26, 1031–1033. (in Chinese) [傅声雷 (2018) 利用新方法和野外实验平台加强土壤动物多样性及其生态功能的研究. 生物多样性, 26, 1031–1033.]

Gardner PP, Watson RJ, Morgan XC, Draper JL, Finn RD, Morales SE, Stott MB (2019) Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. PeerJ, 7, e6160.

Gómez-Rodríguez C, Timmermans MJTN, Crampton-Platt A, Vogler AP (2017) Intraspecific genetic variation in complex assemblages from mitochondrial metagenomics: Comparison with DNA barcodes. Methods in Ecology and Evolution, 8, 248–256.

Gueuning M, Ganser D, Blaser S, Albrecht M, Knop E, Praz C, Frey JE (2019) Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. Molecular Ecology Resources, 19, 847–862.

Hao JF, Zhang XH, Wang YS, Liu JL, Zhi YC, Li XJ (2017) Diversity investigation and application of DNA barcoding of Acridoidea from Baiyangdian Wetland. Biodiversity Science, 25, 409–417. (in Chinese with English abstract) [郝金凤, 张晓红, 王昱淞, 刘金林, 智永超, 李新江 (2017) 白洋淀湿地蝗虫多样性调查及DNA条形码应用研究. 生物多样性, 25, 409–417.]

Hardulak LA, Morinière J, Hausmann A, Hendrich L, Schmidt S, Doczkal D, Müller J, Hebert PDN, Haszprunar G (2020) DNA metabarcoding for biodiversity monitoring in a National Park: Screening for invasive and pest species. Molecular Ecology Resources, 20, 1542–1557.

Hebert PDN, Ratnasingham S, de Waard JR (2003) Barcoding

animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society B: Biological Sciences, 270, S96–S99.

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources, 21, 2190–2203.

Jackson JK, Battle JM, White BP, Pilgrim EM, Stein ED, Miller PE, Sweeney BW (2014) Cryptic biodiversity in streams: A comparison of macroinvertebrate communities based on morphological and DNA barcode identifications. Freshwater Science, 33, 312–324.

Ji YQ, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang XY, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters, 16, 1245–1257.

Kirse A, Bourlat SJ, Langen K, Fonseca VG (2021) Unearthing the potential of soil eDNA metabarcoding—Towards best practice advice for invertebrate biodiversity assessment. Frontiers in Ecology and Evolution, 9, 630560.

Lanzén A, Dahlgren TG, Bagi A, Hestetun JT (2021) Benthic eDNA metabarcoding provides accurate assessments of impact from oil extraction, and ecological insights. Ecological Indicators, 130, 108064.

Lavelle P, Decaëns T, Aubert M, Barot S, Blouin M, Bureau F, Margerie P, Mora P, Rossi JP (2006) Soil invertebrates and ecosystem services. European Journal of Soil Biology, 42, S3–S15.

Linard B, Swenson K, Pardi F (2019) Rapid alignment-free phylogenetic identification of metagenomic sequences. Bioinformatics, 35, 3303–3312.

Liu LY, Cui HF, Tian G (2013) Application of high throughput sequencing in metagenomics. Chinese Medicinal Biotechnology, 8, 196–200. (in Chinese) [刘莉扬, 崔鸿飞, 田埂 (2013) 高通量测序技术在宏基因组学中的应用. 中国医药生物技术, 8, 196–200.]

Mathon L, Valentini A, Guérin PE, Normandeau E, Noel C, Lionnet C, Boulanger E, Thuiller W, Bernatchez L, Mouillot D, Dejean T, Manel S (2021) Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. Molecular Ecology Resources, 21, 2565–2579.

Murali A, Bhargava A, Wright ES (2018) IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. Microbiome, 6, 140.

Oliverio AM, Gan HJ, Wickings K, Fierer N (2018) A DNA metabarcoding approach to characterize soil arthropod communities. Soil Biology and Biochemistry, 125, 37–43.

Pan KW, Zhang L, Shao YH, Fu SL (2016) Thematic monitoring network of soil fauna diversity in China: Exploring the mystery of soils. Biodiversity Science, 24, 1234–1239. (in Chinese with English abstract) [潘开文, 张林, 邵元虎, 傅声雷 (2016) 中国土壤动物多样性监测: 探知土壤中的奥秘. 生物多样性, 24, 1234–1239.]

Phillips HRP, Heintz-Buschart A, Eisenhauer N (2020) Putting soil invertebrate diversity on the map. Molecular Ecology, 29, 655–657.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2—Approximately maximum-likelihood trees for large alignments. PLoS ONE, 5, e9490.

Ratnasingham S, Hebert PDN (2007) BOLD: The barcode of life data system (http://www.barcodinglife.org). Molecular Ecology Notes, 7, 355–364.

Rodgers TW, Xu CCY, Giacalone J, Kapheim KM, Saltonstall K, Vargas M, Yu DW, Somervuo P, McMillan WO, Jansen PA (2017) Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community. Molecular Ecology Resources, 17, e133–e145.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: A versatile open source tool for metagenomics. PeerJ, 4, e2584.

Sales NG, Wangensteen OS, Carvalho DC, Deiner K, Præbel K, Coscia I, McDevitt AD, Mariani S (2021) Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. Science of the Total Environment, 754, 142096.

Shen W, Ren H (2021) TaxonKit: A practical and efficient NCBI taxonomy toolkit. Journal of Genetics and Genomics, 48, 844–850.

Shi LL, Fu SL (2014) Review of soil biodiversity research: History, current status and future challenges. Chinese Science Bulletin, 59, 493–509. (in Chinese with English abstract) [时雷雷, 傅声雷 (2014) 土壤生物多样性研究: 历史、现状与挑战. 科学通报, 59, 493–509.]

Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M (2017) Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. Scientific Reports, 7, 12240.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology, 21, 2045–2050.

Thakur MP, Phillips HRP, Brose U, de Vries FT, Lavelle P, Loreau M, Mathieu J, Mulder C, van der Putten WH, Rillig MC, Wardle DA, Bach EM, Bartz MLC, Bennett JM, Briones MJI, Brown G, Decaëns T, Eisenhauer N, Ferlian O, Guerra CA, König-Ries B, Orgiazzi A, Ramirez KS, Russell DJ, Rutgers M, Wall DH, Cameron EK (2020) Towards an integrative understanding of soil biodiversity. Biological Reviews, 95, 350–364.

Torrell H, Cereto-Massagué A, Kazakova P, García L, Palacios

H, Canela N (2021) Multiomic approach to analyze infant gut microbiota: Experimental and analytical method optimization. Biomolecules, 11, 999.

van der Heyde M, Bunce M, Wardell-Johnson G, Fernandes K, White NE, Nevill P (2020) Testing multiple substrates for terrestrial biodiversity monitoring using environmental DNA metabarcoding. Molecular Ecology Resources, 20, 732–745.

Wang WY, Srivathsan A, Foo M, Yamane SK, Meier R (2018) Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. Molecular Ecology Resources, 18, 490–501.

Yang CX, Ji YQ, Wang XY, Yang CY, Yu DW (2013) Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. Science China: Life Sciences, 56, 73–81.

Zhan LL (2013) Diversity and Influencing Factor of Meso-soil Animal Under Farm Land of Black Soil. PhD dissertation, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun. (in Chinese with English abstract) [战丽莉 (2013) 农田黑土中小型土壤动物多样性特征及其影响因素. 博士学位论文, 中国科学院东北地理与农业生态研究所, 长春.]

Zhang WW, Xie YW, Yang JH, Yang YN, Li D, Zhang Y, Yu HX, Zhang XW (2017) Applications and prospects of metabarcoding in environmental monitoring of phytoplankton community. Asian Journal of Ecotoxicology, 12, 15–24. (in Chinese with English abstract) [张宛宛, 谢玉为, 杨江华, 杨雅楠, 李娣, 张咏, 于红霞, 张效伟 (2017) DNA宏条形码(metabarcoding)技术在浮游植物群落监测研究中的应用. 生态毒理学报, 12, 15–24.]

Zhang ZD, Dong WH, Wei J, Gai YH (2012) Research progresses of soil fauna. Chinese Agricultural Science Bulletin, 28, 242–246. (in Chinese with English abstract) [张志丹, 董炜华, 魏健, 盖玉红 (2012) 土壤动物学研究进展. 中国农学通报, 28, 242–246.]

(责任编委: 傅声雷　责任编辑: 李会丽)

附录　Supplementary Material

附录1　RAPPAS运行时间随数据库使用次数变化情况
Appendix 1　Variation of running time for RAPPAS during usage count of the reference database
https://www.biodiversity-science.net/fileup/PDF/2022252-1.pdf

附录2　过滤前后参考数据库分类预测耗时比较
Appendix 2　Running time comparison between the databases before and after filtering
https://www.biodiversity-science.net/fileup/PDF/2022252-2.pdf

附录3　5款分类预测软件在1,000量级和5,000量级参考数据库应用中的运行时间和内存占用
Appendix 3　Running time and memory usage of five taxonomic assignment tools when applying reference databases with 1,000 sequences and 5,000 sequences respectively
https://www.biodiversity-science.net/fileup/PDF/2022252-3.pdf

附录4　5款分类预测软件在各个类群和条码应用中的准确性具体表现
Appendix 4　Accuracy performance of five taxonomic assignment tools among different groups and different markers
https://www.biodiversity-science.net/fileup/PDF/2022252-4.pdf