生物多样性 2015, **23** (4): 550–555 doi: 10.17520/biods.2015120
Biodiversity Science http: //www.biodiversity-science.net

• 软件介绍 •

批量下载GenBank基因序列数据的新工具——NCBIminer

徐晓婷[1] 王志恒[1*] Dimitar Dimitrov[2] Carsten Rahbek[3,4]

1 (北京大学城市与环境学院生态学系, 北京大学地表过程分析与模拟教育部重点实验室, 北京 100871)
2 (Natural History Museum, University of Oslo, Oslo, Norway)
3 (Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark)
4 (Imperial College London, Grand Challenges in Ecosystems and the Environment Initiative, Silwood Park Campus, Berkshire, UK)

摘要: 核苷酸序列是生物体遗传信息的载体, 是现代生物学和生态学的基础数据。随着测序技术的进步, 大量核苷酸序列被提取并存储在公共数据平台中, 其中GenBank(http://www.ncbi.nlm.nih.gov/genbank/)是目前最大的核苷酸序列数据平台之一。截至2015年2月, 该平台收录核苷酸序列总数已超过1.8亿条、覆盖全球超过30万个物种。但如何从如此海量的数据中准确、快速查找并下载所需数据已成为限制基因数据广泛使用的障碍之一。为此, 我们开发了一款可高效、准确下载GenBank数据的生物信息学软件NCBIminer。NCBIminer可根据用户提供的核苷酸序列名称、数据类型、一或多条初始化参考序列, 查找并下载用户指定的多个物种或类群的特定基因序列数据。该软件下载地址为https://github.com/greengirl/NCBIminer/releases/, 可在Windows、Linux和MAC操作系统下免费使用; 同时, 其操作简单, 用户无需生物信息学背景。为方便该软件的使用, 本文将介绍该软件的工作流程与算法、安装及使用过程中的参数设置等。

关键词: GenBank, 生物信息学, 基因序列, 系统进化, DNA, 核苷酸序列

Using NCBIminer to search and download nucleotide sequences from GenBank

Xiaoting Xu[1], Zhiheng Wang[1*], Dimitar Dimitrov[2], Carsten Rahbek[3,4]

1 *Department of Ecology and Key Laboratory for Earth Surface Processes of the Ministry of Education, College of Urban and Environmental Sciences, Peking University, Beijing 100871*
2 *Natural History Museum, University of Oslo, Oslo, Norway*
3 *Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark*
4 *Imperial College London, Grand Challenges in Ecosystems and the Environment Initiative, Silwood Park Campus, Berkshire, UK*

Abstract: GenBank is the leading public genetic resources database and currently contains over 10^{12} base pairs from about 300,000 formally described species. It offers valuable resources for studies on the evolution of species, genes, and genomes. However, difficulties in GenBank data mining hinder the potential wide application of this tool for big data collection. To address this issue, we introduce new bioinformatics software —NCBIminer. NCBIminer is a freely available, cross-platform, and user-friendly software for mining nucleotide sequences from GenBank. The main purpose of NCBIminer is to download sequences for user required genes and taxonomic groups based on gene names, types, and one or several reference sequences. The program algorithms have been described elsewhere and here, we focus on introducing the details in the usage of the program including how to install, run, and set parameters.

Key words: GenBank, bioinformatics, gene, phylogenetic evolution, DNA, nucleotide sequences

收稿日期: 2015-05-07; 接受日期: 2015-07-09
基金项目: 国家自然科学基金(31470564, 31400467, 31321061)和中国博士后科学基金(2014M550555)
* 通讯作者 Author for correspondence. E-mail: zhiheng.wang@pku.edu.cn

核苷酸序列(nucleotide sequence)是生物体遗传信息的载体, 是现代生物学和生态学研究的基础数据(任保青和陈之端, 2010; 陈之端和李德铢, 2013; 鲁丽敏等, 2014)。随着DNA提取和测序技术的进步, 大量核苷酸序列被提取并存储在公共数据平台中。快速积累的基因数据极大地促进了进化与分子生物学、分子生态学、宏观生态学等相关学科的发展(Driskell *et al.*, 2004; Qiu *et al.*, 2012; Yang *et al.*, 2012; Holt *et al.*, 2013; Xu *et al.*, 2013; Zanne *et al.*, 2013), 使当前的生物学与生态学研究进入了基因时代(Li, 2013)。

GenBank (http://www.ncbi.nlm.nih.gov/genbank/)是目前最大的核苷酸序列数据平台之一, 收录了海量的基因序列数据。截至2015年2月, GenBank收录的基因序列总数已超过1.8亿条, 序列长度总和已超过10^{12}个碱基对, 覆盖全球30多万个物种。随着DNA barcoding等项目的开展, 我国植物的基因数据大量积累(Li *et al.*, 2011; 裴男才, 2015)。利用NCBIminer检索GenBank (http://www.ncbi.nlm.nih.gov/genbank/)显示, 现有基因数据已涵盖我国木本植物约1,090属(占全部1,175属的93%)。GenBank数据形式如附录1所示。从如此海量数据中准确、快速查找和下载所需数据, 是高效利用这些数据的基础, 也是目前这些基因数据难以被广泛使用的障碍之一(Sanderson *et al.*, 2008; Jones *et al.*, 2011)。

具体来说, GenBank数据的广泛使用主要存在以下3个难点:

(1)数据量大, 手工难以处理(Jones *et al.*, 2011; Pearse & Purvis, 2013; Xu *et al.*, 2015)。比如, 在GenBank中查找某一类群的序列时, 返回的结果经常成百上千甚至上万。这些数据可能来自不同的基因片段、不同的种群和不同的研究。通过手工处理如此大量的数据十分耗时、耗力, 有时甚至无法完成。虽然部分研究采用BLAST等序列比对(sequences alignment)工具(Altschul *et al.*, 1990), 但这一方法下载的序列较原始序列短, 且下载结果对初始参考序列非常敏感, 不适用于对多类群基因序列的批量下载。

(2)序列的查找严重依赖于注释信息, 但GenBank中注释信息的质量参差不齐。注释不清甚至注释错误的序列较多, 这些序列难以使用, 且严重影响后续的序列比对的结果。同时, 这种方法只适合少量序列的查找和下载, 当下载序列的数量较大时, 这种方法费时、费力且极易出现错误下载。

(3)虽然人们可利用GenBank的Entrez数据库, 在查找序列时设定所要查找的类群和特定基因, 但查询结果通常鱼龙混杂, 需要花费大量的时间和精力进行数据清理。当类群较大、数据较多时, 数据清理更是难上加难。针对这些问题, 我们开发了一款新软件——NCBIminer (Xu *et al*. 2015; https://github.com/greengirl/NCBIminer/releases), 本文主要介绍NCBIminer的工作流程与基本算法、使用和参数设置等。

1　NCBIminer的工作流程

NCBIminer的主要功能是从GenBank中下载用户指定的基因和类群的核苷酸序列及DNA样品的采集信息等。为此, 我们开发了迭代BLAST算法、BLAST与GenBank Annotation双向检验算法、多查询(query)归并算法和基因变异空间的估算方法。例如, 对一个类群的一个特定的基因片段, NCBIminer首先估算其基因变异空间, 从而建立反映变异空间整体的优化参考序列集。基于优化参考序列集, 结合多查询归并算法, 下载指定类群的基因序列。这些算法是NCBIminer的基础, 保障用户可高效、准确下载所需的基因序列。

NCBIminer的具体工作流程主要分为两步(附录2):

第一步, 根据给定的初始参考序列(集), 为用户指定类群的基因片段确定优化参考序列集。这一步主要分为3个部分。首先, 利用迭代BLAST算法, 根据初始参考序列(集), 在GenBank中查找用户所指定类群和基因片段的全部基因序列。其次, 根据这些序列与初始参考序列的匹配起始位置, 采用欧式距离法计算该类群的基因变异空间, 使之能反映该类群基因变异的整体情况。最后, 根据序列在变异空间内的距离将整个变异空间分为若干子集, 并在每个子集中寻找1个最优序列作为代表, 组成新的最优参考序列集。最优序列的确定采用BLAST与GenBank Annotation双向检验算法。该算法通过对序列的GenBank注释信息进行自动判读, 将拥有准确GenBank注释信息、并与初始参考序列的相似性最高的序列选入最优参考序列集。最优参考序列集通常能够较好地反映该类群的基因变异。

第二步, 用最优参考序列集中的每一条参考序列分别在GenBank中进行BLAST查找, 并利用多查询归并算法将BLAST结果合并作为最后的输出序列。多查询归并算法是将每条参考序列BLAST的结果集进行汇总, 并按照BLAST结果将来自于同一条序列的起始位置进行合并。由于每个参考序列都代表了该类群的一种序列变异特征, 多查询归并算法能够克服BLAST下载序列通常较短的缺点, 获得比较完整的基因序列。

2 NCBIminer的特点

NCBIminer有如下一些特点: (1)可跨平台运行, 包括Windows 32/64 bit, MAC和Linux等系统; (2)操作便捷, 用户不需要具有生物信息学背景和任何编程经验; (3)代码开源(https://github.com/greengirl/NCBIminer/releases/); (4)NCBIminer的主要运算都在GenBank服务器上完成, 因此对运行电脑的计算能力没有特殊要求。与其他GenBank查找和下载工具的比较分析显示, NCBIminer查找更全面, 下载数据的质量更高(Xu *et al.*, 2015), 是一款挖掘GenBank大数据、推动基因数据在生态学中应用的较好的工具。

3 NCBIminer的下载、安装与运行

NCBIminer使用MATLAB (R2014b)的生物信息学工具箱(Bioinformatics Toolbox)和并行运算工具箱(Parallel Toolbox)开发, 使用MATLAB Compiler编译为可执行文件。目前发行的版本为NCBIminer 1.0, 包括Windows, Linux和MAC三种版本。

NCBIminer的安装很容易。首先, 下载和安装MATLAB的运行环境MATLAB Compiler Runtime (MCR, http://www.mathworks.cn/products/compiler/mcr/)。目前NCBIminer只能在MCR 8.4版本下运行, 不兼容其他版本。该运行环境为免费软件; 在Windows操作系统下安装时可能需要关闭360等杀毒软件。然后, 下载NCBIminer (https://github.com/greengirl/NCBIminer/releases/)并运行即可。

在Windows环境下运行NCBIminer, 只需要双击NCBIminer程序图标(Win7系统下需要以管理员身份运行), 然后根据程序提示选择输入NCBIminer的参数配置文件, 例如下载包中包含的Demo1.txt文件。NCBIminer运行进程会在屏幕上显示, 结果会保存在Demo1.txt所在的文件夹。下载结束后, 按键盘任意键将退出程序。

在Linux和MAC操作系统下, 可以通过命令行来调用NCBIminer。具体有以下4步:

在命令行下, 将路径更改到NCBIminer文件夹。输入以下命令行, 运行NCBIminer:

```
./run_NCBIminer.sh <mcr_directory>
```

其中<mcr_directory>是MCR的安装路径。假设MCR安装在/mathworks/home/application/v84文件夹下, 那么命令行应该输入:

```
./run_NCBIminer.sh/mathworks/home/application/v84
```

当命令行显示如下提示信息时, 键入配置文件所在路径:

```
Where is the input file, e.g. '/home/NCBIminer/':
```

出现以下提示时, 键入配置文件名称:

```
input file name, e.g. 'Demo1.txt':
```

程序运行时会显示运行进度。在MAC系统下也可以双击NCBIminer图标运行该程序, 并选择输入文件, 但是这种情况下程序会在后台运行, 用户不能实时跟踪程序运行进度, 只能通过查看输出文件确定程序是否已经运行完毕。

4 NCBIminer参数的含义与赋值

NCBIminer运行时通过调用1个文本配置文件执行用户任务。该文本配置文件包括14个参数, 其中5个为必需参数, 是文本文件的核心内容。这5个参数分别为基因类型(feature type)、基因名称(feature name)、初始参考序列(initial query)、目标类群(taxon list)和输出文件名前缀(output prefix)等。包含这5个基本参数后, NCBIminer就可以正常运行了。

4.1 基本参数

(1) 基因类型。是指GenBank定义的序列类型(附录2)。基因类型通常是对序列的生物学功能的注释, 例如可编码蛋白质序列注释为CDS (coding DNA sequence)。其他基因类型, 如基因注释为gene、核糖体RNA注释为rRNA、内含子注释为intron、常用的核基因ITS1和ITS2序列则多被注释为

misc feature。该参数在建立优化参考序列集时使用。对于GenBank中某一条序列或序列总的某一片段来说, 只有当注释的序列类型与用户提供的序列类型之一符合时, 该序列才有可能成为参考序列。在GenBank中, 不同研究者对相同类型的序列所属基因类型的注释存在差异, 比如5.8s rRNA有时被注释为“gene”, 有时被注释为“rRNA”。因此在配置文件中用户可输入多个基因类型。

(2) 基因名称。是用户需要下载核苷酸序列的名称, 包括其异名。与基因类型参数一样, 该参数也是在建立优化参考序列集时使用。对于GenBank中某一条序列或序列中的某一片段, 只有当注释中的序列名称与用户提供的名称之一符合时, 该序列才有可能成为参考序列。基因名称不区分大小写, 不计空格。

(3) 目标类群。是指用户需要下载的基因所属的生物类群, 格式为“大类群\小类群”。大类群用来定义步骤1中BLAST的查找范围和序列的变异空间, 例如壳斗科(Fagaceae)。小类群表示用户需要下载的目标类群或者物种, 例如栎属(*Quercus*)或蒙古栎(*Quercus mongolica*)。

(4) 初始参考序列。是在步骤1(附录2)中进行迭代BLAST时所使用的参考序列。基于初始参考序列, NCBIminer将通过迭代BLAST算法建立能反映该基因在某类群中变异情况的“序列变异空间”, 并进而在该“序列变异空间”中选择少数序列建立“优化参考序列集”。初始序列的选择对序列的选择和下载过程至关重要。一般来说, 该序列应该来自用户的目标类群或目标类群的近缘种, 且是一条完整的序列(complete sequence)。通过测试分析发现, 由于NCBIminer仅利用初始参考序列建立“序列变异空间”, 因此在进行序列查找和下载时, 初始序列所在的类群对所下载序列的数量和质量影响很小(Xu *et al.*, 2015), 但在保证序列完整性的前提下, 选择与研究类群较为相近的物种的基因序列作为参考序列可以减少BLAST的次数从而提高下载效率。

(5) 输出文件名前缀。用来定义输出文件的名称。输出文件名的设置应该符合操作系统的文件命名规则。NCBIminer输出文件为4个, 如果设置输出文件名前缀为Demo1, 则输出的文件为:

Demo1_log.txt: 日志文件, 记录程序运行进度和结果。

ref_Demo1.fas: 优化参考序列集, 以fasta格式保存。

Demo1.fas: 存储下载到的基因序列, 以fasta格式保存。每条序列的名称由物种名、序列的获取号(accession)以及起始位置确定。

Demo1_table.txt: 表格形式存储的序列下载信息, 是以制表符分隔的文本文件。

4.2 可选参数

NCBIminer配置文件中还包含9个可选参数, 分别为期望值(ExpectValue)、初始参考序列长度的比例(maximum query sequences difference)、序列长度(sequence length)、扩展序列长度(extended length)、允许BLAST返回的序列数(alignments)、GenBank检索范围(Entrezs)、DNA采集信息(location)、目标数据库(database)、等待时间(time out)。

(1) 期望值。是BLAST的参数, 表征了BLAST结果集中的某一序列被随机查找到的可能性。该值越大, BLAST结果中包含短序列和随机序列的概率越大, 程序运行的时间也越长。经过多次下载试验, 对于保守的基因片段, 例如CDS, 该值可设为10^{-10}; 对于变异较大的基因片段, 例如ITS1或ITS2, 该值可设为0.001。默认值是10^{-10}。

(2) 初始参考序列长度的比例。是在估算目标类群的基因变异空间时使用的一个参数。为估算目标类群的基因变异空间, NCBIminer使用层次聚类(hierarchical clustering)方法对BLAST的结果进行聚类分析。该参数则为聚类分析结果进行合并分类时的阈值, 其数值范围为0–1, 默认值为1。该值越小, 聚类分析分的类越多, 因而找到的优化参考序列集越大。一般情况下, 对于保守的基因片段, 该值可设置为1; 而对于变异较大的基因片段, 如ITS1, 该值可设置为0.3。

(3) 序列长度和扩展序列长度。用于定义下载序列的长度。前者定义了最短序列长度, 默认值为(50, inf), Inf表示无穷大(infinity), 也即只下载长度大于50 bp的序列。后者用来对下载序列起始位置和终止位置进行扩展。例如扩展序列长度设为(10, 20), 表示在迭代BLAST算法找到的目标序列起止位点基础上, 起始位点向前(左)推10 bp, 终止位点向后(右)推20 bp。该参数的作用是弥补BLAST工具下载序列时极易丢失序列两端数据的缺陷。同时, 可以

通过该参数下载相邻的序列。该参数的默认值为(0, 0), 表示不对下载结果进行扩展。

(4) 允许BLAST返回的序列数。默认值是5,000, 最大值为100,000。用户可以根据所要下载的类群在GenBank中的序列数量设置合理的值。

(5) GenBank检索范围。可用于更详细界定要查找和下载的基因序列和类群范围, 提高下载序列的准确度和效率, 其书写语法可参考GenBank Entrez帮助文档(http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml#entrez_query)。

(6) DNA采集信息。用于表明用户是否需要下载DNA样品的采集信息。若该值设为1表示需要下载采集信息, 0则表示不需要。默认值为0。

(7) 等待时间。是指当出现网络连接错误时, 程序持续的等待时间。如果超出等待时间, 程序会自动终止。等待时间的默认值是1,800 s。

(8) 目标数据库。是需要下载序列的数据库名称, 默认数据库是核苷酸序列数据库(nucleotide)。用户可以设置多个数据库, 例如“nucleotide, gene”表示从核苷酸数据库和基因数据库中查找和下载序列。

5 配置文件示例

文本配置文件中每个参数名称以半角冒号结尾, 换行后输入参数值。对于运行多个输入值的参数, 每个值占1行。而注释信息以#开始。以5个必需参数为例:

```
feature type: #基因类型
gene
CDS
feature name: #基因名称
rbcL
ribulose-1,5-bisphosphate carboxylase/oxygenase
large subunit
initial query: #初始参考序列, Fasta格式
>gi|194400737|gb|EU713435.1|
ATGTCACACAAACAGGACTAAGCAAATGTGGATTAAGGCTGGT
GTTAAGATATAAATTAACTTATTATACTCCGATATGAAACAAGG
TAGGTATTTGCAGCCTTTCGATAACTCCTCAACCCGATTCCCC-
CGGAAGAA….
taxon list: #目标类群
Campanulaceae\Codonopsis
Campanulaceae\Campanumoea
Output prefix: #文件名前缀
Demo1
```

如果将以上内容保存为文本文件作为NCBI-miner的配置文件, 那么用户运行NCBIminer, 并调用该配置文件, 将下载到rbcL基因序列120条(搜索日期: 2015年1月8日), 耗时约1 min。用户可以根据以上介绍修改配置文件后, 批量下载特定类群的某一基因的所有核苷酸序列。

参考文献

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

Chen ZD (陈之端), Li DZ (李德铢) (2013) On Barcode of Life and Tree of Life. *Plant Diversity and Resources* (植物分类与资源学报), **35**, 675–681. (in Chinese with English abstract)

Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ (2004) Prospects for building the Tree of Life from large sequence databases. *Science*, **306**, 1172–1174.

Holt B, Lessard JP, Borregaard MK, Fritz SA, Araujo MB, Dimitrov D, Fabre PH, Graham CH, Graves GR, Jonsson KA, Nogues-Bravo D, Wang ZH, Whittaker RJ, Fjeldsa J, Rahbek C (2013) An update of Wallace's zoogeographic regions of the world. *Science*, **339**, 74–78.

Jones M, Koutsovoulos G, Blaxter M (2011) iPhy: an integrated phylogenetic workbench for supermatrix analyses. *BMC Bioinformatics*, **12**, 30.

Li DC (2013) Similarity analysis of DNA sequences based on CLZ complexity. *Journal of Computational and Theoretical Nanoscience*, **10**, 481–487.

Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL, Yang JB, Fu CX, Zeng CX, Yan HF, Zhu YJ, Sun YS, Chen SY, Zhao L, Wang K, Yang T, Duan GW, Grp CPB (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences*, *USA*, **108**, 19641–19646.

Lu LM (鲁丽敏), Sun M (孙苗), Zhang JB (张景博), Li HL (李洪雷), Lin L (林立), Yang T (杨拓), Chen M (陈闽), Chen ZD (陈之端) (2014) Tree of Life and its applications. *Biodiversity Science* (生物多样性), **22**, 3–20. (in Chinese with English abstract)

Pearse WD, Purvis A (2013) phyloGenerator: an automated phylogeny generation tool for ecologists. *Methods in Ecology and Evolution*, **4**, 692–698.

Pei NC (裴男才) (2015) Applications of DNA barcoding in

evolutionary ecology. *Biodiversity Science* (生物多样性), **23**, 291–292. (in Chinese)

Qiu Q, Zhang GJ, Ma T, Qian WB, Wang JY, Ye ZQ, Cao CC, Hu QJ, Kim J, Larkin DM, Auvil L, Capitanu B, Ma J, Lewin HA, Qian XJ, Lang YS, Zhou R, Wang LZ, Wang K, Xia JQ, Liao SG, Pan SK, Lu X, Hou HL, Wang Y, Zang XT, Yin Y, Ma H, Zhang J, Wang ZF, Zhang YM, Zhang DW, Yonezawa T, Hasegawa M, Zhong Y, Liu WB, Zhang Y, Huang ZY, Zhang SX, Long RJ, Yang HM, Wang J, Lenstra JA, Cooper DN, Wu Y, Wang J, Shi P, Wang J, Liu JQ (2012) The yak genome and adaptation to life at high altitude. *Nature Genetics*, **44**, 946–949.

Ren BQ (任保青), Chen ZD (陈之端) (2010) DNA barcoding plant life. *Chinese Bulletin of Botany* (植物学报), **45**, 1–12. (in Chinese with English abstract)

Sanderson M, Boss D, Chen D, Cranston K, Wehe A (2008) The PhyLoTA browser: processing GenBank for molecular phylogenetics research. *Systematic Biology*, **57**, 335–346.

Xu X, Wang Z, Rahbek C, Lessard J-P, Fang J (2013) Evolutionary history influences the effects of water–energy dynamics on oak diversity in Asia. *Journal of Biogeography*, **40**, 2146–2155.

Xu XT, Dimitrov D, Rahbek C, Wang ZH (2015) NCBIminer: sequences harvest from Genbank. *Ecography*, **38**, 426–430.

Yang ZY, Ran JH, Wang XQ (2012) Three genome-based phylogeny of Cupressaceae s.l.: further evidence for the evolution of gymnosperms and southern hemisphere biogeography. *Molecular Phylogenetics and Evolution*, **64**, 452–470.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, Beaulieu JM (2013) Three keys to the radiation of angiosperms into freezing environments. *Nature*, **506**, 89–92.

(责任编委: 葛学军　责任编辑: 闫文杰)

附录　Supplementary Material

附录1　GenBank中的序列数据格式。左侧方框中是GenBank定义的基因类型(feature type), 右侧方框中为该序列的相关注释信息。

Appendix 1　Data format for a sequence in GenBank. The items in the left box are feature types defined in GenBank, while those in the right box are GenBank annotation information.

http://www.biodiversity-science.net/fileup/PDF/w2015-120-1.pdf

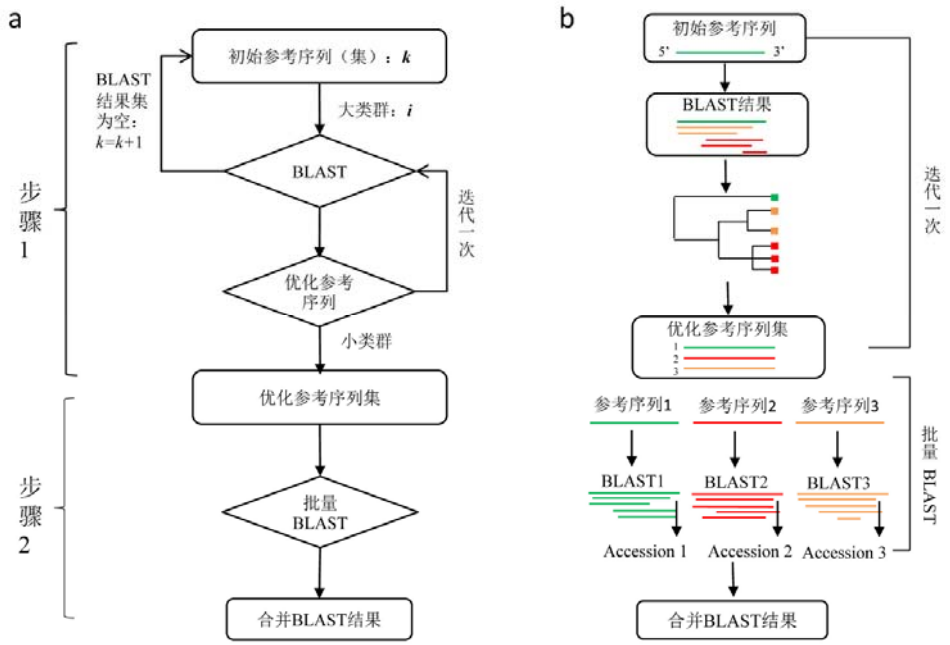
附录2　NCBIminer的工作流程。a为NCBIminer工作的主要流程, b详细解释了优化参考序列集建立和多查询归并算法的步骤。根据Xu *et al.* (2015)修改。

Appendix 2　NCBIminer workflow. a, Major steps of the NCBIminer's work flow; b, The algorithms for the establishment of improved reference sequences and sequence combination of multiple queries. Modified from Xu *et al*. (2015).

http://www.biodiversity-science.net/fileup/PDF/w2015-120-2.pdf

附录**1** **GenBank**中的序列数据格式。左侧方框中是**GenBank**定义的基因类型**(feature type),** 右侧方框中为该序列的相关注释信息。

Appendix 1 Data format for a sequence in GenBank. The items in the left box are feature types defined in GenBank, while those in the right box are GenBank annotation information.



附录2　NCBIminer的工作流程。a为NCBIminer工作的主要流程，b详细解释了优化参考序列集建立和多查询归并算法的步骤。根据Xu *et al.* (2015)修改。

Appendix 2　NCBIminer workflow. a, Major steps of the NCBIminer's work flow; b, The algorithms for the establishment of improved reference sequences and sequence combination of multiple queries. Modified from Xu *et al.* (2015).