2017 年第 2 卷第 3 期

文献 DOI：
10.11922/csdata.180.2016.0017
数据 DOI：
10.11922/sciencedb.117
文献分类：生物科学

收稿日期：2016-07-14
开放同评：2016-08-22
发表日期：2017-07-19

用于生命之树重建的数据集

孟珍[1]，杨拓[2]，刘红梅[3]，黎建辉[1*]，曹巍[1]，刘勇[1]，杲艳平[1]，刘奇[1]，林小光[1]，何星[1]，邵靖[1]，郑波[1]，王学志[1]，周园春[1]，陈之端[2*]，林立[2]，苏俊霞[2]，孙苗[2]，董晓宇[2]，李洪雷[2]，鲁丽敏[2]，张景博[2]，朱新宇[2]，李勇[3]，张寿洲[3*]，董慧[3]，廖一颖[3]，杨蕾蕾[3]，万涛[3]

1. 中国科学院计算机网络信息中心，北京 100190；
2. 中国科学院植物研究所，系统与进化国家重点实验室，北京 100093；
3. 深圳中国科学院仙湖植物园，北京 100093

*通讯作者（Email: lijh@cnic.cn，zhiduan@ibcas.ac.cn，shouzhouz@126.com）

摘要：由中国科学院计算机网络信息中心、中国科学院植物研究所、深圳市中国科学院仙湖植物园“三方两地”共同合作研究建设的“达尔文树”——分子数据分析应用环境（DarwinTree——Molecular Data Analysis and Application Environment），从中国陆地植物发育系统框架的研究出发，逐步推动解决生命之树构建过程中存在的技术难题，探索利用基因和基因组信息构建生命之树的策略和方法，研究和开发 DNA 序列信息自动采集和生命之树自动生成技术（Automatic Reconstruction of The Tree of Life），建立生命之树信息平台及其利用体系，为最终在我国建立具有国际影响的，能很好地兼容物种分类、地理分布、形态性状、化石信息以及 DNA 信息的物种库（Species Bank）创造条件。DarwinTree 旨在为科研人员提供数据和分析并举的工作平台，该平台将承担数据汇集和面向实际科研工作应用的双重作用。本文发布的数据集包括：（1）DarwinTree 基础数据集：来自国际公共序列数据的标记处理得到的分子标记数据及其与任意阶元物种分类名称对应的统计数据集；（2）DarwinTree 自测序数据集：面向中国陆地植物研究的补充测序序列数据；（3）DarwinTree 中国维管植物进化数据集：已构建的中国维管植物属系统发育树的数据（Generic tree of Chinese vascular plants）。

关键词：DarwinTree；系统发育；生物进化；分子序列；基础数据集；自测序数据集；中国维管植物进化数据集

数据库（集）基本信息简介

数据库（集）中文名称	用于生命之树重建的数据集
数据库（集）英文名称	The datasets used for the reconstruction of the tree of life
数据作者	孟珍、杨拓、刘红梅、黎建辉、曹巍、刘勇、杲艳平、刘奇、林小光、何星、邵靖、郑波、王学志、周园春、陈之端、林立、苏俊霞、孙苗、董晓宇、李洪雷、鲁丽敏、张景博、朱新宇、李勇、张寿洲、董慧、廖一颖、杨蕾蕾、万涛
通讯作者	黎建辉（lijh@cnic.cn）、陈之端（zhiduan@ibcas.ac.cn）、张寿洲（shouzhouz@126.com）
数据服务系统网址	http://www.darwintree.cn http://www.sciencedb.cn/dataSet/handle/117

基金项目	国家科技基础条件平台子专题（Y207011108）、深圳市科技研发国际合作资金项目（ZYA201007060093A）、中国科学院知识创新工程青年基金项目（Y114061108）、国家自然科学基金（91224006、31270268 和 61003138）、科技部支撑计划（2014CB954100）、中科院战略先导项目（XDA06010202）、中国科学院计算机网络信息中心一三五规划重点培育方向专项（CNIC_PY-1405）、深圳市科技创新委员会资助（KQC201105310009A）
数据库（集）组成	数据集由三部分数据组成，其一为DarwinTree基础数据集，其二是DarwinTree自测序数据集，其三是DarwinTree中国维管植物进化数据集。

引　言

生命之树（Tree of Life，TOL）是指将所有生物种类（包括现存和已灭绝的）联系在一起并蕴涵海量信息的系统进化树。它可用来阐明生命的起源、生物进化式样、生物各大门类演化和亲缘关系、以及生物多样性的存在方式和变化的动态规律。构建生命之树并充分挖掘和利用其中的信息资源是生命科学面临的又一挑战。目前国际上已经开展了由多个国家、多个科研团队和较多科研人员共同参与的构建 TOL 的研究工作[1]。

从研究积淀上来讲，近 20 年快速积累的基因和基因组信息为生命之树的构建奠定了重要基础。然而目前在 DNA 数据的自动采集和筛选、数据整合、超大树（Supertree）构建、以及信息的进一步挖掘和共享等方面都存在很多技术难题，各国都在寻找利用基因和基因组信息构建生命之树的有效策略和方法[2]。构建超大生命之树有两种不同的途径：（1）依据两个或若干个较小树的重叠部分，把多个已完成的小树整合成超大树；（2）直接对超大数据矩阵进行分析，构建生命之树。但无论哪种途径目前都面临同样的问题，即如何充分利用公共数据库中已有的 DNA 序列信息？如何对这些信息进行有效筛选？如何快速自动生成反映不同生物类群进化历史的生命之树？如何充分挖掘和利用生命之树中蕴涵的巨大信息？

由中国科学院计算机网络信息中心、中国科学院植物研究所、深圳市中国科学院仙湖植物园三方两地共同合作研究建设的“达尔文树”——分子数据分析应用环境（DarwinTree——Molecular Data Analysis and Application Environment）[3-4]，从中国的陆地植物发育系统框架的研究出发，逐步推动解决生命之树构建过程中存在的技术难题，探索利用基因和基因组信息构建生命之树的策略和方法，研究和开发 DNA 序列信息自动采集和生命之树自动生成技术（Automatic Reconstruction of The Tree of Life），建立生命之树信息平台及其利用体系，为最终在我国建立具有国际影响的、能很好地兼容物种分类、地理分布、形态性状、化石信息以及 DNA 信息的物种库（Species Bank）创造条件。DarwinTree 意在为科研人员提供数据和分析并举的工作平台，并承担数据汇集和面向实际科研工作应用的双重作用。DarwinTree 平台按照“统一规划、统一标准、突出重点、分步实施”的原则，从以下几个方面进行建设：DNA 序列信息与其他信息的数据整合技术及物种库构建的框架模型；生命之树自动生成技术及超大树的组装策略；DNA

序列信息的自动采集、评价和标记；基因筛选，及数据矩阵的自动装配；生命之树的信息挖掘和分析系统建设。DarwinTree 遵循的业务流程如图 1 所示。

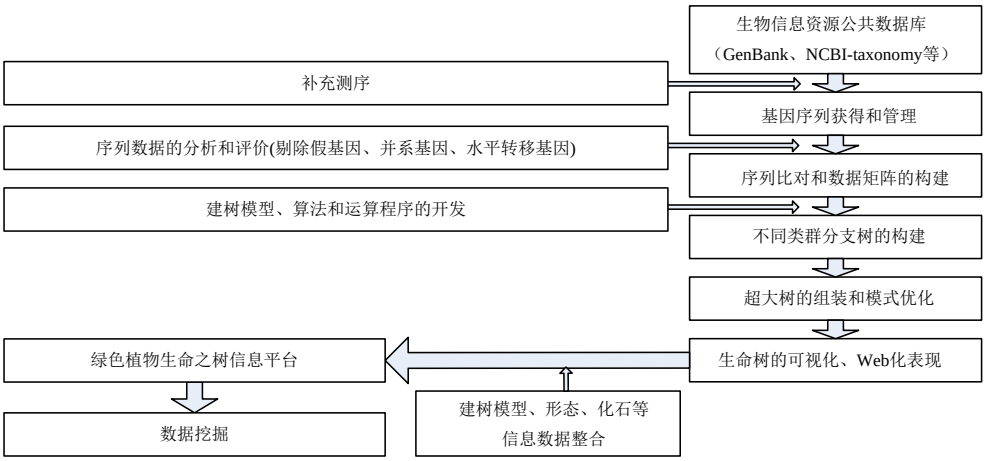


图 1　DarwinTree 现有业务流程

DarwinTree 在数据建设中做了以下工作。第一方面是对基础数据集的标记注释和统计分析。首先通过抽提公共的生物信息学资源库，比如 GenBank[5]、NCBI-taxonomy（http://www.ncbi.nlm.nih.gov/taxonomy）等得到基础数据，随后进行序列数据的分析标记和数据量统计。第二方面是补充测序数据的采样测序和质量控制。终端用户根据前面数据的标记和统计，可以方便检索和下载任意分类阶元、拉丁学名、分子标记（Molecular Marker）数据的统计信息和序列数据。随后，系统生物学团队再从具体问题出发依据研究现状选择物种进行基因补充测序。第三方面是系统发育大树的构建。来自公共数据库和项目团队补充测序的数据首先需进行清洗（即序列的分析和评价），以剔除假基因、并系基因和水平转移基因；然后进行基因的多序列比对和数据矩阵的构建、不同树分支的创建和超大树的组装及模式优化，以形成系统发育大树。最后 DarwinTree 平台整合形态等化石信息，进行生命树的可视化和 Web 化呈现，从而形成绿色植物的工作平台。平台用户只需登录即可进行相关的数据定制和分析工作。

综上所述，在从系统发育进化方法出发的研究中，DarwinTree 生产 3 大方面的数据集：来自公开数据库的测序数据及其统计数据、从具体研究问题出发的科学家补充的测序数据、通过系统发育方法构建的描述物种间进化关系的系统发育树的数据。

本数据论文发布的数据集包括：（1）DarwinTree 基础数据集：来自标记处理国际公共序列数据得到的基因标记（Gene Mark）数据及其与任意阶元物种分类名称对应的统计数据集；（2）DarwinTree 自测序数据集：面向中国陆地植物研究的补充测序序列数据；（3）DarwinTree 中国维管植物进化数据集：已构建的中国维管植物属系统发育树的数据（Generic tree of Chinese vascular plants）。

1 数据采集和处理方法

1.1 DarwinTree 基础数据集

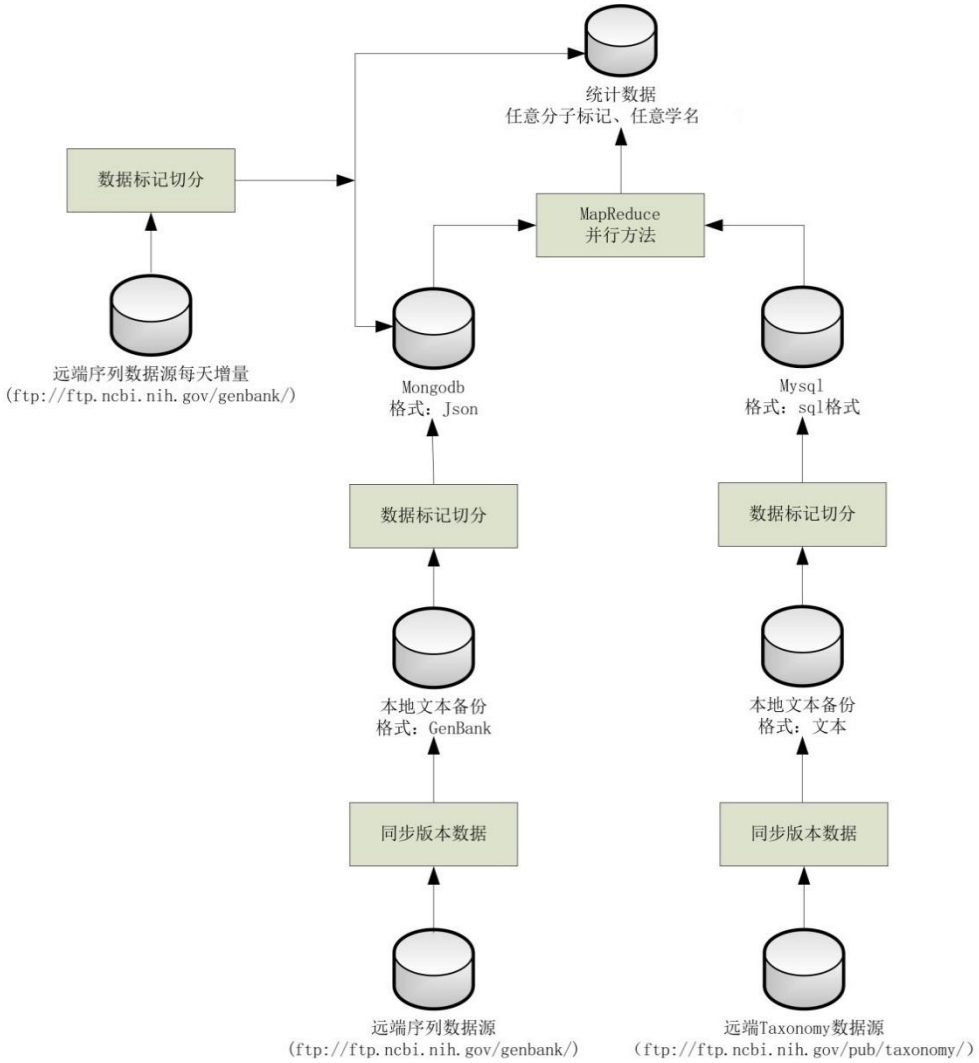


图 2　DarwinTree 基础数据集处理流程

DarwinTree 基础数据集是面向所有生物的分子标记数据及其与任意阶元物种分类名称对应的统计数据集。分子标记序列数据每天同步远端数据源（ftp://ftp.ncbi.nih.gov/genbank）并依据注释信息（Features）进行标记和切分，具体数据处理流程如图 2 所示。数据存储上采用非关系型数据库（NoSQL 数据库）MongDB：由于 NoSQL 对数据库之间的依赖要求减弱，设计将 GenBank 数据元数据（存于表 Meta 中）与序列数据（存于表 Sequence 中）分开存储，并且针对查询需要设计特别的轻量级查询快表（存于表 Query 中），经验证明这一做法能够满足实时查询的需求。在查询检索时，先在 Query 集合中进行，再通过对应的 Meta 和 Sequence 数据还原出原始数据。表 Meta、Sequence、Query、Statistics、Taxonomy 之间通过序列的“accession”进行关联，并且这种关联是非强制性的。各个 Mongodb 的集合（Collection）针对 Accession、Scientific Name、Taxon、Gene 等字段建立索

引，并且数据分片存储（5 个分片）可以实现并行查询。数据采用多副本（3 个副本）存储，保证数据的安全性。这样系统进化研究人员可以检索下载任意分子标记、物种、阶元物种类群的数据集，不用自己再注释切分。

任意阶元物种分类名称对应的基因系列统计数据集是针对任意分子标记（约 1 092 098）的任意物种阶元拉丁学名进行的统计，对约 1 亿 9 千万以“accession”为单位的序列数据按 NCBI-taxonomy 库的学名（Scientific Name，约 985 845）统计，形成约 31 268 448 个统计结果。

1.2 DarwinTree 自测序列数据集

DarwinTree 自测序列数据集是 DarwinTree 分析团队面向中国陆地植物具体研究类群，根据同行研究现状选择物种进行基因补充测序得到的补充数据集。DarwinTree 自测序数据集处理流程如图 3 所示。

数据模块流：在基于生物条形码的物种鉴定的研究过程中，数据模块主要包括采集数据、实验数据、鉴定数据和参考数据 4 部分。研究样本的采集，在编号时要记录样品信息、地理信息、GPS 数据、采集人、采集日期等。在样品信息中注意记录雌雄植株、是否再繁殖、生长阶段、凭证标本类型、凭证标本数、采集部位、DNA 是否采集、生境、习性、野外鉴定以及其他描述信息。在地理信息的采集中要标记样本采集处所属大洲、国家、省/自治区、地区、区县、具体地点、并记录相关的空间 GPS 数据（纬度、经度、海拔）。

数据操作流：用来标示在研究过程中每个数据模块部分要经历的详细数据操作流程。比如在实验数据模块中，就要按照提取样本 DNA、检测 DNA 质量、PCR 扩增序列、检测 PCR 质量、纯化 PCR 反应、测定 DNA 序列、查看序列峰图进行质量判断以决定是否从头进行实验以及对检测合格的峰图进行序列拼接，这几个数据实验操作的步骤依次进行，并记录整个过程中的中间数据（如测序谱图文件、实验的光谱图、电泳图、序列文本文件等）。

数据对接流：即在各个数据模块中，经过判断和质量控制后的数据流进入下一个数据模块进行处理，通过数据文件的传递进行到下一个数据处理模块中。采集数据和实验数据的对接由样本实体的传递来实现；实验数据和物种鉴定的对接通过拼接序列的传递来实现；物种鉴定和数据追溯的对接通过物种鉴定的报告文件来实现。在鉴定数据和参考数据对接中，一方面依赖参考数据 LIB 进行物种鉴定生成鉴定数据，另一方面把经专家确认的鉴定数据和国际上已鉴定的数据生成鉴定 LIB 作为参考数据。

数据追溯流：在物种鉴定后，平台会依据 BLAST 算法的结果生成物种鉴定报告，可以追溯查看采集数据、实验数据和鉴定数据等，也可以追溯数据库如 GenBank、DarwinTree 等。

本文发布的 DarwinTree 自测序列数据集，通过对采样样本的处理和测序并检验质量得到合格基因片段序列，其他采集数据集 A、鉴定数据集 C 和参考数据集 D 不在本文发布。

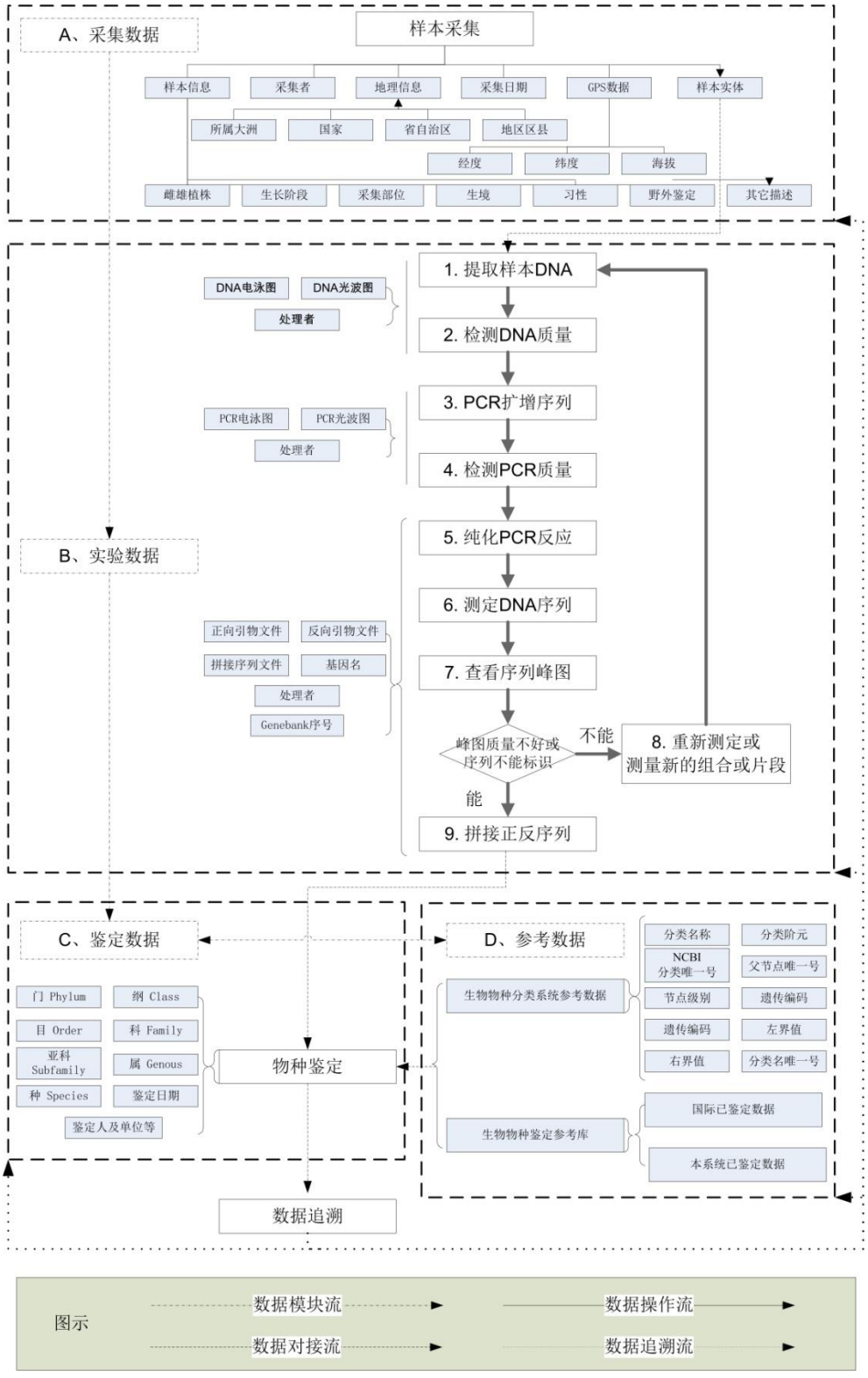


图 3　DarwinTree 自测序数据处理流程

1.3　DarwinTree 中国维管植物进化数据集

通过利用国际公开数据和自测数据相结合，接着进行清洗（即序列的分析和评价），以剔除假基因、并系基因和水平转移基因数据等；然后进行基因的多序列比对和数据矩阵的构建、不同树分支的创建和超大树的组装及模式优化，得到系

统发育树的数据集。

操作中首先使用 MUSCLE[6]进行自动比对得到初步结果，然后在 Bioedit[7]等软件中进行手工校对。若因进化速率差异等原因导致某一条基因序列与其他序列间难以找寻同源性，则将该序列从矩阵中删除，以避免产生进化噪音，影响系统发育树的可靠性及稳定性。由于材料缺乏或测序工作中的问题等，部分物种无法取得所有的基因片段，因此我们采用 supermatrix（超级矩阵）的方式，将所有序列拼接为 supermatrix 代表数据矩阵，并将缺失的数据以“-”表示。完成多序列比对的数据矩阵用于在软件 RAxML[8]中使用最大似然法搜寻最优的系统发育树，并执行 100 次 bootstrap 检验最优系统发育树中每个节点的可靠性。

2 数据样本描述

2.1 DarwinTree 基础数据集

DarwinTree 基础数据集以 JSON 格式存储在 Mongodb 数据库中，本文在样例说明中以数据 Accession AB008781 为例，其远端国际数据可访问 http://www.ncbi.nlm.nih.gov/nuccore/AB008781。

数据样例详见 http://darwintree.cn/common/downloadfile.shtml?id=281（DarwinTree 基础数据集（以 Accession AB008781 为例））。

2.2 DarwinTree 自测序列数据集

DarwinTree 自测序列数据集以 Fasta 格式存储，本文在样例说明中以数据 Abelmoschus_manihot_CPG14101 为例。

数据样例详见 http://darwintree.cn/common/downloadfile.shtml?id=281（DarwinTree 自测序数据集（以 Abelmoschus_manihot_CPG14101 为例））。

2.3 DarwinTree 中国维管植物进化数据集

系统发育树以 newick 格式存储，本文以团队构建的中国维管植物系统树为例。

数据样例详见 http://darwintree.cn/common/downloadfile.shtml?id=281（DarwinTree 中国维管植物进化数据集（以 Generic tree of Chinese vascular plants 为例））。

3 数据质量控制和评估

本文所述质量控制方案是在分析当前生物数据，尤其是基因序列数据在学科发展中的特点，综合数据规则和标准以及数据清洗、质量控制工具的研究现状，结合具体项目实践中的特点和应用，针对基因序列数据建立的。在基础数据集的大批量数据抽提和清洗中，主要涉及对大量数据检索、序列比对和模型匹配及其效率等问题。同时，该方案在面对海量数据的处理，结合数据密集特点进行基于 MapReduce（分布式计算系统）的并行应用设计。在私有数据（自测数据）的质量

控制部分，主要涉及基因序列数据精度的计算、Stop Codons（终止密码子）和 Contaminants（污染物）的排除以及 Trace Files（自测序列文件）数据质量的计算和系列相似性比较等。在中国维管植物属系统发育树数据的质量控制部分采用计算模拟过程稳定可重复策略，即使用数据矩阵可以重复整个构建系统发育树的流程并得到相同结果。

3.1 DarwinTree 基础数据集

在起始大规模数据筛选中，采用基于 BLAST 数据筛选方式[9]，如图 4 所示，初始输入数据来自于根据基因注释信息抽提的并由研究领域内专家确认的“种子”序列文件，经由 BLAST 数据抽提，获得 BLAST 报告进行解析，提取信息如 Accession、序列内容等进行结构化后应用。针对面向整个 International Nucleotide Sequence Database Collaboration（国际核苷酸序列数据库共享联盟，包括 GenBank、EMBL 和 DDBJ）数据的抽提，本方案采用基于 MapReduce 的 bCloudBLAST[10]的分布式计算来提高抽提效率。

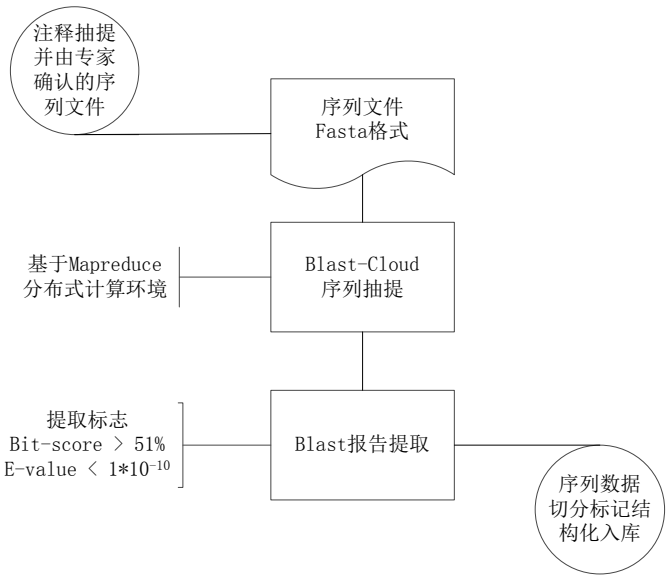


图 4　数据抽提流程

bCloudBLAST[10]采用切分输入文件的方法进行分割任务，每个任务先分别对其输入文件进行 BLAST 运算，最后汇总所有任务的结果并输出。具体实现时，采用 Apache Hadoop（MapReduce 模型的开源实现）来并行地调用 BLAST。Apache Hadoop 的流模式提供了在分布式环境中执行外部应用程序的方法，因此在 Map 中可以很方便地调用第三方程序。bCloudBLAST 采用 NCBI 的实现版本 NCBI BLAST2 为执行版 BLAST 程序，采用 NCBI-blast-lib 作为执行时的 Lib 库。并行计算时，由于每个结点都执行相同的 Map 并产生结果，所以最终的结果可以使用 Hadoop 的分布式文件系统（HDFS）提供的合并命令来得到。整个的流程如图 5 所示，并注意以下几点：

1. 输入文件切分成大小相等的块，并分配到各个节点中；

2. 对于每个块文件进行检查并将一个完整的核酸或蛋白质序列作为一个键值对传送到一个 Map 中；

3. 每一个结点上的 Map 对输入的核酸序列做 BLAST 运算，运算结果写入到本地文件中；

4. 使用 HDFS 的“merger”命令得到最终的结果文件。

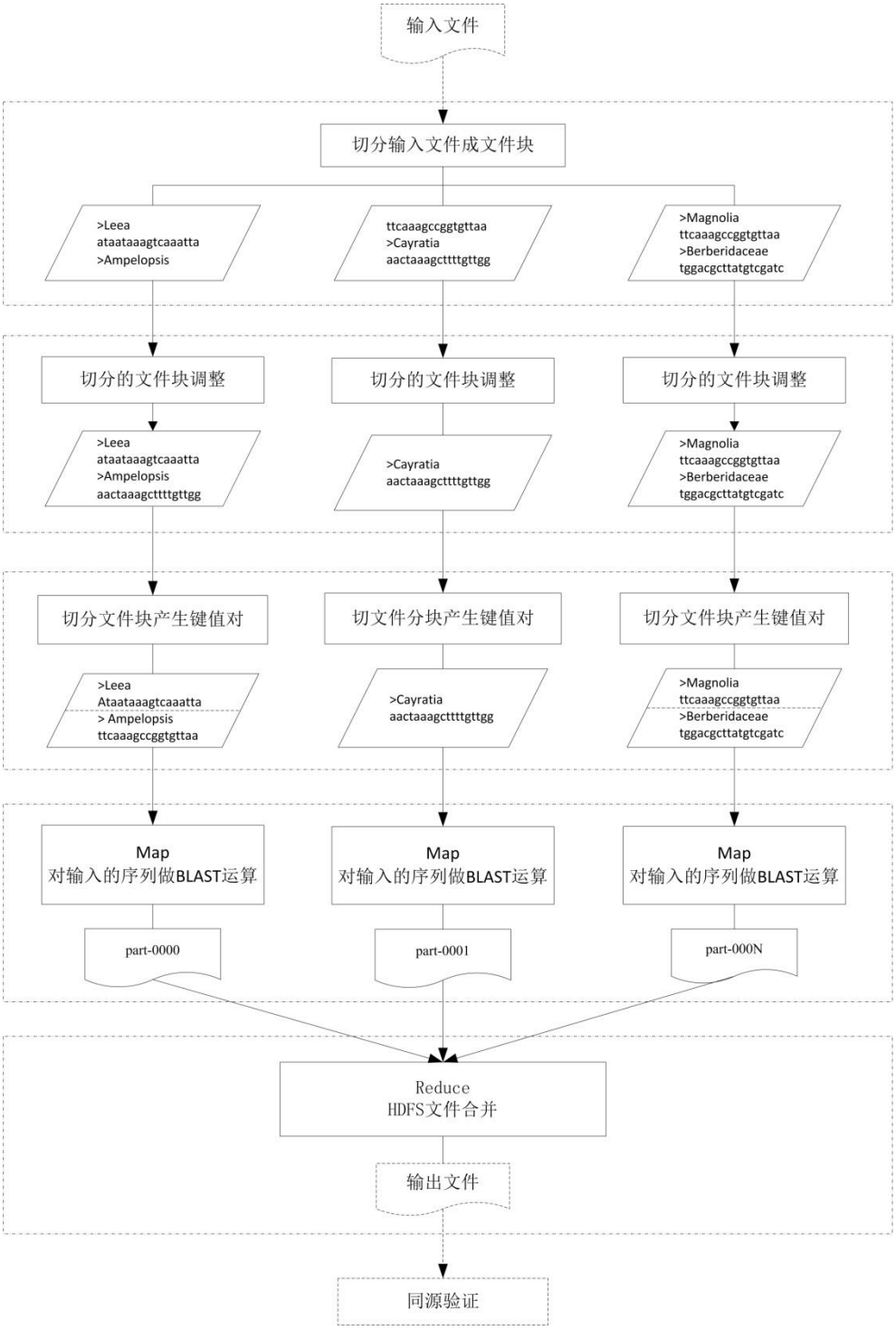


图 5　bCloudBLAST 计算流程

在 Hadoop 中把输入文件分割成大小相同的块时容易出现这种情况，一个核酸或蛋白质序列正好处于边界时容易被分到两个块中。为解决这种情况的产生，可以写一个分割的 Reader 类来处理这种问题，该类把整个序列做为一个键值对的

键传入到 Map 中。

3.2 DarwinTree 自测序列数据集

在自测分子标记序列数据的质量控制方面，采用多个模型算法生成序列质量测评报告给用户和审批人员，以决定是否入库。包括基因序列 N 等非确定值的测算，可能的 Stop Codons 的测算（检测可能的假基因）、Contaminants 的测算、Trace Files 的检测、相似度（bit score）和 E 值（E-value）的计算等。同时，考虑到学科研究的发展和不同数据挖掘目的对数据的精细化要求的不同，方案特别增加了相关的扩展模型接口，并在具体实现中通过文件的配置来完成。其处理流程如图 6 A 所示。其中所有的中间处理结果均保存于一个 Blast2Result 对象当中。

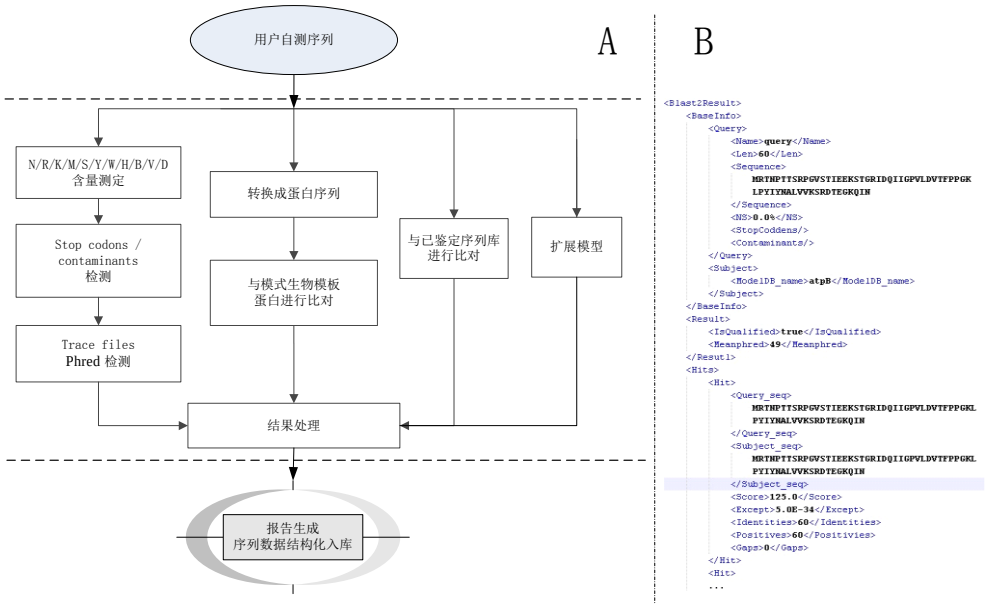


图 6　自测数据质量控制处理流程和报告配置设计

3.2.1 常规验证

1. 统计转化后的序列当中 N 和非确定碱基（R/K/M/S/Y/W/ H/B/V/D）的含量，并计算其在整个序列当中所占的百分比，将计算结果保存。

2. 检测是否有 Stop Codons 和 Contaminants。

在 DNA 序列当中，Stop Codons 指的是以下的三个片段：TAG、TAA、TGA。对转化后的序列进行判断，如果其中包含有这些基因片段，那么说明含终止密码子，可能是假基因。在项目实现时，将所包含的 Stop Codons 以及它在序列当中的位置信息保存在 Blast2Result 对象当中。

Contaminants 这里指的是引物未去除干净的污染，一般指的是一个序列对分前后两段。如果一个序列以某个 Contaminants 的第一段开头而以其第二段结束，那么表明它含有该 Contaminants。在项目实现时，存在一个 Contaminants 库，程序可以通过它来判断该序列是否包含 Contaminants，并将结果保存。

3. 根据该序列的 Trace Files 判断序列质量。

采用 Phred 程序（www.phrap.org）可以读取 Trace Files，将取得的数据如质量

分数等写出到输出文件当中。

3.2.2 序列与模板库进行比对

在此质量控制阶段，先将用户输入的基因序列转化成蛋白序列，然后将蛋白序列与用户所选择的模板库进行比对，根据比对的结果判断该序列与模板库是否同源。模板库当中仅含一条蛋白标准序列，并且一个模板库一般与一个包含多条核酸序列的序列库对应。

1. 转化成蛋白序列：该步骤主要是将用户所输入的要验证序列转化成蛋白序列进行。

2. 序列比对：通过调用 NCBI 的 blastp 程序，来将转化后的序列与所选择的模板库进行比对，并处理返回的结果。其中，先对模板库进行处理，提取出当中的序列并存储在 subject.txt 中，再将其与转化后的序列进行两两比对。

3. 结果处理：通过读取上一步比对结果的数据，将每个命中序列的“Score”、“Expect”、“Identities”、“Positives”以及“Gaps”值等保存在 Blast2Result 中。

3.2.3 序列与序列库进行比对

序列库依据基因和研究物种大类的不同分为若干个。里面的序列来自抽提的公共数据库或是自测序列，但必须经过领域内的专家认证。

3.2.4 结果验证

经过以上处理，序列与模板库的所有处理数据都保存于一个 Blast2Result 对象当中。最后就是要根据保存在这个对象中的数据进行判断，验证该序列是否与所选择的模板库同源，是否可以加入对应的序列库，并生成自测数据质量报告。报告将包括：

1. N 和非确定碱基值验证：将所保存的 N 和非确定碱基与用户输入的 N 和非确定碱基的值（一般缺省取值是 1%）进行比较，如果这个值较用户输入的值大，那么表明该序列不能通过验证。

2. Stop Codons 与 Contaminants 验证：如果该序列当中包含有 Stop Codons 或者 Contaminants，那么该序列不能通过验证。如果是包含有 Contaminants，那么需要提示用户将这些 Contaminants 去掉。

3. Trace Files 验证：根据前面对 Trace Files 的检验，确定该序列的质量等级，低质量（Meanphred＜30）、中等质量（Meanphred=30～40）、高质量（Meanphred＞40）。

4. BLAST 比对分值验证，将 Blast2Result 对象当中的 Score、Expect、Identities 等等与用户输入数值进行比较，来判断该序列是否能通过验证。

5. 生成报告序列入库，根据以上验证，可以得到该序列与模板库是否同源且质量可靠，并生成相关报告返回用户和审核员以决定是否加入序列库，报告将以 XML 格式配置，如图 6 B 所示。

基础缺省值（用户可根据需要做配置）包括：BaseInfo 部分存储的是序列与

模板库的名称序列长度等基本信息以及待验证序列的 N 和非确定碱基等值、Stop Codons 和 Contaminants 等；Result 中，IsQualified 表明序列是否通过如 Expect、Score、以及 Identities 等的验证，Meanphred 中存储的是以 Phred 程序从 Trace Files 中取得的序列质量分数；Hits 中存储的是命中序列段（可能有多段），Hit 中是每段的信息以及其各项如 Score 等的数值。用户通过这个报告最终决定是否将该序列加入序列库。

3.3 DarwinTree 中国维管植物进化数据集

在中国维管植物属系统发育树数据的质量控制部分采用计算模拟过程稳定可重复策略，即使用数据矩阵可以重复整个构建系统发育树的流程并得到相同结果。

综上所述，在本质量控制方案中，数据清洗方面采用基于基因注释的数据抽提与基于 BLAST 相似性比对抽提相结合的方案，保证一方面根据关键注释信息从国际公共数据库中较全面地提取过滤相关的序列信息，另一方面应用 BLAST 同源比对来控制序列的质量，剔除原始基因注释错误的序列。自测序列质量控制方面基于 Blastn 的比对打分和基于 Blastp 的模板比对相结合对自测数据的质量控制，同时报告序列整体质量，控制污染序列和假基因等入库。系统发育树数据的质量控制通过使用数据矩阵可以重复整个系统发育树的构建流程并得到相同结果的方法，来保证数据质量稳定。

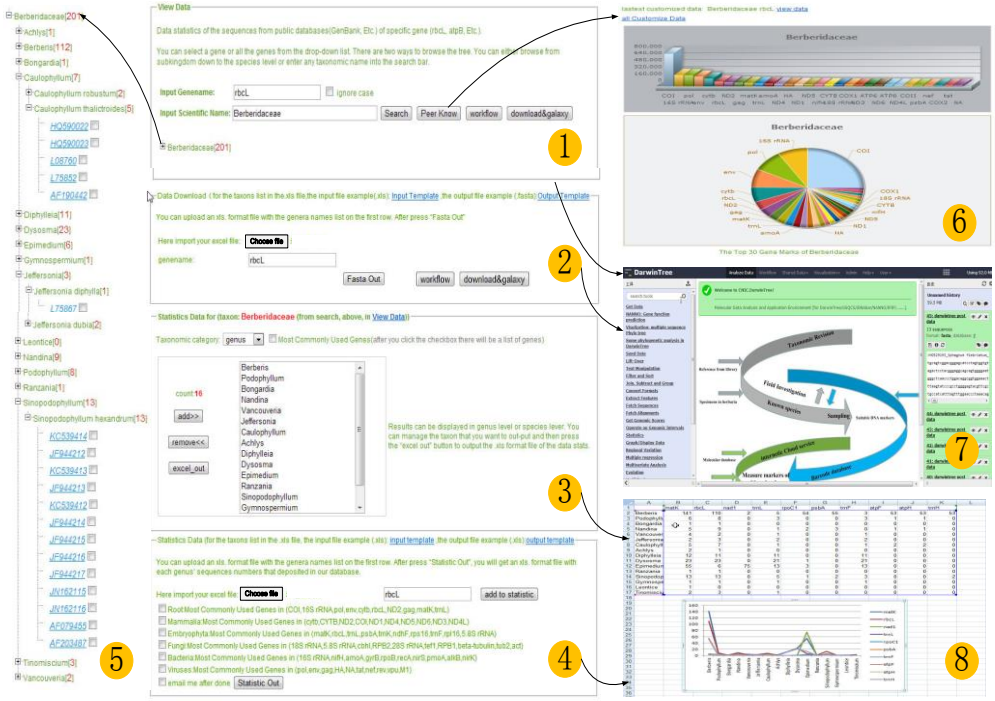


图 7　DarwinTree 数据浏览

注：①数据检索；②序列数据浏览下载（根据用户自己整理的名录）；③统计数据浏览下载（根据公共的生物分类系统）；④统计数据浏览下载（根据用户自己整理的名录）；⑤序列数据浏览下载（根据公共的生物分类系统，任意生物学名、分子标记）；⑥“Peer Known”（任意生物学名相关的前 30 位分子标记公开序列数据统计状况）；⑦检索数据集可以直接提交到分析平台分析；⑧任意生物学名、分子标记的统计数据浏览示例。

4 数据使用方法和建议

整体数据的更新、浏览、下载、后续分析工作流都可以通过以下网址进行使用：http://darwintree.cn。数据集的元数据信息参见：http://www.sciencedb.cn/dataSet/handle/117。数据集使用帮助文档参见：http://darwintree.cn/common/moredocs.shtml?type=1。图 7 所示是数据浏览出发的使用举例。

另外，其他典型数据应用入口有：中国维管植物进化数据集大树数据应用服务：http://www.darwintree.cn/flora/index.shtml；应用系列数据标准库的物种鉴定服务：http://darwintree.cn/common/identification.shtml。

数据作者分工职责

孟珍（1982—），女，山东人，高级工程师，研究方向为生物信息学、生物大数据技术。主要承担工作：DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据库设计、分析和组织。

黎建辉（1973—），男， 湖北人，研究员，研究方向为大数据技术。主要承担工作：DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据统筹管理。

王学志（1979—），男，湖南人，副研究员，研究方向为大数据技术。主要承担工作：DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据技术研发。

周园春（1975—），男，江西人，研究员，研究方向为大数据技术。主要承担工作：参加 DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据统筹管理。

曹巍（1986—），男，四川人，工程师，研究方向为计算机软件，主要承担工作：DarwinTree 基础数据集的数据库研发。

刘勇（1987—），男，山东人，工程师，研究方向为计算机应用技术，主要承担工作：DarwinTree 基础数据集的数据分析筛选。

杲艳平（1986—），女，江苏人，工程师，研究方向为计算机应用技术，主要承担工作：DarwinTree 基础数据集的数据分析筛选。

刘奇（1986—），男，湖南人，工程师，研究方向为计算机应用技术，主要承担工作：DarwinTree 基础数据集的数据分析筛选。

林小光（1983—），男，重庆人，助理研究员，研究方向为计算机软件，主要承担工作：DarwinTree 基础数据集的数据分析筛选。

何星（1983—），男，湖南人，工程师，研究方向为计算机软件，主要承担工作：DarwinTree 基础数据集的数据分析筛选。

郑波（1985—），男，河南人，工程师，研究方向为计算机软件，主要承担工作：DarwinTree 中国维管植物进化数据集数据库的技术研发。

邵靖（1987—），男，贵州人，工程师，研究方向为计算机软件与理论，主要承担工作：DarwinTree 中国维管植物进化数据集数据库的技术研发。

陈之端（1964—），男，山东人，研究员，研究方向为系统植物学。主要承担工作：DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据统筹管理。

杨拓（1988—），男，天津人，博士，研究方向为系统植物学。主要承担工作：DarwinTree 中国维管植物进化数据集数据库的设计和数据提供。

林立（1986—），男，上海人，博士，研究方向为系统植物学。主要承担工作：DarwinTree 中国维管植物进化数据集的数据库设计和数据提供。

苏俊霞（1974—），女，山西人，副教授，研究方向为植物分类和系统学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

孙苗（1983—），男，陕西人，博士后，研究方向为系统植物学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

董晓宇（1984—），女，江苏人，助理研究员，研究方向为植物分类学与系统植物学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

李洪雷（1985—），男，山东人，博士后，研究方向为系统植物学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

鲁丽敏（1986—）女，河南人，助理研究员，研究方向为系统植物学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

张景博（1985—），男，内蒙古人，博士后，研究方向为系统植物学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

朱新宇（1963—）男，河北人，教授，研究方向为生物信息学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

李勇（1967—），男，内蒙古人，教授级高级工程师，研究方向为植物学。主要承担工作：DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据统筹管理。

张寿洲（1964—），男，陕西人，研究员，研究方向为植物分类学。主要承担工作：DarwinTree 基础数据集、自测序列数据集和中国维管植物进化数据的数据统筹管理。

刘红梅（1979—），女，河南人，副研究员，研究方向为蕨类植物系统学。主要承担工作：DarwinTree 基础数据集的数据库设计和 DarwinTree 自测序列数据集的数据提供。

董慧（1981—），女，河北人，高级工程师，研究方向为昆虫分类学。主要承担工作：DarwinTree 基础数据集的数据库设计。

廖一颖（1985—），女，湖南人，助理研究员，研究方向为植物分子系统地理学，主要承担工作：DarwinTree 自测序列数据集的数据提供。

杨蕾蕾（1988—），女，山东人，助理工程师，研究方向为植物系统学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

万涛（1983—），男，江西人，高级工程师，研究方向为植物细胞学。主要承担工作：DarwinTree 自测序列数据集的数据提供。

参考文献

[1] Benton M, Ayala F. Dating the tree of life[J]. Science, 2003, 300: 1698–1700.

[2] Ciccarelli F, Doerks T, Mering C, et al. Toward Automatic Reconstruction of a Highly Resolved Tree of Life[J]. Science, 2006, 311(5765): 1283.

[3] Meng Z, Dong H, Li J, et al. Darwintree: A Molecular Data Analysis and Application Environment for Phylogenetic Study[J]. Data Science Journal, 2015, 14: 1–10.

[4] Meng Z, Lin X, He X, et al. Construction of the Platform for Phylogenetic Analysis[M]. New York: Springer Science+ Business Media, 2011: 507–514.

[5] Benson D, Karsch-Mizrachi I, Lipman D, et al. GenBank[J]. Nucleic Acids Research, 2011, 39: D32–D37.

[6] Edgar R. MUSCLE: a multiple sequence alignment method with reduced time and space complexity[J]. BMC Bioinformatics, 2004, 5: 113.

[7] Hill T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT[J]. Nucleic Acids Symp., 1999: 95–98.

[8] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models[J]. Bioinformatics, 2006, 22: 2688–2690.

[9] Meng Z, Xiao X, Li J, et al. Cloud-GSQCT: a parallel approach to screen gene sequences for phylogenetics analysis[C]// International Conference on Computer Science and Information Processing. 2012: 660–663.

[10] Meng Z, Li J, Zhou Y, et al. bCloudBLAST: an Efficient MapReduce Program for Bioinformatics Applications[C]// International Conference on Biomedical Engineering and Informatics, BMEI 2011, 4: 2085–2089.

论文引用格式

孟珍, 杨拓, 刘红梅, 等. 用于生命之树重建的数据集[J/OL]. 中国科学数据, 2017, 2(3). DOI: 10.11922/csdata.180.2016.0017.

数据引用格式

孟珍, 杨拓, 刘红梅, 等. 用于生命之树重建的数据集[DB/OL]. Science Data Bank. DOI: 10.11922/sciencedb.117.

The datasets used for the reconstruction of the tree of life

Meng Zhen[1], Yang Tuo[2], Liu Hongmei[3], Li Jianhui[1*], Cao Wei[1], Liu Yong[1], Gao Yanping[1], Liu Qi[1], Lin Xiaoguang[1], He Xing[1], Shao Jing[1], Zheng Bo[1], Wang Xuezhi[1], Zhou Yuanchun[1], Chen Zhiduan[2*], Lin Li[2], Su Junxia[2], Sun Miao[2], Dong Xiaoyu[2], Li Honglei[2], Lu Limin[2], Zhang Jingbo[2], Zhu Xinyu[2], Li Yong[3], Zhang Shouzhou[3*], Dong Hui[3], Liao Yiying[3], Yang Leilei[3], Wan Tao[3]

1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, P. R. China;

2. State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Beijing 100093, P. R. China

3. Fairylake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen 518004, P. R. China

*Email: lijh@cnic.cn, zhiduan@ibcas.ac.cn, shouzhouz@126.com

Abstract: DarwinTree (http://www.darwintree.cn) was initiated to construct a molecular data analysis and application environment for the phylogenetic study and compiled collaboratively by three organizations from Chinese Academy of Sciences, which are Computer Network Information Center (CNIC), Institute of Botany (IB) and Shenzhen Fairylake Botanical Garden (SZBG). DarwinTree is expanding its development in a global scale to build an international research alliance. It has initiated the first international campaigns with the University of Florida for the study on the phylogeny of angiosperms. Recently, DarwinTree provides an integrated bioinformatics platform that supports all phases of the analytical pathway for phylogenetic study from data collection including gene data acquisition and management to phylogenetic tree reconstruction including sequences alignment and data matrix construction, to different branches reconstruction, to assembly of large trees and mode of optimization, to visualization of the tree of life and web-based rendering, and to specific application service and data mining. The datasets included in this published paper are: (1) the basic datasets in DarwinTree: the Gene Mark Data labeled from international public sequence data and the statistical datasets with any scientific name and any mark name; (2) the sequencing datasets in DarwinTree: the complementary sequencing data for China land plants; (3) Generic tree of Chinese vascular plants datasets.

Keywords: DarwinTree; Phylogeny; Biological evolution; Molecular sequence; Basic datasets; Complementary sequencing data; Generic tree of Chinese vascular plants datasetss