

## 陆地植物系统发育研究的工作平台构建

孟珍<sup>11</sup>, 陈之端<sup>22</sup>, 黎建辉<sup>11</sup>, 刘红梅<sup>22</sup>, 何星<sup>11</sup>, 林小光<sup>11</sup>, 张寿洲<sup>33</sup>, 李勇<sup>33</sup>, 胡良霖<sup>11</sup>, 周园春<sup>11</sup>

(1. 1. 中国科学院计算机网络信息中心科学数据中心, 北京 100190; 22. 中国科学院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; 33. 中国科学院 中国科学院深圳仙湖植物园, 广东 深圳 518004)

**摘要:** 在讨论利用基因和基因组信息构建生命之树的历史推进、有效策略和方法的基础上, 针对生命之树的构建进行业务流程和应用设计分析, 构建面向陆地植物的系统发育平台。平台实现基因数据从国际数据库的自动获取、清洗与自测数据的提交、整理功能, 给出系统树的流程化构建, 整合数据抽提、多重序列比对、编辑清洗、分模型构树、组装评估、可视化编辑等系列分析算法和模型, 经若干交互界面, 得到系统树的自动生成、辅助实验决策。

**关键词:** 系统发育; 生命之树; 数据抽提; 自动建树; PALPP

## Construction of Work Platform Construction for Phylogenetic Analysis Research of Land Plants

MENG Zhen<sup>11</sup>Zhen<sup>11</sup>, CHEN Zhi-duan<sup>22</sup>duan<sup>22</sup>, LI Jian-hui<sup>11</sup>hui<sup>11</sup>, LIU Hong-mei<sup>22</sup>mei<sup>22</sup>, HE Xing<sup>11</sup>Xing<sup>11</sup>, LIN Xiao-guang<sup>11</sup>guang<sup>11</sup>, ZHANG Shou-zhou<sup>33</sup>zhou<sup>33</sup>, LI Yong<sup>33</sup>Yong<sup>33</sup>, HU Liang-lin<sup>11</sup>lin<sup>11</sup>, ZHOU Yuan-chun<sup>11</sup>chun<sup>11</sup>

(1. Scientific Data Center, Computer Network Information Center, Chinese Academy of Sciences I., Scientific Data Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; 22. State Key Laboratory of Systematic and Evolutionary botany Botany, Institute of Botany, Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100093, China; 33. FairyLake Botanical Garden, Chinese Academy of Sciences, Chinese Academy of Sciences, Shenzhen 518004, China)

**【Abstract】** Based on discussing the history of advancement to building the Tree Of Life(TOL) using genetic and genomic information, effective strategies and methods for the construction of the tree of life, this paper carries out business process analysis and application design. It implements a phylogenetic analysis platform for the land plants based on this analysis. The platform extracts molecular data from the international public databases in batch, which is automated acquisition, cleaning function for users to understand the situation of peer data. The process of phylogenetic reconstruction includes several public modes and tools, such as batch extraction, multiple sequence alignment, cleaning & editing, tree reconstruction, phylogeny evaluation and visualization. All these procedures demand a number of interactive interfaces for phylogenetic tree automatic generation and decision-making aids experiment.

### 1 概述

生命之树(Tree Of Life, TOL)是能将所有生物种类(包括现存的和灭绝的)联系在一起的蕴涵着巨量信息的系统进化树, 可用来阐明生命的起源、生物进化式样、各大门类生物演化和亲缘关系、以及生物多样性的生存方式和动态变化规律。构建生命之树并充分挖掘和利用其中的信息是生命科学面临的又一挑战。美国已于 2002 年正式启动了一个 15 年的 TOL 国家研究计划。欧盟国家也已多次酝酿并将很快启动其 TOL 计划, 以英国邱园为中心, 从 2002 年至今连续在 Patras、Paris、London 和 Brussels 召开 4 次会议, 商讨欧洲的 TOL 项目<sup>[1]</sup>。2002 年美国密西根大学华裔学者仇寅龙博士向中国科学院植物研究所洪德元院士写信建议尽快在中国启动 TOL 项目。在最近的几年中, 中国学者经过多次讨论认为在我国开展 TOL 项目是完全必要的, 也是非常适时可行的。在 2003 年 10 月召开的香山会议上, 来自微生物、植物、动物、及进化生物学各领域的中外学者具体讨论了在我国实施该项计划的细节; 此后, 又进一步讨论形成了“生命之树——中国国家行动计划纲要”, 意在逐步推进 TOL 在中国的进程。

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<sup>[2]</sup>。构建超大型的生命之树有 2 种不同的途径: (1)依据 2 个或若干个较小树的重叠部分把多个业已完成的小树进行整合, 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的筛选, 如何能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

**基金项目:** 中国科学院“十一五”重大专项基金资助项目“数据应用环境建设与服务”(0846061372, 0846061108, 0846061208)

**作者简介:** 孟珍(1982-), 女, 工程师、硕士, 主研方向: 生物信息学, 数据挖掘; 陈之端、黎建辉、李勇, 研究员、博士; 刘红梅, 助理研究员助研、博士; 何星、林小光、胡良霖, 工程师、硕士; 张寿洲、周园春, 副研究员研、博士; 李勇, 研究员、博士; 胡良霖, 工程师、硕士; 周园春, 副研究员、博士李勇, 研究员、博士; 胡良霖, 工程师、硕士; 周园春, 副研、博士

**收稿日期:** 2010-046-124 **E-mail:** zhenm99@cnic.cn

命之树中蕴涵的巨大信息。

由中国科学院植物研究所、中国科学院深圳仙湖植物园和中国科学院计算机网络信息中心“三方两地”，共同合作研究建设的陆地植物系统发育平台(PALPP)从中国的陆地植物发育系统框架的研究出发，逐步推动生命之树构建过程中存在的技术难题解决，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 TOL 中国国家计划的启动和实施做准备，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、形态性状、化石信息、以及 DNA 信息的物种库(Species Bank)创造条件。PALPP 意在为科研人员提供数据和分析并举的工作平台，承担数据汇集和面向实际科研工作应用的双重作用。

## 2 平台应用设计

陆地植物系统发育平台按照“统一规划、统一标准、突出重点、分步实施”的原则，从以下几 5 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(Marker Gene)筛选，及数据矩阵(Data Matrix)的自动装配；生命之树的信息挖掘和分析系统建设。PALPP 遵循的业务流程如图 1 所示。

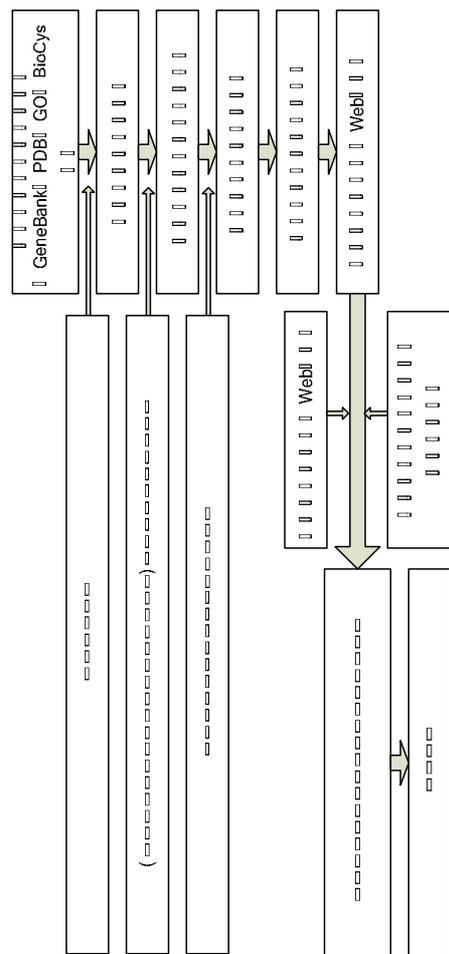
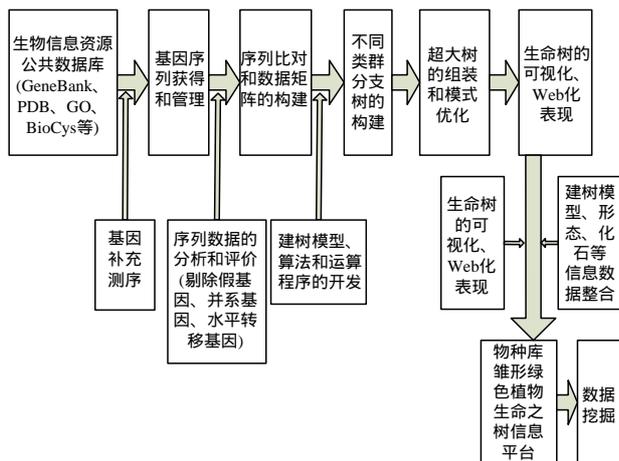


图 1 PALPP 业务流程

### 2.1 平台业务流程

PALPP 建设中通过分析公共的生物信息学资源库，比如 GeneBank<sup>[3]</sup>、Gene Ontology(Gene Ontology Consortium, 2004) 等，抽提得到基础数据，并根据同行研究现状选择物种进行基因补充测序；随后进行清洗即序列的分析和评价部分，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、进行不同树分支的创建和超大树的组装及模式优化；最后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。PALPP 建设注重建树模型、算法和运算程序的应用和开发积累。

### 2.2 平台建设框架

PALPP 系统整体建设框架如图 2 所示，工作包括数据库体系设计、模型算法的运行环境建设、起始大批量数据整理、数据获取、数据清洗、生命树构建、数据平台整合共 7 个方面的内容。系统集成 BioEdit<sup>[4]</sup>、Clustalx<sup>[5]</sup>、modeltest<sup>[6]</sup>、mrbayers<sup>[7]</sup>、phyml<sup>[8]</sup>，以及 ncbi-sequin(version 5.26, 2004)、ATV<sup>[9]</sup>等软件包的功能，并对分析中耗时较多的分析程序和环节(如 Alignment Multiple)进行在单服务器多核、多服务器 Cluster 计算和深腾 7000 超级计算的任务调度<sup>[10]</sup>。用户可以通过各分析模块进行精细化分析，也可以在工作流中定制使用。



图2 平台建设整体框架

系统整合 BioEdit<sup>[4]</sup>、Clustalx<sup>[5]</sup>、modeltest<sup>[6]</sup>、mrbayers<sup>[7]</sup>、phyml<sup>[8]</sup>，以及 ncbi-sequin(version 5.26, 2004)、ATV<sup>[9]</sup>等软件包的功能，并对分析中耗时较多的分析程序和环节(如 Alignment Multiple)进行在单服务器多核、多服务器 Cluster 计算和深腾 7000 超级计算的任务调度。用户可以通过各分析模块进行精细化分析，也可以在工作流中定制使用。

### 3 平台进展

PALPP 目前实现面向整个中国陆地植物体系，针对 rbcL、atpB 等基因数据从国际公认数据库(GeneBank、DDBJ、EMBL)的自动获取、清洗等功能，以方便平台用户实时了解同行的数据状况，PALPP 目前实现面向整个中国陆地植物体系在 rbcL、atpB 等基因数据从国际公认数据库(GeneBank、DDBJ、EMBL)的自动获取、清洗以方便平台用户实时了解同行的数据状况，并拥有自测基因数据的提交、整理功能以保证科研成果未发表前的数据挖掘，从而实现公共、私有数据的结构化、整合化。平台实现系统树的流程化构建，整合数据抽提、多重序列比对、编辑清洗、分模型构树、组装评估、可视化编辑等系列公认的分析算法和模型，通过若干交互界面，保证系统树的自动生成、辅助实验决策。平台用户依托中科院网络信息中心的数据环境和计算环境，只需登录定制相关的研究范围、分析模块和参数就可进行数据应用和挖掘工作。

PALPP 的界面运用 JavaServer Page 和 Java Applet 技术。普通用户、特权用户和管理用户以及超级用户均可通过页面 (<http://phylo.csdb.cn:8080/palpp/index.jsp>) 进入系统，依据相关权限进行诸如数据更新、数据挖掘、工作流定制、模块化分

析等工作。图 3 举例标明普通用户在 PALPP 上的工作流程和数据分析工作。A 为数据定制选择；B 为定制数据的分类查看；C 为模块精细化分析举例；D 为树文件可视化。

(1)A 到 B 过程所示为用户登录平台定制相关的数据范围并查看同行的研究现状。

(2)A 所示为用户进行分析时的操作，A1 用户选择已有的数据、A2 选择相应的参数(基因名称、分类阶元、代表序列数据、添加自测序列)、如果选择一键式计算(A3 所示的 One-Key-Run)就可得到分类结果呈现(D 所示)并保留各分析阶段(C 所示)的中间文件；用户选择相应参数后，如选择分步骤计算(A3 所示的 Run-Extraction)就可以按照各分析阶段(C 数据抽提、多重序列比对、编辑清洗、分模型构树、组装评估、可视化编辑)依模块进行精细化分析。

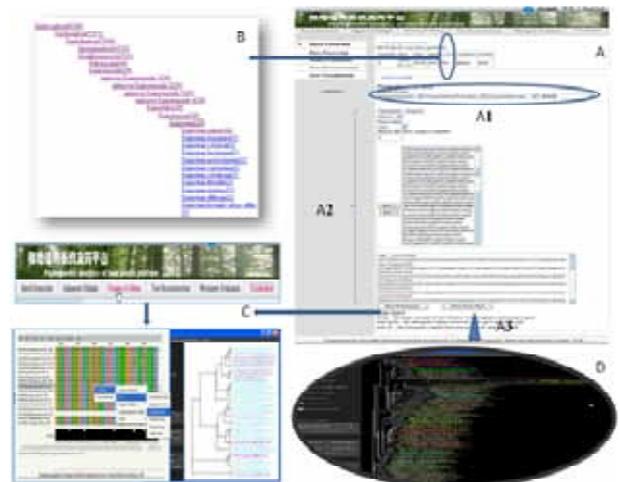


图3 普通用户的工作流程和数据分析工作

在图 3 中，A 为数据定制选择；B 为定制数据的分类查看；C 为模块精细化分析举例；D 为树文件可视化。

(1)A 到 B 过程所示为用户登录平台定制相关的数据范围并查看同行的研究现状。

(2)A 所示为用户进行分析时的操作，A1 用户选择已有的数据，A2 选择相应的参数(基因名称、分类阶元、代表序列数据、添加自测序列)，如果选择一键式计算(A3 所示的 One-Key-Run)就可得到分类结果呈现(D 所示)并保留各分析阶段(C 所示)的中间文件；用户选择相应参数后，如选择分步骤计算(A3 所示的 Run-Extraction)就可以按照各分析阶段(C 数据抽提、多重序列比对、编辑清洗、分模型构树、组装评估、可视化编辑)依模块进行细化分析。

### 4 结束语

平台逐步扩展应用范围建设从陆地植物到全体植物，再到可面向整个生物的工作平台。平台是开放性的，用户可以定制已有的数据范围；也可以通过提交研究的物种范围、所需基因、数据来源等参数来预定数据；用户可以应用平台已有的模型算法也可以定制自己的工作流。随着专家用户的需求牵引、数据类型的丰富、数据交互的完善、数据管理的加强，平台会逐步实现科研数据的交汇共享和工作平台的职能作用，实现系统发育。研究学者只需一根网线就可以完成全部的数据管理和数据挖掘工作。

同时,在整个 TOL 中的框架中,TOL 计划的总目标是重建所有生物物种的进化历史,或者说要让所有的生物有机体都在生命树上找到自己的位置。当生命之树最终在种的水平上包括所有的生物有机体时,也许将来对生物可以进行类似条形码式的鉴定(Bar coding)。到那时,人们在野外可利用手提式物种分析仪(handhold species analyzer),通过瞬间测定一个或少数几个基因序列来鉴定已命名过的物种或发现新的物种。尽管目前这还只是幻想,但 TOL

研究计划在理论上和具体实施途径上展现了一个美好的前景。

### 参考文献

- [1] Benton M J, Ayala F J. Dating the Tree of Life[J]. Science, 2003, 300(5626): 1698-1700.
- [2] Ciccarelli F D, Doerks T, von Mering C, et al. Toward Automatic Reconstruction of a Highly Resolved Tree of Life[J]. Science, 2006, 311(5765): 1283-1287.
- [3] Benson D A, Boguski M S, Lipman D J, et al. GenBank[J]. Nucleic Acids Reserch., 1999, 27(1): 12-17.

(上接第 271 页)

## 6 结束语

本文提出了一种改进的二维 DOA 估计算法——改进的时空 DOA 矩阵法。该算法通过构造  $X$  轴上的平移不变子阵列,产生 2 个 DOA 矩阵,利用这 2 个 DOA 矩阵的角度兼并曲线的差异,其中一个矩阵的角度兼并曲线平行  $\alpha$  轴,另外一个矩阵的角度兼并曲线平行  $\beta$  轴,任何 2 个信号不可能同时在 2 个矩阵上发生角度兼并,解决了原时空 DOA 矩阵方法的角度兼并问题。同时,该算法还保持了原时空 DOA 矩阵法无需二维谱峰搜索和参数配对等优点。由于本文在构造  $X$  轴上的平移不变子阵列时,充分地利用了原时空 DOA 估计方法的阵元,使 2 个子阵可以重复利用阵元,改进的时空 DOA 矩阵法仅比原时空 DOA 矩阵法多 2 个阵元,基本无冗余阵元和孔径损失。仿真结果证明了上述结论。

### 参考文献

- [1] 薛晓峰,王永良,张永顺,等. 二维 MUSIC 算法分维处理及其阵列结构的研究[J]. 空军工程大学学报: 自然科学版, 2008, 9(5): 33-37.
- [2] 郭跃,王宏远,陈思捷,等. 阵列测向中的精确高速并行谱峰搜索算法[J]. 微电子学与计算机, 2007, 24(12): 50-54.
- [3] 范达,张莉,吴瑛. 利用两次奇异值分解实现二维 ESPRIT 参数配对[J]. 通信学报, 2002, 23(11): 80-85.
- [4] 殷勤业,邹理, Newcomb R. 一种高分辨率二维信号参数估计方法——波达方向矩阵法[J]. 通信学报, 1991, 4(12): 1-7.
- [5] 金梁,殷勤业. 时空 DOA 矩阵方法[J]. 电子学报, 2000, 28(6):

- [4] Hill Hall T A. BioEdit: a User-friendly Biological Sequence Alignment Editor for Windows 95/98/NT[J]. Nucleic Acids Symposium Series., 1999, 41(1999): 95-98.
- [5] Thompson J D, Clustal W. Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice[J]. Nucleic Acids Research, 1994, 22(22): 4673-4680.
- [6] Posada D, Crandall K A. Modeltest: Testing the Model of DNA Substitution[J]. Bioinformatics, 1998, 14(9): 817-818.
- [7] Huelsenbeck J P, Ronquist F. MRBAYES: Bayesian Interface of Phylogenetic Trees[J]. Bioinformatics, 2001, 17(8): 754-755.
- [8] Guindon S, Gascuel O. A Simple, Fast and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood[J]. Systematic Biology, 2003, 52(5): 696-704.
- [9] Zmasek C M, Eddy S R. ATV: Display and Manipulation of Annotated Phylogenetic Trees[J]. Bioinformatics, 2001, 17(4): 383-384.
- [10] 毛国勇, 张晓斌, 谢江. 面向生物信息学的网格问题求解平台[J]. 计算机工程, 2010, 36(11): 253-255.

编辑 任吉慧

8-12.

- [6] 金梁,殷勤业. 时空 DOA 矩阵方法的分析与推广[J]. 电子学报, 2001, 29(3): 300-303.

编辑 索书志