bio-101 www.bio-protocol.org/e1010661 DOI:10.21769/BioProtoc.1010661

PhyloSuite 在多基因系统发育分析中的应用

Application of PhyloSuite to Phylogenetic Analysis Using Concatenated Sequences

张东 [1, 2]，李文祥 [2]，高芳銮 [3, *]，王桂堂 [2, *]

[1] 生态学创新研究院，兰州大学，兰州；[2] 中国科学院水生生物研究所，武汉；[3] 植物病毒研究所，福建农林大学，福州

*通讯作者邮箱：raindy@fafu.edu.cn; gtwang@ihb.ac.cn

引用格式：张东，李文祥，高芳銮，王桂堂. (2021). PhyloSuite 在多基因系统发育分析中的应用. *Bio-101* e1010661. Doi: 10.21769/BioProtoc.1010661.

How to cite: Zhang, D., Li, W.X., Gao, F. L. and Wang, G.T. (2021). Application of PhyloSuite to Phylogenetic analysis using concatenated sequences. Bio-101 e1010661. Doi: 10.21769/BioProtoc.1010661. (in Chinese)

摘要：随着 DNA 测序技术的发展，基于单基因的系统发育分析已无法满足系统基因组学研究的需求，而多基因数据集可以增加不同分类单元的解析度，越来越多的系统生物学研究采用多基因系统发育分析。但相比较于成熟的单基因系统发育分析，多基因系统发育分析相对复杂且操作繁琐。近期开发的 PhyloSuite 可以针对基于基因组、转录组以及细胞器基因组 (线粒体基因组、叶绿体基因组等) 等多基因进行联合分析，可以实现从序列提取、序列比对、数据集分区、最佳替代模型选择、系统发育关系重建以及系统发育树美化等流程化操作，简单易用。本文将以单殖吸虫 (Monogenea) 的线粒体基因组 12 个蛋白编码基因建树为示例，介绍如何应用 PhyloSuite 进行多基因联合建树。

关键词：多基因系统发育，PhyloSuite，数据集分区，PartitionFinder，IQ-tree

研究背景

系统发育分析是系统生物学研究中一项常见的分析内容，在揭示不同生物类群（物种）间的进化关系中起着重要的作用。目前，基于单基因序列的系统发育分析方法已非常成熟，但由于不同基因的进化速率不同，仅从单基因序列获得的系统发育关系可能不可靠

或解析力相对有限 (Gontcharov *et al*., 2004)。采用联合多基因数据进行系统发育分析，可以增加系统发育信号以提高不同分类单元的解析度 (Hillis 1996; Philippe *et al*., 2011)，现已被越来越多的系统生物学研究采用。

多基因联合系统发育分析流程主要包括数据获取、基因提取、多重序列比对、比对序列裁切、多基因序列串联 (concatenation)、最适分区策略及模型选择、系统发育树构建等步骤，分析过程需要用到多款不同软件，操作相对繁琐。新近开发的分析平台 PhyloSuite (Zhang *et al*., 2020) 整合了 PartitionFinder2 (Lanfear *et al*., 2017)、IQ-TREE (Nguyen *et al*., 2014) 和 MrBayes (Ronquist *et al*., 2012) 等系统发育分析所需的相关软件，可以联合上下游分析软件，自动整理输入和输出文件，大大降低了多基因系统发育分析的门槛。同时，针对基因组、转录组以及细胞器基因组 (线粒体基因组、叶绿体基因组等) 等多基因联合分析，PhyloSuite 还进行了一系列优化，可以支持流程化操作 (只需输入文件并配置好参数，即可一键完成系统发育分析)。此外，该分析平台还可以结合 iTOL 快速实现系统发育树的美化。本文将以单殖吸虫 (Monogenea) 线粒体基因组的 12 个蛋白编码基因序列为示例，应用 PhyloSuite 进行多基因联合建树。

运行环境及下载地址

PhyloSuite 为基于 Python 编写的程序，已编译可运行于 Windows、Mac OS 和 Linux 操作系统的对应版本。为避免程序需要的插件缺失，推荐用户下载带插件的版本。当前 PhyloSuite 最新版本为 1.2.2。

1. Windows 系统：

   64 位系统：PhyloSuite_v1.2.2_Win64_with_plugins.rar

   32 位系统：PhyloSuite_v1.2.2_Win32_with_plugins.rar

2. Mac OS：PhyloSuite_v1.2.2_Mac_with_plugins.zip

3. Linux 系统：PhyloSuite_v1.2.2_Linux.tar.gz

实验步骤

一、序列下载与准备

1. 批量下载序列

   根据分类从 NCBI 的 Nucleotide 数据库搜索并下载整个单殖吸虫纲已发表的线

粒体基因组数据，如图 1 所示。

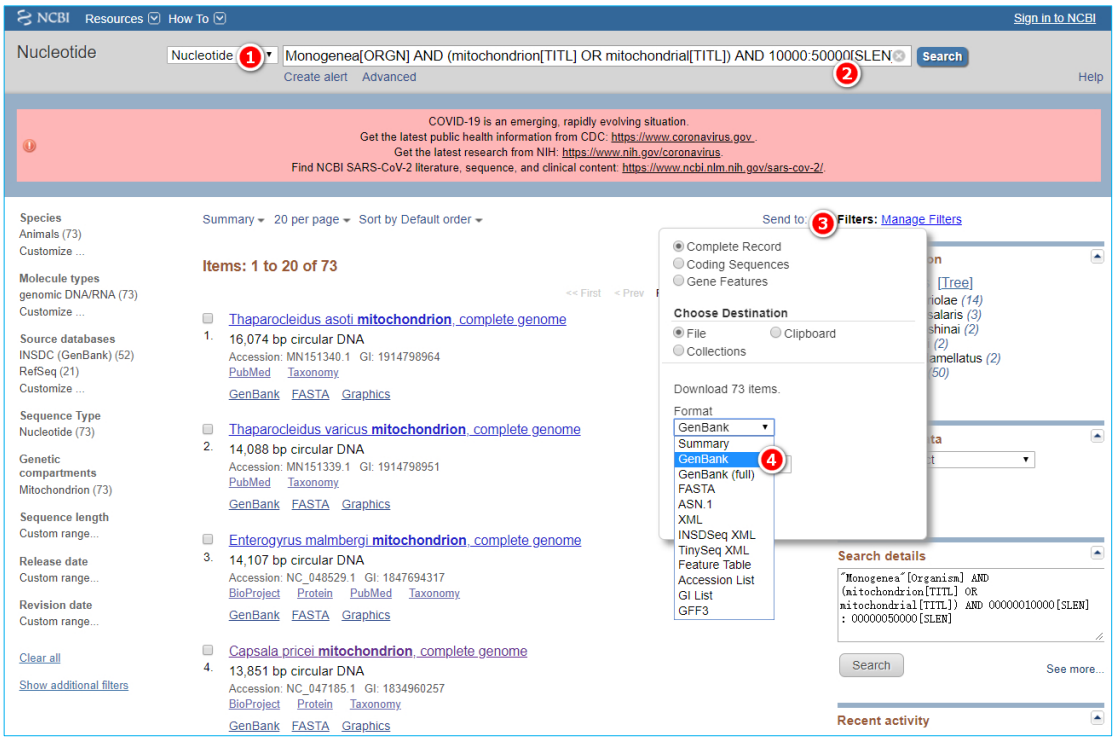

图 1. NCBI 序列数据下载界面

① 进入 NCBI 官网 (https://www.ncbi.nlm.nih.gov/)，选择 Nucleotide 数据库。

② 输入 Monogenea[ORGN] AND (mitochondrion[TITL] OR mitochondrial[TITL]) AND 10000:50000[SLEN] 进行搜索，该词条可以分解为 3 个部分，第一部分是限定分类群为单殖吸虫 (Monogenea[ORGN])，第二部分是限定 title 里出现了 mitochondrion 或者 mitochondrial ((mitochondrion[TITL] OR mitochondrial[TITL]))，第三部分是限定序列的长度在 10000 到 50000 之间 (10000:50000[SLEN])，这些限定可以根据自己的需求调整。

③ 选择 Send to 按钮保存搜索出来的所有序列。

④ 依次选择 Complete Record → File → GenBank 进行保存。

注：图中①、②… 代表对应的文字描述顺序，下同。

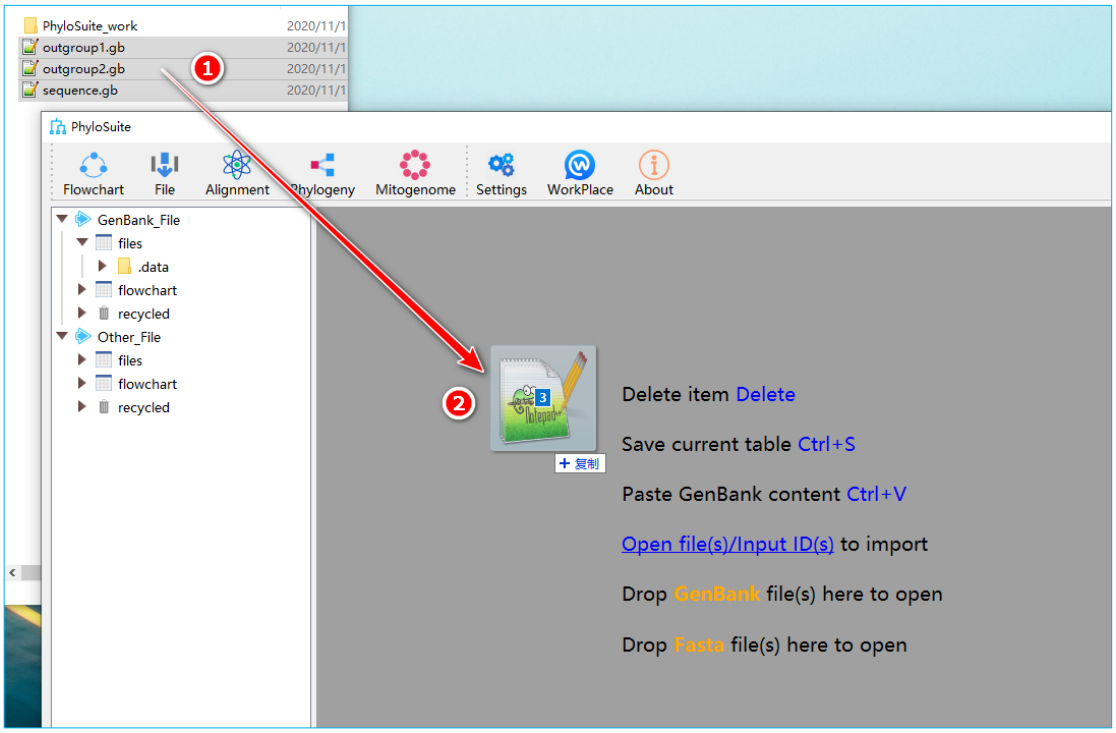

图 2. 导入下载的序列

2. 导入已下载序列

将下载的序列 (*.gb) 文件导入 PhyloSuite (图 2)，操作方法如下：

① 首先在 PhyloSuite 界面选择左侧菜单 GenBank_File 下任意选定一个工作文件夹 (files 或 flowchart)。

② 将下载好的序列文件 (*.gb) 拖入图 2 所示区域 (PhyloSuite 展示区)，程序自动完成序列导入。

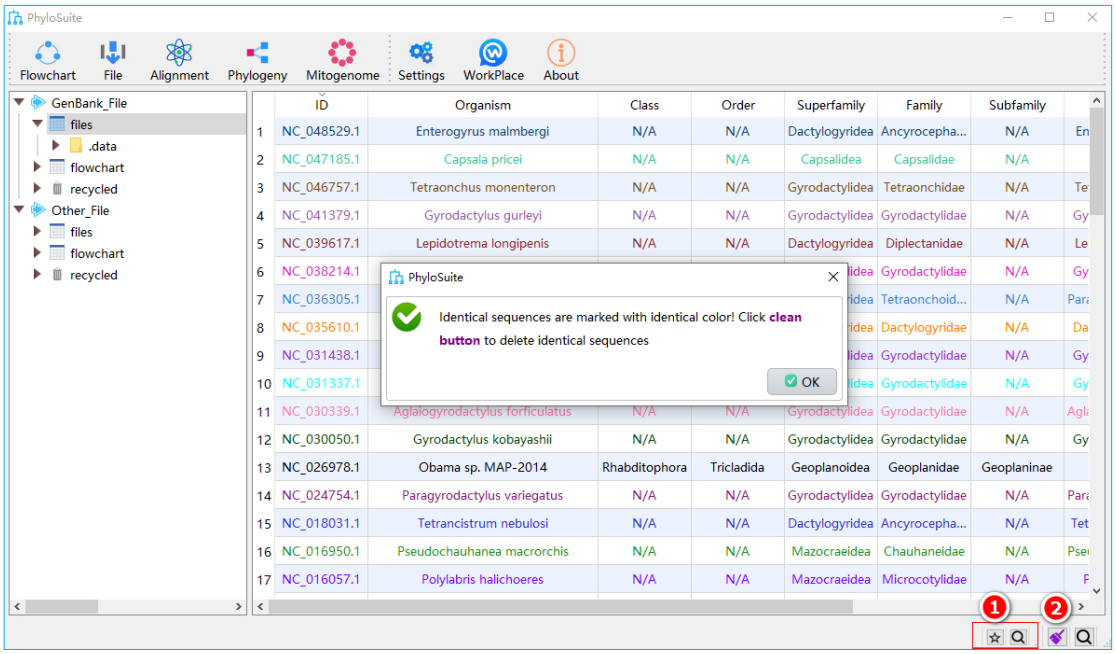

图 3. 冗余序列清洗

3. 清除冗余序列

导入序列后，可以通过右下角的序列筛查功能对冗余序列进行清除 (图 3)：

① 点击软件界面右下角的五角星按钮，会提示将相同的序列标记为一样的颜色。

② 此时五角星按钮变成了紫色的扫帚形状，只需要再点击一次这个按钮，重复的序列将会被清除 (NC 开头的序列将会被优先保留)。

上一步完成以后，同一物种还可能存在多条序列，可先对物种名排序，然后手工删除多余序列即可。分类信息可双击单元格修改或通过选中序列右键 → Get taxonomy (NCBI) 来自动获取。

二、基因提取

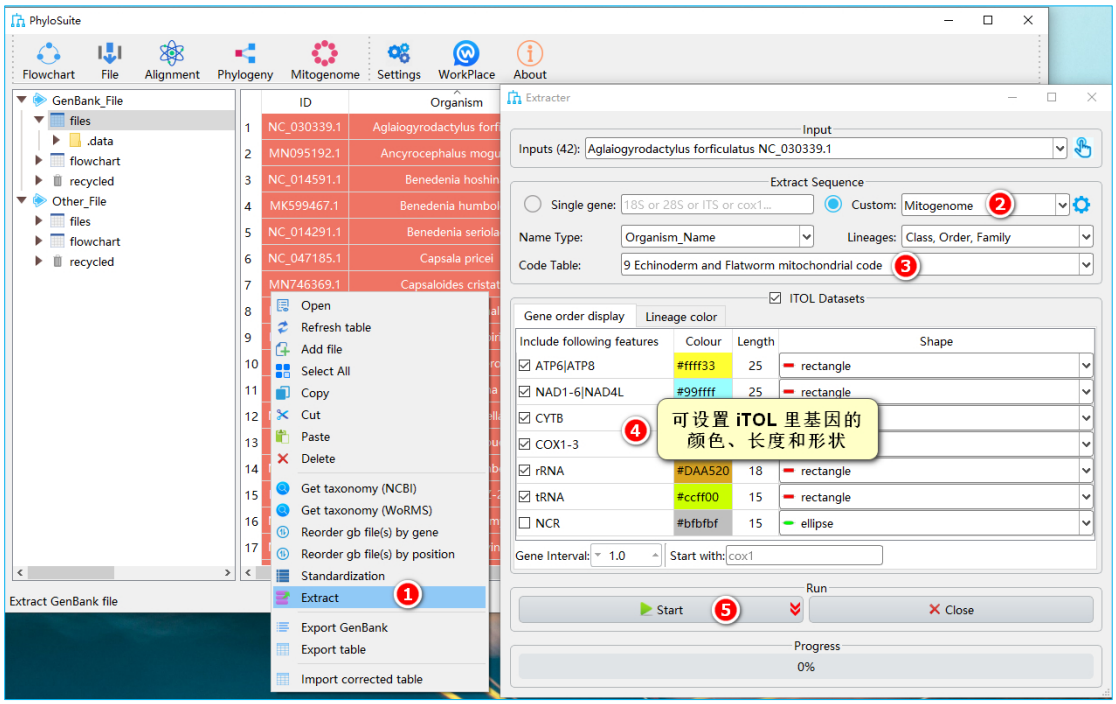

图 4. 线粒体基因组基因数据提取

导入下载序列并清除冗余序列后，可以进行基因提取操作，如图 4 所示：

① 按下组合键 Ctrl+A 全选导入的序列后，点击右键"Extract"后弹出提取界面。

② 序列类型：在"Extractor"界面中，在"Custom"下拉菜单中选择对应的序列类型。本示例数据为线粒体基因组，故选择"Mitogenome"。

③ 密码子表：选择适用于单殖吸虫的第 9 套密码子 (The Echinoderm and Flatworm Mitochondrial Code)，用于正确识别终止密码子以及翻译核苷酸序列。

④ iTOL 美化参数设置：在 ITOL Datasets 标签下，可以设置基因顺序展示图中不同基因的形状、颜色以及长度等，也可以定义不同分类的颜色。

⑤ 参数设置比，点击"Start"按钮开始提取基因数据。

三、多重序列比对及其优化

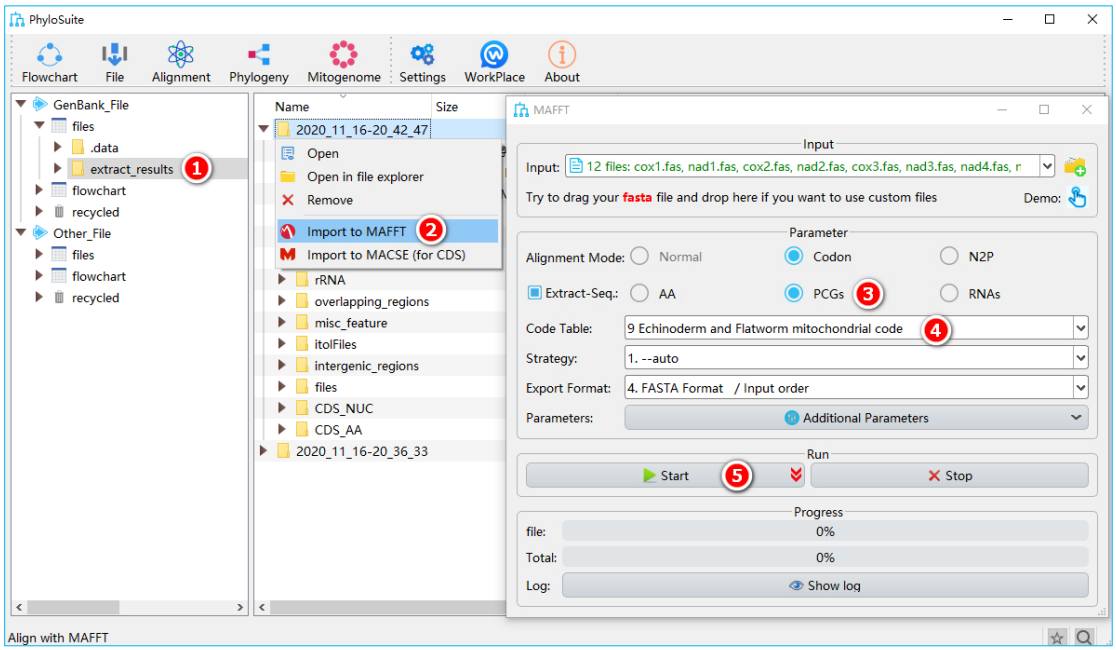

图 5. 基于 Codon 模式的多重序列比对

1. 多重序列比对

上一部提取的基因序列会自动保存在"extract_results"中以提取开始时间命名的文件夹内，可以导入 MAFFT 进行多重序列比对，如图 5 所示。

① 选中上一步的 extract_results 文件夹中的结果文件夹。

② 鼠标右键选定该结果文件夹 → Import to MAFFT。

③ Extract-Seq 选框将自动选中 (此模式仅导入提取结果至 MAFFT 时自动启用)，切换 AA (氨基酸)/PCGs (蛋白编码基因)/RNAs (tRNA 和 rRNA 基因) 按钮将自动导入对应序列，并选中相应的比对模式，如此处选中 PCGs，Codon 模式将自动启用。

④ 同样地，密码子表选择适用于单殖吸虫的第 9 套密码子。

⑤ 点击"Start"按钮开始多重序列比对。

2. 基于密码子算法的多重序列比对优化（可选）

与 MEGA (by Codon) (Kumar *et al.*, 2018) 和 TranslatorX (Abascal *et al.*, 2010) 等软件的比对原理相似，PhyloSuite 中 MAFFT 的 Codon 比对方法也是先将核苷

酸序列翻译为氨基酸进行比对，完成后再回译为对应的核苷酸序列，但是这种方式并没有使用真正的密码子比对算法，因此最好用 MACSE (Ranwez *et al.*, 2018; Ranwez *et al.*, 2011) 的密码子比对算法进行优化，如图 6 所示。

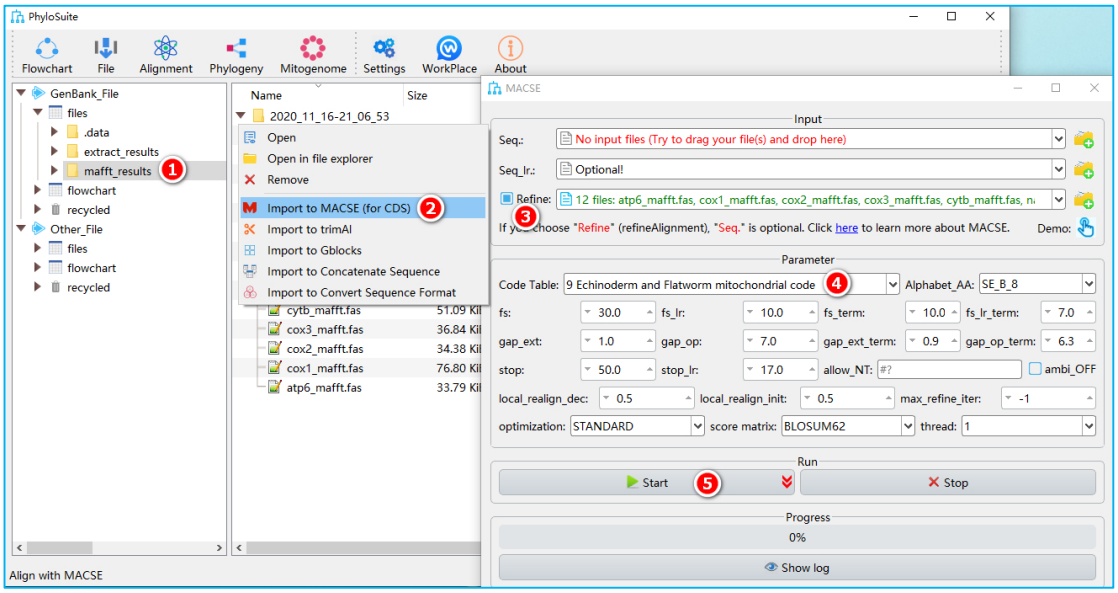

图 6. 应用 MASCE 进行多重序列比对的优化

① 选中上一步的 mafft_results 中的结果文件夹。

② 右键结果文件夹 → Import to MACSE (for CDS)。

③ Refine 选框将自动选中，MAFFT 比对结果被自动导入，代表对它们进行优化。

④ 密码子表选择适用于单殖吸虫的第 9 套密码子。

⑤ 点击"Start"按钮开始优化。

注：由于应用 MACSE 进行多重序列比对速度比较慢，在 MAFFT 的 Codon 模式比对基础上，再应用 MACSE 优化，可以大量节省多重序列比对时间。由于在结果文件中 MACSE 会使用"！"或""符号标记检测到的移码突变，而这些符号会影响下游的分析，因此除 MACSE 原本的输出文件以外（带"_NT"和"_AA"字样），PhyloSuite 还将额外生成替换这些特殊字符为"？"的文件（带"removed_chars"字样），这些替换后的文件将被用于下游分析。*

DOI:10.21769/BioProtoc.1010661

四、多重比对序列裁切优化

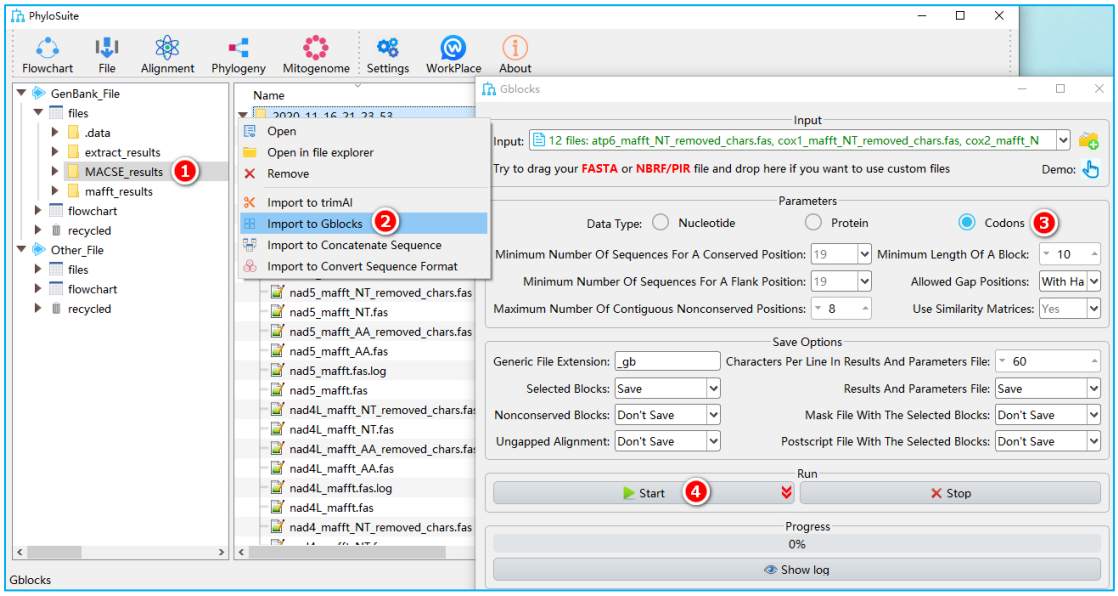

图 7. 多重序列的裁切优化

对于信息位点足够多的序列，经裁切处理后可以获得优化后的比对序列 (保守区)。但对于信息位点较少的短序列，本步骤（图 7）可以忽略。

① 选中上一步的 MACSE_results 中的结果文件夹。

② 右键结果文件夹 → Import to Gblocks，MACSE 比对结果将被自动导入。

③ 选择 Codons 模式。

④ 点击"Start"按钮开始裁切优化。

五、多基因序列串联

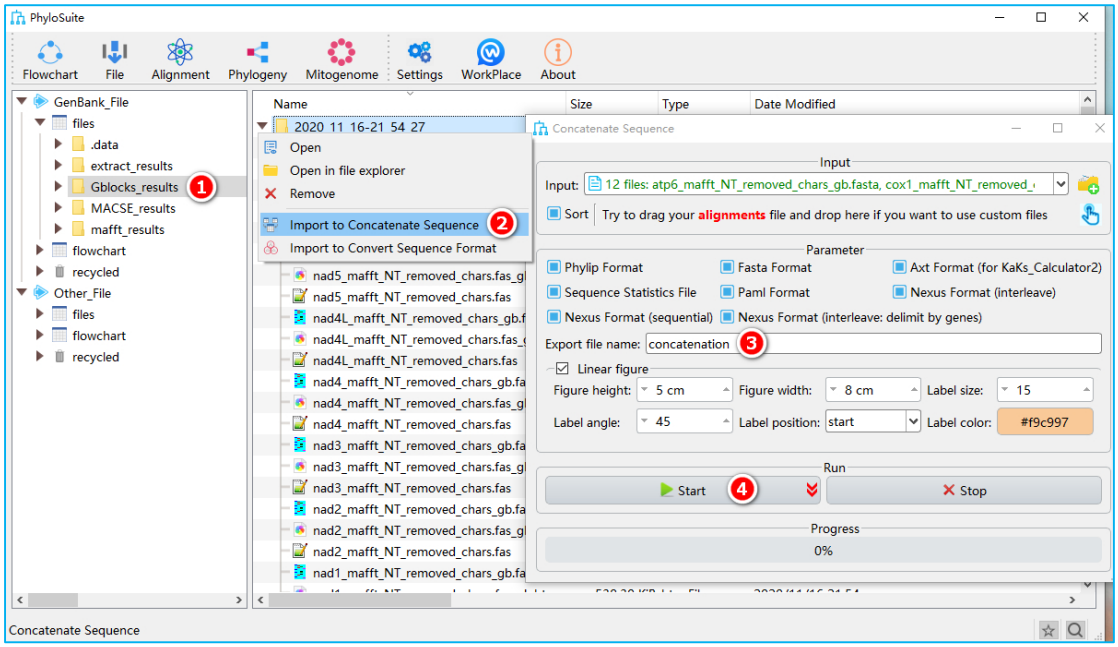

图 8. 多基因序列串联

经裁切优化后的多基因比对序列，可以应用 PhyloSuite 中的"Concatenate sequence"功能进行多基因序列串联，如图 8 所示：

① 选中上一步的 Gblocks_results 中的结果文件夹。

② 右键结果文件夹 → Import to Concatenate Sequence，Gblocks 结果将被自动导入。

③ 设置串联后序列的名字，默认为 concatenation。

④ 点击 "Start"按钮开始处理。

六、数据集分区及最优模型选择

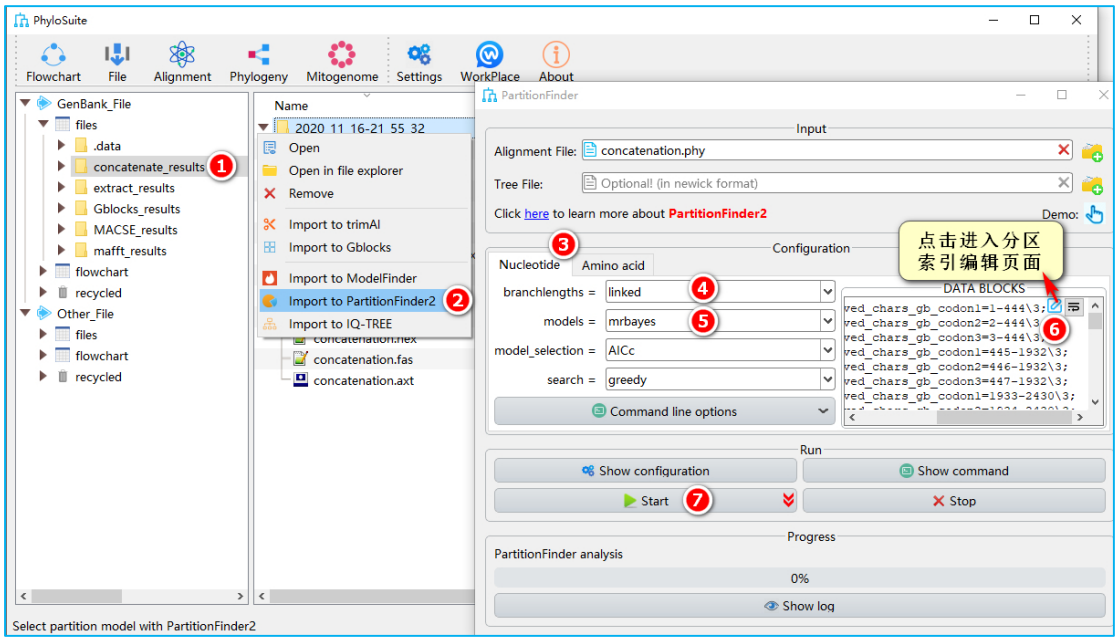

图 9. 数据集分区及模型选择

根据数据集分区的对象不同，多基因系统发育分析通常有基于基因的数据集分区和基于密码子的数据集分区两种方式(Nylander *et al.*, 2004; Shapiro *et al.*, 2006)。前者是将具有相似进化特征的基因数据汇合形成多个数据子集，而后者是提取密码子第 1 位、第 2 位和第 3 位序列汇合为不同数据子集。

① 选中上一步的 concatenation_results 中的结果文件夹。

② 右键结果文件夹 → Import to PartitionFinder2，串联结果将被自动导入，基因索引将会自动配置。

③ 选择 Nucleotide 模式。

④ Branchlengths 参数建议选 linked，因为在 unlinked 模式下，MrBayes 建树将会每一个分区生成一棵树，且不易收敛。

⑤ models 选项主要根据后续建树软件进行选择，如后续使用 MrBayes 进行多基因系统发育分析，此选项选择 MrBayes 即可，这时程序只计算 MrBayes 可支持的替代模型。IQ-TREE（最大似然法）的最优分区策略及最优替代模型由其自带的 ModelFinder 完成（见第八步）。

⑥ 双击文本框或点击编辑按钮可以在分区索引编辑页面进行检查修改。

⑦ 参数配置完毕，点击"Start"按钮开始数据分区及最优模型选择。

七、贝叶斯法 (BI) 重建多基因系统发育树

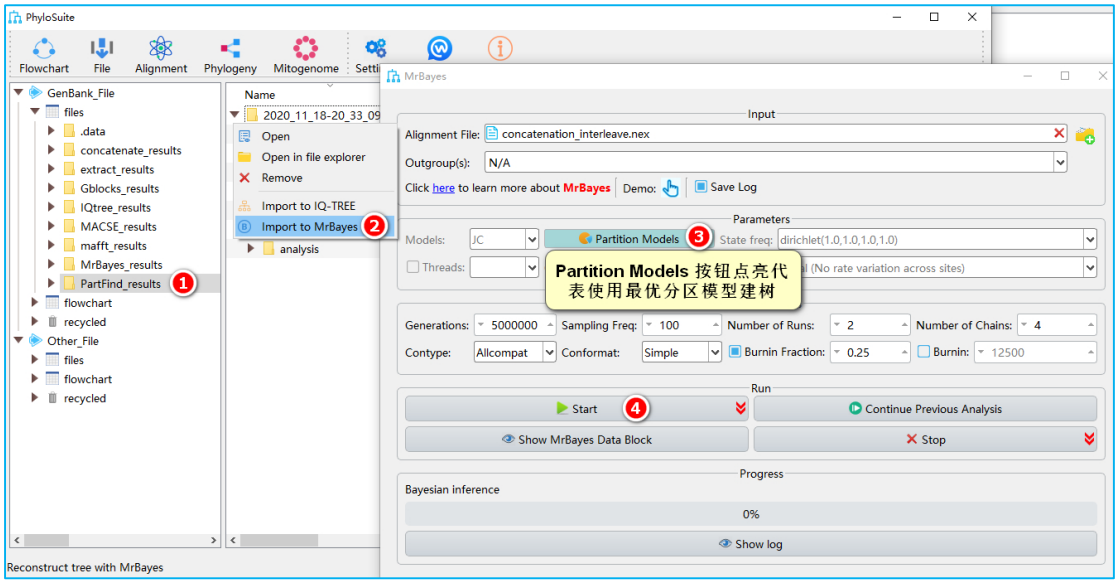

图 10. 贝叶斯法重建多基因系统发育树

应用 MrBayes 基于贝叶斯算法进行分析，可以将上一步数据集分区及最优模型选择的结果导入到 MrBayes 中，如图 10 所示：

① 选中 PartFind_results 中的结果文件夹。

② 右键结果文件夹 → Import to MrBayes，最优分区结果将自动导入并点亮 Partition Models 按钮。

③ 双击 Partition Models 按钮可查看分区配置 (图 11)。

④ 确认分区配置参数正确后，点击"Start"按钮开始贝叶斯分析。

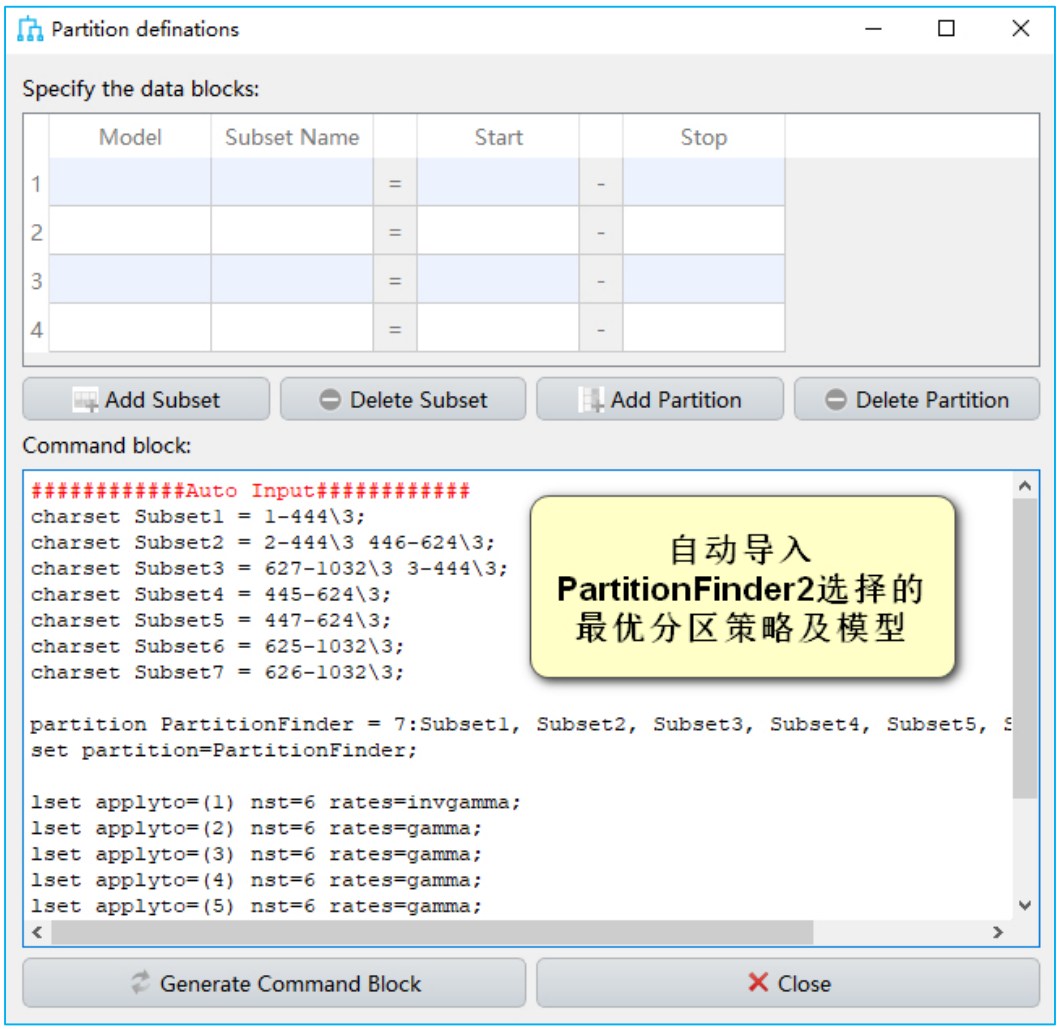
Partition defination
Specify the data blocks:
Model
Subset Name
Start
Stop
1
2
3
4
Add Subset
Delete Subset
Add Partition
Delete Partition
Command block:
############Auto Input############
charset Subset1 = 1-444\3;
charset Subset2 = 2-444\3 446-624\3;
charset Subset3 = 627-1032\3 3-444\3;
charset Subset4 = 445-624\3;
charset Subset5 = 447-624\3;
charset Subset6 = 625-1032\3;
charset Subset7 = 626-1032\3;
partition PartitionFinder = 7:Subset1, Subset2, Subset3, Subset4, Subset5,
set partition=PartitionFinder;
lset applyto=(1) nst=6 rates=invgamma;
lset applyto=(2) nst=6 rates=gamma;
lset applyto=(3) nst=6 rates=gamma;
lset applyto=(4) nst=6 rates=gamma;
lset applyto=(5) nst=6 rates=gamma;
自动导入
PartitionFinder2选择的
最优分区策略及模型
Generate Command Block
Close

图 11. 数据集分区模块

八、最大似然法 (ML) 重建系统发育树

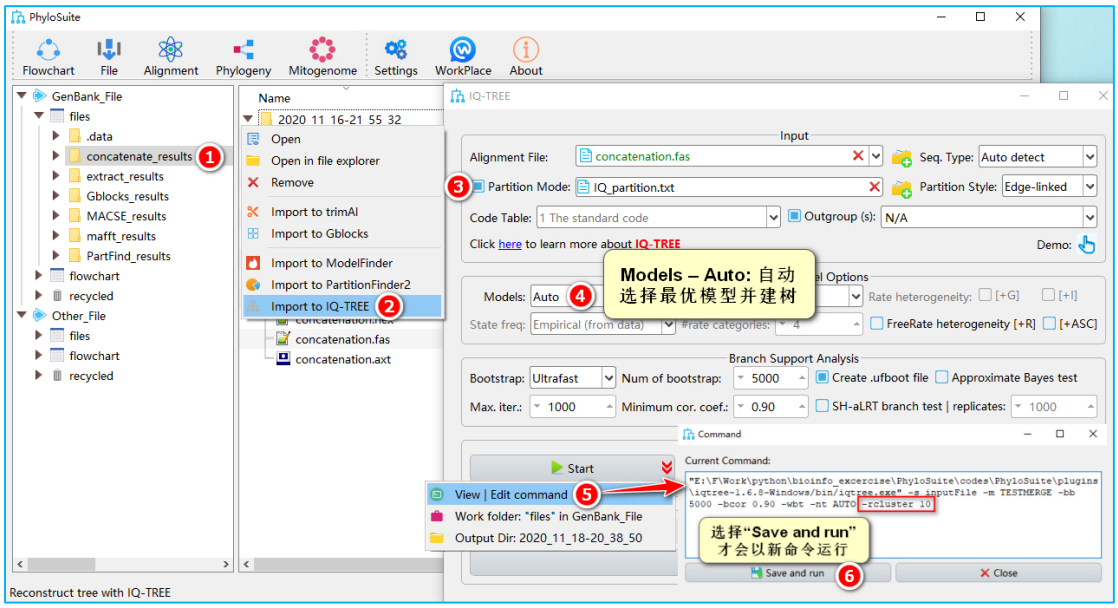

图 12. 最大似然法重建多基因系统发育树

应用 IQ-tree 基于最大似然法重建多基因系统发育树，可以将数据集分区及最优模型选择的结果导入到 IQ-TREE 中，如图 12 所示：

① 选中 concatenation_results 中的结果文件夹。

② 右键结果文件夹 → Import to IQ-TREE，串联结果 (concatenation.fas) 及分区文件 (IQ_partition.txt) 将被自动导入。

③ 勾选 Partition Mode，开启分区建树模式。

④ Models 选择 Auto 代表自动选择最优 (分区) 模型并建树。

⑤ 点开 Start 按钮右边的红色箭头 → 弹出下拉菜单 → View | Edit command，在命令末尾添加 -rcluster 10 命令 (注意要与旧命令以空格隔开)，该命令可通过只检查 top 10% 的分区合并策略以提升运算速度，适用于大数据。

⑥ 点击 Save and run 以新命令运行。

九、系统发育树美化

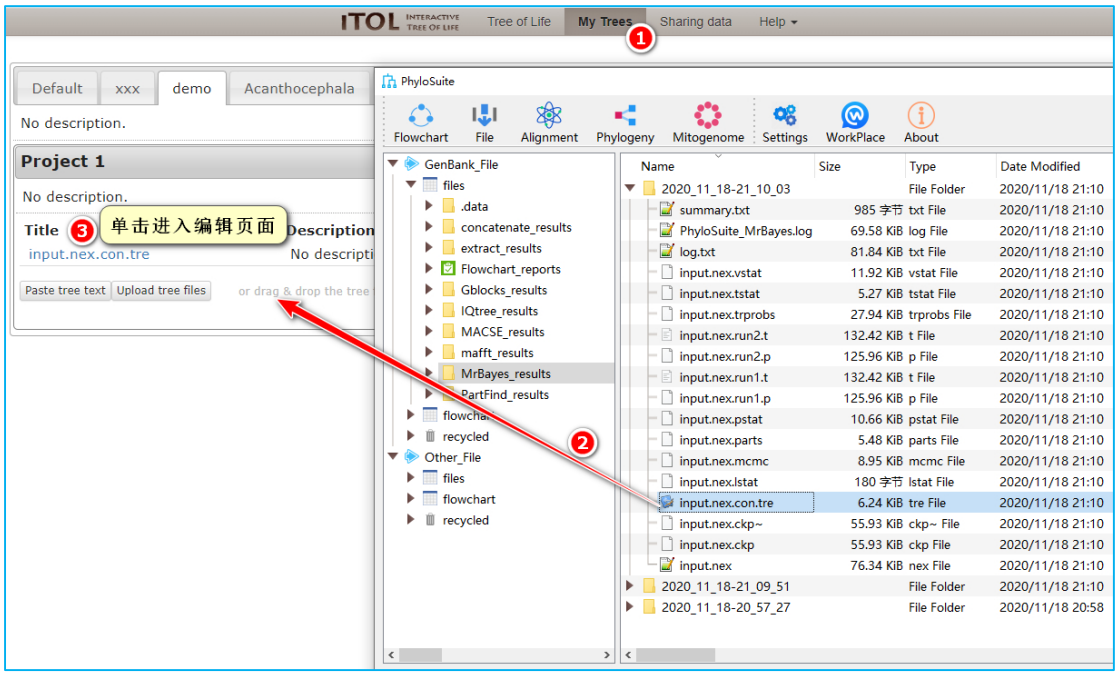

图 13. 系统发育树导入 iTOL 美化

iTOL 是一款集显示、操作和注释系统发育树于一体的的在线网络工具，PhyloSuite 中整合了 iTOL 参数模块，多基因系统发育树的美化只需要简单的几步操作 (如图 13 和图 14)：

① 打开 iTOL 主页 (https://itol.embl.de/)，登录自己的账号，然后选择 My Trees。

② 以 MrBayes 为例，选中 MrBayes_results 结果文件夹，从建树结果中找到 *.con.tre 文件拖入图示区域即自动导入（注意 iTOL v6 的导入界面有所变化，需要点开 Tree upload，然后将树拖拽至 Tree files 处即可完成导入）。

③ 单击该树，进入编辑界面。

④ 选中 extract_results 中的结果文件夹 (图 14)。

⑤ 从结果文件夹中找到 itolFiles 文件夹。

⑥ 将需要用于美化的文件拖拽导入 iTOL 界面即可完成美化。最后根据自己喜好稍作调整后，即可将美化好的系统发育树以图片格式 (如 SVG) 导出。

DOI:10.21769/BioProtoc.1010661

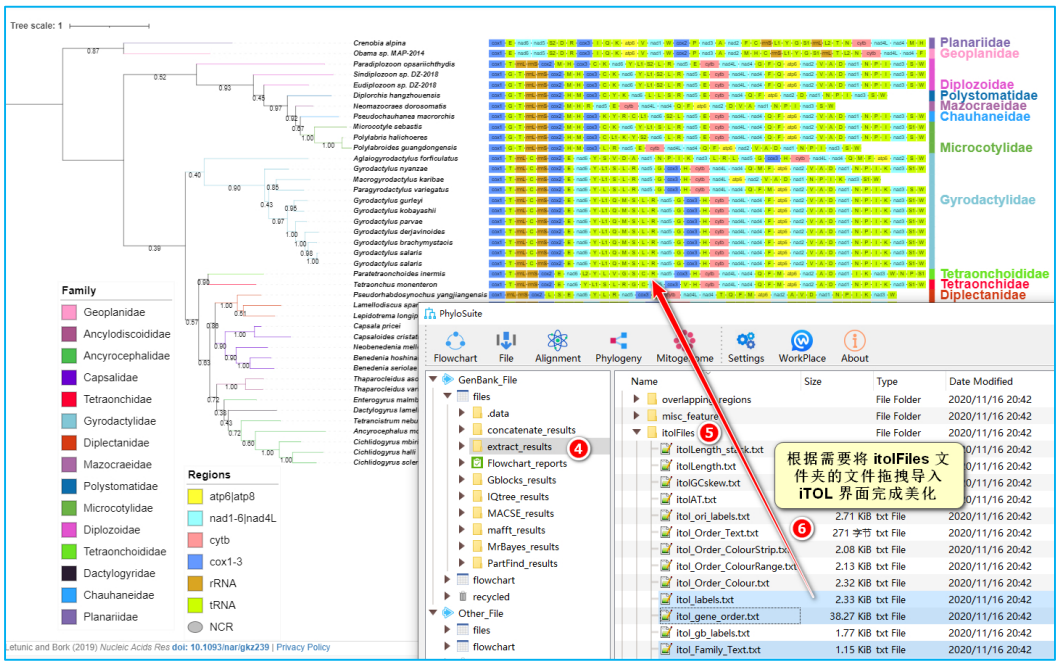

图 14. iTOL 中系统发育树美化参数设置

PhyloSuite 的视频教程见：https://www.bilibili.com/video/BV11h411X7LL/

注意事项

1. PhyloSuite 中的 MAFFT 程序不支持含有中文字符文件或路径，分析时需要使用英文命名文件和路径。
2. 应用多基因数据集进行分析，需要考虑数据集之间的不相合性 (incongruence)，因此分析前最好进行不相合长度差异检验 (Incongruence length difference test , ILD Test) 或同质性检验 (Partition homogeneity test, PHT) (Farris, 1994; Swofford, 1998)。不论 ILD 或 PHT 检验，均以 $p = 0.05$ 作为数据集能否联合的阈值，但也有相关研究提出，即使 $p < 0.05$ 甚至 $p < 0.01$ 时，数据集照样可以进行联合分析，因此具体问题需要结合具体情况进行分析 (黄原，2012)。

致谢

特别感谢广大 PhyloSuite 用户的反馈意见使得程序不断完善。

竞争性利益声明

作者声明没有利益冲突

参考文献

1. 黄原. (2012). *分子系统发生学*. 科学出版社. 北京.
2. Abascal, F., Zardoya, R. and Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* 38: W7-13.
3. Farris J. S., Källersjö, M., Kluge, A. G. and Bult, C. (1994). Testing significance of incongruence. *Cladistics* 10: 315-319.
4. Gontcharov, A. A., Marin, B. and Melkonian, M. (2004). Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and *rbc*L sequence comparisons in the Zygnematophyceae (Streptophyta). *Mol Biol Evol* 21: 612–624.
5. Hillis, D. M. 1996. Inferring complex phylogenies. *Nature* 383:130-131.
6. Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35: 1547-1549.
7. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. and Calcott, B. (2017). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol* 34: 772-773.
8. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. and Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol Biol Evol* 32: 268-274.
9. Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P., Nieves-Aldrey, J. L. (2004). Bayesian phylogenetic analysis of combined data. *Syst Biol* 53:47-67
10. Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *Plos Biol* 9: e1000602.
11. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. and Delsuc, F. (2018). MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol Biol Evol* 35: 2582-2584.

12. Ranwez, V., Harispe, S., Delsuc, F. and Douzery, E. J. P. (2011). MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *Plos One* 6: e22594.
13. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61: 539-542.
14. Shapiro, B., Rambaut, A., Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences, *Mol Biol Evol* 23: 7-9.
15. Swofford, D. L. (1998). *Phylogenetic analysis using parsimony (*and other methods)*. Version 4. Sinauer Associates. Sunderland, Mass.
16. Zhang, D., Gao, F., Jakovlic, I., Zou, H., Zhang, J., Li, W. X. and Wang, G. T. (2020). PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol Ecol Res* 20: 348-355.