bio-101 www.bio-protocol.org/e1010617 DOI:10.21769/BioProtoc.1010617

使用 BEAST 2 开展贝叶斯支端定年分析

Using BEAST 2 for Bayesian Tip Dating

罗阿蓉 [1,*]，张驰 [2]，朱朝东 [1]

[1]动物进化与系统学院重点实验室，中国科学院动物研究所，北京；[2]脊椎动物演化与人类起源重点实验室，中国科学院古脊椎动物与古人类研究所，北京

*通讯作者邮箱：luoar@ioz.ac.cn

引用格式：罗阿蓉，张驰，朱朝东. (2021). 使用 BEAST 2 开展贝叶斯支端定年分析. Bio-101 e1010617. Doi: 10.21769/BioProtoc. 1010617.

How to cite: Luo, A., Zhang, C. and Zhu, C. D. (2021). Using BEAST 2 for Bayesian Tip Dating. Bio-101 e1010617. Doi: 10.21769/BioProtoc. 1010617. (in Chinese)

摘要：生物类群的历史分异时间长期是生物学研究的基础和热点问题。相比经典的贝叶斯节点定年，新近提出的贝叶斯支端定年在理论上具有诸多优势。特别配合石化生灭过程模型的使用，贝叶斯支端定年有望能更准确地推断类群的历史分异时间。借助 BEAST 2 等软件包，本文较为详细地展示贝叶斯支端定年的基本操作步骤和要点，希望对该方法的使用者提供一定借鉴。

关键词：贝叶斯支端定年，BEAST 2，分异时间，分子钟

研究背景

生物类群的历史分异时间一直是生物学研究的基础和热点问题。但由于逝去历史的诸多不确定性，如何准确推断分异时间长期以来也面临各种挑战。伴随分子数据的涌现、化石信息的积累、以及演化模型 (如分子钟) 的发展，贝叶斯定年法近些年得到了广泛关注和应用。该方法将置换模型、演化速率、类群分异过程等纳入一个整体的统计学分析框架中，将化石信息纳入先验模型，从而估算类群的绝对历史分异时间。

常用的贝叶斯定年方法可以分为两类，一类是节点定年 (node dating)，另一类是支端定年 (tip dating)，其主要区别在于如何利用化石的信息。节点定年法把化石信息转化为概率分布来校准系统发生树上的部分内部节点，继而估计其它内部节点的分异时间

(e.g., Drummond *et al.*, 2006)。这一类方法通常应用于仅包含现生类群的分析。支端定年法则直接利用化石信息，把化石和现生类群同时作为系统发生树的支端进行分析，从而估计树中内部节点的分异时间 (Pyron, 2011; Ronquist *et al.*, 2012)。

支端定年相比节点定年在理论方面存在诸多优势。比如：首先，支端定年可以充分利用化石和现生类群的形态数据，从而可以开展全证据支端定年 (total-evidence tip dating) 分析。由于一般无法获得化石的分子数据，因此化石的系统发生位置由形态数据决定。其次，支端定年可以潜在利用与所研究类群相关的所有化石记录，无需像节点定年仅挑选与某节点有关的最古老的化石记录。第三，支端定年只需根据化石自身年代信息设置时间校准先验，无需对某节点的历史分异时间特别挑选统计分布以描述其时间先验。特别伴随描述类群分异及化石采样过程的石化生灭过程 (fossilized birth-death process, FBD) 模型的提出 (Stadler, 2010)，支端定年得到了广泛应用。具体分析还可根据实际情况针对各历史时间段设置不同分异和采样速率，以及根据现生类群取样情况设置多样化采样策略等 (Gavryushkina *et al.*, 2014; Zhang *et al.*, 2016)。多项研究提示：贝叶斯支端定年在推断类群历史分异时间方面具有良好的准确性和精确性 (如 Gavryushkina *et al.*, 2014; Zhang *et al.*, 2016; Luo *et al.*, 2020)。

借助 BEAST 2 软件包 (Bouckaert *et al.*, 2019) 及 Tracer (Rambaut *et al.*, 2018)、FigTree (http://tree.bio.ed.ac.uk/software/figtree/) 等软件，本文将展示贝叶斯支端定年的基本操作步骤和要点信息。

仪器设备

BEAST 2、Tracer、Figtree 在 Windows、Linux、macOS 等主流计算机操作系统下均可运行，界面基本一致。由于篇幅所限，本文仅展示 macOS Big Sur (11.0.1) 系统下的操作界面和流程。

软件版本信息及下载地址

BEAST 是根据马尔可夫链蒙特卡罗 (MCMC) 算法开展贝叶斯进化分析的开源免费软件。其虽然主要用于推断生物类群的历史分异时间，但也可以用于构建系统发生树、重建类群祖先性状、估计种群大小、实施模型选择等等 (Bouckaert *et al.*, 2019)。和 BEAST 1 不同，BEAST 2 采用崭新的架构编写，突出模块化，可以通过加载模块或软件包从而

拥有多种功能。BEAST 2 软件包中除主要实施 MCMC 分析的 BEAST 程序外，还拥有 BEAUti、LogComiber、TreeAnnotator 等多个程序可以在前期参数设置、后期数据整理等方面提供帮助。

BEAST 2 软件包可以从官方网站 https://www.beast2.org 下载最新版本。其安装步骤请参照下载文件夹中的 README.txt 文件。本文采用撰写时的最新版本 BEAST v.2.6.3。BEAST 2 的正常运行依赖于 Java v8 或更高版本。因此，需首先安装 Java 以保证 BEAST 2 的正常运行。Tracer 和 FigTree 在本文用于将 BEAST 分析结果可视化，从而易于对结果进行判定和解析。其分别可以从 https://www.beast2.org/tracer-2/和 http://tree.bio.ed.ac.uk/software/figtree/下载最新版本。本文采用 Tracer v1.7.1 和 FigTree v1.4.4。

实验步骤

1. 实验数据

出于演示目的，本文采用计算机模拟的数据，其源于 Luo *et al.* (2020)，包含 50 个现生类群 (或物种) 和 7 个化石类群 (或物种)。分子数据 (基因序列) 部分包含 5 个分区 (partition)，分别为 1_1st，1_2nd，1_3rd，1_4th，1_5th，化石物种无分子数据，由"?"替代。形态数据部分由"0"或"1"代表化石和现生类群的离散性状特征 (图 1，附件信息 https://github.com/ArongLuo/Protocol-)。

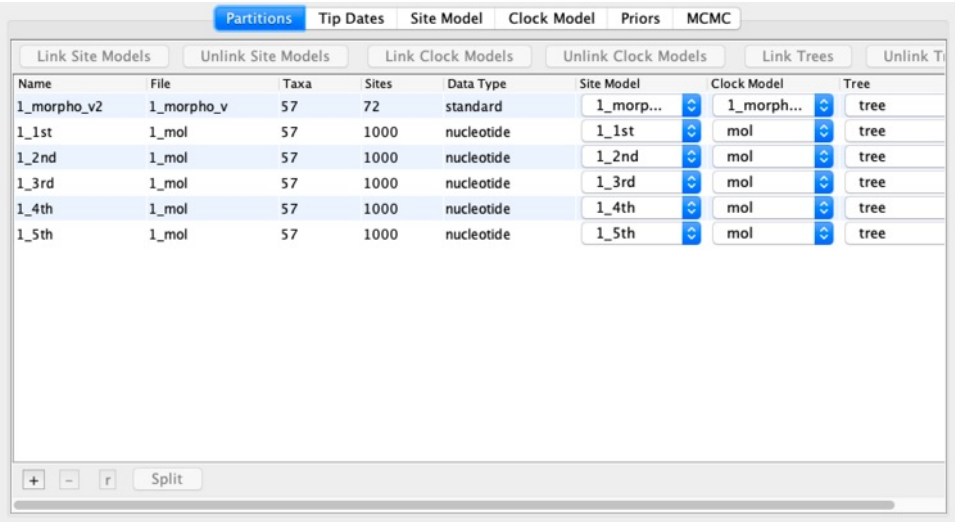

图 1. BEAUti 加载数据及初步设置后所示

2. BEAUti 参数设置

2.1 在 BEAST 2 安装目录 (如/Applications/BEAST 2.6.3/) 中双击 BEAUti 图标，打开 BEAUti。

2.2 基本的全证据支端定年分析依赖模块 SA (sampled ancestor) 和 MM (morphological model)，所以需要首先加载安装。在 File 中单击 Manage Packages，在软件包列表中查找 SA 和 MM 并进行安装 (图 2)。

2.3 在 BEAUti 中通过 Import Alignment 和 Add Morphological Data 分别打开分子和形态数据。假设所有的数据源于相同的类群演化过程，因此对 6 个数据分区 (分子数据 5 个分区和形态数据) Link Trees，并命名为 tree；假设形态和分子数据具不同的进化速率，则只对分子数据的 5 个分区 Link Clock Models，并分别对分子和形态部分命名为 mol 和 1_morpho_v。由于描述各分区的碱基置换模型 (或特征变换模型) 不同或参数相对独立，则对各分区 Unlink Site Models (图 1)。

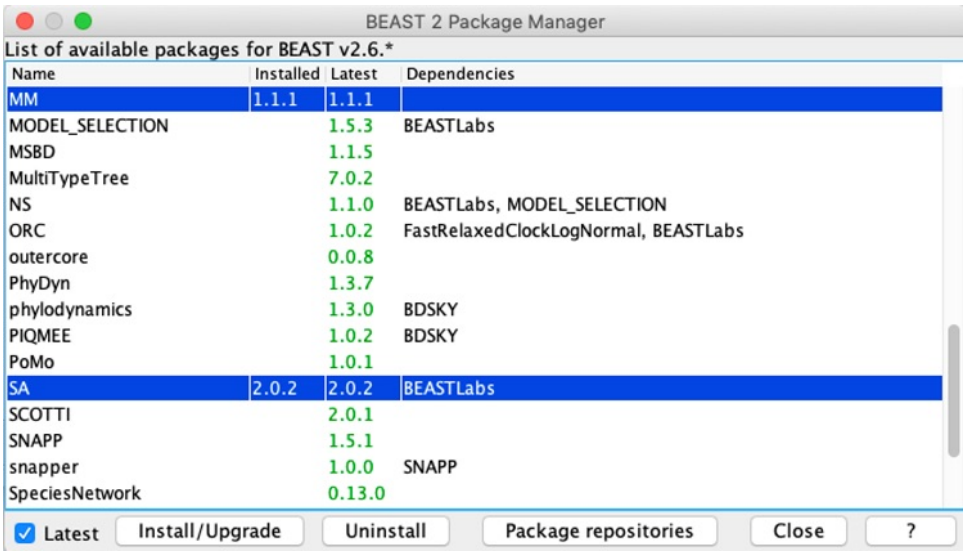

图 2. BEAST 2 软件包列表及所需安装的 SA 和 MM

2.4 点击 BEAUti 主选项 Tip Dates，选中 Use tip dates。在准备该数据文件时，已经将序列名称"_"后设置为类群的历史时间，如 t9_0 代表现生物种 t9，et50_84 代表存在于 84 百万年左右的化石物种 et50。在时间单位为年的假设下，选择 Before the present，并实际以百万年为真实时间单位。根据序列名称特点，在

此通过 Auto-configure 增加各物种的时间信息 (图 3)。值得注意的是，类群的时间也可以手动填写或修改，后续对速率先验的设置需与真实的时间单位保持一致。

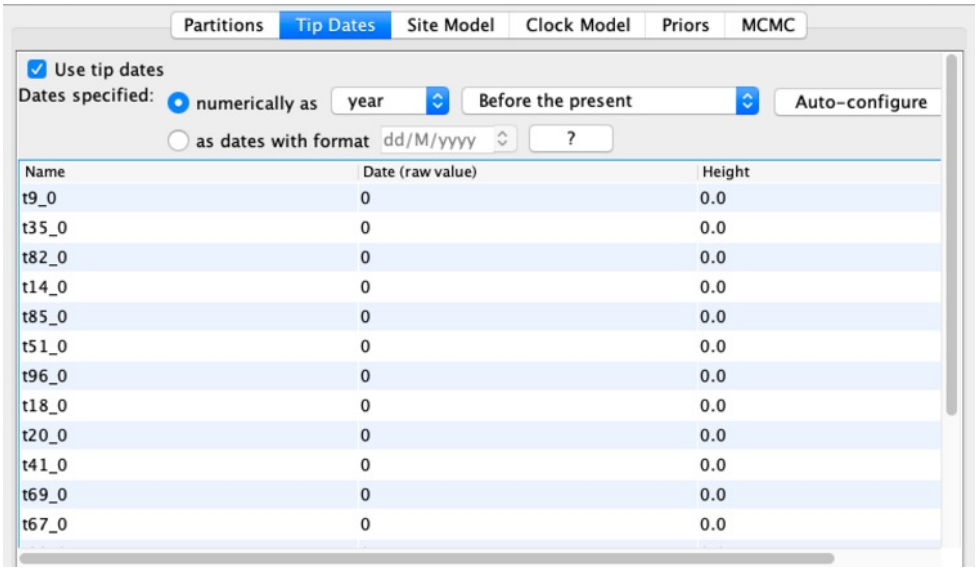

图 3. 在 BEAUti 中设置各物种的历史时间信息

2.5 点击 BEAUti 主选项 Site Model 开始位点变换模型设置。在此，分子数据各分区均设为 HKY + G，形态数据采用 MK 模型 (图 4)。

2.6 点击 BEAUti 主选项 Clock Model 开始演化速率设置。对形态和分子数据分别采用宽松钟模型，可具体设为 Relaxed Clock Log Normal。该模型假设演化速率服从独立的对数正态分布。

2.7 点击 BEAUti 主选项 Priors 开始先验设置。多数参数可使用其默认先验，但 1) 需特别选择 FBD 模型作为支端定年的树先验，并可通过 Condition On Rho Sampling 使树先验受限于已知的现生类群取样百分比 Rho，本数据分析中 Rho 为 1 (图 5)；2) 需考虑 FBD 模型所涉参数的先验设置，如无明确信息，可对 diversificationRate、samplingProportion、turnover 分别采用信息较弱的统计分布如 exponential (10)、beta (1,1)、beta (1,1)；3) FBD 模型一般基于类群根部时间或起源时间，在此对类群起源时间设置先验 Uniform (94,300)，其中 94 是所有 7 个化石物种历史时间的最大值。4) 在已选宽松分子钟模型基础上设置速率先验，将 ucldMean 设置为 uniform (10e-6, 1)，ucldStdev 保留默认。值得注

意的是，由于各类群年龄在 2.4 中是以百万年为真实单位，速率在此以每百万年为单位。

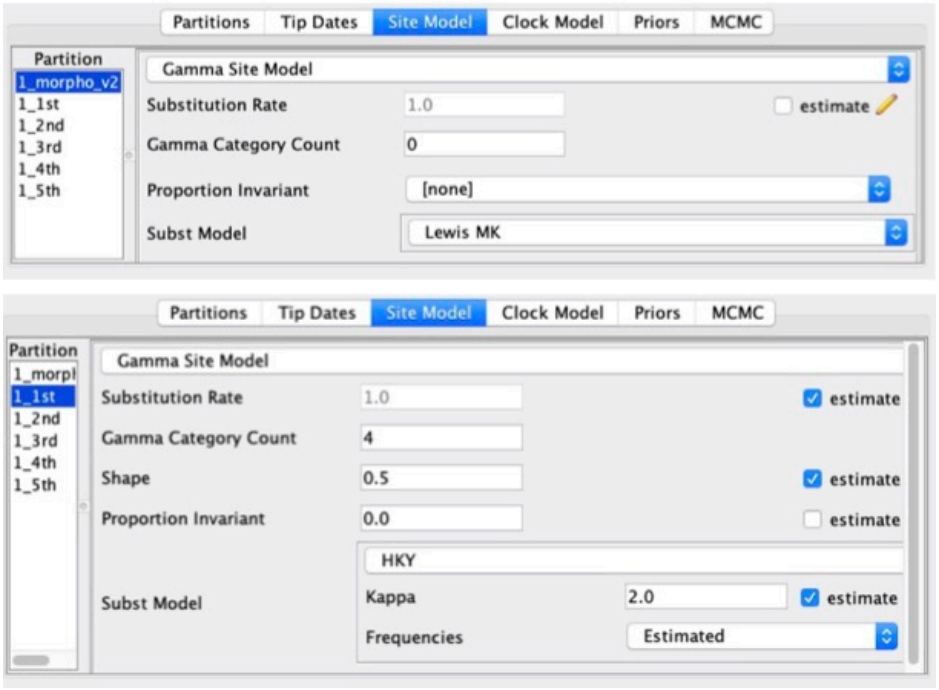

图 4. 在 BEAUti 中对形态和分子数据分别设置碱基置换模型 HKY + G 和特征变换模型 Mk

2.8 点击 BEAUti 主选项 MCMC 开始 MCMC 参数设置，可包括代数、记录频率、文件名称等。在此设 Chain Length 为 100 百万代，树文件和文本文件的 Log Every 是 5000。上述各项设置完成后，即可保存.xml 文件。

3. BEAST 数据运行

在 BEAST 2 安装目录中双击 BEAST 图标，打开 BEAST 后在弹出页面通过 Choose File 加载之前保存成功的.xml 文件，点击 Run。如无意外，BEAST 即开始 MCMC 分析 (图 6)。为了保证结果的可靠性，一般需要对同一个数据至少独立运行两次以保证后验概率得到收敛。

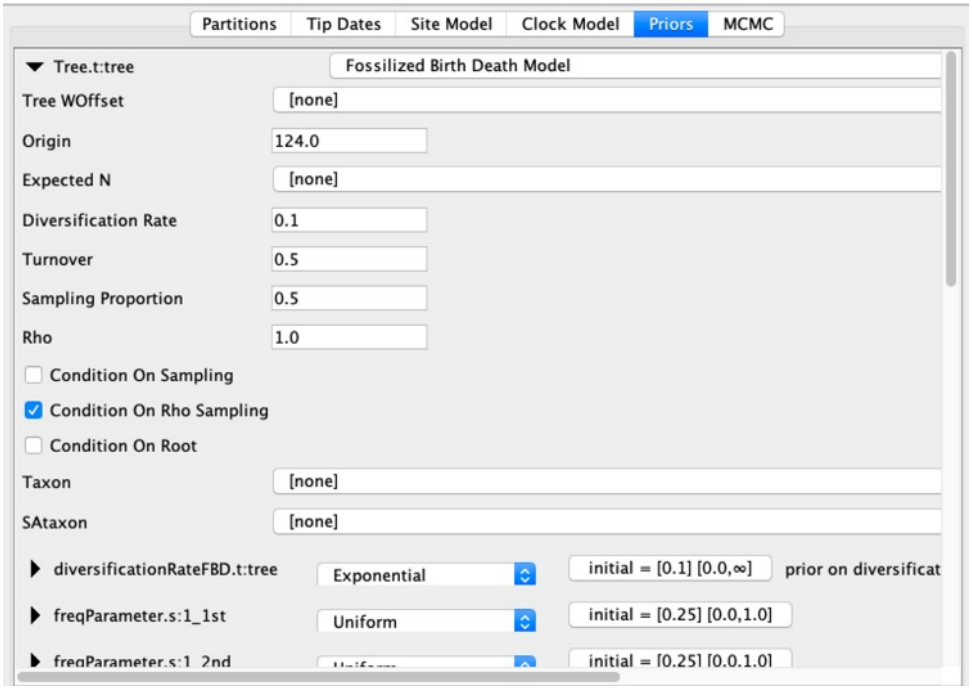

图 5. 在 BEAUti 中设置 FBD 模型为树先验

```
  molecular clock of mitochondrial DNA. Journal of Molecular Evolution
  22:160-174.

================================================================================
Start likelihood: -173058.78054302
Writing file /Users/arongluo/Desktop/BEAST protocol/r1/1_beast.log
Writing file /Users/arongluo/Desktop/BEAST protocol/r1/1_beast.trees
         Sample      posterior ESS(posterior)     likelihood           prior
              0   -173058.7805              N   -172929.0722      -129.7083 --
           1000   -113008.5693         2.0      -112802.7319      -205.8374 --
           2000    -83135.6228         3.0       -82927.2273      -208.3955 --
           3000    -68314.4524         4.0       -68097.1133      -217.3390 --
           4000    -60610.8458         3.6       -60386.4956      -224.3502 --
           5000    -56937.3845         3.2       -56710.4705      -226.9139 --
           6000    -54878.1731         3.3       -54640.1768      -237.9963 --
           7000    -52801.5694         3.5       -52556.6828      -244.8866 --
           8000    -51991.9818         3.7       -51746.5324      -245.4493 --
           9000    -51165.7487         3.7       -50910.9533      -254.7954 --
          10000    -50516.3561         3.9       -50246.9821      -269.3739 --
          11000    -50114.1590         4.2       -49844.1188      -270.0401 12m5
          12000    -49827.3272         4.3       -49578.5333      -248.7939 12m3
          13000    -49503.5872         4.5       -49253.8469      -249.7402 12m2
          14000    -49175.2614         4.6       -48937.1362      -238.1252 11m5
          15000    -48793.7161         4.8       -48553.0869      -240.6292 11m5
```

图 6. BEAST 成功运行界面

4. 结果分析

BEAST 一个独立运行结果一般包含三个文件：.log 文件，.trees 文件，.state 文件。结果分析中一般仅需考虑.log 文件和.trees 文件。在此，r1 和 r2 两个文件夹 (两次独立运行) 分别有 1_beast.log 和 1_beast.trees。

4.1 在 Tracer 安装目录 (一般为/Applications/) 双击 Tracer 图标，通过 File 下拉菜单 Import Trace File 分别加载 r1 和 r2 中的 1_beast.log 文件。加载后，即可查看 posterior、prior 等参数的各种统计值如 Mean，Median，95% HPD 等。特别的，可通过查看参数的 ESS 数值来获悉各参数在 MCMC 分析中的有效独立取样大小；理想情况下，ESS 值一般大于 200。在逐一查看.log 文件时，也可以同时选中两个.log 文件比较两者的分析结果，也可以点击 Combined 检查两者的结果是否收敛。通过 Tracer 分析，如果判定两次独立分析结果比较理想 (图 7)，即可以开展后续分析。

4.2 在 BEAST 2 安装目录中双击 LogCombiner 图标打开 LogCombiner。将 File type 设定为 Tree Files，点击"+"加载 r1 和 r2 中两个.trees 文件并可将 Burnin percentage 设为 25 (与 Tracer 默认 Burnin percentage 为 10 不同)。对 Output File 设置路径和命名后，点击 run 即可将两个.trees 文件合并为一个文件。

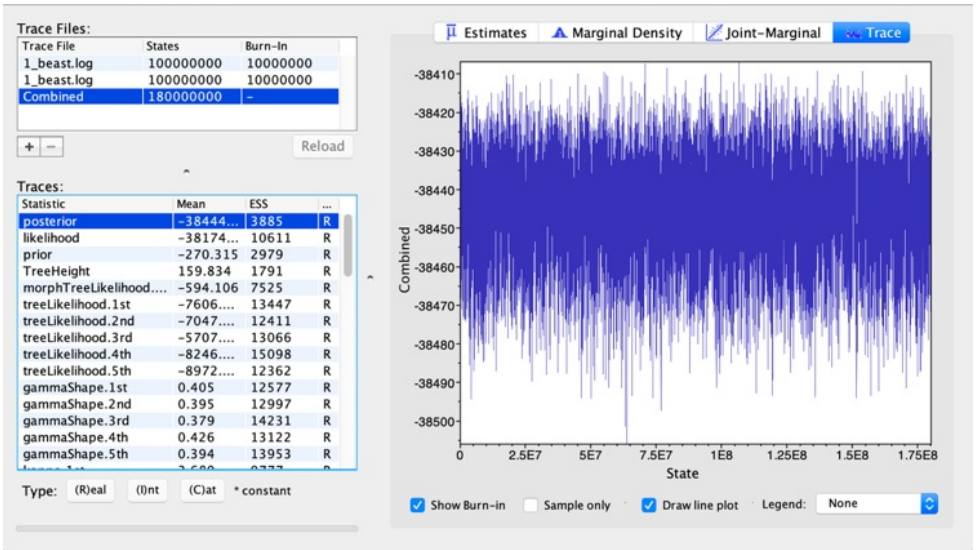

图 7. Tracer 提示 BEAST 两次独立运行情况良好

4.3 在 BEAST 2 安装目录中双击 TreeAnnotator 图标打开 TreeAnnotator。如果需要获得 Maximum clade credibility tree，一般将 Node heights 设置为 Median

heights，在对 Input Tree File 和 Output File 分别设置后，点击 Run 即可开始分析。由于在树文件合并时已经对 Burnin 有所设置，在此只需保留 Burnin percentage 为 0。

4.4 在 FigTree 安装目录双击 FigTree 图标，通过 File/Open 加载 4.3 产生的 Maximum clade credibility tree 文件。通过 FigTree 各选项可将类群的系统发生关系和历史分异时间进行展示。如图 8 可较清晰呈现估算的类群根部时间 95%置信区间、化石类群的系统发生位置等等。该步骤当然也可以用其它软件如 DensiTree 进行实现。

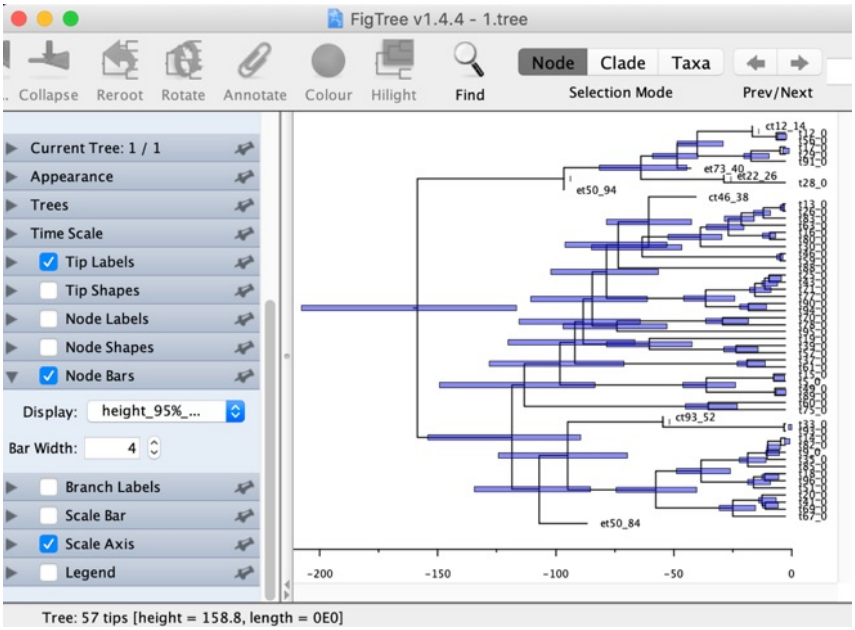

图 8. FigTree 软件显示类群的系统发生关系和历史分异时间

小结与建议

本文向读者展示了使用 BEAST 2 软件包进行贝叶斯全证据支端定年分析的基本流程。值得注意的是，贝叶斯支端定年分析也可在化石和现生类群形态特征数据缺失的情况下实施 (Heath *et al.*, 2014)。该情况下，类似节点定年，一般需要对化石的系统发生位置设置拓扑限制；后验概率则会在兼顾拓扑限制同时，根据 FBD 先验统计化石系统发生位置的各种可能。具体操作可参考：https://taming-the-beast.org/tutorials/FBD-tutorial/。这种 (非全证据) 的贝叶斯支端定年特别适用于分子数据不断涌现但形态特征数据相对稀缺的系统基因组学时代。

另需注意的是，虽然支端定年存在诸多理论优势，但由于其涉及形态特征和化石数据，所以在实际应用中也可能面临不少挑战。比如，化石的形态特征数据会存在不完整或片碎化等特点，容易对化石的系统发生位置推断造成偏倚；用于描述形态特征变换的 Mk/Mkv 模型过于简单，往往不能充分反映形态特征的复杂特点；形态特征是否遵循形态钟仍有待研究，等等。所以，对于贝叶斯支端定年的实际分析结果，特别是依据形态数据推断的化石系统发育位置，仍需结合其它证据 (如古生物学证据) 等综合考虑。

竞争性利益声明

本文作者无利益纷争。

致谢

作者首先感谢 bio-protocol 为本文提供了发表平台，同时也特别感谢同行专家提出的宝贵修改意见。

参考文献

1. Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kuhnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Muller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C. H., Xie, D., Zhang, C., Stadler, T. and Drummond, A. J. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLOS Comput Biol* 15: e1006650.
2. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLOS Biol* 4: e88.
3. Gavryushkina, A., Welch, D., Stadler, T. and Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLOS Comput Biol* 10: e1003919.
4. Heath, T. A., Huelsenbeck, J. P. and Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc Natl Acad Sci USA* 111: E2957-E2966.

5. Luo, A., Duchene, D. A., Zhang, C., Zhu, C. D. and Ho, S. Y. W. (2020). A simulation-based evaluation of tip-dating under the fossilized birth-death process. *Syst Biol* 69: 325-344.
6. Pyron, R. A. (2011). Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Syst Biol* 60: 466-481.
7. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol* 67: 901-904.
8. Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L. and Pasnitsyn, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst Biol* 61: 973-999.
9. Stadler, T. (2010). Sampling-through-time in birth-death trees. *J Theor Biol* 267: 396-404.
10. Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A. and Ronquist, F. (2016). Total-evidence dating under the fossilized birth-death process. *Syst Biol* 65: 228-249.