利用聚类推断多重共识树和超树：综述

Vladimir Makarenkov[1*]、Gayane S. Barseghyan[1] 和 Nadia Tahiri[2]

[1] 蒙特利尔魁北克大学计算机科学系，邮政信箱8888号，市中心分局，蒙特利尔，魁北克省，H3C 3P8，加拿大[2] 谢布鲁克大学计算机科学系，德拉大学大道，谢布鲁克，魁北克省 J1K 2R1，加拿大。通讯作者：makarenkov.vladimir@uqam.ca

摘要： 系统发育树（即进化树、可加树或*X*-树）在物种进化的建模和表示过程中发挥着关键作用。通常用物种系统发育树对特定物种群的基因组进化进行建模，以表示垂直遗传的主要模式。然而，每个基因的进化历程各不相同，可以用其自身的基因树来表示，而该基因树可能与整体物种树的表示存在显著差异。在进化研究中，共识树和超树已被广泛用于整合各个基因树所包含的系统发育信息。然而，如果现有基因树彼此差异很大，那么所得的共识树或超树可能包含许多与高阶内部节点相对应的未解析子树，也可能只是一个星状树。如果现有基因树受到不同网状进化事件的影响，例如水平基因转移、杂交或遗传重组，就可能出现这种情况。因此，利用聚类推断多个备选共识树或超树的问题变得十分重要，因为这种方法可以将进化模式相似的基因树重新划分到不同的簇中（例如，代表经历了相同水平基因转移或重组事件的基因的基因树）。我们批判性地回顾系统发育树聚类领域的最新进展和方法，探讨这些方法的数学性质，并介绍多重共识树和超树方法的主要优势与局限。在应用部分，我们展示如何利用多重超树聚类方法，根据进化模式对氨酰-tRNA合成酶基因树进行聚类。

关键词： 聚类、聚类有效性指标、共识树、*k*-均值、*k*-中心点、系统发育树、Robinson-Foulds 距离、超树。

引言

系统发育（即系统发育树或进化树）一词由海克尔于1866年提出，他将其定义为“类比于个体发育，即生物个体发育史的古生物学发展史”。系统发育树表示关于某一物种群体进化的假说，这些物种通常与树叶相关联。

在数学中，系统发育树称为加性树或*X*-树（因为其叶节点通常与物种集合*X*相关联； Barthélemy 和 Guénoche 1991）。现在我们来介绍一些与系统发育树相关的必要数学定义。系统发育树 T 中两个顶点x 和 y 之间的距离$\delta(x,y)$ 定义为连接 x 和 y 的唯一 T 中的路径上的边长之和。这样的路径记作 (x,y). 叶是度为一的顶点。通常，叶代表一个现生物种（或分类单元）。

定义 1。设 X 是一个含有 n 个分类单元的有限集。X 上的差异度 d 是定义在 $X \times X$ 上的非负函数，满足对任意 x、$y \in X$： $d(x,y) = d(y,x) \geq d(x,x) = 0$。

定义 2。若对任意 x、y、z、$w \in X$ 均有：$d(x,y) + d(z,w) \leq max\{d(x,z) + d(y,w); d(x,w) + d(y,z)\}$，则称 X 上的差异度 d 满足四点条件。

定义3。对于有限集 X，系统发育树（即可加树或 X-树，也就是叶节点根据物种有限集 X 标记的树）是一个有序对 (T, φ)，由一棵树 T 和映射 $\varphi: X \to V$ 组成，其中 T 的顶点集为 V；该映射满足：对于所有度数不超过二的 $x \in X$，均有 $x \in \varphi(X)$。如果 φ 是从 X 到 T 的叶节点集的双射，且每个内部顶点的度数均为三，则该系统发育树为二叉树。

联系四点条件与差异度可由系统发育树表示这一性质的定理如下：

定理1（*Zarestskii*、*Buneman*、*Patrinos* 与 *Hakimi*、*Dobson*）。任何在 $X \times X$ 上满足四点条件的差异度（其中 X 是物种的有限集），都可以由一棵系统发育树 T 表示，使得对于 X 中任意 x、y，$d(x,y)$ 等于 T 中连接叶节点 x 和 y 的路径长度。这种差异度称为树度量。此外，这棵树是唯一的。

图1给出了一个关于由五个分类单元组成的集合 X 的树度量及其对应系统发育树的示例。

遗憾的是，现实中的进化距离（或差异度）很少满足四点条件。因此，需要采用近似算法，根据给定的进化距离矩阵推断树度量矩阵（Gascuel 2005）。最著名的基于距离的近似算法包括：

邻接法（Saitou 1988）、非加权组平均法（Sokal and Michener 1958）、FITCH（Felsenstein 1997）和 MW（Makarenkov and Leclerc 1996, 1999）。

	x2	x3	x4	x5
x1	6	6	4	2
x2		2	4	6
x3			4	6
x4				4

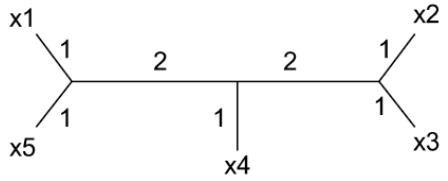

图1. 五个分类单元构成的集合 *X* 上的树度量示例（左），以及对应的系统发育树（可加树或 *X*树）（右）。

生物学家常常需要比较系统发育树，以发现支配给定物种集合的不同进化历史。比较系统发育树有多种度量方法，其中最常用的包括 Robinson and Foulds拓扑距离（*RF*）（Robinson and Foulds 1981）、最小二乘距离（*LS*）、二分划分差异度（*BD*）（Boc等人，2010年）以及四元组距离（*QD*）（Bryant 等，2000）。在这篇文献综述中，我们将主要探讨基于 Robinson-Foulds 距离的方法。两棵树之间的 Robinson and Foulds拓扑距离（Robinson and Foulds 1981），是将一棵系统发育树转换为另一棵所需的节点基本操作（收缩和扩展）的最小次数。它也等于一棵树中存在而另一棵树中不存在的分裂（或二分划分）的数量。所比较的两棵系统发育树必须具有相同的分类单元集合。两棵系统发育树在拓扑结构上越接近，*RF*距离的值就越小。通常有必要将*RF*距离除以其可能的最大值（对于具有*n*个叶节点的二叉系统发育树，该最大值为 2*n*-6），从而对其进行归一化。对于具有*n*个叶节点的两棵树，*RF*距离的计算可在*O*(*n*)时间内完成（Day 1985，Makarenkov 1997，Makarenkov and Leclerc 2000）。

系统发育树重建方法通常不会只输出一棵系统发育树，而会给出一组不同的树（Gascuel 2005）。此外，针对不同基因推断出的系统发育树往往彼此不同。除了采用内在标准（例如使用自举分数）之外，并没有绝对标准可以判定哪棵树优于其他树。因此，最好寻求这些树的共识表示，使其中一致的部分相对于不一致的部分清晰呈现。由此得到的表示称为共识树。传统的共识方法会生成一棵代表所有输入树的系统发育树（Bryant 2003）。Adams 提出了最早的共识方法之一（Adams 1972）。此后，各种各样的方法相继问世。如何使用这些方法一直备受争议（Bryant 2003, Dong et al. 2010）。

共识树的主要类型如下：严格共识树（Sokal 和 Rohlf 1981，Moon 和 Eulenstein 2017）、多数规则共识树（Margush 和 McMorris 1981）、Nelson 共识树（Nelson 1979），以及

扩展多数规则共识树（Felsenstein 1985）。下面简要回顾这些共识树各自的主要特征。

严格共识树（或 Nelson 的分支图）是通过仅考虑所有被比较的树中完全相同的树分裂（即由树内部边诱导的二分划分）推断得出的。系统发育树中相互冲突的部分在严格共识树中以多分叉表示。

有时，为了纳入不一定存在于所有树中的二分划分，采用比严格共识树所用标准宽松的标准会更方便。比较一组拓扑结构各异的系统发育树时，可以寻找在所有被比较的树中出现频率最高的单系群（通常出现在超过 50% 的树中）。由此得到的树就是多数规则共识树。

扩展多数规则共识树包含所有多数二分割，并依次加入其余与之相容的二分划分，从给定树集中出现频率最高的二分划分开始。此过程持续进行，直至得到完全解析的（即二叉）树。扩展多数规则共识树在分子生物学中最常用，因为在迄今讨论的三类共识树中，它的解析度始终最高。

Nelson 共识树包含一组权重最大的相容二分划分。其构建方法是在包含全部二分划分的相容图中寻找一个权重最大的团，这是 NP 困难问题（Nelson 1979，Bryant 2003）。

遗憾的是，在许多实际情况下，作为共识树重建方法输入的系统发育树可能彼此差异很大。例如，当输入树代表不同基因的进化，而这些基因受到多种网状进化事件的影响时，就可能出现这种情况；这些事件包括水平基因转移、杂交或基因内／基因间重组、古老的基因重复或基因丢失（Makarenkov 和 Legendre 2000，Mirkin 等，2003，Bapteste 等，2004）。这些进化事件可能只发生在输入基因树中的某个子群中。因此，用其自身的共识树来表示这一子群似乎更为合适。然而，传统的共识树方法针对给定的一组输入基因系统发育树，只提供一棵候选树，而不考虑其中可能存在的子群（或簇）（Maddison 等，2007）。

图 2 展示了四棵七叶系统发育树的示例：T_1, 、T_2, 、T_3, 和 T_4。在此，包含两棵多数规则共识树的解——T_{12} 和 T_{34}, ——似乎比传统的共识解更合适；后者仅包含一棵多数规则共识树 T_{1234}, ，也就是说，这里是一棵星状树（完全没有内部边的树）。

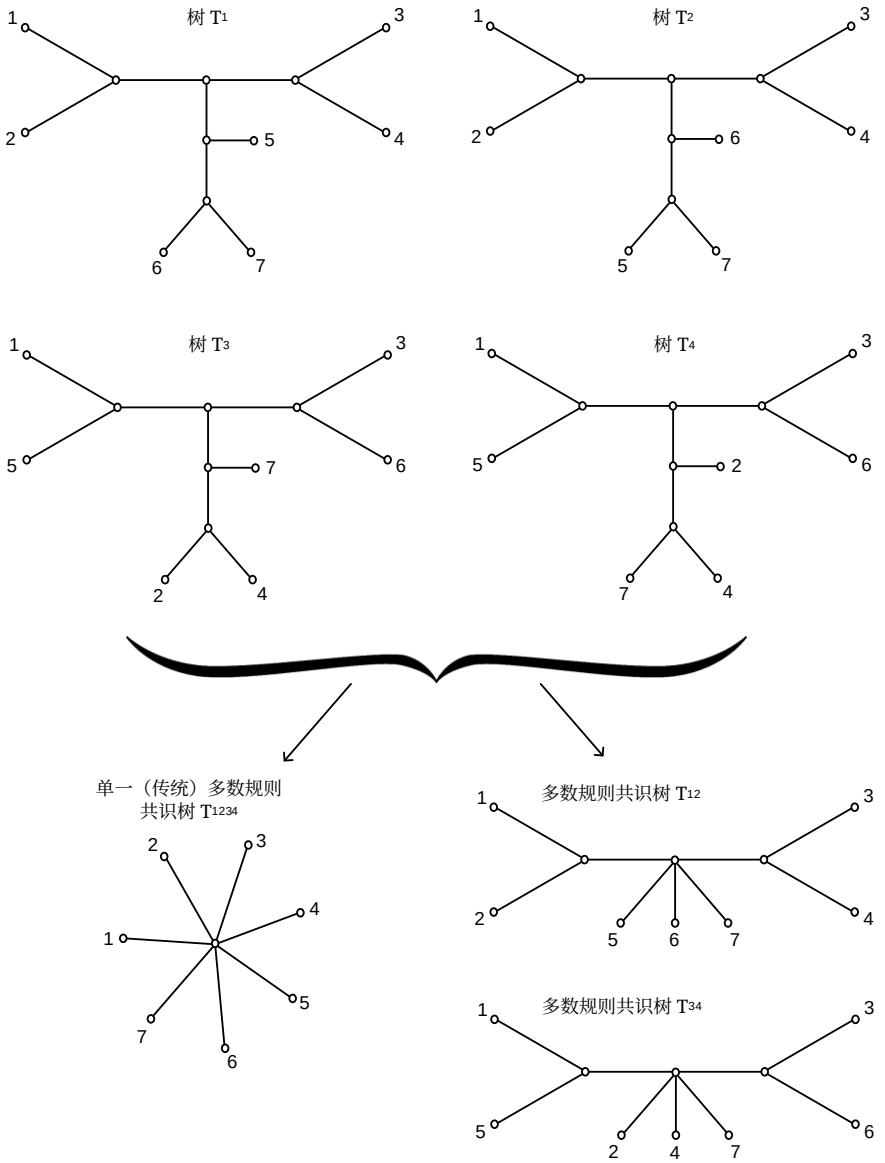

图 2. 定义在同一组七个叶节点上的四棵系统发育树 T_1, T_2, T_3, 和 T_4 。它们的单一（传统）多数规则共识树是一棵星状树 T_{1234}。分别为拓扑结构相近的树对 T_1 和 T_2, ， 以及 T_3 和 T_4, 构建的多数规则共识树为 T_{12} 和 T_{34}, 。

在许多进化研究中，待合并的基因树所定义的分类单元集合各不相同，但彼此部分重叠（例如，参见生命之树项目；Maddison 等，2007）。所有纳入考量的基因都已完成测序的可能性极低，

相同的物种集合。为了协调这类树，应采用超树重建方法（Bininda-Emonds 2004，Wilkinson 等人 2007，McMorris 和 Wilkinson 2011，Warnow 2018）。超树将给定的一组较小（即部分）树进行整合；这些树的分类单元仅部分重叠，整合后得到涵盖所有给定树中分类单元的完整超树。

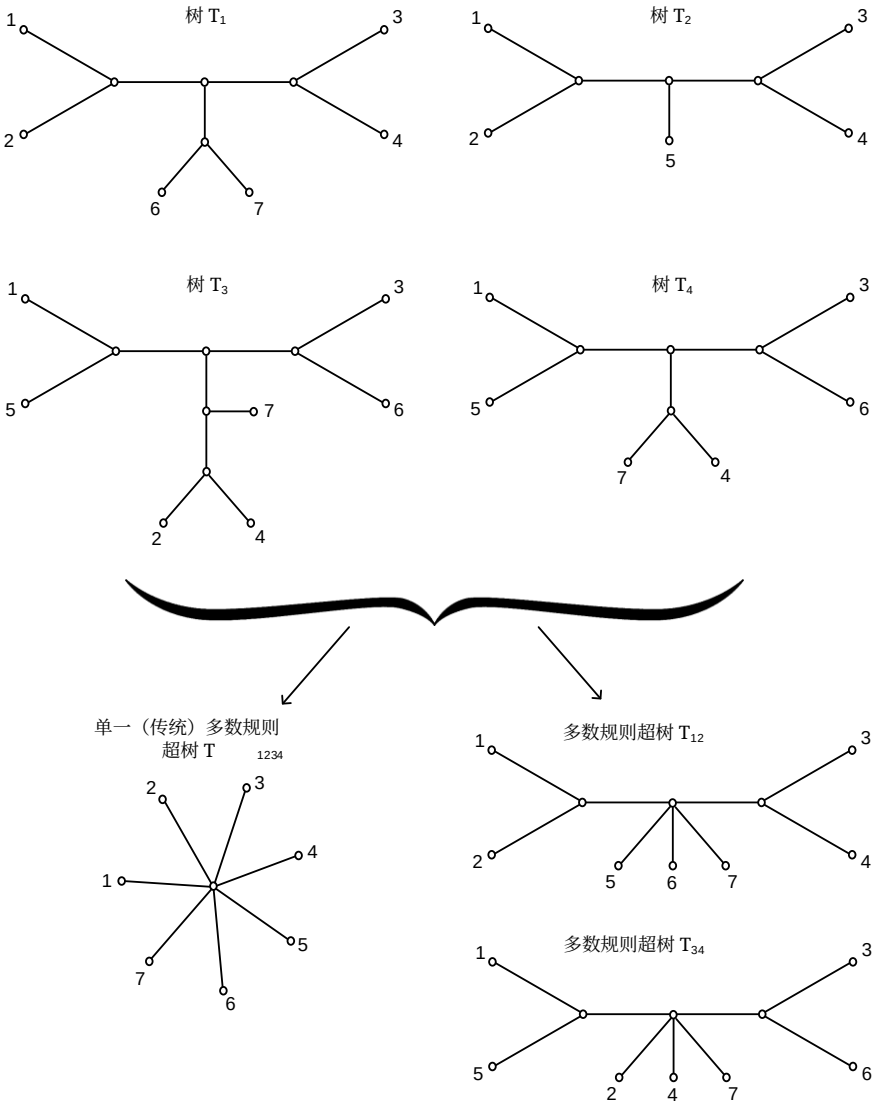

图 3。 四棵系统发育树 T_1、T_2、T_3、和 T_4 分别定义在七个分类单元的不同但彼此重叠的集合上。它们的单一（传统）多数规则超树是一棵星状树 T_{1234}。对于拓扑结构相近的树对 T_1 和 T_2，以及 T_3 和 T_4，分别构建了多数规则超树 T_{12} 和 T_{34}。

最知名的超树推断方法是简约法矩阵表示（MRP）（Baum 1992，Ragan 1992），该方法对给定的部分树进行类似矩阵的聚合。超树重建方法通常用于大型基因组生物的系统发育分析（Mank 等人 2005，Bininda-Emonds 等人 2007，Faurby 等人 2016，Kimbal 等人 2019）。对于原核生物等小型基因组生物，人们已采用多种基因组系统发育分析方法。特别是，超树分析为原核生物的进化提供了新的见解，而许多其他方法无法解决相关问题（Daubin 等人 2001）。近期，Makarenkov 等人（2021）和 Tahiri 等人（2022）采用超树系统发育分析来刻画 SARS-CoV-2基因的进化。

与共识树的情况类似，在许多实际情形下，应推断多个保守超树，以最佳表示给定基因树组的进化。图 3 展示了一个例子，其中四棵系统发育树 T_1, T_2, T_3, 和 T_4 分别定义在七个分类单元的不同但相互重叠的集合上。在此，与由传统超树方法得到的单个多数规则超树 T_{1234}，（即此处的一棵星状树）相比，由两棵多数规则超树 T_{12} 和 T_{34}, 组成的解更为恰当。

构建多重共识树的思想最初由 Maddison（Maddison 1991）提出。他发现，对于某些输入树子集，其共识树可能存在很大差异，而且通常比刻画整个输入树集的单一传统共识树具有高得多的分辨率。人们已经开发出许多方法，用于基于广为人知的聚类算法（例如 *k*-均值和 *k*-medoids）对系统发育树进行分类。我们将在“方法”部分讨论这些方法的主要特征。

划分是一种聚类方法，用于将给定的元素（或分类单元）集合划分为有意义的元素（对象或实体）组，这些组称为簇（或类别）（Mirkin 1996，Mirkin 2005）。划分的目标是依据给定的相似性度量，找出相似元素组成的组。根据元素的特征（或变量）集合对其进行分组时，主要有以下四种划分方法：1）重心，即*k*-均值算法（Lloyd 1957，MacQueen 1967），其中*k*表示簇的数量；2）几何中位数，即*k*-中位数（Bradley et al. 1997）；3）包含出现频率最高的众数的中心，即*k*-众数（Huang 1998）；4）基于中心点的方法，其中中心点是指簇中使其与其他所有簇元素之间的距离之和最小的元素，即*k*-中心点（Kaufman and Rousseeuw 1990）。在本文献综述中，我们将主要关注*k*-均值和*k*-中心点算法，因为它们已广泛用于树聚类（参见“方法”部分）。这两种算法的运行速度都很快：*k*-均值的时间复杂度为$O(I \times K \times M \times N)$，其中$I$是*k*-均值内循环的迭代次数，$K$是簇的数量，$M$是描述给定元素集合的特征数量，$N$是元素数量；而*k*-中心点的时间复杂度为$O(I \times K \times M \times (N-K)^2)$。值得注意的是，与*k*-均值相比，*k*-中心点算法对离群值的敏感性要低得多。欧几里得度量、曼哈顿度量和闵可夫斯基度量是*k*-

均值和*k*-中心点目标函数中最常用的度量（Mirkin 2005，de Amorim and Mirkin 2012，de Amorim and Makarenkov 2016）。然而，在树聚类中，应改用 Robinson–Foulds 拓扑距离或其他树距离，并将系统发育树作为簇元素。

方法

系统发育岛（Maddison 1991）是一种根据树的分支长度以及输入树之间存在差异所需的分支重排次数，对一组树进行划分的方法。作者考察了以下三种分支重排类型：NNI（最近邻交换）、SPR（子树剪枝-再嫁接）和 TBR（树二分-重连）（Swofford and Olsen 1990, Swofford 1991）。在 NNI 重排中，一个演化支（即子树）只能移动到邻近分支；在 SPR 中，它可以移动到邻近或远处的分支；而在 TBR 中，它也可以移动到邻近或远处的分支，同时还可以对该演化支重新定根。开发这种方法是为了利用树搜索算法寻找最简约树；也就是说，它从多个起始点开始搜索，以找到多个孤岛。Maddison 将长度为 L 的树岛正式定义为由 n 棵树组成的集合，并满足以下三个条件：（1）所有树的长度均为 <L；（2）岛中的每棵树都通过一系列树与其他每棵树相连，这些树的长度均为 <L，且序列中相邻的树之间仅相差一次重排；（3）所有满足条件 1 和 2 的树都包含在该集合中。可以使用 PAUP*（Swofford 1991）和 Henning86（Farris 1988）中可用的树搜索工具进行多次搜索来发现多个孤岛，每次搜索均从一棵不同的树开始。正如对树间划分度量的分析所示（例如 Robinson-Foulds 距离或划分度量），同一孤岛内的树通常比不同孤岛间的树更为相似。作者据此得出结论，不同孤岛上的树可能会对性状演化产生不同影响。

最小化信息损失的特征树（Stockham 等人，2002）是系统发育分析中单一共识树后处理方法之外的另一种途径。该方法基于流行的聚类算法开发而成，即 k-均值算法和凝聚聚类（Maddison，1991）。Stockham 等人提出的方法运用特征树概念来最小化信息损失。该方法可用于提高输出共识树的分辨率，并提供更多有关候选树分布情况的细节。该方法的主要局限在于，输入的系统发育树必须包含相同的物种集合（共识情形），且无法处理同质数据的情形（即簇数 K = 1）。Stockham 等人（2002）所研究方法的目标函数如下：

$$OF = \sum_{k=1}^{K} \sum_{i=1}^{N_k} RF^2(T_k^{st}, T_{ki}), \qquad (1)$$

其中，K 为簇数，N_k 为簇 k 中树的数量，RF^2 为树 T_{ki}（即簇 k 中的第 i 棵树）与树 T_k^{st}（即簇 k 的严格共识树）之间的 Robinson-Foulds 拓扑距离的平方。

多极共识（MPC）方法（Bonnard 等人，2006）旨在找出一组规模较小的树，其中包含所有支持度高于预设阈值的分裂。在给定要展示的分裂后，必须尽可能减少多极共识中的树的数量。与多数规则共识相比，该方法可以展示更多次要的进化信号。与 Maddison（1991）和 Stockham 等人（2002）的方法一样，MPC 方法总是生成多个共识树作为解，而不会生成单个共识树。Bonnard 等人（2006）采用了一种称为贪心着色算法的启发式着色方案，该方案包含两个主要步骤：（1）为顶点排序；（2）按顺序逐一处理顶点，为每个顶点分配第一个尚未分配给与其相关的已着色顶点的颜色。MPC 方法至少在两个方面有别于其他树聚类方法：1）它更加简约，因为输入树中出现的每个非核心分裂仅表示一次；2）它不需要预先对输入树进行聚类。因此，MPC 的时间复杂度相对于输入分裂的数量是多项式级的，但相对于输入树的数量则是线性的。

TreeOfTrees方法（Darlu 和 Guénoche，2011年）可以比较从多组比对后的基因序列中获得的*X*-树拓扑结构。该方法的主要目标是利用自助法抽样以及加权或非加权共识，检测具有相同进化历史的基因。树集合之间的比较基于多种树度量，最终得到一棵以基因树标记的唯一树（即提出了一种层次树聚类方法）。为了评估输入基因树之间一致性的稳健性，采用重抽样程序，由此构建“基因树之树”，同时提供：基因树接近程度的简单树表示，以及树之树每个二分划分的自助法值。树之树的每个叶对应一个单独的基因（或代表其进化过程的自助法系统发育树）。树拓扑之间的比较首先将每棵输入树转换为成对距离矩阵，计算分隔两个分类单元的边数，或使用路径长度度量。所得的树距离矩阵可以无歧义地确定树拓扑。共识树 T 通过枚举属于输入树集合的所有二分划分来构建。Darlu 和 Guénoche 提出了加权共识方法，其定义采用以下权重函数：

$$w(B_i) = \sum_{T_k \in S} \tau_k, \tag{2}$$

其中，B_i 是二分划分 i，T_k 是簇 k 的簇树，S 是包含二分划分 B_i的输入树子集，而τ_k 是树 T_k的质量度量。随后，作者按以下公式，将输入树 T_k 的权重定义为其包含的内部边的权重之和：

$$\Omega(T_k) = \sum_{B_i \in T_k} w(B_i). \quad (3)$$

多重共识树（Guénoche 2013）是一种树聚类方法，旨在判断输入基因树中是否存在单一共识，并通过划分方法检测分化基因。如果给定的基因树彼此完全一致，它们应当与单一共识树相容；否则，则可以识别出对应于不同遗传模式的多个共识树。多重共识树方法针对一组树划分优化一个广义得分，以判断给定的基因树集合是否具有同质性。作者仅考虑无根 X 树，并着重于以下共识策略：将一棵 X 树表示为其二分划分的集合，每个二分划分对应树的一条内部边。移除每条内部边会产生一个分裂，进而将分类单元集合 X 二分划分。每个二分划分$B_i = X_i \cup X_i'$ 的权重，是树谱中包含该二分划分的 X 树数量 N_i。作者将一棵 X 树 T 相对于由 N 棵树构成的树谱π 的权重定义如下：

$$W_\pi(T) = \sum_{B_i \in T_m} w(B_i) = \sum_{B_i \in T_m} N_i, \quad (4)$$

其中，$w(B_i)$ 是 T 的每个二分划分 B_i 的权重，N_i 是内部多数边的数量（即满足以下条件 $N_i > \frac{N}{2}$ 的边），而 T_m 是仅保留其多数边后的树 T。每个二分划分 B_i 的权重是包含该二分划分的树集合中 X 树的数量 N_i。

作者将评分（4）推广到树集合 P_π 分为 k 个类别的情形，定义如下：

$$W^k(P_\pi) = \sum_{i=1,\ldots,k} p_i \times W_{\pi_i}\left(T_i^{maj}\right), \quad (5)$$

其中，P_π 是树集合 π 的一个划分，分为 k 个类别（$\pi_1, \ldots, \pi_k$），各类别分别包含 $\{p_1, \ldots, p_k\}$ 棵树；T_i^{maj} 是与第 i 个类别相对应的多数共识树。

树岛（Silva 和 Wilkinson，2021）是一种基于任意合适的树间两两距离度量的方法，它将“岛”的概念扩展到任意树集或树多重集，例如通过贝叶斯方法或自助法生成的树集或树多重集，并便于事后寻找树岛。当最简约树的严格共识相对缺乏分辨率时，这种方法可能很有用；不过，它依赖分析程序（Silva 和 Wilkinson 使用的是 PAUP*），不仅要识别树岛的数量，还要确定最简约树的组成。树岛等不同的树子集，与其他数据探索方式相辅相成；后者尝试划分树集，以获得更好的概要并加深对进化的理解。然而，这种方法对于大型系统发育树分布的用途有限，因为它以大量树间两两比较取代了距离的计算。

使用 k-中心点算法推断多重共识树（Tahiri 等人，2018）是一种快速方法，可根据一组在相同物种集上定义的系统发育树推断多重共识树。该方法基于 k-中心点算法的划分算法，将给定的树集划分为多个树簇。著名的轮廓系数和 Caliński-Harabasz 聚类有效性指标已针对使用 k-中心点算法进行树聚类作了调整，以确定最合适的簇数。该方法可用于识别基因树群组：群组内的基因树具有相似的进化历史，而不同群组之间的进化历史则各不相同。此方法适用于大型基因组和系统发育数据集的分析。

与 Stockham 等人（2002）使用的目标函数（见公式 1）相比，Tahiri 等人（2018）使用多数规则共识树而非严格共识树，并使用未平方的 RF 距离而非平方距离。因此，要最小化的直接目标函数如下：

$$OF = \sum_{k=1}^{K} \sum_{i=1}^{N_k} RF(T_k^{maj}, T_{ki}), \quad (6)$$

其中，RF 表示树 T_{ki}（即簇 k 中的第 i 棵树）与 T_k^{maj}（簇 k 的多数规则共识树）之间的 Robinson-Foulds 距离。不过，计算多数规则共识树或扩展多数规则共识树至少需要 O(nN) 的时间，其中 n 是每棵树的叶节点数（分类单元或物种数），N 是树的数量。

因此，Tahiri 等人（2018）在其基于 k-中心点算法的方法中使用了以下目标函数：

$$OF_{med} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} RF(T_k^{m}, T_{ki}), \quad (7)$$

其中，T_k^m 是簇 k 的中心点，定义为 k 中所有其他树之间的 RF 距离之和最小化。与基于公式（6）的目标函数相比，这一版本的目标函数速度快得多，因为它无需在聚类算法的每个基本步骤中重新计算多数规则共识树。该方法的运行时间为 $O(nN^2+rK(N-K)^2I)$，其中 $O(nN^2)$ 是预先计算所有输入树两两之间 RF 距离矩阵（大小为 $(N\times N)$）所需的时间，K 是簇的数量，I 是 k-中心点算法内部循环的迭代次数，r 是 k-中心点算法使用的不同随机初始点数量（通常需要数百个不同的随机初始点才能获得良好的聚类结果；Mirkin 2005）。

使用 k-均值算法推断多重共识树和超树（Tahiri 等人，2022）是一种新方法，用于推断多棵备选共识树和超树，以最佳呈现给定基因树集合中的主要进化模式。该方法基于广受欢迎的 k-均值聚类算法和 Robinson-Foulds拓扑距离，将给定的树集合划分为

一个树簇（适用于同质数据），或多个树簇（适用于异质数据）。作者展示了如何在使用 k-均值算法进行树聚类时，采用广受欢迎的 Caliński-Harabasz、轮廓系数、Ball 和 Hall 及 Gap 聚类有效性指标。利用 Robinson-Foulds 距离平方根的欧几里得性质，可以定义一个快速高效的目标函数，如下所示：

$$OF_{EA} = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} RF(T_{ki}, T_{kj}), \quad (8)$$

基于公式 (8) 的树聚类算法的时间复杂度为 $O(nN^2+rNKI)$。

此外，作者还确立了公式 (6) 中定义的一般目标函数的一些有趣性质，并将其用于聚类过程中。具体而言，该目标函数 *OF* 的下界和上界由下文的定理 2 给出：

定理 2（Tahiri 等人，2022）。对于给定的簇 k，其中包含 N_k 棵系统发育树（即加性树或 X 树），以下不等式成立：

$$\frac{1}{N_k - 1} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} RF(T_{ki}, T_{kj}) \leq \sum_{i=1}^{N_k} RF(T_k^{maj}, T_{ki}) \leq \frac{2}{N_k} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} RF(T_{ki}, T_{kj}), \quad (9)$$

其中，N_k 是簇 k 中树的数量，T_{ki} 和 T_{kj} 分别是簇 k 中的树 i 和 j，T_k^{maj} 是簇 k 的多数规则共识树。

在同一篇论文中，Tahiri 等人展示了如何将他们的方法扩展到超树聚类情形。在超树聚类的背景下，我们假设给定的 N 棵无根系统发育树可能包含不同但彼此重叠的叶节点集合。在这种情况下，公式 (6) 所示的原始目标函数 OF 可重新表述如下：

$$OF_{ST} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} RF_{norm}(ST_k, T_{ki}) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\frac{RF(ST_k, T_{ki})}{2n(ST_k, T_{ki}) - 6} \right), \quad (10)$$

其中 K 是簇 k 的数量，$RF_{norm}(ST_k, T_{ki})$ 表示簇 k 中第 i 棵树（记为 T_{ki}）与该簇的多数规则超树（记为 ST_k）之间的归一化 Robinson-Foulds 拓扑距离，该超树被缩减为一棵包含 T_{ki} 所有共同叶的子树。此处通过将 RF 距离除以其最大可能值来进行归一化（即 $2n(ST_k,T_{ki})-6$，其中 $n(ST_k,T_{ki})$ 是 ST_k 和 T_{ki} 中共同叶的数量）。此处进行 RF 距离归一化，是为了使每棵树对聚类的贡献相同。显然，只有当 ST_k 和 T_{ki} 中共同叶的数量大于 3 时，才能使用公式 (10)。

公式 (8) 的类似形式可用于超树聚类，从而避免在 k-均值算法的每一步都重新计算超树。可使用以下目标函数：

$$OF_{ST_EA} = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} \left(\frac{RF(T_{ki}, T_{kj})}{2n(T_{ki}, T_{kj}) - 6} + \alpha \times \frac{n(T_{ki}) + n(T_{kj}) - 2n(T_{ki}, T_{kj})}{n(T_{ki}) + n(T_{kj})} \right), \quad (11)$$

其中，$n(T_{ki})$ 是树 T_{ki} 中叶节点的数量，$n(T_{kj})$ 是树 T_{kj} 中叶节点的数量，$n(T_{ki},T_{kj})$ 是树 T_{ki} 和 T_{kj} 中共同叶的数量，而 α 是惩罚（调节）参数，取值介于 0 和 1 之间，用于避免将共同叶比例较低的树分到同一个簇中。

Tahiri 等人（2022）进行的模拟表明，他们新提出的树聚类方法比本节前面讨论的 Stockham 等人（2002）、Tahiri 等人（2018）和 Bonnard 等人（2006）的方法速度更快，整体效率也更高。

适用于树聚类的聚类有效性指标

本节将介绍如何在采用 k-均值算法的树聚类中使用常见的 Caliński-Harabasz、轮廓系数、Ball 和 Hall 及 Gap 聚类有效性指标。

适用于树聚类的 *Caliński*-Harabasz 聚类有效性指标

我们在此考察的第一个聚类有效性指标是 Caliński-Harabasz 指数（Caliński 和 Harabasz 1974）。该指标有时也称为方差比准则，定义如下：

$$CH = \frac{SS_B}{SS_W} \times \frac{N-K}{K-1}, \quad (12)$$

其中，SS_B 是*w*中组内评估指标的表示，K 为簇数，N 为元素数（本例中即树的数量）。最佳簇数对应于 CH 的最大值。

在 CH 的传统版本中，考虑欧几里得距离时，SS_B系数通过使用 L_2-范数进行评估：

$$SS_B = \sum_{k=1}^{K} N_k \|m_k - m\|^2, \quad (13)$$

其中，m_k (k = 1 ... K) 是簇 k 的质心，m 是给定数据集 X 中所有元素的总体均值（即质心），N_k 是簇 k 中元素的数量。在欧几里得距离的背景下，SS_W 指数可使用以下两个等价表达式计算：

$$SS_W = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \|x_{ki} - m_k\|^2 = \sum_{k=1}^{K} \frac{1}{N_k} \left(\sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} \left\|x_{ki} - x_{kj}\right\|^2 \right), \quad (14)$$

其中 x_{ki} 和 x_{kj} 分别表示（Caliński and Harabasz 1974）。

为了在树聚类中使用公式（13）和（14）的类似形式，Tahiri 等人（2022）采用了给定树集的质心这一概念。在树聚类算法中，中位树（Barthélemy 和 Monjardet 1981；Barthélemy 和 McMorris 1986）发挥着质心的作用。中位数程序（Barthélemy 和 Monjardet 1981）定义如下。对于具有相同叶集合 S 的树集 Π = {T_1, ..., T_N}，其中位树集合 Md(Π) 是所有定义在 S 上的树 T 构成的集合，且满足：$\sum_{i=1}^{N} RF(T, T_i)$ 最小。若 N 为奇数，则 Π 的多数规则共识树 Maj(Π) 是 Md(Π) 中唯一的元素。若 N 为偶数，则 Md(Π) 由 Maj(Π) 以及一些分辨率更高的树组成。

Tahiri 等人（2022）提出，在类 k 均值树聚类中，使用基于欧几里得距离性质的一些公式来定义 SS_B 和 SS_W。这些公式不需要在 k-均值算法的每次迭代中计算多数（或扩展多数）规则共识树。具体而言，它们将项 $\|x_{ki} - x_{kj}\|_2$在公式（14）中替换为 $RF(T_{ki}, T_{kj})$，从而得到 SS_W的公式：

$$SS_W = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} RF(T_{ki}, T_{kj}), \tag{15}$$

其中，T_{ki} 和 T_{kj} 分别是簇 k 中的树 i 和树 j。

同样，在欧几里得距离的情况下，公式如下：

$$SS_B + SS_W = \frac{1}{N}\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|x_i - x_j\|^2\right), \tag{16}$$

其中，x_i 和 x_j 是 X 中两个不同的元素（Caliński 和 Harabasz 1974）。因此，

e

按如下方式计算：

$$SS_B = \frac{1}{N}\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} RF(T_i, T_j)\right) - SS_W, \tag{17}$$

其中，T_i 和 T_j 分别是树集 Π 中的第 i 棵树和第 j 棵树，SS_W 根据公式 (15) 计算。

根据 Robinson-Foulds 距离平方根的欧几里得性质，公式（15）和（17）给出了计算目标函数 OF_{EA} 的指标 SS_B 和 SS_W 的精确公式，该目标函数由公式（8）定义。有趣的是，目标函数 OF_{EA} 也可用作公式（6）所定义目标函数的近似（显然，树簇的质心不一定是该簇的共识树；此外，它也不一定是系统发育树）。

适用于树聚类的 *Ball-Hall* 指数

本综述中另一个值得考虑的相关准则是 Ball-Hall 指数。1965 年，Ball 和 Hall（BH）提出了 ISODATA 程序，用于衡量对象组相对于均方根距离（即组内距离）的平均离散度。与 CH 指数不同，BH 指数可用于寻找仅包含一棵共识树的解。Tahiri 等人（2022）将 BH 准则调整用于 k-均值算法的树聚类，由此得到以下公式：

$$BH = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} RF(T_k^{maj}, T_{ki}). \tag{18}$$

此外，可以使用以下公式来避免计算多数规则树：

$$BH = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k^2}\sum_{i=1}^{N_k-1}\sum_{j=i+1}^{N_k} RF(T_{ki}, T_{kj}). \tag{19}$$

适用于树聚类的轮廓指数

本文考虑的下一个常用准则是轮廓宽度指数（SH）（Rousseeuw，1987）。传统上，簇 k 的轮廓宽度定义如下：

$$s(k) = \frac{1}{N_k}\left[\sum_{i=1}^{N_k}\frac{b(i)-a(i)}{max(\,a(i), b(i)}\right], \tag{20}$$

其中，N_k 是属于簇 k 的元素数量，a(i) 是元素 i 与属于簇 k 的所有其他元素之间的平均距离，b(i) 则是在所有不同于 k 的簇 *k'* 中，i 与簇 *k'* 的所有元素之间的平均距离的最小值。

在树聚类的情况下，可分别使用公式 (21) 和 (22) 计算 a(i) 和 b(i)：

$$a(i) = \frac{\sum_{j=1}^{N_k} RF\left(T_{ki}, T_{kj}\right)}{N_k}, \tag{21}$$

$$b(i) = \min_{1\le k'\le K, k'\ne k}\frac{\sum_{j=1}^{N_{k'}} RF\left(T_{ki}, T_{k'j}\right)}{N_{k'}}, \tag{22}$$

其中，$T_{k'j}$ 是簇 k' 中的第 j 棵树，且 $k' \neq k$，$N_{k'}$ 是簇 *k'* 中的树的数量。

最佳簇数 K 对应于 SH 的最大平均值，计算方法如下：

$$SH = \overline{s}(K) = \sum_{k=1}^{K} \frac{[s(k)]}{K}. \quad (23)$$

由公式 (23) 定义的 SH 指数的取值范围介于 -1 和 +1 之间。

针对树聚类调整的 *Gap* 统计量

这里讨论的最后一个评价准则是 Gap 统计量（Tibshirani 等人（2001））。与 BH 指数一样，Gap 统计量允许得出仅包含一棵共识树的解。Tibshirani 等人（2001）提出的公式基于欧几里得距离的性质。在树聚类的背景下，Tahiri 等人（2022）通过定义总簇内距离 D_k，来表征属于同一簇 k 的树之间的凝聚程度，具体如下：

$$D_k = \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} RF(T_{ki}, T_{kj}). \quad (24)$$

平均 V_K 之和可使用以下公式计算：

$$V_K = \sum_{k=1}^{K} \frac{1}{2N_k} D_k. \quad (25)$$

对于具有 K 个簇的给定聚类解，Gap 统计量传统上定义如下：

$$Gap_N(K) = E_N^*\{\log(V_K)\} - \log(V_K), \quad (26)$$

其中，E_N^* 表示参考分布 n，我们的方法使用以下公式（Tibshirani 等人（2001））计算 $log（V_K）$ 的期望值：

$$E_N^*\{\log(V_K)\} = \log\left(\frac{Nn}{12}\right) - \left(\frac{2}{n}\right)\log(K), \quad (27)$$

其中，n 为树叶的数量。Gap 统计量的最大值对应最佳聚类结果。

进化数据应用示例

氨酰-tRNA合成酶（氨酰-tRNA合成酶）是一类将相应的氨基酸连接到其匹配的转运 RNA 上的酶。长期以来，生物学家一直对氨酰-tRNA合成酶的结构与功能关系感兴趣（Woese等人（2000），Godwin 等人（2018））。研究发现，氨酰-tRNA合成酶在翻译过程中发挥着核心作用，这表明其进化历史与遗传密码的进化历史可能密切相关（Woese 等人（2000））。这些信息使氨酰-tRNA 合成酶基因结构域分析成为生命之树推断的关键组成部分（Bullwinkle 和 Ibba（2014），Unvert 等人（2017））。Woese等人考察了每一种用于

生物细胞的 20 种标准氨酰-tRNA合成酶的进化谱，并据此构建了蛋白质的进化历史，将其划分为 5 组（非极性脂肪族R基团、非极性芳香族R基团、极性不带电R基团、带正电R基团和带负电R基团）。在开展这项著名的氨酰-tRNA合成酶分析时，Woese等人考察了来自 3 个主要域（古菌、真核生物和细菌）的共 72 个物种，这些物种可表示为相关系统发育树的叶节点。
在本研究中，我们使用了 Woese 等人最初构建的 36 棵氨酰-tRNA合成酶系统发育树（即氨酰-tRNA合成酶基因树）。这些树包含不同但彼此重叠的叶节点集合（总共涉及 72 个不同物种）。这些树以及我们的程序可在以下 GitHub 仓库地址获取：https://github.com/TahiriNadia/KMeansSuperTreeClustering。我们将这 36 棵树作为 *KMeansSuperTreeClustering* 算法（Tahiri 等人，2022）的输入。超树聚类算法采用以下选项运行：使用 Caliński-Harabasz（Caliński 和 Harabasz 1974）聚类有效性指标选取最佳簇数（实验中簇数范围为 2 至 10），并将惩罚参数 α 设为 1。

在这些设置下，我们的算法发现，这些数据的最佳解对应于划分为 2 个簇。每个树簇都可以用其各自的超树表示。第一个得到的簇包含 19 棵树，共涉及 61 个不同物种；第二个得到的簇包含 17 棵树，共涉及 56个物种。两个树簇对应的超树（见图 3 和图 4）使用 CLANN 程序（Creevey和McInerney，2005年）推断得到。在 CLANN 中，我们使用最相似超树（dfit）方法（Creevey等人，2004年）和 MRP准则。该准则采用基于简约性准则的矩阵表示。接下来，我们推断了各簇中最常见的水平基因转移（水平基因转移），这些转移体现了两个树簇中所含系统发育树的进化。为此，我们使用了 Boc 等人（2010）提出的水平基因转移检测方法。该方法通过协调物种系统发育树和基因系统发育树来进行分析。在本研究中，两个得到的超树充当基因树，而物种系统发育树则依据 NCBI 的分类系统（见 https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi）；图 4 和图 5 中以实线边表示这些树。这些超树并未完全解析（即第一棵超树（见图 4）包含 9 个度大于 3 的内部节点，而第二棵超树（见图 5）包含 10 个度大于 3 的内部节点）。我们使用 T-Rex 网站上提供的水平基因转移算法版本（ Boc 等人，2012年）以及 Armadillo 1.1（Lord等人，2012年）工作流平台，识别水平基因转移事件的各种情景，从而使每棵物种树与对应的超树相协调。所有这些树的根均置于将细菌演化支与真核生物及古菌演化支分开的边上。第一棵超树中发现了两次常见的水平基因转移，第二棵超树中则发现了四次。我们的结果表明，大多数氨酰-tRNA合成酶经历了双向进化。所得结果与 Dohm 等人（2006）和 Sharaf 等人（2019）的研究结果一致：氨酰-tRNA合成酶的大多数 tRNA合成酶都有两个版本，一个是细胞质型，另一个是线粒体型。

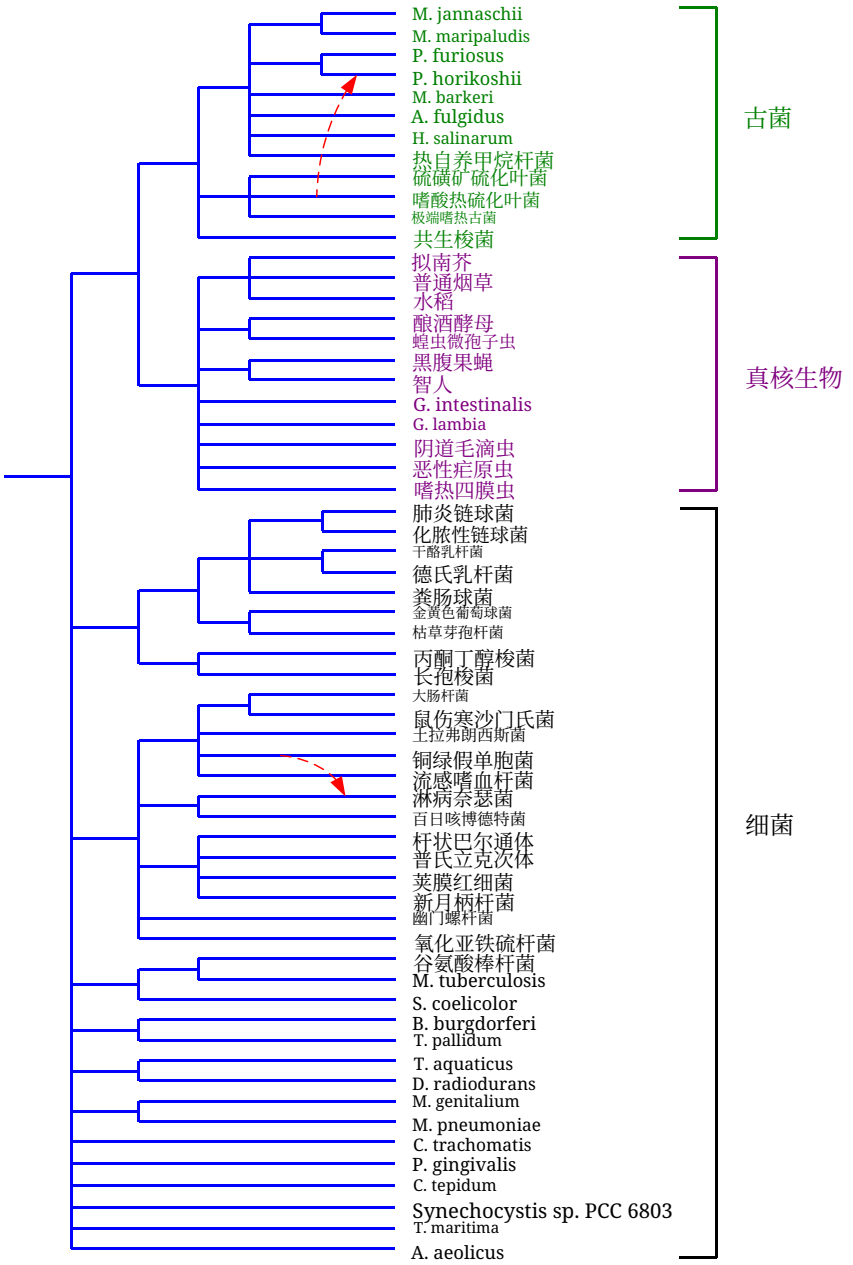

图4。根据NCBI分类学分类构建的物种树（完整边），对应于来自19棵氨酰-tRNA合成酶系统发育树的第一个簇中的61个物种。使用Boc等（2012）的HGT-Detection程序发现了两次水平基因转移（箭头所示）。

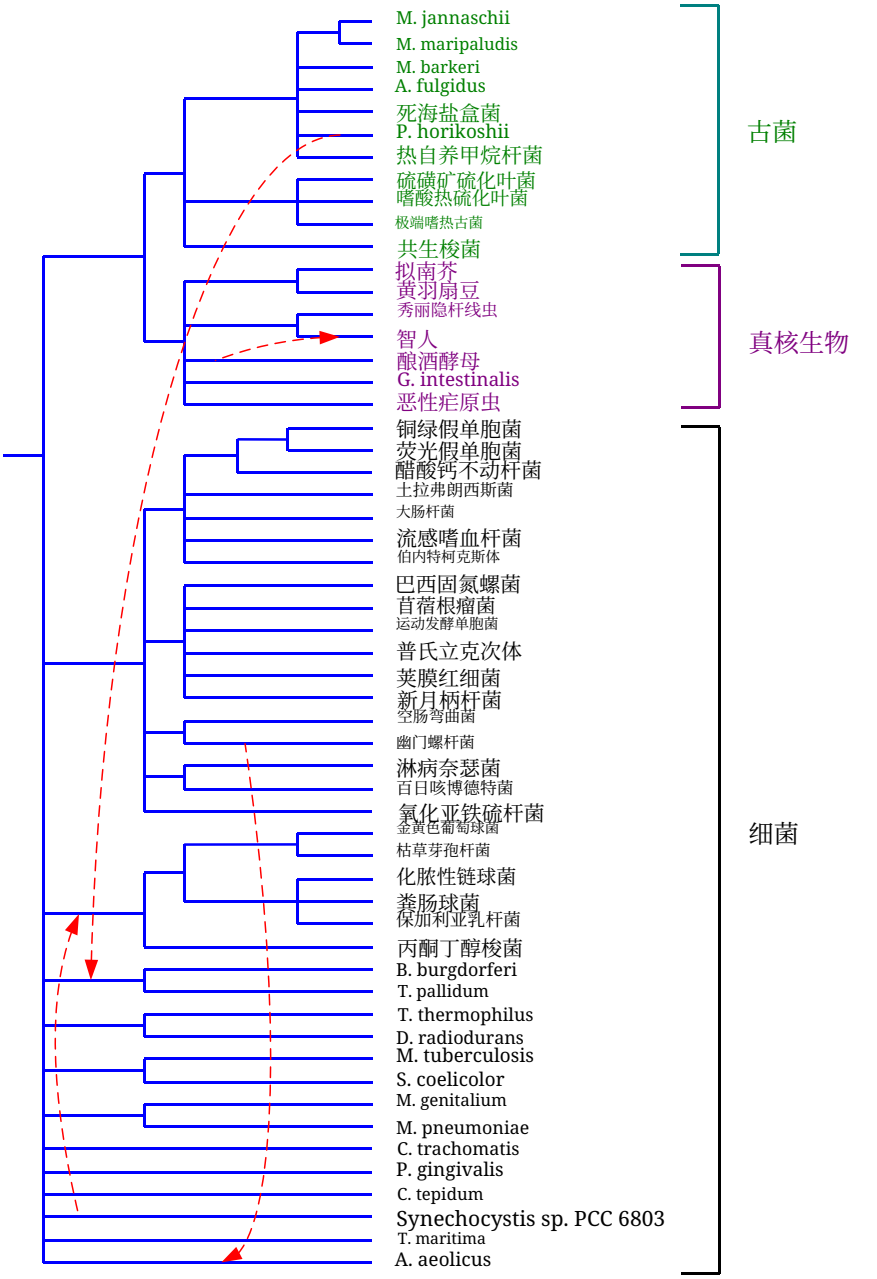

图5. 与 NCBI分类学分类相对应的物种树（完整边），根据第一簇中的17棵氨酰-tRNA合成酶系统发育树为56个物种构建。使用 Boc等（2012）的水平基因转移检测程序，发现了四次水平基因转移（由箭头标示）。

结论

本文综述了从给定的系统发育树集合（即加性树、进化树或 *X*-树）中推断多个备选共识树和超树的最先进系统方法。大多数被综述的论文介绍了通过对树拓扑结构进行 *k*-均值或 *k*-中心点聚类来运行的算法。在共识树聚类问题中，所有树都应定义在同一组分类单元（即与树叶节点相关联的物种）上；而在超树聚类问题中，树可以定义在不同但相互重叠的分类单元集合上。在许多情况下，多个共识树和超树比传统的单个共识树和超树更能代表相关的进化模型。多个共识树和超树的分辨率通常远高于通过传统方法推断出的单个共识树或超树（Maddison 1991）。因此，多个共识树和超树有可能保留给定基因树集合中更多可信的信息。聚类似乎是推断多个共识树和超树的一种直观而自然的解决方案。树聚类在进化研究中具有直接的实际应用。它可以帮助识别受到相同水平基因转移、杂交、基因内/基因间重组事件影响的基因集合，或识别在进化过程中经历了相同古老基因复制和基因丢失的基因集合（Makarenkov 等，2004；Bapteste 等，2004；Diallo 等，2006；Boc 和 Makarenkov，2011）。

自生命之树推断项目启动以来（Maddison 等，2007），探讨超树理论的研究数量大幅增加。本文所述方法可用于推断生命之树的多个备选子树，因为其中存在许多尚未解决的演化支（即内部节点度数较高的子树）。从实际角度看，构建多个备选超树比构建多个备选共识树更为重要，因为目前大多数可用的基因树并非基于完全相同的分类单元集合。不过，据我们所知，近期 Tahiri 等人（2022）的研究仍是唯一探讨这一重要问题的工作。该研究作者展示了 Robinson-Foulds拓扑距离（原始距离或标准化距离）以及*k*均值划分算法的一些显著特性如何用于实现极具前景的树聚类效果。最后，在应用部分，我们展示了如何将此方法用于对著名的氨酰-tRNA合成酶（aaRS）系统发育数据集中的系统发育树进行聚类；该数据集最初由 Woese等人（2000）描述。

进一步研究的一个有趣方向，是在聚类算法的目标函数中使用其他一些常用的树距离。其中，需要提及分支评分距离（Kuhner 和 Felsenstein，1994）和四元组距离（Bryant 等，2000）；与 Robinson-Foulds 距离的平方根一样，它们也具有欧几里得性质。

参考文献

Ball GH，Hall DJ（1965）ISODATA，一种新颖的数据分析与模式分类方法。门洛帕克：斯坦福研究院 Bapteste E，Boucher Y，Leigh J 等（2004）系统发育重建与基因侧向转移。Trends Microbiol 12(9):406-411 Barthélemy JP，Guénoche A（1991）树与邻近表示。奇切斯特：J. Wiley Barthélemy JP，McMorris FR（1986）。n树的中位数程序。J Classif 3(2):329-334 Barthélemy JP，Monjardet B（1981）聚类分析与社会选择理论中的中位数程序。Math Soc Scie 1(3):235-267 Baum BR（1992）将树作为组合系统发育推断数据集的一种方式，以及组合基因树的必要性。Taxon 41(1):3-10 Bininda-Emonds OR（编）（2004）系统发育超树：结合信息以揭示生命之树。施普林格科学与商业媒体 Bininda-Emonds OR，Cardillo M，Jones KE 等（2007）现生哺乳动物的延迟崛起。Nature 446:507-512 Boc A，Diallo AB，Makarenkov V（2012）T-REX：用于推断、验证和可视化系统发育树与网络的网络服务器。Nucleic Acids Research，40(W1):W573-W579 Boc A，Makarenkov V（2011）朝向准确识别嵌合基因与部分水平基因转移迈进。Nucleic Acids Research 39 (21):e144 Boc A，Philippe H，Makarenkov V（2010）利用二分划分差异度推断并验证水平基因转移事件。系统生物学 59(2):195-211 Bonnard C，Berry V，Lartillot N（2006）系统发育树的多极共识。系统生物学 55(5):837-843 Bradley PS，Mangasarian OL，Street WN（1997）通过凹函数最小化进行聚类。Adv Neural Inf Process Syst 9:368-374 Bryant D，Tsang J，Kearney PE 等（2000）计算进化树之间的四元组距离。SIAM J Appl Math. 9(11):285-286 Bryant D（2003）系统发育学共识方法的分类。离散数学与理论计算机科学 DIMACS 系列 61:163-184 Bullwinkle TJ，Ibba M（2014）涌现与进化。Top Curr Chem 344:43-87 Caliński T，Harabasz J（1974）用于聚类分析的树突法。Commun Stat-Theor M 3(1):1-27 Creevey CJ，Fitzpatrick DA，Philip GK 等（2004）类树系统发育是否仅存在于原核生物的末端? Proc R Soc Lond B Biol Sci 271(1557):2551-2558 Creevey CJ，McInerney JO（2005）CLANN：通过超树分析研究系统发育信息。生物信息学 21(3):390-392 Darlu P，Guénoche A（2011）用于评估基因树之间一致性的TreeOfTrees方法。J Classif 28:390-403 Daubin V，Gouy M，Perrière G（2001）利用超树方法进行细菌分子系统发育研究。Genome Inform 22:155-164

Day WH（1985）用于比较带标签叶节点的树的最优算法。J Classif 2:7–28；de Amorim RC、Mirkin B（2012）闵可夫斯基度量、特征加权及K均值聚类中的异常簇初始化。Pattern Recognit 45 (3):1061-1075；de Amorim RC、Makarenkov V（2016）在采用分布式质心的加权K均值中应用子聚类和Lp距离。Neurocomputing 173:700-707；Diallo AB、Makarenkov V、Blanchette M（2006）。寻找最大似然插入缺失情景。载于：Bourque, G., El-Mabrouk, N.（编）《比较基因组学》。RCG 2006。《计算机科学讲义》，第4205卷，171–185；Dohm JC、Vingron M、Staub E（2006）包括亮氨酸特异性亚型在内的氨酰-tRNA合成酶中的水平基因转移。J Mol Evol 63(4):437-447；Dong J、Fernández-Baca D、McMorris FR（2010）构建多数规则超树。Al-gorithms Mol Biol 5(1):2；Farris JS（1988）Hennig86，1.5版。由作者发行，纽约杰斐逊港站；Faurby S、Eiserhardt WL、Baker WJ 等（2016）棕榈科植物（棕榈科）的全证据物种级超树。Mol Phylogenet Evol 100:57–69；Felsenstein J（1985）系统发育树的置信界限：一种使用自助法的方法。Evol 39(4):783-791；Felsenstein J（1997）利用成对距离推断系统发育树的交替最小二乘法。Syst Biol 46(1):101–111；Gascuel O（2005）《进化与系统发育数学》。牛津（英国）：牛津大学出版社，第121-142页；Godwin RC、Macnamara LM、Alexander RW 等（2018）含有核心替换的tRNAMet的结构与动力学。ACS Omega 3(9):10668-10678；Guénoche A（2013）多重共识树：一种分离分化基因的方法。BMC生物信息学 14(1):46；Haeckel E（1866）《生物的一般形态学》[《生物的一般形态学》]。柏林：G. Reimer；Huang Z（1998）针对类别值大型数据集聚类的k均值算法扩展。数据挖掘与知识发现 2:283–304；Kaufman L、Rousseeuw PJ（1990）围绕中心点划分（PAM程序）。Wiley概率与统计丛书，第68–125页。Kimball RT、Oliveros CH、Wang N 等（2019）鸟类的系统基因组超树。多样性；Kuhner MK、Felsenstein J（1994）在进化速率相同和不同的情况下对系统发育算法进行模拟比较。分子生物学与进化 11(3):459-468；Lloyd SP（1957）二进制分组编码。贝尔实验室技术杂志 36(2):517-535；Lord E、Leclercq M、Boc、Diallo AB、Makarenkov V（2012）Armadillo 1.1：用于设计和开展系统发育分析与模拟的原创工作流平台。公共科学图书馆·综合。7(1):e29903；MacQueen J（1967）多变量观测值分类与分析的一些方法。载于《第五届伯克利数学统计与概率研讨会论文集》1(14):281-297；Maddison DR（1991）最简约树中多个孤岛的发现及其重要性。Syst Biol 40(3):315-328

Maddison DR、Schulz KS、Maddison WP（2007）生命之树网络项目。Zootaxa 1668:19-40；Makarenkov V、Leclerc B（1996）树度量的循环序及其在系统发育树重建与拟合中的应用。《数学层级与生物学》，第183-208页。Makarenkov V（1997）。《树距离的组合性质：算法与应用》。博士论文，巴黎，法国社会科学高等研究院。Makarenkov V、Leclerc B（1999）根据加权最小二乘准则拟合树度量的算法。J Classif 16 (1):3-26；Makarenkov V、Leclerc（2000）利用循环序比较加性树。J Comput Biol 7(5):731-744；Makarenkov V、Legendre P（2000）利用网状连接改进差异矩阵的加性树表示。载于：《数据分析、分类及相关方法》。施普林格，柏林，海德堡，35-40；Makarenkov V、Legendre P、Desdevises Y（2004）利用网状网络建模系统发育关系。Zool Scr 33 (1):89-96；Makarenkov V、Mazoure B、Rabusseau G 等（2021）SARS-CoV-2基因的水平基因转移与重组分析有助于发现其近缘种并揭示其起源。BMC Ecol Evo 21:5；Mank JE、Promislow DEL、Avise JC（2005）辐鳍鱼亲代抚育进化的系统发育视角。Evol 59:1570-1578；Margush T、McMorris FR（1981）共识n树。B Math Biol 43(2):239-244；McMorris FR、Wilkinson M（2011）保守超树。Syst Biol 60(2):232-238；Mirkin B（1996）《数学分类与聚类》。克鲁维尔学术出版社，1206；Mirkin B（2005）《数据挖掘聚类：一种数据恢复方法》。Chapman and Hall/CRC，910；Mirkin B、Fenner TI、Galperin MY 等（2003）用于计算基因组进化、末次普遍共同祖先以及水平基因转移在原核生物进化中的主导作用之简约进化情景的算法。BMC Evol Biolo 3(1):1-34；Moon J、Eulenstein O（2017）利用严格共识法综合大规模物种树。J Bioinform Comput Biol 15(3):1-17；Nelson G（1979）支序分析与综合：原则与定义，并附阿当松《植物科志》（1763-1764）的历史注记。Syst Zool 28:1-21；Ragan MA（1992）基于树的矩阵表示进行系统发育推断。Mol Phy-logenet Evol 1(1):53-58；Robinson DF、Foulds LR（1981）系统发育树的比较。Math Biosci 53(1-2):131-147；Rousseeuw PJ（1987）轮廓系数：用于解释和验证聚类分析的图示辅助工具。J Comput Appl Math 20:53-65；Saitou N（1988）最大似然法用于分子系统发育的性质与效率。J Mol Evol 27:261-273；Sharaf A、Gruber A、Jiroutová K 等（2019）色藻类中氨酰-tRNA合成酶的特征分析。Genes 10(8): 582。

Silva AS、Wilkinson M（2021）论树岛的定义与识别及大型岛屿偏差的缓解。系统生物学 70(6):1282-1294；Sokal RR、Michener CA（1958）一种评估系统关系的统计方法。堪萨斯大学科学通报 38:1409-1438；Sokal RR、Rohlf FJ（1981）Syst Zool 30:309-325；Stockham C、Wang LS、Warnow T（2002）通过聚类对系统发育分析进行基于统计的方法后处理。生物信息学 18(1):285-293；Swofford DL（1991）PAUP：基于简约法的系统发育分析，3.0q版。伊利诺伊州自然历史调查局，尚佩恩，伊利诺伊州；Swofford DL、Olsen GJ（1990）系统发育重建。载于：Hillis DM、Moritz C（编）《分子系统学》，Sinauer出版社，桑德兰，马萨诸塞州，第411-501页；Tahiri N、Willems M、Makarenkov V（2018）一种利用 k-中心点算法推断多个共识树的新型快速方法。BMC进化生物学 18(1):48；Tahiri N、Fichet B、Makarenkov V（2022）利用 k-均值算法和 Robinson-Foulds 距离构建备选共识树和超树，《生物信息学》（即将发表），btac326；Tibshirani R、Walther G、Hastie T（2001）利用 Gap 统计量估计数据集中的簇数。英国皇家统计学会B辑 63(2):411-423；Unvert KE、Kovacs FA、Zhang C 等（2017）亮氨酰-tRNA合成酶在真核生物物种形成过程中的进化。《美国本科生研究》14:69-83；Warnow T（2018）超树构建：机遇与挑战。arXiv 预印本，https://arxiv.org/abs/1805.03530；Wilkinson M、Cotton JA、Lapointe FJ 等（2007）共识情境下超树方法的性质。系统生物学 56(2):330-337；Woese CR、Olsen GJ、Ibba M 等（2000）氨酰-tRNA合成酶、遗传密码与进化过程。《微生物学与分子生物学评论》64(1):202-236