第1章 问题建模和描述

1.1 基本定义

本节将给出在研究进化树结构的过程中所需要的一些重要概念的定义。

定义 1.1.1 **有根二叉进化树**（简称进化树）是一棵叶节点被集合X中的元素所标记的满二叉树，即除叶节点没有子节点外，其余节点都有且仅有两个子节点，记作T。将T的所有边的集合记为E_T。将X称作该进化树的**标识集**。

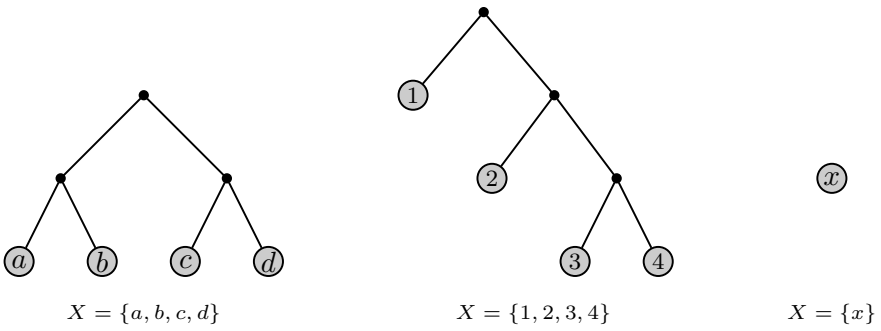

图 1–1 有根二叉进化树示例

定义 1.1.2 如果两棵进化树具有相同的标识集，并且**同构**，则认为这两棵进化树相等，记作$T_1 = T_2$。

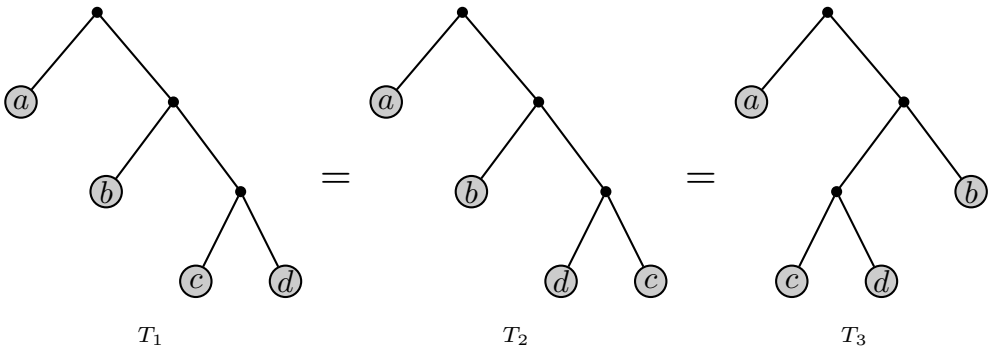

图 1–2 相等的3棵进化树

定义 1.1.3 将一棵二叉树删去所有只有一个子节点的内部节点以及所有未被标记的叶节点，使其变成一棵进化树的操作称为**收缩**。

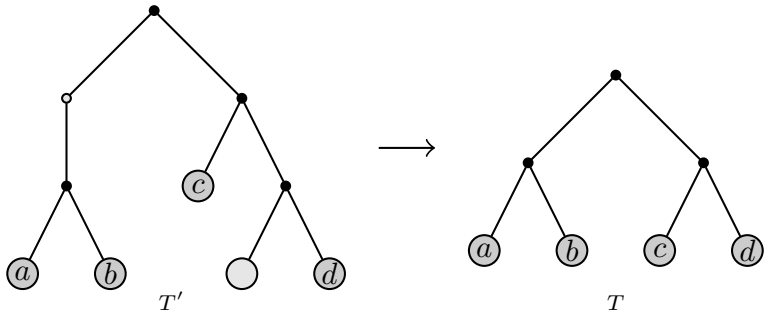

图 1–3 将T'收缩得到T

定义 1.1.4 **子树剪切再接**（rooted subtree prune and regraft，简称rSPR）对于一棵进化树T，剪去任意一条边得到一棵子树t，将t嫁接到余下子树$T-t$的一条边上，并对操作后的树进行一次**收缩**操作，获得一棵新的进化树，该过程称为一次**rSPR操作**。对于两棵进化树T_1, T_2，将其中一棵树转化为另一棵所需的最少SPR操作次数称为T_1, T_2的**rSPR距离**，记作$d(T_1, T_2)$。

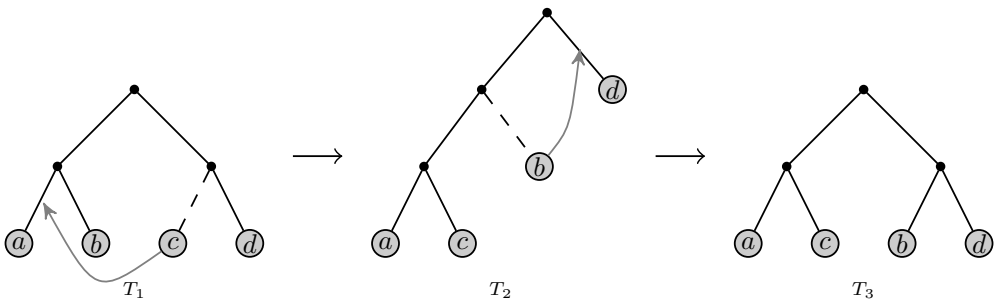

图 1–4 将T_1转化为T_3需要至少2次SPR操作，$d(T_1, T_3) = 2$

定义 1.1.5 **进化树森林**（简称森林）是由若干棵进化树组成的集合。记作$F = \{t_1, t_2, ..., t_n\}$。将一棵进化树$T$删去若干条边$E$，并对每棵独立的子树进行一次收缩操作得到的森林$F$，记作$F = T - E$。同理，将一个森林$F$删去若干条边$E$，并对每棵独立的子树进行一次收缩操作得到的森林$F'$，记作$F' = F - E$。

定义 1.1.6　最大一致森林（maximum-agreement forest，简称MAF）对于两棵进化树T_1, T_2，若存在$E_1 \subset E_{T_1}$，$E_2 \subset E_{T_2}$，使得$F = T_1 - E_1$并且$F = T_2 - E_2$，我们称F为T_1, T_2的**一致森林**。将T_1, T_2的所有一致森林中，包含最少进化树个数的森林成为**最大一致森林**，其包含的进化树个数记为$m(T_1, T_2)$。

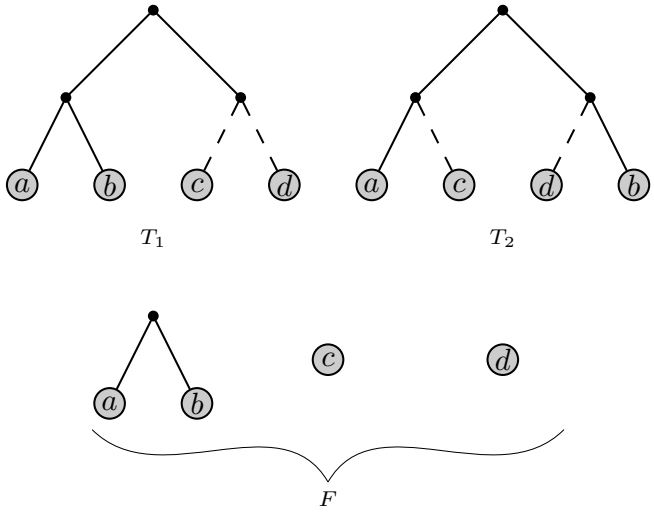

图 1–5 将T_1和T_2删去对应边后获得最大一致森林F，$m(T_1, T_2) = 3$

Bordewich和Semple[?]已经发现两棵进化树的rSPR距离与MAF的大小存在着等价关系，他们证明了如下定理：

定理 1.1.1　对于两棵进化树T_1, T_2，存在

$$d(T_1, T_2) = m(T_1, T_2) - 1$$

根据定理1.1.1我们可以将求解两棵进化树的rSPR距离转化为求解它们的最大一致森林的最优化问题，而因为该问题是NP完全问题，我们可以将其进一步转化为可以用固定参数算法求解的判定性问题，具体描述如下：

第2章 算法思想与设计

第3章 数据结构设计及算法实现

第4章 实验分析

第5章 总结及展望