arXiv:2302.10015v1 [q-bio.PE] 20 Feb 2023

基于扩展μ-表示法*比较果园网络

Gabriel Cardona[a]、Joan Carles Pons[a]、*、Gerard Ribas[b]、Tom´ás Mart´ínez Coronado[a]

[a]*Department of Mathematics and Computer Science – University of the Balearic Islands, Ctra. Valldemossa, km. 7.5, ES-07122, Palma, Spain*

[b]*Higher Polytechnic School – University of the Balearic Islands, Ctra. Valldemossa, km.*7.5*, ES-07122, Palma, Spain*

摘要

系统发育网络是系统发育树的推广形式，用于对网状进化事件进行建模。尽管系统发育树的比较已得到充分研究，且已有多种高效的比较方法，但系统发育网络的情况则大不相同。

已知某些系统发育网络类别，主要是树子网络，能够通过其 μ表示法进行高效分类；该表示法本质上是对每个节点，计算其到每个叶节点的路径数量。在本文中，我们引入了网络的扩展 μ表示法，其中也将到网状节点的路径数量纳入考虑。这一修改使我们能够区分果园网络，并在此类网络的空间上定义一种合理的度量，而且该度量还可以高效计算。

果园网络类别不仅具有生物学意义（其中一个此类网络可以解释为一棵带有涉及共存生物额外弧的树），还是数学意义上最一般的类别之一；对于这类网络，这种表示法或许可以存在，因为对定义稍作放宽，就会导致一个图同构完全问题。

关键词：系统发育网络、果园网络、网络同构、 μ-表示 2000 MSC：05C85、92D15

*由MCIN/AEI资助的PID2021-126114NB-C44号资助项目（10.13039/501100011033）以及“欧洲区域发展基金：建设欧洲的一种方式”提供支持。

*通讯作者

引言

系统发育网络将系统发育树推广到用于描述复杂的进化关系，这些关系无法很好地用树来拟合。具体而言，系统发育网络可以包含网状节点，即具有多个入弧的节点，用于描述现有操作分类单元（简称 OTU）之间的相互作用，并由此产生新的操作分类单元。这些相互作用可能包括重组、杂交或基因水平转移（以及其他类型），在植物、细菌和病毒等生物中日益明显，凸显了树表示的局限性 [1, 2]。在本文中，我们只考虑二元系统发育网络。这排除了以下两种情况：无法确定物种形成事件发生的先后顺序（用生物学术语来说，即“软”多分叉），以及在一次网状节点事件中有两个以上不同的操作分类单元发生相互作用。

系统发育网络的研究是一个不断发展且十分活跃的领域。然而，若研究目标是提出能够进行计算处理的对象，或表示具有生物学意义的情景，那么从最广义的角度研究系统发育网络并不合适。为实现这些目标，近年来文献中出现了系统发育网络的多个子类（近期综述见 [3] ）。本文将聚焦于果园（系统发育）网络这一子类 [4], ，它们也称为摘樱桃网络 [5], ；这类网络既能表示具有生物学意义的情景，也构成了一个在计算和数学方面都颇为重要的网络类别。

从生物学角度看，果园网络有一种自然的解释：这类网络可以理解为一棵带有水平弧（即涉及共存操作分类单元）的树，这些水平弧表示网状事件 [6]。关于这种时间一致标记（文中称为水平基因转移一致标记）的形式化描述，见[6] 中的定义 6。请注意，这种标记与类似语境中长期使用的其他时间一致性定义相似但并不相同，例如 [7] 中的定义。这种生物学解释也表明，果园网络是基于树的网络的一个子类 [8], ，其中额外的弧不必是水平弧。从数学角度看，果园网络是指通过反复应用约简（每次约简都会移除称为樱桃和网状樱桃的特定原子结构），最终可约简为平凡网络的网络。还请注意，果园网络包含树子网络 [9], ，而树子网络是研究最广泛的网络类别之一。

本文将探讨果园网络的比较问题。首先，要判断两个（原则上不同的）网络是否相同（更正式地说，是否同构）。其次，与此相关的是，量化这两个网络之间的差异（更正式地说，计算它们之间的某种距离）。

比较网络在实际应用中十分必要，因为采用不同的重建方法或使用不同的 DNA 样本，都可能得到不同的网络。通常的比较策略是为每个网络关联一些其同构类的不变量，这些不变量易于比较，然后在不变量空间中进行比较。由于判断两个一般的系统发育网络是否同构这一问题是图同构完全 [10], 的，人们认为，不可能存在既能区分任意系统发育网络、又能高效计算的不变量。因此，对于每一组不变量，都存在某一类系统发育网络，在该类网络上这些不变量能够区分网络（也就是说，不同网络具有不同的不变量）。这些不变量可以是网络中包含的某类子结构（比一般网络更简单），例如树 [11] 或三元组 [12, 13, 14, 15, 16]；也可以是叶节点 [14, 17, 18], 之间的距离，或与网络节点相关的某些数据 [9, 7, 4]。

属于这一最后类别的不变量之一是网络的 μ-表示，它已知可以区分树子代 [9]和半二叉、树兄弟且时间一致的 [19] 网络。该不变量使我们能够定义网络之间的 μ-距离；对于固定分类单元集合上的上述类别网络（同构类）所构成的空间而言，这是一种可靠距离，并将系统发育树上著名的 Robinson-Foulds 距离推广到了网络。网络的 μ-表示是其各节点的 μ-向量所构成的（多重）集合；这些向量分别统计从所考察节点到每个叶节点的路径数量。这些数据本质上与网络的祖先谱相同；祖先谱能够对无堆叠（即不存在由一条弧连接的网状节点对）果园网络进行分类， [20],但不能分类一般的果园网络（请注意， [20] 中的作者给出了反例，反驳了 [4,定理 2.2]）。

本文展示了如何通过修改 μ-向量（即同时统计从节点到任意网状节点的路径总数），将（二叉）果园网络区分开来，且无需施加其他限制。

本文结构如下。第 1 节给出全文所用的基本定义，第 2 节回顾果园网络的定义。第 3 节正式定义与网络相关联的（扩展） μ-表示，第 4 节证明该表示足以对果园网络进行分类，并用它在这类网络上定义一种合理的度量距离。最后，第 5 节讨论一些结语，并展示本文结果的一种实现。

1. 预备知识

本节给出全文将用到的一些定义。

对于任意正整数 n，我们用 $[n]$ 表示集合 $\{1,\dots,n\}$，用 $[n]^*$ 表示集合 $\{0,1,\dots,n\}$。

我们将讨论有向无环图 $N=(V,A)$。给定一个节点 $u\in V$，我们用 $\operatorname{indeg} u$（分别用 $\operatorname{outdeg} u$）表示以 u 为终点（分别为起点）的弧的数量。给定两个节点 $u,v\in V$，如果存在一条起点为 u、终点为 v 的弧（或从 u 到 v），记作 $uv\in A$，则称 u 是 v 的父节点，或称 v 是 u 的子节点。如果存在一条从 u 到 v 的有向路径，则称 u 是 v 的祖先，或称 v 是 u 的后代。注意，这也包括 $u=v$ 的情形，因为我们允许长度为 0 的平凡路径。

我们用 $M_N(u,v)$ 表示 N 中从 u 到 v 的路径集合，并用 $m_N(u,v)=|M_N(u,v)|$ 表示这类路径的数量；特别地，对于 N 中的每个节点 u，都有 $m_N(u,u)=1$。如果上下文中图的含义明确，我们将省略下标，直接写作 $M(u,v)$ 和 $m(u,v)$。

我们称有向图中的节点 u 为初等节点，当其入度为 $u=$、出度为 $u=1$ 时。对它进行化简，即将其（连同关联弧）移除，并连接它唯一的父节点和唯一的子节点。

系统发育网络，或简称网络，$N=(V,A,\varphi)$ 是定义在分类单元集合 X 上的有向无环图 (V,A)，其中没有平行弧，并且任意节点 $u\in V$ 均属于以下类型之一：

(i) 根节点，入度为 $u=0$，出度为 $u=1$（且此类节点至多有一个）；或 (ii) 叶节点，入度为 $u=1$，出度为 $u=0$；或 (iii) 树节点，入度为 $u=1$，出度为 $u=2$；或 (iv) 网状节点，入度为 $u=2$，出度为 $u=1$，

以及标记函数 φ，它是 X 与叶节点集合之间的双射。

系统发育树（或简称树）是不含网状节点的系统发育网络。

此后，我们将分类单元集合与叶节点集合视为相同，并假设 $X\subseteq V$，这意味着我们可以从网络的定义中删去 φ。给定一个系统发育网络 N，我们用 $V_H(N)$ 表示其网状节点集合，用 $V_T(N)$ 表示由叶节点或树节点构成的节点集合。此外，我们用 $\rho(N)$ 表示 N 的根节点。若上下文已明确网络所指，我们也可以分别简写为 V_H、V_T 和 ρ。

若 X 是单元素集合 $X = \{x\}$，我们用 I_x 表示这样一个网络：其唯一的节点是根节点和叶节点 x，两者由一条弧连接。

从现在起，我们始终假设 X 由正整数组成，因此对于某个 n， $X \subseteq [n]$ 。

最后，按照系统发育学中的惯例，我们称以 X 为底集的两个网络$N_1 = (V_1, A_1)$ 和 $N_2 = (V_2, A_2)$ 同构，记作 $N_1 \cong N_2$，如果存在一个双射 $f\colon V_1 \to V_2$ ，使其在 X 上为恒等映射，并且当且仅当 $uv \in A_1$ 时有 $f(u)f(v) \in A_2$。

2. 果园网络

本节总结 [5], 中的一些定义和结果，但请注意，我们使用的部分记号取自 [4], ，其余则是新引入的。

设 $N = (V, A)$ 是以 $X \subseteq [n]$ 为底集的网络，并设 $(i, j) \in X \times X$，其中 $i \neq j$。此外，分别用 p_i、p_j 表示 N 中叶节点 i 和 j 的父节点。若 $p_i = p_j$，则称 (i, j) 为一对樱桃；若 p_i 是网状节点，p_j 是树节点，且 p_j 是 p_i 的父节点之一，则称其为一对网状樱桃。无论 (i, j) 是一对樱桃还是一对网状樱桃，我们都称 (i, j) 在 N 中是可约的。

如果 (i, j) 在 N 中是可约的，那么 (i, j) 在 N 中的约简（记作 $N^{(i,j)}$）如下所示:

- 如果 (i, j) 是一个樱桃，则移除叶节点 i （及其入弧），并简化 p_i，它现在是一个初等节点。
- 如果 (i, j) 是一个网状樱桃，则删除弧 $p_j p_i$，然后简化 p_i 和p_j，它们现在都是初等节点。

显然，成为樱桃和成为网状樱桃的条件互不相容，这意味着该操作定义良好。

给定一个长度为$k \geq 1$的整数对序列 $S = (s_1, \ldots, s_k)$，为简洁起见，我们将其写作 $S = s_1 \cdots s_k$，其中 $s_t = (i_t, j_t)$ 且 $i_t, j_t \in [n]$，我们称 S 在 N 中是可约的，当且仅当:

- s_1 在 N 中是可约的。

对于每个 $t \in \{2, \ldots, k\}$， s_t 在 $(\ldots (N^{s_1})^{s_2} \ldots)^{s_{t-1}}$ 中都是可约的。

在这种情况下，我们将 N 关于 S 的约简定义为 $(\ldots (N^{s_1})^{s_2} \ldots)^{s_k}$，并记作 N^S。

此外，如果存在某个 $i \in X$ 使得 $N^S = I_i$ ，我们就称 S 为完全的；如果存在这样一个完全序列，则称 N 为果园网络。

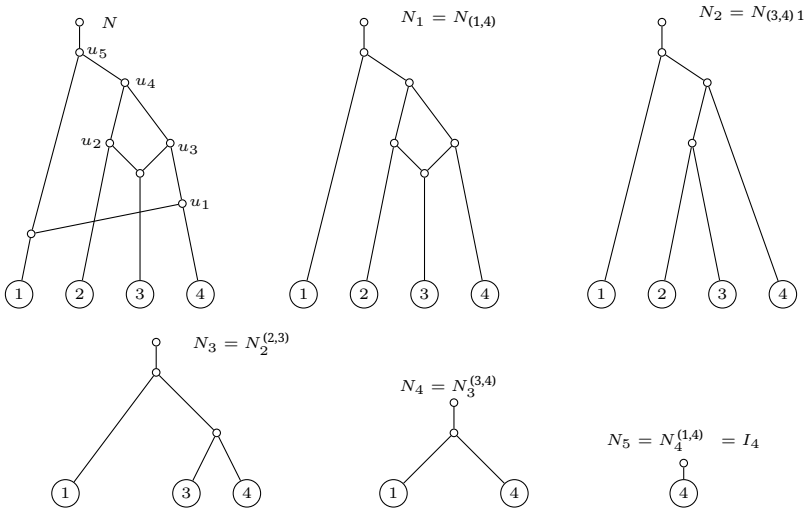

图 1：一个果园网络 N 以及通过（樱桃或网状樱桃）约简得到的中间网络集合，直至到达 N^S，即相对于完备序列 $S = (1, 4)(3, 4)(2, 3)(3, 4)(1, 4)$ 对 N 的约简。注意，例如，(1, 4) 这一对在 N 中是网状樱桃，而 (2, 3) 这一对在 N_2 中是樱桃。

利用完全可约序列对果园网络进行分类的基本结果如下，该结果改编自 $[5, \text{定理推论 } 1]$。

定理1. 设 S 是两个果园网 N 和 N'的一条完整约简序列。那么，$N \cong N'$。

然而请注意，果园网络的完整约简序列通常并不唯一。

图 1 展示了一个果园网络，以及一系列将其约简为平凡网络的约简。请注意，我们可以取这个平凡网络，逆转已执行的约简，并且仅利用已执行的约简序列所包含的信息，恢复初始网络。这基本上就是定理1的证明。

3. 网络的扩展 μ-表示

本节将 μ-表示的定义从 [9] 扩展，并定义一种约简操作，用于约简此类数据中的某些构件。下一节将介绍这一扩展如何使我们能够刻画果园网络，这与原始定义能够比较树子网络的方式类似。

原始 μ-表示与祖先谱

我们首先回顾 μ-表示法 [9] 和祖先谱 [4], 的定义，并略作修改，以使符号与本文其余部分保持一致。

设 $N = (V, A)$ 是定义在 $X \subseteq [n]$ 上的网络。对于每个节点 $u \in V$ 和 $i \in [n]$，我们将 $\mu_i(u)$ 定义为从 u 出发并以叶节点 i为终点的路径数，即

$$\mu_i(u) = m(u, i).$$

显然，若 $i \notin X$，则 $\mu_i(u) = 0$。u 的 μ-向量为

$$\mu(u) = (\mu_1(u), \ldots, \mu_n(u)).$$

若上下文中无法明确判断网络，我们将通过写作 $\mu^N(u) = (\mu_1^N(u) \ldots, \mu_n^N(u))$ 来指明。

N 的 μ-表示法是以下（多重）集合

$$\boldsymbol{\mu}(N) = \{\mu(u) \mid u \in V_T\}.$$

请注意，在 [9], 中，所有节点都纳入了考虑，也就是说，$\boldsymbol{\mu}(N)$ 还包含网状节点的 μ-向量；但由于此类节点的 μ-向量与其唯一子节点的向量相同，因此这两种表示法是等价的。

现在任取一个固定的树节点排列，例如 $v_1, \ldots, v_t$。对于每个 $i \in X$，考虑其祖先元组 $\sigma(i) = (m(v_1, i) \ldots, m(v_t, i))$，并令 $\Sigma_N = \{(i, \sigma(i)) \mid i \in X\}$ 为该网络的祖先谱。

请注意，我们可以任意重新排列树节点（即，对节点应用置换后会得到同构网络），这相当于对所有祖先元组的各项应用同一个置换；但是祖先元组的序列不能重新排列（即，对叶节点进行置换会得到非同构网络）。因此，从数学上讲，我们得到的对象是一个（有序的）序列，其元素是（有序的）序列，而对后者，对称群会同时作用于所有序列。

这个对象等价于 μ-表示，因为二者都统计树节点与叶节点之间的路径数；但在 μ-表示的情形下，该对象更简单，因为我们得到的是一个集合，其元素为（有序的）序列，且不再有其他作用。

网络的 μ表示（或等价地，其祖先谱）已知能够刻画不同类型的系统发育网络，即树子代 [9], 半二叉树兄弟时间一致 [19], 以及无堆叠果园 [20] 网络。

扩展 μ向量

我们通过将 $\mu_0(u)$定义为从 u 出发并终止于 N的网状节点的路径数，扩展 μ-向量的定义，即

$$\mu_0(u) = \sum_{h \in V_H} m(u, h).$$

显然，如果 N 没有网状节点，那么 $\mu_0(u) = 0$。那么， u 的扩展μ-向量为向量

$$\mu(u) = (\mu_0(u), \mu_1(u), \ldots, \mu_n(u)).$$

与之前一样，如果上下文中无法明确网络是哪一个，我们将通过写作 $\mu^N(u) = (\mu_0^N(u)\ \mu_1^N(u) \ldots, \mu_n^N(u))$来指明。

N 的扩展 μ-表示是（多重）集合

$$\boldsymbol{\mu}(N) = \{\mu(u) \mid u \in V_T\}.$$

表 1 展示了图 1 所示网络 N 的扩展 μ-表示。

u	$\mu(u)$
1	(0, 1, 0, 0, 0)
2	(0, 0, 1, 0, 0)
3	(0, 0, 0, 1, 0)
4	(0, 0, 0, 0, 1)
u_1	(1, 1, 0, 0, 1)
u_2	(1, 0, 1, 1, 0)
u_3	(2, 1, 0, 1, 1)
u_4	(3, 1, 1, 2, 1)
u_5	(4, 2, 1, 2, 1)

表 1：图 1 中网络 N 的扩展 μ-表示。

有时我们会省略“扩展”一词，直接称为 μ-向量和 μ-表示。如果指的是原始定义 [9] （即不计算通往网状节点的路径数），我们会明确说明。另请注意，我们将 $\boldsymbol{\mu}(N)$定义为一个多重集，其中每个向量μ 出现的次数，等于具有该 μ-向量的内部树节点的数量。稍后我们会看到，在我们关注的情形中，不同的树节点不会有相同的 μ-向量，因此可以将其视为集合。目前，我们仍将其视为多重集。

对于任意 $S \subseteq [n]^*$，令 δ_S 为一个向量，其 i分量（从 0 到 n编号）在 $i \in S$ 时为 1，在 $i \in [n]^* \setminus S$时为 0。为简化记号，对于任意 j_1、……、$j_k \in [n]^*$，我们定义 $\delta_{j_1,\ \ldots\ldots,\ j_k} = \delta_{\{j_1,\ldots,j_k\}}$；特别地，如果 $j \in [n]^*$，则 $\delta_j = \delta_{\{j\}}$。注意，$\boldsymbol{\mu}(I_i) = \{\delta_i\}$，其中回顾一下，$I_i$ 是仅有一个叶节点 i的平凡树。

下面的命题表明，网络$N = (V, A)$在 $X \subseteq [n]$ 上的 μ-表示可以按顺序计算，例如，按高度递增的顺序访问各节点（高度是从该节点出发的最长路径的长度，并可在 $O(|V|)$时间内计算）。因此，此类网络的 μ-表示可以在$O(|V|n)$时间内计算。

命题 2. 设 $u \in V$。则：

- 若 u 是叶节点 i，则 $\mu(u) = \delta_i$。
- 若 u 是树节点，则 $\mu(u) = \mu(v)+\mu(w)$，其中 v、w 是 u的子节点。
- 若 u是网状节点，则 $\mu(u) = \mu(v)+\delta_0$，其中 v是 u唯一的子节点。

证明。显然，从叶节点出发的路径只有平凡路径，且终止于同一个叶节点。这证明了第一个断言。

若 u 是一个树节点，其子节点为 v 和 w，则对于每个叶节点 i，$M(u, i)$ 中的路径集合与 $M(v, i)$ 和 $M(w, i)$ 中的路径的不交并之间存在一个双射，分别通过在路径前添加或删除弧 uv和 uw建立。由于 u 不是网状节点，这一结论同样适用于通向网状节点的路径。

若 u是网状节点，且 v 是其唯一子节点，则对于每个叶节点 i，$M(u, i)$ 中的路径集合与 $M(v, i)$ 中的路径集合之间存在一个双射，可通过在路径前添加或删除弧 uv建立。至于从 u 通向网状节点的路径，除那些在从 v 出发的路径前添加弧 uv 而得到的路径外，还会有一条从 u出发（并以其为终点）的新路径（平凡路径）。□

“是……的后代”这一关系赋予网络的节点集合偏序集结构。节点的 μ-向量也具有这样的结构，即对于每个 $i \in [n]^*$，当且仅当 $\mu_i(v) \leq \mu_i(u)$ 时，$\mu(v) \leq \mu(u)$。下面的技术引理将说明这两个偏序之间的关系，后文将用到它。

引理 3. 设 u、v 是网络 N 的两个节点。

- 如果 v 是 u 的后代，那么 $\mu(v) \leq \mu(u)$。
- 此外，如果 u 和 v 是不同的树节点，那么 $\mu(u) \neq \mu(v)$。

证明。第一个陈述可由命题 2 轻易推出。

对于第二个陈述，假设存在一条非平凡路径，从树节点 u 通往另一个树节点 v，且 $\mu(u) = \mu(v)$。由陈述的第一部分可知，这条路径上的所有节点都具有相同的 μ 向量。设 $w \neq u$ 为这条路径上最先经过的节点（并注意到 w 是 u 的子节点）。如果 w 是网状节点（注意，由于 v 是树节点，因此 $w \neq v$），考虑其唯一的子节点 w'；于是，$\mu_0(w) = \mu_0(w') + 1 > \mu_0(w')$，由此得到 $\mu_0(u) \geq \mu_0(w) > \mu_0(w') \geq \mu_0(v)$，这与假设矛盾。如果 u 是树节点，且 w' 是 u 中不同于 w的那个子节点，那么 $\mu(u) = \mu(w) + \mu(w')$，同样会导出矛盾，因为 $\mu(w') \neq 0$。 □

μ-表示中的樱桃和网状樱桃

现在，我们仿照为网络的叶节点对定义的樱桃和网状樱桃，将其定义应用于向量（多重）集中的索引对，并证明这些定义是合理的，因为对一个网络取 μ-表示，会将前者转化为后者。

设 $\boldsymbol{\mu}$ 是由长度为 $n+1$ 的向量构成的（多重）集，这些向量的分量均为非负整数，并由 $0, 1, \ldots, n$ 索引。若满足以下条件，我们称 (i, j) 是 $\boldsymbol{\mu}$ 的樱桃：

1. $\delta_{i,j} \in \boldsymbol{\mu}$（在多重集中出现一次），且 2. 对每个 $\mu = (\mu_0, \mu_1, \ldots, \mu_n) \in \boldsymbol{\mu} \setminus \{\delta_i, \delta_j\}$，都有 $\mu_i = \mu_j$。

命题 4. 设 $i, j \in [n]$。那么，(i, j) 是 N 的樱桃，当且仅当它是 $\boldsymbol{\mu}(N)$ 的樱桃。

证明。假设（i, j）是 N的一个樱桃，并令 p为其共同父节点。显然，$p \in V_T$ 和 μ（p）$= \delta_{i,\ j}$属于 $\boldsymbol{\mu}$（N）。

假设 $p' \neq p$ 是一个满足 μ（p'）$= \delta_{i,\ j}$的树节点。由于从根节点 ρ到 j 的所有路径都必须经过 p，因此 p' 要么是 p的后代，要么是其祖先。由引理 3可知 $p = p'$，这与假设矛盾。

对于每个树节点 u，通过添加和移除弧 pi（分别为 pj），$M(u, i)$（分别为 $M(u, j)$）中的路径集合与 $M(u, p)$ 中的路径集合之间存在双射；因此 $\mu_i(u) = m(u, i) = m(u, p) = m(u, j) = \mu_j(u)$。最后，若 u 是一个不同于 i, j 的叶节点，则显然 $\mu_i(u) = \mu_j(u) = 0$。

反过来，假设 (i,j) 是 $\boldsymbol{\mu}(N)$ 中的一个樱桃，并设 $p \in V_T$ 满足 $\mu(p) = \delta_{i,\ j}$。由于 $\delta_0(p) = 0$，从 p 出发无法到达任何网状节点，这意味着以 p 为根的子网络必定由树节点构成，因此它是一棵不含初等节点的树，且其唯一的叶节点为 i, j（因为对于其他每个 $k \in [n]$，$\delta_k(p) = 0$）。因此，这棵子树必定是一个樱桃。

□

我们称 (i,j) 是 $\boldsymbol{\mu}$ 的网状樱桃，当且仅当：

1. $\delta_{0,\ i,\ j} \in \boldsymbol{\mu}$（在多重集中重数为 1）；并且 2. 对每个 $\mu = (\mu_0, \mu_1, \ldots, \mu_n) \in \boldsymbol{\mu} \setminus \{\delta_i, \delta_j\}$，都有 $\mu_0 \geq \mu_i \geq \mu_j$。

命题5. 设 $i, j \in [n]$，那么 (i,j) 是 N 的网状樱桃，当且仅当它是 $\boldsymbol{\mu}(N)$ 的网状樱桃。

证明。假设 (i,j) 是 N 的网状樱桃，并设 p, q 分别是 i, j 的父节点，这意味着 p 是一个网状节点，q 是一个树节点，且存在一条弧 qp。显然，$\mu(q) = \delta_{0,\ i,\ j}$。

假设 $q' \neq q$ 是一个树节点，且 $\mu(q) = \delta_{0,\ i,\ j}$。由于从 ρ 到 j 的所有路径都必须经过 q，因此 q' 要么是 q 的后代，要么是其祖先。由引理 3 可知 $q = q'$，这与假设矛盾。

令 u为 N的任意一个树节点。首先注意到，通过添加和移除弧 pi，$M(u,i)$ 中的路径与 $M(u,p)$ 中的路径构成双射，因此 $m(u,i) = m(u,p)$。由于 p是一个网状节点，因此可得$\mu_0(u) \geq m(u,p) = m(u,i) = \mu_i(u)$。其次注意到，$M(u,j)$ 中的所有路径都必须经过 q，而 $M(u,i)$ 中的路径则可以经过 q，也可以经过 p除 q以外的另一个父节点。由此可得$\mu_i(u) = m(u,i) \geq m(u,j) = \mu_j(u)$，最后，条件 $\mu_0(u) \geq \mu_i(u) \geq \mu_j(u)$ 成立。

如果 $u \neq i, j$是 N的叶节点，显然 $\mu_0(u) = \mu_i(u) = \mu_j(u) = 0$，并且条件 $\mu_0(u) \geq \mu_i(u) \geq \mu_j(u)$ 成立。

反过来，假设 (i,j) 是 $\boldsymbol{\mu}(N)$ 的一个网状樱桃，并令$u \in V_T$ 满足 $\mu(u) = \delta_{0,\ i,\ j}$。由于从 u 可到达的叶节点为i, j，以 u 为根的 N 的子网络必定是一棵树，其初等节点在 N中均为网状节点。由于 $\mu_0(u) = 1$，从 u 到网状节点只有一条路径，这意味着从 u可到达的网状节点至多有一个。因此，以 u 为根的子网络必定是一个樱桃，其中一片叶节点上方插入了一个初等节点。令 p, q 分别为 i, j的父节点。那么，要么 $u = q$ 且存在一条弧 qp，要么 $u = p$ 且该弧为 pq。在第二种情况下，令 v 为 N根节点的唯一子节点，它必然属于 V_T；$M(v,i)$中的所有路径都必须经过

p，而以 j 为终点的路径则可以经过 p ，也可以经过 q 的另一个父节点（不同于 p ）（且这类路径至少存在一条）；由此可得 $\mu_j(v) > \mu_i(v)$，这与假设矛盾。因此，存在一条弧qp ，且 (i,j) 是一个网状樱桃。 □

推论 6. 如果 $\boldsymbol{\mu}$ 是某个果园网络的 μ-表示，那么 (i,j) 是 $\boldsymbol{\mu}$ 的樱桃和网状樱桃这两个条件互斥。

如果 (i,j) 在 $\boldsymbol{\mu}$ 中是樱桃或网状樱桃，我们就称它为可约。采用这一记号，命题 4 和 5 可以写成如下形式。

推论 7. 设 $i,j \in [n]$。那么，(i,j) 在 N 中可约，当且仅当它在 $\boldsymbol{\mu}(N)$ 中可约。

μ-表示的约简

现在，我们将约简的概念从网络扩展到向量（多重）集。与之前一样，这一扩展与网络的 μ-表示的计算相兼容。

设 $\boldsymbol{\mu}$ 是一个长度为 $n+1$ 、各项均为非负整数的向量（多重）集，并假设 (i,j) 是 $\boldsymbol{\mu}$ 的樱桃。我们将 $\boldsymbol{\mu}$ 关于 (i,j) 的樱桃约简定义为通过以下运算从 $\boldsymbol{\mu}$ 得到的（多重）集：

- 从 $\boldsymbol{\mu}$ 中移除 δ_i 和$\delta_{i,\ j}$；
- For every other $\mu = (\mu_0, \mu_1, \ldots, \mu_n) \in \boldsymbol{\mu}$, set $\mu_i = 0$.

更正式地说，它是以下（多重）集合：

$$\{\mu - \mu_i\delta_i \mid \mu = (\mu_0, \mu_1, \ldots, \mu_n) \in \boldsymbol{\mu} \setminus \{\delta_i, \delta_{i,j}\}\}.$$

设 $\boldsymbol{\mu}$ 是一个由长度为 $n+1$ 的向量组成的（多重）集合，这些向量的分量均为非负整数，并假设 (i,j) 是 $\boldsymbol{\mu}$ 的一个网状樱桃。我们将相对于 (i,j) 对 $\boldsymbol{\mu}$ 进行网状樱桃约简定义为通过以下运算从 $\boldsymbol{\mu}$ 得到的（多重）集合：

- 从 $\boldsymbol{\mu}$中移除$\delta_{0,\ i,\ j}$；
- 对于其他每个 $\mu = (\mu_0, \mu_1, \ldots, \mu_n) \in \boldsymbol{\mu}$，从 μ_0 中减去 μ_i，并将 μ_j 加到 μ_i。

更形式化地说，它是这个（多重）集合：

$$\{\mu - \mu_i\delta_0 - \mu_j\delta_i \mid \mu = (\mu_0, \mu_1, \ldots, \mu_n) \in \boldsymbol{\mu} \setminus \{\delta_{0,i,j}\}\}.$$

如果有序对 (i, j) 在 $\boldsymbol{\mu}$ 中不是同时为樱桃和网状樱桃（请注意，根据推论 6，在处理果园网络的 μ-表示时，这种情况总是成立），我们将其简单地称为相对于 (i, j) 的 $\boldsymbol{\mu}$ 约简，并记作 $\boldsymbol{\mu}^{(i,j)}$。还请注意，修改 μ-向量的过程原则上可能导致重复元素（即使 $\boldsymbol{\mu}$ 中没有重复元素），但由于我们将 $\boldsymbol{\mu}$ 视为多重集，因此可以将这些重复计入其中。如前所述，在我们所关注的情形中，这种情况永远不会发生。

命题 8. 设 (i, j) 是网络 N 中的一个可约对。那么，

$$\boldsymbol{\mu}(N^{(i,j)}) = \boldsymbol{\mu}(N)^{(i,j)}.$$

证明。首先假设 (i, j) 是 N 中的一个樱桃，这意味着 (i, j) 是 $\boldsymbol{\mu}(N)$ 中的一个樱桃。为简便起见，记 $\tilde{N} = N^{(i,j)}$。

注意，p（i 和 j的父节点）属于 V_T，并且是唯一一个在抑制节点 i后变为初等节点的节点。显然，$\mu_p(j) = \delta_i$不会对 $\boldsymbol{\mu}(N)$有所贡献。对于 N的任意其他节点 u，我们有 u也属于 $\tilde{N}$。此外，$\mu_k^N(u) = m_{\tilde{N}}(u, k) = m_N(u, k)\mu_k^N(u)$（对每个$k \in [n] \setminus \{i\}$都成立），因为 $M_N(u, k)$中的路径与 $M_{\tilde{N}}(u, k)$中的路径一一对应；而 $\mu_i^N(u) = m_{\tilde{N}}(u, i) = 0$，因为 i不再是N的叶节点。因此，N 的 μ-表示与 N 的μ-表示相对于(i, j)的约简相一致。

现在假设 (i, j) 是 N 中的一个网状樱桃，这意味着 (i, j) 是 $\boldsymbol{\mu}(N)$ 中的一个网状樱桃。与之前一样，我们记作 $\tilde{N} = N^{(i,j)}$。

注意，p, q（分别是 i, j 的父节点）在移除连接它们的弧后会变成初等节点。然而，p 是 N 中的一个网状节点，因此其 μ-向量不会对 $\boldsymbol{\mu}(N)$ 产生贡献。于是，q 属于 V_T，且其在 N 中的 μ-向量为 $\delta_{0,i,j}$，它是唯一一个对 $\boldsymbol{\mu}(N)$ 的贡献会在 $\boldsymbol{\mu}(N)$ 中消失的节点。

对于 u 中的任意其他节点 N，我们都有 u 也属于 N。$M_N(u, i)$ 中的路径可以分为两类：一类经过 q（与 $M_N(u, j)$ 双射），另一类经过 p 的另一个父节点。移除弧qp后，前一类将不再是 N 中的路径，而后一类将与 $M_{\tilde{N}}(u, i)$ 中的路径双射。因此，我们得到 $\mu_i^N(u) = m_{\tilde{N}}(u, i) = m_N(u, i) - m_N(u, j) = \mu_i^N(u) - \mu_j^N(u)$。注意

˜还要注意， p 是 N 中唯一一个在 N中消失的网状节点，并且˜ $M_N(u,p)$ 中的路径与 $M_N(u,i)$ 中的路径双射。因此， $\mu_0^N(u) = \mu_0^N(u) - m_N(u,p) = \mu_0^N(u) - \mu_i^N(u)$。因此，与前一种情况一样，˜我们要从 $\boldsymbol{\mu}(N)$ 得到 $\boldsymbol{\mu}(N)$ 所需执行的运算，与计算 $\boldsymbol{\mu}(N)^{(i,j)}$ 时所需执行的运算完全相同。 □

4. 使用 $\boldsymbol{\mu}$-表示对果园网络进行分类

在上一节中，我们介绍了与网络相关联的 μ-表示，以及能够消除樱桃和网状樱桃的约简。我们只是在偶然情况下提到了果园网络。本节将说明如何利用 μ-表示刻画果园网络（在同构意义下），并定义它们之间的可靠距离度量。

果园网络中 μ-向量的唯一性

我们此前已经宣布（但未给出证明）：尽管网络的 μ-表示被定义为多重集，但对于果园网络，不存在重复的 μ-向量，因此可以将其视为集合。现在我们证明这一结论。

命题9. 设 N是一个果园网络，且 $u, v \in V_T(N)$。若 $\mu(u) = \mu(v)$，则 $u = v$。

证明. 若 u 或 v 是叶节点，则结论显然成立。

设 S 是 N的一个完全可约序列，并考虑中间网络 $N_k = (\ldots(N^{s_1})^{s_2}\ldots)^{s_k}$。在此约简过程中，某个时刻 u 和 v 最终都会被简化，因为它们变成了初等节点。假设 u、v 属于 N_k，但 u 不属于 N_{k+1}。若令 $(i,j) = s_{k+1}$，那么根据 s_{k+1} 是樱桃还是网状樱桃，要么有 $\mu^{N_k}(u) = \delta_{i,j}$，要么有 $\mu^{N_k}(u) = \delta_{0,i,j}$。现在，由于 $\boldsymbol{\mu}(N_k) = (\ldots(\boldsymbol{\mu}(N)^{s_1})^{s_2}\ldots)^{s_k}$，我们有 $\mu^{N_k}(v) = \mu^{N_k}(u)$。若 $u \neq v$，则 δ_i、$_j$ （或 δ_0、$_i$、$_j$）在 $\boldsymbol{\mu}(N_k)$中的重数至少为 2，这与假设 s_{k+1} 是 $\boldsymbol{\mu}(N_k)$的一个可约对相矛盾。 □

推论10。果园网络 N 的 μ-表示 $\boldsymbol{\mu}(N)$ 是一个集合。

可约序列与果园网络的同构

现在我们将说明，迄今为止引入的定义和结果如何使我们能够证明，网络的 μ-表示可以唯一确定其同构类。为此，我们将网络的完全可约序列这一概念转化为 μ-表示，并利用定理1的分类结果。

给定一个长度为 $k \geq 1$ 的整数对序列 $S = s_1 \cdots s_k$，其中 $s_t = (i_t, j_t)$，若满足以下条件，我们称 S 在 $\boldsymbol{\mu}$ 中可约:

- s_1 在 $\boldsymbol{\mu}$ 中可约。
- 对每个 $t \in 2, \ldots, k$， s_t 在 $(\ldots(\boldsymbol{\mu}^{s_1})^{s_2}\ldots)^{s_{t-1}}$ 中可约。

在这种情况下，我们将 $\boldsymbol{\mu}$ 关于 S 的约简定义为 $(\ldots(\boldsymbol{\mu}^{s_1})^{s_2}\ldots)^{s_k}$，并记作 $\boldsymbol{\mu}^S$。

定理11. 一个整数对序列在果园网络 N 中可约，当且仅当它在 $\boldsymbol{\mu}(N)$ 中可约。此外，在这种情况下，

$$\boldsymbol{\mu}(N^S) = \boldsymbol{\mu}(N)^S.$$

证明。由推论 7 和命题 8 的迭代应用可得。

□

我们称 S 是 $\boldsymbol{\mu}$ 中的可约序列，若对于某个整数 i， $\boldsymbol{\mu}^S = \{\delta_i\}$ ，则称其为完全的。

推论12。 N 的可约序列是完全的，当且仅当它对于 $\boldsymbol{\mu}(N)$ 也是完全的。

定理13。设 N、N' 是 X、$X' \subseteq [n]$ 上的两个果园网络。那么，$N \cong N'$ 当且仅当 $\boldsymbol{\mu}(N) = \boldsymbol{\mu}(N')$。

证明。该陈述的“仅当”部分显然。反过来，假设 $\boldsymbol{\mu}(N) = \boldsymbol{\mu}(N')$，并设 S 是 N 的一个完全可约序列。那么，由推论12可知， S 也是 $\boldsymbol{\mu}(N)$ 的一个完全可约序列；又由于 $\boldsymbol{\mu}(N) = \boldsymbol{\mu}(N')$，再次应用推论12可得， S 也是 N' 的一个完全可约序列。最后，应用定理1可得 $N \cong N'$。

□

本节最后指出， μ-表示比原始 μ-表示（或等价地说，祖先谱）具有更强的区分能力。例如，图 2 中的网络（改编自

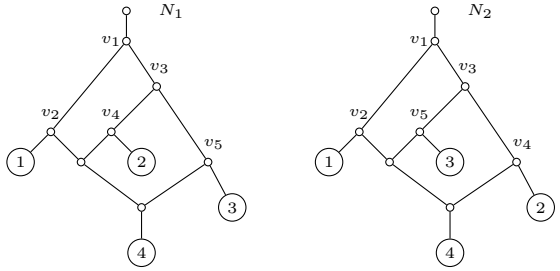

图 2：具有相同 *original* μ-表示但具有不同 *extended* μ-表示的非同构网络。

[20, 图 2]）具有相同的原始 μ-表示（或祖先谱），但扩展的 μ-表示不同。事实上，内部树节点的排序使得所有叶节点的祖先元组都相同，但 $\mu^{N_1}(v_4)$ $=(2, 0, 1, 0, 1)$，$\mu^{N_1}(v_5)$ $=(1, 0, 0, 1, 1)$，而 $\mu^{N_2}(v_4)$ $=(1, 0, 1, 0, 1)$，$\mu^{N_2}(v_5)$ $=(2, 0, 0, 1, 1)$。

果园网络类上的可靠距离

作为一个附带结果，我们发现可以利用 μ-表示，在果园网络（同构类）的集合上定义一种可靠距离度量。我们回顾一下，定义这种距离相当于给出一个映射，为每一对网络 N_1, N_2 赋予一个实数 $d(N_1, N_2)$，并满足以下性质：

1.（正性） $d(N_1, N_2) \geq 0$；2.（分离性） $d(N_1, N_2) = 0$ 当且仅当 $N_1 \cong N_2$；3.（对称性） $d(N_1, N_2) = d(N_2, N_1)$；4.（三角不等式） $d(N_1, N_3) \leq d(N_1, N_2) + d(N_2, N_3)$。

定理14。映射

$$d_\mu(N_1, N_2) := |\boldsymbol{\mu}(N_1) \triangle \boldsymbol{\mu}(N_2)|,$$

在分类单元集合包含于 $[n]$ 的果园网络同构类集合上定义了一种距离。

证明。由于该映射基于集合对称差的基数，只需验证分离性，这由定理13给出。

□

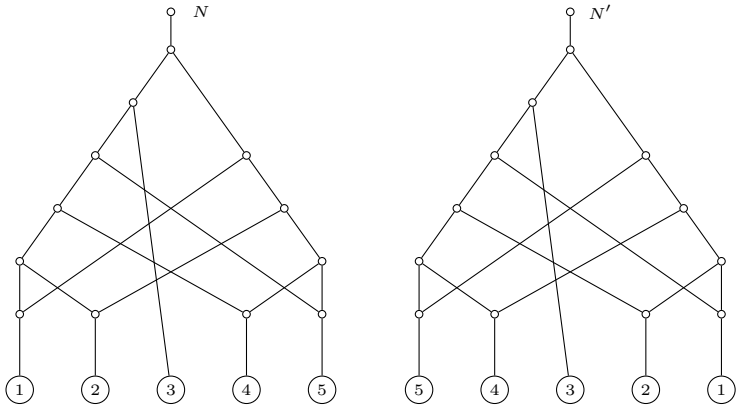

图 3：具有相同 μ-表示的非同构网络。

请注意，与原始的 μ-距离 [9], 的情况一样，刚刚定义的距离 d_μ 在限制到系统发育树时，给出的就是经典的 Robinson-Foulds 距离 [21]。还要注意，距离通常定义在具有相同分类单元集合的系统发育网络（或树）之间，但只要不同网络的分类单元集合都包含于某个共同的超集中， d_μ 也可用于这些网络之间。

5. 结论

本文提出了网络的 μ-表示的一种扩展，该扩展能够区分任意果园网络并为其定义距离，且计算效率高。不过请注意， μ-表示无法区分更一般类型的网络。例如，图 3 中的两个网络是树兄弟 [19], 栈自由 [22], 可分离 [23], FU-稳定 [24] 且树基 [8], 的，它们具有相同的μ-表示，但并不同构。易于计算和比较的不变量无法区分任意网络，这一点并不令人意外，因为正如前文所述，即便对于树兄弟时间一致网络这一（限制颇多的）情形，比较它们也是图同构完全问题。

我们已实现使用可约序列对果园网络进行分解与重构、计算网络的 μ-表示、根据其 μ-表示重构果园网络（若原网络是果园网络，则可恢复原网络），以及计算网络之间的 μ-距离。这一实现

已收录于 Python 软件包 phylonetwork 2.2 版，该软件包可从 https://pypi.org/project/phylonetwork/ 获取。我们还制作了上述功能的演示，可通过 https://github.com/gerardet46/OrchardMu 访问。

可通过其中提供的链接 https://mybinder.org/ 在线运行。

参考文献

[1] W. Martin，镶嵌细菌染色体：通往基因组树之路上的挑战，《Bioessays》21（1999）99–104。[2] C. R. Linder，L. H. Rieseberg，重建植物网状进化模式，《美国植物学杂志》91（2004）1700–1708。[3] S. Kong，J. C. Pons，L. Kubatko，K. Wicke，显式系统发育网络的类别及其生物学和数学意义，《数学生物学杂志》84（2022）47。[4] P. L. Erd˝os，C. Semple，M. Steel，一类可由祖先谱重建的系统发育网络，《数学生物科学》313（2019）33–40。[5] R. Janssen，Y. Murakami，关于摘樱桃与网络包含，《理论计算机科学》856（2021）121–150。[6] L. van Iersel，R. Janssen，M. Jones，Y. Murakami，果园网络是带有额外水平弧的树，《数学生物学通报》84（2022）1–21。[7]B. M. Moret，L. Nakhleh，T. Warnow，C. R. Linder，A. Tholse，A. Padolina，J. Sun，R. Timme，系统发育网络：建模、可重建性与准确性，IEEE/ACM计算生物学与生物信息学汇刊 1（2004）13–23。[8] A. R. Francis，M. Steel，哪些系统发育网络不过是带有额外弧的树？，《系统生物学》64（2015）768–777。[9] G. Cardona，F. Rossell´o，G. Valiente，树子代系统发育网络的比较，IEEE/ACM计算生物学与生物信息学汇刊 6（2008）552–569。[10]G. Cardona，M. Llabr´es，F. Rossell´o，G. Valiente，树兄弟时间一致系统发育网络的比较是图

同构完全性，《科学世界期刊》2014（2014）254279。doi:10.1155/2014/254279.[11] S. Willson，正则网络可以由其树唯一构造，IEEE/ACM《计算生物学与生物信息学汇刊》8（2010）785–796。[12] P. Gambette，K. T. Huber，关于有界层级系统发育网络的编码，《数学生物学杂志》65（2012）157–180。doi:10.1007/s00285-011-0456-y.[13] L. Van Iersel，V. Moulton，三网编码树子代和第 2 层系统发育网络，《数学生物学杂志》68（2014）1707–1729。[14] G. Cardona，M. Llabres，F. Rossello，G. Valiente，系统发育网络的度量 II：节点度量和三元组度量，IEEE/ACM《计算生物学与生物信息学汇刊》6（2009）454–469。doi:10.1109/TCBB.2008.127.[15] G. Cardona，M. Llabr´es，F. Rossell´o，G. Valiente，环绕树的比较，IEEE/ACM《计算生物学与生物信息学汇刊》8（2011）410–427。doi:10.1109/TCBB.2010.60.[16] C. Semple，G. Toft，三网编码果园系统发育网络，《数学生物学杂志》83（2021）1–20。[17] G. Cardona，M. Llabr´es，F. Rossell´o，G. Valiente，树子代时间一致杂交网络中的路径长度，《信息科学》180（2010）366–383。doi:10.1016/j.ins.2009.09.013.[18] M. Bordewich，C. Semple，根据分类群间距离确定系统发育网络，《数学生物学杂志》73（2016）283–303。doi:10.1007/s00285-015-0950-8.[19] G. Cardona，M. Llabr´es，F. Rossell´o，G. Valiente，一类树兄弟系统发育网络的距离度量，《生物信息学》24（2008）1481–1488。doi:10.1093/bioinformatics/btn231.[20] A. Bai，P. L. Erd˝os，C. Semple，M. Steel，使用祖先谱定义系统发育网络，《数学生物科学》332（2021）。doi:10.1016/j.mbs.2021.108537.[21]D. F. Robinson，L. R. Foulds，系统发育树的比较，《数学生物科学》53（1981）131–147。doi:10.1016/0025-5564(81)90043-2。

[22] C. Semple, J. Simpson，何时系统发育网络仅仅是两棵树的合并?《数学生物学通报》80（2018）2338–2348。[23] J. C. Pons, T. M. Coronado, M. Hendriksen, A. Francis，一类新型系统发育网络的多项式不变量，PLOS ONE 17（2022）e0268181。[24] K. T. Huber, V. Moulton, M. Steel, T. Wu，系统发育树和网络的折叠与展开，《数学生物学杂志》73（2016）1761–1780。