arXiv:2304.11826v1 [q-bio.PE] 24 Apr 2023

基因家族历史理论

Marc Hellmuth 和 Peter F. Stadler

摘要 大多数基因都属于由进化相关基因组成的更大基因家族。基因家族的历史通常涉及基因的复制和丢失，以及向其他生物体中的水平转移。详细重建基因家族的历史，即相对于其 underlying 物种的系统发育树精确确定进化事件的时间，尽管这对于深入研究基因家族各成员的适应和功能进化具有重要基础作用，却一直是一个具有挑战性的课题。此外，直系同源基因的识别是本文所讨论的更一般问题中的一个尤其重要的子问题。在过去几年中，大量数学研究成果相继出现，将直系同源关系、基因之间最佳匹配的形式化概念以及水平基因转移紧密联系起来。本章旨在概述其中一些关键的数学洞见，并讨论其对实际应用的影响。具体而言，我们重点关注无树方法，即在不使用关于 underlying 树的先验知识或用于推断系统发育树的统计模型的情况下，推断直系同源关系或水平基因转移，以及基因树、物种树和它们之间协调关系的方法。相反，初始步骤旨在提取基因之间的二元关系。

Marc Hellmuth：瑞典斯德哥尔摩大学理学院数学系，SE - 106 91 斯德哥尔摩。Peter F. Stadler：德国莱比锡大学计算机科学系生物信息学组及跨学科生物信息学中心，H¨artelstraße 16-18，D-04107 莱比锡。德国马克斯·普朗克数学科学研究所，Inselstraße 22，D-04103 莱比锡。德国弗劳恩霍夫细胞疗法与免疫学研究所，Perlickstraße 1，D-04103 莱比锡。奥地利维也纳大学理论化学研究所，W¨ahringerstraße 17，A-1090 维也纳。丹麦哥本哈根大学非编码 RNA 技术与健康中心，Grønneg˚ardsvej 3，DK-1870 腓特烈斯贝。美国新墨西哥州圣菲研究所，1399 HydePark Rd，Santa Fe，NM 87501。电子邮件：marc.hellmuth@math.su.se，studla@bioinf.uni-leipzig.de

关键词：直系同源基因、旁系同源基因、基因家族、蛋白质家族、水平基因转移、最佳匹配、系统发育、无树方法。

1 引言

在典型的基因组中，大多数基因都属于较大的同源基因家族，即共享共同祖先的基因。基因家族的进化历史涉及物种形成（此时一个基因会传递给分化出的每个谱系）、基因组内的复制、丢失，有时还包括水平转移，无论是单个基因的转移，还是杂交事件所导致的转移。这些事件会产生同源关系：如果两个基因最近的共同祖先是物种形成事件（或复制事件），则它们分别称为直系同源基因（或旁系同源基因）。此外，异源同源基因指的是因水平转移而彼此分离的基因。这类事件往往会影响基因功能。物种形成后，由于环境约束不同，基因所受的选择压力可能发生变化；基因复制后，旁系同源拷贝可能发生亚功能化或新功能化 [1]，基因丢失 [2]后也可能如此。因此，亲缘关系较近的基因组中的直系同源基因通常具有大致相同的功能。相比之下，旁系同源基因的功能往往相关但明显不同 [3–6], ，尽管也不乏例外 [7]。因此，准确区分直系同源基因和旁系同源基因是基因组功能注释的一项关键任务。可靠地鉴定直系同源基因对于比较基因组学分析也至关重要[8]。此外，一对一的直系同源基因是分子系统发育学中首选的性状 [6, 9]。

因此，准确了解一个基因家族的进化历史，是对其进行比较分析以及理解其功能谱系演化的基础。然而，基因家族历史无法直接测量，而必须根据当今的序列数据推断出来，也就是说，必须根据基因之间以及承载这些基因的基因组之间相似性的测量结果进行推断。这需要一个形式化框架，在该框架中可以研究基因与物种相互交织的历史，并评估计算方法的正确性、局限性和准确性。本章概述了进化情景这一框架，该框架由描述基因和物种系统发育关系的系统发育树（或更一般地说，系统发育网络），以及将基因系统发育嵌入物种系统发育中的协调映射共同构成。本文将重点讨论树状进化的情形，并仅简要说明如何将其推广到网络。

进化情景定义了用于标注基因系统树中顶点的事件类型，例如物种形成或复制，并将边标注为水平转移。这些进化情景反过来构成了基因之间不同同源关系（如直系同源关系、旁系同源关系或异源同源）的形式化定义基础 [10–12], 参见图 1 以了解示例。这三种同源关系并不是唯一值得关注的关系。然而，（互为）最佳匹配关系描述了来自不同物种且在进化上亲缘关系最接近的所有基因对 [13–16]。相比之下

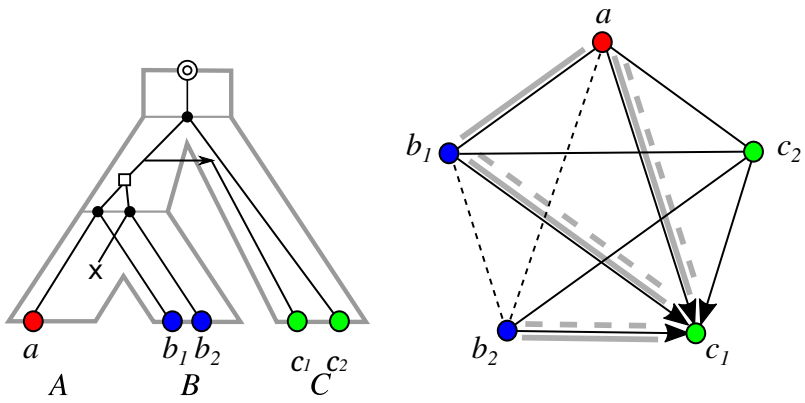


图 1 进化情景（左）由一棵物种树 S （管状灰色轮廓）组成，基因树 T （黑色）通过协调映射嵌入其中，该映射将基因树的每个顶点映射到物种树的一条边或一个顶点上（灰色水平线）。协调映射确定了物种树上的事件类型：基因复制 事件（□）位于 S 的边内；水平基因转移边（箭头“→”）上的事件则使直接后代位于 S的不同分支中。物种形成 事件（•）位于 S的内部顶点。这些事件的后代分布在物种树的每个分支中，但由于基因丢失 （×），其中一些后代可能无法存续至今。现存基因是T 的叶节点，它们被映射到S 的叶节点上（图底部所示的叶节点按其所在物种着色，即 $\sigma(a) = A$ （红色）、 $\sigma(b_1) = \sigma(b_2) = B$ （蓝色），以及 $\sigma(c_1) = \sigma(c_2) = C$ （绿色）。）。

事件直系同源关系包括所有由细黑色侧线（无箭头）连接的基因对，这些基因对对应于颜色明显不同的基因对，且其最近共同祖先是物种形成事件。旁系同源关系包括所有由细黑色虚线连接的基因对，这些基因对的最近共同祖先是复制事件。Fitch（异源同源）关系包括由细黑色箭头边连接的基因对。后期分化时间关系包括分化时间低于相应物种的分化时间的基因对（灰色虚线）；互为最佳匹配关系则包括来自不同物种 X 和 Y 的所有基因对（x，y），其中不存在更近的亲缘基因，即不存在用于 y 的 $x' \in X$，也不存在用于 $y' \in Y$ 的 x（灰色实线）。

除了直系同源关系和旁系同源关系（其定义取决于最近共同祖先是物种形成事件还是基因复制事件）之外，还可以直接根据基因组序列数据估计最佳匹配。同样，后期分化时间（*LDT*）关系包括仅在其中基因所在的两个物种发生分化之后才分化的基因对 [17–19]，它能够为异源同源关系提供信息；异源同源关系记录了两个基因在其进化历史中是否被一个水平基因转移事件分隔。尽管无法定向测量异源同源关系，后期分化时间关系可以通过实际数据分析获得。

重建基因家族历史的问题可以通过不同方式来处理。基于树的方法首先使用成熟的系统发育方法分别推断基因树和物种树。这样，计算协调映射的问题就成为一个独立的优化问题，通常以最小化丢失和水平基因转移事件的数量为目标。这种方法的一个棘手问题是，它严重依赖于准确的树重建。然而，

对于历史复杂的基因家族而言，很难获得可靠的基因树 [20]。一种补救方法是联合推断基因树和物种树，参见例如 [21, 22]。另一种方法是无树方法：直接根据序列相似性数据估计二叉关系，例如互为最佳匹配关系，然后进一步分析这些关系，在不显式构建树的情况下提取直系同源关系和旁系同源关系。目前看来，基于树的方法与无树方法在准确性方面并无根本差异 [23, 24]。 [25],中评述的水平基因转移推断方法可分为三大类：(1) 基于树的方法，根据某种代价函数计算最优协调 [26–28],；参数方法利用基因组特征，即特定于某个物种或物种群体的序列特征，来识别水平插入的遗传物质 [29, 30], ；所谓隐式方法或间接方法，则利用基因之间的距离作为水平基因转移的指标：如果该距离与相应物种的进化距离相比非常小或非常大 [17, 31]。事实上，这些方法本质上是在估计 LDT 关系 [18]。

本章旨在介绍演化情景的形式化框架，并回顾一些描述二元关系之间关系的关键数学结果。这些二元关系可以从相似性数据中测量得到，而同源关系则是进化生物学所关注的对象。此外，我们还将探讨一个问题：基因家族历史的多少信息，以及哪些信息，实际上编码在这些二元关系之中。我们将看到，对这些二元关系的稳健估计已经对基因树、物种树和协调施加了严格约束。

2 情景

2.1 记号

图 G 是由顶点集 V (G) :=V 和边集 E(G) := E 组成的元组 (V, E)。图可以是无向的（边是 V 的二元子集），也可以是有向的（边是 V $\times$V 的子集）。本文所考虑的所有图都是无环的，即如果 {x,y} 或 (x,y) 是一条边，那么 x $\neq$ y。对于有向图 G，我们用 G^{sym} 表示其底层无向图，其中 V (G^{sym}) =V (G)。一般而言，我们使用“图”一词指代无向图；否则，我们会明确写作“有向图”。此外，对于无向图中的边 {x, y} ，或者在有向图中 (x,y) 和 (y,x) 均为边的情形，我们采用简化记号 xy。将无向图与对称有向图视为同一对象，还使我们能够利用有向图与无向图之间的子图关系。我们广泛使用配备顶点着色 σ 的图 (G σ)。如果对于所有相邻顶点 x 和 y，都有 σ(x) $\neq$ σ(y)，则称 (G σ) 是适当着色的。

有根树 自然地描述基因之间以及物种之间的系统发育关系。一棵顶点集为$V = V(T)$、边集为$E = E(T)$的有根树T包含唯一的顶点 0_T，称为根，用于指示最早的状态

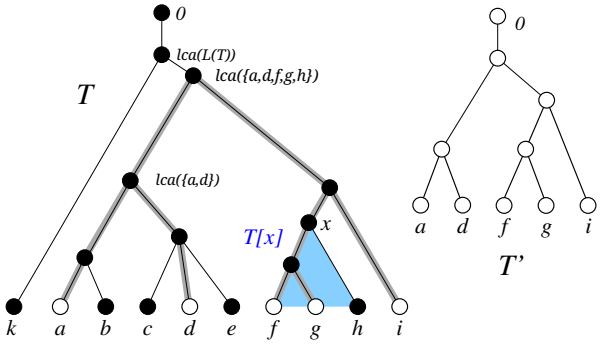


图 2一棵有根的系统发育树的根 0 位于所有叶节点的最近共同祖先 lca($L(T)$)之上。叶集为T' 上的树{a、d、f、g、i}由 T显示。灰色轮廓展示了T'嵌入T的方式。以内部节点x 为根的子树 $T(x)$ 以蓝色阴影表示。图取自本书第一版 *[32]*。

正在考虑中。T 中从 0_T 出发的每条路径都隐含一个时间顺序，并确定如下祖先—后代关系：如果 y 位于从根到 x $\neq$y 的唯一路径上，则 y 是 x 的祖先，而 x 是 y 的后代。在这种情况下，我们写作 x $\prec_T$ y。对于 T 的边，我们写作 e =xy，并约定 y $\prec_T$ x。对于一条边 e =xy，我们称 x 为 y 的父节点，称 y 为 x 的子节点。将祖先关系扩展到 V (T) ∪E(T) 将很有用。对于顶点 x ∈V (T) 和边 e =uv ∈E(T)，我们规定：当且仅当 x $\preceq_T$v 时，x $\prec_T$e；当且仅当 u $\preceq_T$ x 时，e $\prec_T$ x。此外，对于 T 中的边 e =uv 和 f =ab，我们规定：当且仅当 v $\preceq_T$ b 时，e $\preceq_T$ f。通常，$\xi \preceq_T \zeta$ 对所有 ξ、ζ ∈V (T) ∪E(T) 都等价于 $\xi \prec_T \zeta$ 或 $=\zeta$。如果 $\xi \preceq_T \zeta$ 和 $\zeta \prec_T \xi$ 均不成立，我们称 ξ 和 ζ 是 $\preceq_T$-不可比较的。注意，T 的根 0_T 是唯一的 $\prec_T$-极大顶点。

集合 L(T) 中的所有 $\prec_T$极小顶点构成 T 的叶节点。以 x ∈V (T) 为根的子树 T (x) 定义为由顶点集 {y ∈V (T) |y $\preceq_T$x}诱导的子图。对于任意子集 A ⊆ V (T)，最近共同祖先 lca(A) 是作为所有 y ∈A 的祖先的 $\preceq_T$- 极小顶点 w。特别地，对于所有 x ∈V (T)，有 lca({x}) = x。为简便起见，我们用 lca(x,y) 代替 lca({x, y})。此外，我们可以使用下标 "$_T$" 来表示 lca$_T$ 是相对于树 T 取得的。进一步地，我们令 ρ_T := lca(L(T)），并注意到 $\rho_T = 0_T$ 当且仅当根 0_T 至少有两个子节点。树 T 的内部顶点集 V^0(T) :=(V (T)\(L(T)∪{O_T}))∪{ρ_T} 由 V (T)\ L(T) 构成；当 $0_T \neq \rho_T$ 时不包含根。有根树是系统发育的，如果其所有内部顶点都至少有两个子节点；如果其所有内部顶点恰好有两个子节点，则称为二叉的。下文中我们始终假设 T 是系统发育的，但不一定是二叉的。

系统发育树 T 的精化是指具有相同叶集的系统发育树 T'，使得T可由T'通过边收缩得到。将限制树 $T|_{L'}$定义为T 在L' 上的限制，即从树T 中删除$L\setminus$之外的所有顶点及其关联边，并进一步抑制除根之外所有度为二的内部顶点后，得到的叶集为L' 的系统发育树。若某个子集$L' \subseteq L$上的系统发育树T' 与T'重合$T|_{L'}$，则称它被T所显示（或等价地称T 显示 T'），见图 2。

有根三元组是包含三个顶点的二叉有根树。叶集为 x、y、z，且满足 lca(x, y) $\prec$ lca(x,z) =lca(y,z) 的有根三元组记作 xy|z。三元组 xy|z 通常可直接从序列数据中推导出来，并反映出这样一个事实：与 z 相比，分类单元 x 和 y 的进化亲缘关系更近。因此，确定给定的三元组集合 $\mathscr{R}$ 是否一致（即是否存在一棵显示集合 $\mathscr{R}$中所有三元组的树 T）是很有意义的。Aho 等人 [33]设计了一种名为 BUILD 的多项式时间算法，该算法要么构造一棵唯一确定的有根树 Aho($\mathscr{R}$)，显示$\mathscr{R}$；要么判定 $\mathscr{R}$ 不一致。在某些情况下，我们也可能掌握有关目标树未显示的三元组的信息。若存在一棵树 T，显示 $\mathscr{R}$中的所有三元组，但不显示 $\mathscr{F}$中的任何三元组，则两个三元组集合组成的有序对 $(\mathscr{R},\mathscr{F})$ 是一致的。多项式时间算法 MTT（“受限于树的混合三元组问题”） [34] 可在多项式时间内检验三元组集合对 $(\mathscr{R},\mathscr{F})$ 的一致性。

植根树。 感兴趣的进化事件也可能早于所有物种或基因的最近共同祖先。对此，可以考虑植根树，即满足 $0_T \neq \rho_T$ 的树 T。在这种情况下， 0_T 也称为种植根。在植根系统发育树中， 0_T 和 ρ_T 通过边$0_T\rho_T \in E(T)$相连。可以将种植根视为代表“外群”，而 $L(T)$则代表“内群”，即所研究的物种或基因。植根树 T始终展示有根树 $T(\rho_T)$，该树通过收缩边 $0_T\rho_T$ 获得。

定年树。 对于每棵（植根）树 T，我们都可以关联一个时间映射 τ_T：$V(T) \to \mathbb{R}$，使得$x \prec_T y$意味着 $\tau_T(x) < \tau_T(y)$。很容易验证，对于每棵树，都存在这样的时间映射。更进一步，对于每棵树，都可以在线性时间内构造一个时间映射，使树的叶节点被分配到特定的时间点（例如，将所有叶节点都映射到时间点 0，以表示如今仍存的分类单元） [19]。通常很难，而且往往不可能为具有进化意义的事件 [35, 36]获得可靠、准确的“时间戳”$\tau_T(x)$。我们的目的并不需要如此详细的信息。理论结果只取决于相对时序是否存在，即是否知道某个事件早于、晚于另一个事件，或与另一个事件同时发生；而这种信息通常更容易从数据中提取 [37, 38]。

在本章中，我们假设所有树均为有根的系统发育树，除非另有明确说明。

2.2 协调

基因在基因组，即物种中进化。为了理解基因与物种之间的关系，我们需要描述基因树 T 如何与物种树 S“匹配”。这一思想通过协调映射 的概念得以形式化，该映射将基因树的顶点，即基因进化史中的进化事件，定位到物种树中。我们将进化事件建模为精确的

时间点。当然，这是一种近似。实际上，即使是基因复制这样的瞬时事件，也需要一定时间才能在种群中传播并固定下来。同样，物种形成也是一个基于种群的过程，需要一定时间 [39]，还可能涉及不完全谱系排序等其他影响 [40, 41]。尽管如此，在宏观进化的背景下，将事件视为时间点仍是合理的，因为此时种群效应被完全忽略。

为了从形式化的角度研究协调，我们首先仅从一些易于论证的最低要求入手：

定义 1。设 T 和 S 为树。T 与 S 的一个协调是一个三元组（T，S μ），其中 μ：V（T）$\to V$（S）$\cup E$（S）是满足以下条件的映射：
(R0) $\mu(x) = 0_S$ 当且仅当 $x = 0_T$ (R1) $\mu(x) \in L(S)$ 当且仅当 $x \in L(T)$ (R2) 如果 $y \prec_T x$ 且 $\mu(x)\ \mu(y) \in V(S)$，则 $\mu(x) \neq \mu(y)$ (R3) 如果 $y \prec_T x$，则 $\mu(x) \not\preceq_S \mu(y)$

条件（R0）用于识别种植根，条件（R1）确保 T 的叶节点，即现存基因，位于 S 的叶节点中，即位于现存物种中。条件（R2）确保时间上不同的物种形成事件不会映射到 S 中的同一个顶点，即同一个物种形成事件。最后，条件（R3）通过禁止 T 中 x 的后代被映射到 S 中 μ(x) 的祖先，确保时间一致性的一种弱形式。因此，如果 y $\prec_T$ x，那么要么 μ(y) $\preceq_S \mu$(x)，要么 μ(x) 和 μ(y) 是 $\preceq_S$-不可比较的。为完整起见，我们指出，许多作者更倾向于使用如下形式的协调映射 γ：V (T) $\to$ V(S)，即将顶点映射到顶点 [26, 42, 43]。不过，这很容易转换：如果 μ (x) =uv，那么 γ(x) = v，即所有内容都被映射到 S 中该边的较低端点。反向转换则稍微复杂一些。不过，从本质上说，只需在 γ(x) = v 且存在一个 x′ $\prec_T$ x，使得 γ(x′) = γ(x) = v 时，将 μ(x) = uv 设定为即可。

进化中的一个重要区别是垂直遗传与水平遗传之间的差异。垂直遗传意味着，基因的后代由其后代物种携带。相比之下，水平遗传，即水平转移，是指将基因转移到不同谱系中。形式上，可以通过区分基因树 T 中的垂直遗传边和水平基因转移（HGT）边来描述这种情况：

定义 2。设 (T,S μ) 是一个协调。若边 $xy \in E(T)$ 满足：$\mu(x)$ 与 $\mu(y)$ 是 $\preceq_S$-不可比较的，则该边称为**水平基因转移边**。若对于所有 $xy \in E(T)$，$\mu(x)$ 与 $\mu(y)$ 都是 $\preceq_S$-可比较的，则称该协调为**无水平基因转移**。

作为 (R3) 的直接推论，协调 (S,T μ) 无水平基因转移，当且仅当它满足

(R4) 若 $y \prec_T x$，则 $\mu(y) \preceq_S \mu(x)$。

事实上，（R4）指出，T 的所有边都对应垂直遗传。

通常，我们会先验地知道每个基因存在于哪个物种中。也就是说，给定一个映射 σ：$L(T) \to L(S)$，它将每个现存基因对应到其所在的现存物种。

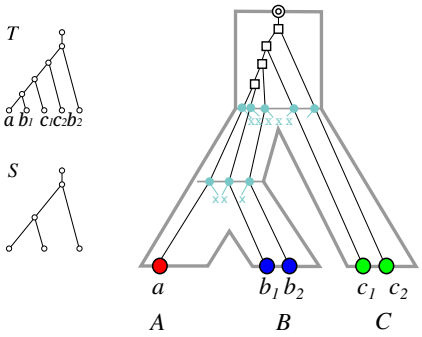


图 3 σ-协调示意图 $\mathscr{S} = (T, S, \mu_0, \sigma)$，其中 μ_0 由公式（1）定义。基因按照其所在物种进行着色，即 $\sigma(a) = A$（红色）、$\sigma(b_1) = \sigma(b_2) = B$（蓝色），以及 $\sigma(c_1) = \sigma c_2) = C$（绿色）。请注意，图 1 中的基因树 T 以及通过移除×-边并压缩其两个相邻顶点而得到的基因树的可观测部分 彼此重合。因此，图 1 和当前图展示了针对相同树T 和S的不同协调。特别是，这些协调中的基因丢失数量差异很大：图 1 中只有 1 次，而 μ_0中有 8 次。

定义 3. 设 T和 S是两棵树，且 $\sigma: L(T) \to L(S)$。那么 (T, S, μ, σ) 是一个σ-协调，如果 (T, S, μ) 是一个协调，并且 $\mu_{|L(T)}$满足 $\mu_{|L(T)} = \sigma$。

该领域早期的一项重要观察是，即使无需考虑水平基因转移，每棵基因树也都可以与任意物种树进行协调[44, 45]:

定理 1. 对于任意两棵树 T 和 S，以及任意映射 $\sigma: L(T) \to L(S)$，存在一个无水平基因转移 σ-协调 (T, S, μ, σ)。

要证明这一点，只需考虑通过如下方式定义的协调映射 μ_0

$$\begin{aligned} &\mu_0(0_T) := 0_S; \\ &\mu_0(x) := \sigma(x) \text{ for all } x \in L(T); \text{ and} \\ &\mu_0(x) := 0_S \rho_S \text{ for all inner vertices } x \in V^0(T). \end{aligned} \tag{1}$$

每个协调映射 μ 都可用于推断基因丢失事件。为此，考虑边 xy $\in$V (T)，对于给定的协调映射μ，有 μ(y) $\prec_S$ μ(x)，且存在顶点 u $\in$ V (S)，使得 μ(y) $\preceq_S$ u $\preceq_S$ μ(x)，因此 u $\notin$L(S)。在物种形成 u $\in$ V^0(S)期间，基因组中每个现存基因的后代都会传递给每个后代谱系，即每条边 uw $\in$E(S)。如果在从 w 延伸出的物种中找不到该基因的任何后代，那么该基因必定已在整个物种子树 S(w) 中消失。最简约的解释是，沿边 uw $\in$E(S) 已发生了一次丢失事件。图 3 表明，对于协调 μ_0，这种推理会得出不切实际的大量基因丢失事件：例如，最早一次物种形成时存在的五个旁系同源基因总共产生 10 个基因，其中一半随后又丢失了。一般而言，这种协调 μ_0 不会是合理的生物学解释，因为它意味着基因丢失数量过大。这一观察促使人们提出了若干替代评分函数 F(μ) := F(T, S, μ, σ)，用于给定树 T 和 S 以及给定叶分配 σ的协调映射，并对复制和／或基因丢失的数量进行计数。有关详细讨论，请参阅 [46]及其中的参考文献。

这里我们简要考虑无水平基因转移情形下的这一优化问题。假设给定一个叶分配 σ。任务是寻找一种能优化给定评分函数的 σ-协调。根据定理 1，我们知道解始终存在。由于任何顶点 $x \in V(T)$ 都必须映射到其后代所在的所有物种的最近共同祖先处或其上方，因此每个解都满足

$$\mathrm{lca}_S(\sigma(L(T(x)))) \preceq_S \mu(x) \tag{2}$$

所谓的 最近共同祖先协调 μ^* 会在可能的情况下达到等式。更准确地说，我们定义 μ^* 如下。对于每个 $x \in V(T)$，令 $w(x) := $ 最近公共祖先$_S(\sigma(L(T(x))))$，并令 u 为 $w(x)$ 在 S 中的父节点；在植根树中该父节点始终存在。

$$\mu^*(x) := \begin{cases} w(x) & \text{whenever there is no } y \text{ with } y \prec_T x \text{ and } w(x) = w(y) \\ uw(x) & \text{otherwise.} \end{cases}$$

最低共同祖先协调 μ^* 是根据多种代价度量得出的最简约协调，包括复制次数以及复制和基因丢失次数 [46–48]。相应的最大似然和贝叶斯方法分别见 [49] 和 [50], 。如果将水平基因转移事件纳入并加以惩罚，仍可在多项式时间内计算最简约协调 [26, 42]。然而，正如我们将在下一节看到的，这些协调并非总是在生物学上可行，另见 [51]。

2.3 宽松情景

正如 [26, 42] 所述，上一节中协调的定义不足以确保 (T, S, μ) 能被解释为沿线性时间轴发生的一系列事件。为此，我们将进化情景视为定年树的协调：

定义 4. 一个（宽松）情景 $\mathscr{S} = (T, S, \mu, \tau_T, \tau_S)$由一棵定年基因树（$T$ τ_T）、一棵定年物种树（S τ_S）以及一个协调映射 $\mu\colon V(T) \to V(S) \cup E(S)$组成，并满足

(S0) $\mu(\mathrm{x}) = 0_S$ 当且仅当 $\mathrm{x} = 0_T$，(S1) $\mu(\mathrm{x}) \in \mathrm{L(S)}$ 当且仅当 $\mathrm{x} \in \mathrm{L(T)}$，
(S2) $\tau_S(\mu(\mathrm{x})) = \tau_T(\mathrm{x})$，对所有满足 $\mu(\mathrm{x}) \in \mathrm{V(S)}$ 的 $\mathrm{x} \in \mathrm{V(T)}$ 均成立；
以及 (S3) $\tau_S(\mathrm{v}) < \tau_T(\mathrm{x}) < \tau_S(\mathrm{u})$，对所有满足 $\mu(\mathrm{x}) = \mathrm{uv} \in \mathrm{E(S)}$ 的 $\mathrm{x} \in \mathrm{V(T)}$ 均成立。

公理 (S0) 和 (S1) 分别与协调定义中的 (R0) 和 (R1) 相同。其余两个公理 (S2) 和 (S3) 则规定了时间一致性。与上一节中的协调一样，我们不假设已给定叶映射 $\sigma\colon L(T) \to L(S)$，尽管在实际应用中通常会给定该映射。

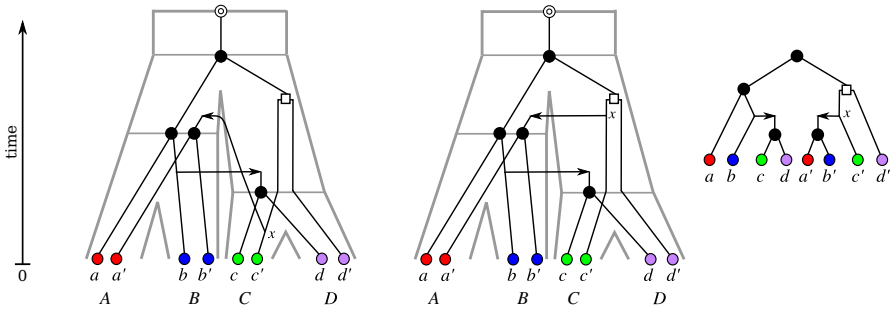


图 4 一个事件标注基因树（左图），其中我们将叶节点 v 的事件标签 ⊙ 替换为带有颜色 σ(v) 的顶点，以指示 v 所在的物种。特别地，对于 (x,x′,X) ∈ {(a,a′,A),(b,b′,B),(c,c′,C),(d,d′,D)}，我们有 σ(x)σ′ = X。图中展示了映射到同一物种树 S 的两个协调结果（左图和中图）。中图所示的协调是时间一致的，尤其会产生一个宽松情景。设包含 c,c′ 和 d,d′ 的物种分别为 c 和 d，并令 u := 最近公共祖先$_S$(c,d)。这两个协调结果的差异仅在于对 HGT-边的起点 x ∈ V(T) 的放置选择：将其放在边 uc 上，或放在边 ρ_Su 上。然而，在第一种情况下，我们有 y $\prec_T$ x，但 t_T(x) < t_T(y)。因此，右图所示的协调无法产生 T 的有效时间映射，因而不是时间一致的。

定义5。 元组（T、S μ、τ_T、τ_S、σ）是一个宽松 σ-情景，当且仅当（T、S μ、τ_T、τ_S）是一个宽松情景，并且 $\sigma\colon L(T) \to L(S)$ 满足 $\sigma = \mu_{|L(T)}$。

宽松情景可以看作明确给定时间轴的协调。不难证明，对于给定的宽松情景 (T,S,μ,τ_T,τ_S)，三元组 (T,S,μ) 确实是一个协调（参见 [19] 中的引理 2 和 3）。然而，反过来一般并不成立，也就是说，并非总能找到时间戳将协调转化为宽松情景；例如参见 [52] 中的图 2。这引出了以下内容

定义6。 协调（T、S μ）是时间一致的，如果存在定年映射τ_T和 τ_S，使得（T、S μ、τ_T、τ_S）是一个宽松情景。

在 $V(S)$ [26, 43] 或 $V(S) \cup V(T)$ [52, 53]上引入了多种不同的辅助图，以概括时间约束。每种情况下，研究都证明，协调（T、S μ）时间一致，当且仅当相应的辅助图无环。此外可以证明，在无水平基因转移协调这一特殊情况下，这些辅助图均无环；也就是说，每个无水平基因转移协调都会自动满足时间一致性，并且可以扩展为一个宽松情景。

在一般情况下，额外的时间一致性要求也使最大简约调和的计算变得困难。如 [26],所示，该问题随后变为 NP难。

2.4 事件标签

从生物学角度来看，基因树 T 的内部顶点 $V^0(T)$ 表示进化事件。在无水平基因转移情景中，即满足 (R4) 的协调（T,S,μ），只有两种本质上不同的事件：基因复制和物种形成事件。由于物种形成事件也由$V^0(S)$ 的内部顶点表示，因此，根据其在协调映射 μ 下的像来区分复制和物种形成是很自然的。这自然引出了以下定义 [13]:

定义 7. 设 (T,S,μ) 是一个无水平基因转移协调。那么，T 上的事件标记是映射 $\ell_\mu\colon V(T)\to\{\ ,\odot,\bullet,\square\}$，定义如下：

$$\ell_\mu(u)=\begin{cases}\circledcirc & \text{if } u=0_T, \text{ i.e., } \mu(u)=0_S\\ \odot & \text{if } u\in L(T), \text{ i.e., } \mu(u)\in L(S)\\ \bullet & \text{if } \mu(u)\in V^0(S) \text{ (speciation)}\\ \square & \text{else, i.e., } \mu(u)\in E(S) \text{ (duplication)}\end{cases}$$

前两种情况以及 $\odot\circledcirc$区分了T的种植根和叶节点。其余两种情况将映射到物种树内部顶点的物种形成事件与$u\in V(T)$ 相对应。相反，如果 $\mu(u)\in E(T)$，则它不对应物种形成，因此对应一个基因复制，记作 $\square$。然而，到目前为止所考虑的协调 (T,S,μ) 所满足的条件，并不足以确保$x\in V^0(T)$ 且 $\ell_\mu(x)=\bullet$ 表示一个具有生物学意义的物种形成事件。还需要对进化情景施加额外约束。

这或许可以通过假设存在一棵二叉基因树 T 来得到最佳解释。在此情况下，如果 $\ell_\mu(x)=\bullet$，则 x 的两个子节点 v' 和 v'' 应通过 μ 映射到两个不同的谱系中。否则，如果 v' 和 v'' 被映射到同一谱系中，则没有明确的历史痕迹能够证明 x 是一个物种形成顶点。换言之，如果 x 的两个子节点 v' 和 v'' 的映像 $\mu(v')$ 和 $\mu(v'')$ 是 $\preceq_S$-不可比较的，则 $\ell_\mu(x)=\bullet$ 在生物学上是合理的。此外，与公式 2 类似，并为了适应最简约协调映射（最低共同祖先协调），可以假设 $\mu(x)=\mathrm{lca}_S(\mu(v'),\mu(v''))$。上述讨论自然可以推广到一般的、不一定是二叉的基因树，具体如下：

（R5）假设对于某个 $x\in V$，$\mu(x)\in V(S)^0$，则

（i）对于 x 在 T中的至少两个不同子节点 v',v''，$\mu(x)=$最近公共祖先$_S(\mu(v'),\mu(v''))$。
（ii）对于 x 在 T中的任意两个不同子节点 v' 和v''，$\mu(v')$ 与 $\mu(v'')$ 在 S 中不可比较。

如 [13, 52], 所示，公理 (R0)-(R5) 等价于文献 [54–58]中常用的公理。具体而言，对于任意无水平基因转移 σ协调（T,S,μ,σ），都存在一个适用于T和S [58]的最低共同祖先协调。后一观察可以轻易推广到涉及水平基因转移的 σ协调，并通过

考察限制在 T 的无水平基因转移子树上的最低共同祖先映射 [58]。因此，此处采用的公理集自然对应于最低共同祖先映射，进而对应于大多数简约协调。此外， [13, 引理2] 表明，对于 x 的一对不同子节点 v′ 和 v″ ，若 L(T (v′))∩ L(T (v″)) ≠/0，则对所有满足 (R5) 的无水平基因转移σ协调，均有 ℓ_μ(x) = □ 。这进而促使我们考虑极值事件标记 ℓ，它假定这些是仅有的复制，因此，只要对 x 的所有子节点对都有 (L(T (v′)))∩μ(L(T (v″)) =/0，就赋予ℓ (x) = ●。需要特别指出的是，极值标记 ℓ仅根据 T 的信息定义，不依赖物种树 S 或协调映射μ是否存在。因此，无法预先保证极值事件标记能够通过实际的生物学情景实现。

3 最佳匹配图与直系同源关系

3.1 定义与刻画

许多用于确定直系同源关系的组合方法都从互为最佳（blast）命中开始。这里，我们将最佳匹配视为一种基本的进化概念，并以“最佳命中”作为其在序列数据上的近似。因此，我们相对于一个底层系统发育树来考察最佳匹配，尽管这棵树通常是未知的。因此，我们以一棵基因树T 及一个映射σ：$V(T) \to Y$为起点，其中Y是一个物种集合。将 σ视为一种着色会很方便：按照现存基因所在的物种为树T 的叶节点着色。

定义 8. 设 (T σ) 是一棵叶着色树。若对于物种 σ(y′) = σ(y) 中的所有叶节点 y′，均满足 σ(x) ≠ σ(y) 且 lca(x, y) $\preceq_T$lca(x, y′)，则叶节点 y ∈L(T) 是叶节点 x ∈L(T) 的最佳匹配。若 y 是 x 的最佳匹配，且 x 是 y 的最佳匹配，则叶节点 x, y ∈L(T) 互为最佳匹配。

请注意，一个基因x 可能有两个或更多个同色的（互为）最佳匹配$r \neq \sigma(x)$。一些直系同源关系检测工具，例如 ProteinOrtho [59], ，会明确尝试从序列数据中提取所有互为最佳匹配。

如果存在一棵叶标记树T σ，使得当且仅当y 是x 在T σ中的最佳匹配时，(x, y)才是 G σ中的一条有向边，那么有向顶点标记图G σ就是一张最佳匹配图（BMG）。给定T σ，我们用 $\mathfrak{B}(T$ $\sigma)$表示它的最佳匹配图。

在刻画最佳匹配图时，一个关键观察是：两个颜色（物种）上的某些顶点子集，会对任何可能解释某个 BMG 的叶标记树所显示的三元组施加约束。

定义 9. 设 a、b、b′是着色有向图 (G σ) 中两两不同的顶点，满足 σ(a) ≠ σ(b) = σ (b′)，且 (a, b) ∈E(G)。则三元组 ab|b′称为信息性三元组，如果 (a, b′) ∉E(G)；如果 (a,b′) ∈E(G)，则称为禁用三元组。

集合 $\mathscr{R}(G$ $\sigma)$ 和 $\mathscr{F}(G$ $\sigma)$ 分别表示所有信息性三元组和禁用三元组的集合。

[13] 结果表明，如果 $(G$ $\sigma)$ = $\mathfrak{B}(T$ $\sigma)$，那么 T显示

$\mathscr{R}(G\ \sigma)$ 中的所有三元组，且不包含 $\mathscr{F}(G\ \sigma)$ 中的任何三元组。集合 $\mathscr{R}(G\ \sigma)$ 和$\mathscr{F}(G\ \sigma)$ 也为最佳匹配图提供了便捷的刻画。

定理 1 适当着色有向图$(G\ \sigma)$是最佳匹配图，当且仅当满足以下条件之一。

(i) $\mathscr{R}(G\ \sigma)$ *一致，并且* $\mathfrak{B}(Aho(\mathscr{R}(G\ \sigma))\ \sigma) = (G\ \sigma)$ *[13, 60](ii)* $(\mathscr{R}(G\ \sigma)\ \mathscr{F}(G\ \sigma))$ *一致，并且* $(G\ \sigma)$ *无同色汇点，即对于每个* $x \in V(G)$ *和每个颜色* $s \neq \sigma(x)$*，都存在一个顶点* y*，使得* $(x, y) \in E(G)$*，且* $y \in N(x)$*，其中* $\sigma(y) = s$*。 [61]*

特别地，最佳匹配图可以在多项式时间内被识别。

有趣的是，每个最佳匹配图 $(G\ \sigma)$ 都与一个唯一的 最小解析树 (T^*, σ) 相关联。也就是说，(i) $\mathfrak{B}(T^*, \sigma) = (G\ \sigma)$，并且 (ii) 如果 $\mathfrak{B}(T\ \sigma) = (G\ \sigma)$，那么$T$显示$T^*$。因此，最小解析树$T^*$捕获了由最佳匹配提供的、关于基因树 T 的可靠系统发育信息。最佳匹配图 $(G\ \sigma)$ 的最小解析树正是树$Aho(\mathscr{R}(G\ \sigma))$，并且可以在多项式时间内构建。

虽然最小分辨树可以作为框架，在不对进化历史作出超出数据实际提供的信息的假设的前提下，涵盖系统发育信息，但在许多情况下，人们还希望找到可视为分辨程度最高的历史的二叉树。二叉树尤其值得关注，因为真正的多分叉很可能十分罕见，也就是说，大多数多分叉都是可用数据分辨率不足所致 [62–64]。然而，并非每个最佳匹配图都能由二叉树解释（参见 [65, 图 6A]）。可由二叉树解释的最佳匹配图，其特征是作为诱导子图不包含某种由四个顶点构成的特定彩色图，称为沙漏图 [65]。请注意，解释最佳匹配图的二叉树不一定唯一；不过，它们都展示$Aho(\mathscr{R}(G\ \sigma))$，并且可以在多项式时间内构造 [66]。

可由二叉树解释的最佳匹配图也可以用三元组集合方便地刻画。考虑对信息三元组集合的如下扩展：

$$\mathscr{R}^B(G,\sigma) := \mathscr{R}(G,\sigma) \cup \left\{ bb'|a : ab|b' \in \mathscr{F}(G,\sigma) \text{ and } \sigma(b) = \sigma(b') \right\} \qquad (3)$$

如 [66], 在定理 1 中用 $\mathscr{R}^B(G\ \sigma)$ 替换 $\mathscr{R}(G\ \sigma)$ 所示，可得到对可由二叉树解释的最佳匹配图的刻画。此外，如果 $(G\ \sigma)$ 是一个可由二叉树解释的最佳匹配图，那么 $T_B := Aho(\mathscr{R}^B(G\ \sigma))$ 具有如下性质：二叉树 $(T\ \sigma)$ 满足$\mathfrak{B}(T\ \sigma) = \mathfrak{B}(T_B\ \ \sigma) = (G\ \sigma)$，当且仅当 T是 T_B 的精化。我们在下面的定理中总结这些结果

定理 2 对于每个最佳匹配图 $(G\ \sigma)$，以下陈述等价：

(i) $(G\ \sigma)$ 可由二叉树解释。*(ii)* $(G\ \sigma)$ 是无沙漏图 *[65]*。*(iii)* $\mathscr{R}^B(G\ \sigma)$ 是一致的 *[66]*。

在这种情况下，可以在多项式时间内构造一棵能够解释$(G\ \sigma)$的二叉树。特别地，$BMG\ (G\ \sigma)$ 可由树$(Aho(\mathscr{R}^B(G\ \sigma))\ \sigma)$ *[66]*的每个精化所解释。

在实践中，通过比较基因序列的相似性，可以经验性地获得$(G\ \sigma)$。因此，$(G\ \sigma)$ 很可能不是最佳匹配图，而是与真实的最佳匹配图存在假阳性边和假阴性边两种差异。针对最佳匹配图的弧修改问题旨在纠正此类错误。与大多数图编辑问题一样，该问题是 NP完全的 [61, 66]。不过，可以设计出高效的启发式算法，以可接受的准确度解决该问题 [67]

3.2 无水平基因转移时的直系同源关系

Walter Fitch [10]将直系同源关系定义为源自物种形成事件的同源关系。尽管后来的讨论在存在水平基因转移时对这一简单概念作了限定（见下文），但在无水平基因转移的情形下，这一概念很明确。本节中，我们始终假设演化情景和协调均不存在水平基因转移。

定义 10。设 (T,t) 是一棵事件标记基因树，且 $x,y \in L(T)$ 是两个不同的基因。若 x 和 y 是直系同源基因，则 $t(\mathrm{lca}_T(x,y)) = \bullet$；若 $t(\mathrm{lca}_T(x,y)) = \square$，则它们是旁系同源基因。

我们用 $\Theta(T,t)$ 表示顶点集为 $L(T)$、边为 $xy \in E(\Theta(T,t))$ 的直系同源图，其中 x 和 y 互为直系同源基因。注意，相应的“旁系同源图”就是 $\Theta(T,t)$ 的补图。此外，$\Theta(T,t)$ 是对称的，但不一定具有传递性（边 xy 和 yz 并不意味着 xz 也是一条边）。直系同源图具有简单的结构：

定理 3（[68]）如果且仅如果图 G 是余图，那么 G 是某个事件标记树 (T,t) 的直系同源图 (T,t)，即 $G = \Theta$。

余图的一个经典刻画如下：G 是余图，当且仅当它不包含 P_4，即不包含四顶点路径作为诱导子图 [69]。

给定一个协调 $(T, S\ \mu)$，因此，直系同源关系由协调映射蕴含，因为根据上面的定义 7，μ 指定了事件标记 ℓ_μ。直系同源关系对 μ 的依赖至关重要。在 μ_0 的极端情况下，我们有 $\mu_0(V(T^0)) \subseteq E(S)$，因此所有事件都被分类为复制。相反，LCA 协调映射将尽可能多的 $u \in V^0(T)$ 映射到 S 的内部顶点，因此可以预期会得到大量直系同源基因对。

许多直系同源关系检测算法都以这样的观察为出发点：两个直系同源基因 x 和 y 也是互为最佳匹配。因此，我们尤其关注如何仅从最佳匹配数据中获取有关直系同源关系的信息。

用 $\mathfrak{B}_{sym}(T\ \sigma)$ 表示最佳匹配 $\mathfrak{B}(T\ \sigma)$ 中由互为的（即双向的）边构成的子图，并用 $\hat{}\ \ell_T$ 表示上一节介绍的 T 的极值事件标记，可得以下关键结果：

定理 4（[13]）设 $(T, S\ \mu, \sigma)$ 是满足（R5）的无水平基因转移 σ-协调。则 $\Theta(T\ \ell_\mu) \subseteq \Theta(T, \hat{}\ \ell_T) \subseteq \mathfrak{B}_{sym}(T\ \sigma) \subseteq \mathfrak{B}(T\ \sigma)$。

定理 4 尤其表明，使用互为最佳匹配图$\mathfrak{B}_{sym}$（T σ）来近似直系同源关系，不会导致假阴性直系同源分配。不过，一般而言， Θ（T ℓ_μ）是$\mathfrak{B}_{sym}$（T σ）的真子图，也就是说，互为最佳匹配图可能包含假阳性直系同源判定。特别是， $\mathfrak{B}_{sym}$（T σ）不一定是余图，因此可能包含诱导子图P_4s。

如果对于 x 的某些子节点 v′ 和 v″，L(T (v′)) ∩L(T (v″)) $\neq$/0，那么对于所有满足 (R5) 的无水平基因转移 σ-协调，都有 ℓ_μ(x) = □（参见 [13, 引理2]）。[65, 引理10]刻画了最佳匹配图中的错误直系同源分配，并给出了更强的结果。特别是，该引理表明，当且仅当对于 x 的某些子节点 v′ 和 v″，L(T (v′)) ∩L(T (v″)) $\neq$/0 时，$\hat{}$ ℓ_μ(x) = □ 。利用 Θ(T ℓ_μ) $\subseteq$ Θ(T, $\hat{}\ell_T$) 这一事实，我们可以将该结果改写为以下更便于使用的形式：

定理 5 对于每棵树 $(T$ $\sigma)$ 以及 T 的任意两个叶节点 x 和 y，以下陈述等价。

(i) 在<*strong*>任意</*strong*>无水平基因转移 σ-协调 $(T, S$ $\mu, \sigma)$中，x 和 y 是旁系同源基因。*(ii)* 在$\mathrm{lca}_T(x,y)$ 处存在两个子节点 v' 和 v''，使得 $L(T(v')) \cap L(T(v'')) \neq 0$。*(iii)* 当 x 和 y 互为最佳匹配时，xy 是“假直系同源基因”。

请注意，定理 5 取决于底层树的结构（T σ），而且能够解释给定最佳匹配图的树不一定唯一。因此，可能存在一棵树（T', σ），使得 $\mathfrak{B}$（T σ） = $\mathfrak{B}$（T', σ），但对于该树而言，边 xy是相对于（T σ）的假阳性直系同源分配，而相对于（T', σ）则不是。因此，我们特别关注最佳匹配图（G σ）中的双向边 xy：对于满足 $\mathfrak{B}$（T σ） =（G σ）的每一棵树（T σ），它们都是假直系同源分配。这类边称为（G σ）中的明确假直系同源分配，并且可以在重叠沙漏链中识别出来。详细信息请参阅 Def.14 和 Def.16，[65] 中对所有“hug”对进行了详细规定，我们得到

定理6（[65,定理11 & 12]） 在最佳匹配图（G σ）中，边 xy 是无歧义假直系同源指派，当且仅当 x 和 y 是一对*“hug”*对。计算最佳匹配图中所有无歧义假直系同源指派的集合可在多项式时间内完成。

无 *HUG* 图 $\mathfrak{N}(G$ $\sigma)$ 是从 $(G^{sym}$， $\sigma)$ $\subseteq(G$ $\sigma)$ 中移除所有属于 “hug” 对的边xy后得到的子图。特别地， $\mathfrak{N}(\mathfrak{B}(T$ $\sigma))$ 包含每个满足 (R5) μ 的无水平基因转移 σ-协调所对应的直系同源图，作为其子图（参见 [65,推论 5]）：

$$\Theta(T, \ell_\mu) \subseteq \Theta(T, \hat{\ell}_T)) \subseteq \mathfrak{N}(\mathfrak{B}(T, \sigma)) \subseteq \mathfrak{B}_{sym}(T, \sigma) \tag{4}$$

如 [65]、 $\mathfrak{N}$（$\mathfrak{B}$（T σ））所示，实际上是一个直系同源图。然而，由于仅移除了明确的假直系同源关系， $\mathfrak{N}$（$\mathfrak{B}$（T σ））中仍可能包含假直系同源基因。也就是说，互为最佳匹配的 xy 在真实演化情景（T ℓ）中可能是旁系同源基因，但在与之 一致的替代演化情景中却显示为直系同源基因

与最佳匹配数据相符。事实上，总是可以将物种形成事件“上移”，并用一次复制及其后的丢失来替代它。这样一来，总是可以在没有直系同源基因的情况下解释最佳匹配。因此，无 HUG 图 $\mathfrak{N}(\mathfrak{B}(T,\sigma))$ 提供了“最简约”的解释，因为只要某种直系同源关系与最佳匹配数据一致，它就会预测这种关系，而不是采用另一种由复制事件及互补丢失构成的解释。因此，$\mathfrak{N}(\mathfrak{B}(T,\sigma))$ 背后的协调在概念上至少与最近共同祖先协调相关。不过，对这种联系进行更深入的考察将留待未来研究。

实际上，可以通过估计一个表示最佳（blast）命中数据的初始物种标记图（$\tilde{G},\sigma$），然后将这些初始数据编辑为最佳匹配图（BMG），即符合定义 8 的图（G,σ），来获得准确的直系同源分配。通常，这一步是 NP 困难的 [61, 70]，因此需要高效的启发式算法。随后，我们可以从最佳匹配图（G,σ）中在多项式时间内移除所有“hug”对，从而得到无 HUG 图 $\mathfrak{N}$（G,σ）；该图不含明确的假直系同源基因，同时也是一个直系同源图。

由于直系同源图是余图，因此还存在一种更直接但准确性较低的方法来估计直系同源图。为此，可以从初始估计的最佳（blast）命中（$\tilde{G},\sigma$）中提取互为最佳命中（$\tilde{G}^{sym},\sigma$），并找出与G^{sym}相比边插入和边删除次数最少的余图$\tilde{H}$。这个余图编辑问题同样是 NP 困难的 [71]。然而，如果G^{sym} 与余图并不太相似，该问题仍然具有˜可处理性。近年来，这项任务已经有了快速启发式算法 [72–74]。

根据定理 1，最佳匹配图可以在多项式时间内识别，而树 T_{Aho}，$\sigma) := (Aho(\mathscr{R}(G,\sigma)),\sigma)$ 是能够解释最佳匹配图 G,σ 的唯一最小解析树。然而，一般来说，无法为 T_{Aho}赋予事件标记 ℓ，使得 $\Theta(T_{Aho},\ell) = \mathfrak{N}(G,\sigma)$。相反，可以扩充 T_{Aho}，得到一棵唯一确定的树 ℓ_{T^*}，并为其赋予极值事件标记 ℓ_{T^*}。带事件标记的树 $(T^*,\hat{\ell}_{T^*})$ 满足

$$\Theta(T^*,\hat{\ell}_{T^*}) = \mathfrak{N}(\mathfrak{B}(T,\sigma)) \tag{5}$$

并且可以在多项式时间内得到（参见 [65, 定理10]）。极值标记 $\hat{\ell}_{T^*}$ 仅根据 T^* 的信息定义，不依赖于物种树 S 或协调映射 μ 的存在。

总而言之，因此可以利用最佳匹配图（G,σ）的信息，在多项式时间内获得关于直系同源关系的数学上可靠的估计，以及由此得到的基因树（T,ℓ），使得 $G' = \Theta(T,\ell)$。

一个耐人寻味的观察是，标记为基因树（T,ℓ）的事件，无论协调映射的具体细节如何，都会对物种树产生影响。为此，令 $\mathscr{S}(T,\ell,\sigma)$ 表示满足以下条件的所有三元组 $\sigma(x)\sigma(y)|\sigma(z)$ 的集合：(i) $\sigma(x)$,$\sigma(y)$ 和 $\sigma(z)$ 是两两不同的物种；(ii) $\ell(\mathrm{lca}(x,z)) = \bullet$，即，$xy|z$ 的根是一个物种形成事件。

命题1（[58, 75]）。对于给定的事件标记树（T,ℓ）和物种映射 σ，以下两个陈述等价。

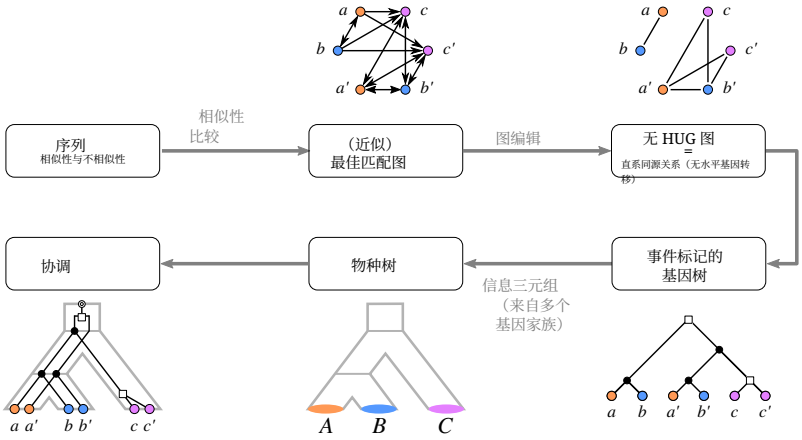


图 5 无树直系同源检测以及随后重建事件标记基因树 T、物种树 S和无水平基因转移 σ协调（T,S μ,σ）的概念性工作流程，假设未发生水平基因转移。在此σ 我们有 , ∈ { , ′ , , ′ , , ′ } (x) (x') X for (xx',X) (aa,A) (bb,B) (cc,C)。利用序列相似性数据获得最佳匹配图的初始估计。随后将该估计校正为数学上合理的最佳匹配图（G σ），并提取无 HUG 图 $\mathfrak{N}(G$ $\sigma)$。无 HUG 图不包含任何假阴性直系同源指派，也不包含明确的假阳性直系同源指派。特别是， $\mathfrak{N}(G$ $\sigma)$ 是一个直系同源图，基于这些信息可以重建事件标记基因树。该事件标记基因树还传达了有关物种树的信息。将来自多个基因家族的后者信息整合起来，可以可靠地估计物种树。结合事件标记基因事件标记基因，这意味着一次协调，从而得到完整的基因家族历史。

（i）存在一棵物种树S，使得存在一个无水平基因转移的 σ-协调（T，S μ，σ）且满足 $\ell = \ell_\mu$。（ii）三元组集合 $\mathscr{S}$（T ℓ，σ）是一致的。

请注意，命题 1 中的物种三元组集合取决于 T 的事件标记 ℓ，但并不取决于协调映射 μ 的具体细节，只要 μ 能够产生正确的事件标记即可。已有研究考察了若干类问题：如何以最优方式编辑给定的无向图，使其成为同时满足附加约束的直系同源关系（例如，所得的事件标记基因树能够与某棵（未知的）物种树进行协调），例如见 [76]。尽管后者中的大多数问题都是 NP完全 [77, 78]，但某些与校正同源关系相关的问题可以在多项式时间内求解；这些同源关系仅提供关于直系同源基因和非直系同源基因的部分信息。

尽管命题 1 相当技术性，但它具有重要的实际意义。例如，通过余图编辑直接从互为最佳匹配数据中获得的（T ℓ）估计，可以提供一组物种三元组。将这些数据汇总于大量基因家族中，确实能够提供足够的信息，以推断完全解析的物种树 [79]。

3.3 直系同源基因簇

直系同源基因通常概括为 直系同源群簇（COGs） [80]。然而，我们已经看到，直系同源基因构成余图，因此直系同源关系通常并不是传递关系。COGs 只是一种近似，尤其适用于基因家族历史中仅包含少量复制的情况。一种特殊情形是一对一的直系同源关系，即每个基因在其他每个物种中都有唯一的直系同源基因。在这种情况下，直系同源关系变为传递关系，并且直系同源图 Θ 化简为若干个团的无交并 [81]。这些簇（例如由 OMA [81],计算得到的簇）是诱导完全子图，即完整直系同源图中的团。

大多数其他用于计算簇正交基因的方法都允许并列直系同源基因，即簇并不局限于直系同源图的团，因此可能包含直系同源基因和旁系同源基因 [80]。人们使用了各种各样的聚类算法，从序列相似性数据中提取簇正交基因。这类簇正交基因的定义必然取决于用于衡量其规模与严格程度之间权衡的严格性参数。从理论角度来看，传递性聚类 [82] 很有趣，因为它在概念上类似于协图编辑：在这里，通过向初始直系同源关系估计中插入和删除边，将其编辑为一个传递图，即划分为多个簇正交基因。在 [83], 中，计算最大 T´uran（完全多部）图。这些图构成协图的一个特殊类别，并且容纳所谓的内旁系同源基因 [24, 84],，即在每个谱系中最近一次物种形成事件之后产生的重复基因。相应地，将一个基因集合进行划分，使外旁系同源基因（即源于早于所有物种形成事件的复制的基因对）被置于不同簇中，也具有重要意义。这些基因对对应于直系同源图的连通分量 Θ。

4 Fitch 图与水平基因转移

4.1 定义与表征

Fitch [11]将“自其共同祖先以来的历史中，至少有一个基因的遗传物质经历了种间（水平）转移的两个基因 x 和 y定义为异源同源基因。”因此，在树T中，x、$y \in L(T)$ 中的两个叶节点是异源同源基因，当且仅当连接x 和y 的唯一路径在T 中包含一条水平基因转移边。根据定义 2，水平基因转移边的子集$H_\mu \subseteq E(T)$以及异源同源关系，显式取决于协调（T、S μ）。与前几节一样，我们希望在不先计算协调（T、S μ）的情况下推断异源同源关系。同样，我们通过研究由协调或宽松情景所隐含的图的性质来处理这一问题。

让我们先考虑不受水平基因转移事件影响的情景部分。从T 中删除所有水平基因转移边，可得到一个森林，并将$L(T)$ =划分为若干部分

$L_1 \cup L_2 \cup \ldots \cup L_k$，使得限制树$T^{|L_i}$中不包含水平基因转移边。因此，可以分别对各个基因集L_i 执行上述针对无水平基因转移系统的分析。请注意，一般而言，这并不等同于简单地考察由L_i诱导的经验最佳匹配图子图（$G[L_i], \sigma_{|L_i}$）。一般而言，具有$V(G_i) = L_i$的经验最佳匹配图（G_i, σ）中，可能会出现劣于由T 以及关于水平基因转移边的额外信息所隐含的（$G[L_i], \sigma_{|L_i}$）中那些最佳匹配的最佳匹配。我们将在下文重新讨论水平基因转移边的推断，以及将基因集划分为极大无水平基因转移子集的问题。

从机制上说，水平基因转移本质上是一个有方向性的事件。水平转移的“拷贝”与继续垂直传递的“原始基因”之间有着明确的区别。因此，水平基因转移边是否位于从 lca(x, y) 到 y 的路径上、从 lca(x, y) 到 x 的路径上，或同时位于这两条路径上，具有重要意义。用数学语言可以将其表述如下。

定义 11. 设 T 为一棵有根树，且 $H \subseteq E(T)$ 是水平基因转移边的集合。那么，(有向) Fitch 图 $\mathfrak{F}(T, H)$ 的顶点集为 $L(T)$；若从 $\mathrm{lca}_T(x, y)$ 到 y 的路径包含一条边 $e \in H$，则 $(x, y) \in E(\mathfrak{F})$。
若某个树 T 及边集$H \subseteq E(T)$ 满足 $G = \mathfrak{F}(T, H)$，则有向图 G 是一个(有向) Fitch 图。
[85] 文献中给出了 Fitch 图的一种刻画，即通过八个三顶点禁子图来刻画。

Fitch 图还可以用其“闭互补邻域”来进行一种出人意料地简单的刻画。对于一个有向图，考虑

$$\overline{N}[x] := \{y \in L \mid (x, y) \notin E(G)\}, \tag{6}$$

由于此处考虑的所有图都不含自环，因此有 $x \in N[x]$。我们记 $\mathscr{N}(G) := \{\overline{N}[y] \mid y \in L\}$。

定理 7 ([86]) *G 是 Fitch 图，当且仅当 (i) $\mathscr{N}(G)$ 具有层次结构型，即对于 N_1, $N_2 \in \mathscr{N}(G)$，有 $N_1 \cap N_2 \in \{N_1, N_2, \emptyset\}$；并且 (ii) 对于每个 $N \in \mathscr{N}(G)$ 以及每个 $y \in N$，都有 $|N[y]| \leq |N|$。*
如果 G 是 Fitch 图，那么 $\mathscr{C}(G) := \overline{\mathscr{N}(G)} \cup \{L\} \cup \{\{x\} \mid x \in L\}$，也就是说，将闭互补邻域扩展为单元素集合及 L 本身后，便在 L 上构成一个层次结构，并对应于一棵唯一确定的有根树 T_G。这棵树 T_G 是 G 的唯一最小分辨树，而 T_G 中的水平基因转移边集合 H_G 则唯一确定，它由 T_G 中所有满足 $x,y \in V^0(T_G)$ [85, 86] 的边 xy 组成。因此，有向 Fitch 图也能提供有关基因树结构的信息。二元组 (T_G, H_G) 还可直接得出 L 划分为无水平基因转移子集的结果。

有向 Fitch 图 G^{sym} 的对称版本称为无向 *Fitch* 图，包含所有双向边 $xy \in E(G)$。因此，当且仅当 x 与y 之间的路径包含一条水平基因转移边时，$xy \in E(G^{sym})$。无向 Fitch 图已在 [87] 中得到研究，并且与完全多部图相一致。因此，无向 Fitch 图完全由其独立集定义。这意味着，无向 Fitch 图无法传达太多有意义的系统发育信息，因为它们完全由分类单元的极大子集 L_i 决定，而这些子集中的分类单元之间没有经历任何水平基因转移事件。

从给定的直系同源图和 Fitch 图中寻找一个共同基因树的问题已在 [88] 中被研究。此外，给定一个叶标记树（T σ）和水平基因转移边 $H \subseteq E(T)$，一个在精神上类似于命题 1 的结果确定了一个三元组集合；在满足（R4）、（R5）和 $H = H_\mu$ [52, 58] 的任何 σ-协调（T, S μ, σ）中，这些三元组必须由物种树 S 显示。在 [53], 中，给出了一个多项式时间算法：该算法以（T σ）和 H 为输入，并在存在时构造物种树S 以及时间一致的协调映射 μ。

4.2 LDT 图

遗憾的是，迄今为止尚未出现能够直接从序列数据推断 Fitch 图的实用方法。一个有前景的方法始于这样一个观察：如果两个基因 x 和y 的亲缘关系比物种 σ (x) 和 $\sigma(y)$ 更近，那么x和y 必然是异源同源基因 [17]。 [31]中介绍了一种用于识别此类基因对的统计上可靠的方法。尽管这些间接方法已经取得了相当大的成功，参见 [25]中的综述，但直到最近，人们才开始研究其数学性质：

定义 12（[18]）。设（T τ_T）和（S τ_S）是两棵日期树，并且有 σ：L（T）$\to L$（S）。则无向图 $\mathfrak{L}$（T，S τ_T，τ_S，σ）的顶点集为L（T），且ab是边，当且仅当τ_T（最近公共祖先$_T$（a，b））$< \tau_S$（最近公共祖先$_S$（σ（a） σ（b））。
顶点着色图 (G σ) 是一个后发散时间图（LDT 图），当且仅当存在一对定年树的元组 (T τ_T) 和 (S τ_S)，以及映射 σ: L(T) $\to$L(S)，使得 (G σ) $= \mathfrak{L}$(T, S τ_T, τ_S, σ)。LDT 图有一个简单的刻画。为此，考虑根三元组的集合 $\mathfrak{S}$(G σ)，其中 σ(x)、 σ(y) 和σ(z) 两两不同，xy、yz $\in$E(G)，且 xy $\notin$E(G)。

定理 2（[18]）。图（G σ）是 *LDT* 图，当且仅当它是一个恰当着色的余图，且三元组集合 $\mathfrak{S}$（G σ）是相容的。
文中还描述了一种多项式时间算法，可为给定的 LDT 图构造一个松弛的 σ-情景，见 [18]。LDT 图的编辑问题为 NP完全，这一点并不令人意外。

对于任何满足 (G σ) $= \mathfrak{L}$(T, S τ_T, τ_S, σ) 的宽松 σ-情景 (T, S $\mu, \tau_T, \tau_S, \sigma$)，LDT 图 (G σ) 始终是无向 Fitch 图$\mathfrak{F}_{sym}$(T, H$^\mu$) 的子图。因此，LDT 图不可能包含假阴性的异源基因对，而 LDT 图的补图 G (G σ) 特别包含所有未被水平基因转移事件分隔的基因对之间的边。为了应用针对无水平基因转移数据开发的方法，我们需要找出给定基因集 L 的极大无水平基因转移子集划分 L $= \mathrm{L}_1 \cup \mathrm{L}_2 \cup \ldots \cup \mathrm{L}_k$。实际上，可以通过在 G 上应用所谓的簇删除问题的解来完成，即删除 G 中的最小边集，使所得图 G* 成为完全图 [89]的不相交并。将 L_i 定义为 G* 中第 i 个团的顶点集，即可得到 L 的极大无水平基因转移子集划分。如果 G（或等价地 G）$\bar{}$是余图，簇删除可以

通过贪心算法 [90] 在线性时间内求解。因此，LDT 图可用于获得 $L(T)$ 到无水平基因转移子集的划分。此类划分与给定基因树的相容性，以及利用此类划分推断水平基因转移边的方向，已成为 [91, 92] 中的研究主题

4.3 存在水平基因转移时的直系同源关系

大多数关于直系同源关系的数学结果，都是在无水平基因转移的情形下获得的。在存在水平基因转移时，源自一个物种形成事件的两个基因的后代甚至最终可能位于同一个物种中，在那里它们表现为旁系同源基因。这促成了对直系同源关系“正确”定义的不同提议。文献 [12]提出了异源同源亚型的一种分类；该分类与 [11], 一致，将术语直系同源基因和旁系同源基因保留用于 x 与 y 之间的路径不包含水平基因转移事件的情形。与 LDT 图类似，“等分化时间”（EDT）图刻画了这样一个事实：两个基因 x 和 y 的分化时间与它们所在物种的分化时间相匹配。与 LDT 图类似，若存在两棵定年树（T τ_T）和（S τ_S）以及 σ，使得 L(T) $\to$L(S)，则顶点着色无向图（ G σ ）是 EDT 图，当且仅当 ab $\in$E(G) 恰好当 τ_T (lca$_T$(a, b)) $=$ τ_S(lca$_S$(σ(a) σ(b))。在这种情况下，我们也写作（G σ） $=$Œ(T, S τ_T, τ_S, σ)。一般而言，EDT 图识别是 NP 困难的 [19],；然而，如果能够获得 LDT 图的信息，则它变为多项式时间可解。为了解释这一点，请注意，EDT 图与 LDT 图的并集的补图包含所有满足 τ_T(lca$_T$(a, b)) $>$ τ_S(lca$_S$(σ(a) σ(b)) 的边 ab；这确定了所谓的先验分化时间图，我们可以应用 [19, 推论。 8]

容易验证，在满足 (R5) 的无水平基因转移 σ-协调中（T, S μ, σ），公理 (R5)(i) 意味着：如果两个基因 x 和 y 的最近共同祖先 lca$_T$(x, y)，与 x 和 y 所在的两个物种的最近共同祖先 lca$_S$(σ(x) σ(y)) 重合，那么这两个基因就是直系同源基因。在这种情况下，直系同源关系与 EDT 图之间显然存在关联。下面我们总结一些研究结果，这些结果表明，在文献讨论过的存在水平基因转移的情形下，EDT 图也与不同的直系同源关系概念密切相关。对水平基因转移存在时直系同源关系定义的分歧，源于这样一个事实：通常，由物种形成事件产生的基因对可能会因水平基因转移而分离，从而成为异源同源基因。同时，它们最终甚至可能位于同一物种中，因此表现为旁系同源基因。

定义 13. 设 $\mathscr{S} = (T, S\ \mu, \tau_T, \tau_S, \sigma)$ 是一个宽松的 σ-情景。那么，两个不同的顶点 $x,y \in L(T)$ 是

- 弱准直系同源基因 当 μ(最近公共祖先$_T$ (x,y)) $\in V^0(S)$。
- 弱直系同源基因 当 μ(最近公共祖先$_T$(x,y)) $\in V^0(S)$，并且在 T 中 x 与 y 之间的路径上不存在水平基因转移边。
- 严格准直系同源若 μ(最近公共祖先$_T$ (x、 y)) $=$最近公共祖先$_S$(σ(x) σ(y))。

•严格直系同源基因 若 μ(最近公共祖先$_T(x, y)) =$最近公共祖先$_S(\sigma(x)\ \sigma(y))$, 并且在T中x 与y 之间的路径上不存在水平基因转移边。

无向图 $\Psi_w(\mathscr{S})$、 $\Theta^w(\mathscr{S})$、 $\Psi_s(\mathscr{S})$ 和 $\Theta^s(\mathscr{S})$ 分别具有顶点集$L(T)$; 当且仅当基因x 和y 在 $\mathscr{S}$ 中分别是弱准直系同源基因、弱直系同源基因、严格准直系同源和严格直系同源基因时，才存在边 xy 。摘要见表1。

表1 给定一个宽松的 σ 情景 $\mathscr{S}$ 时，存在水平基因转移事件情况下直系同源关系的不同概念汇总。

协调条件	水平基因转移无关	排除水平基因转移
μ(最近公共祖先$_T(x, y)) \in V^0$ (S)	$\Psi^w(\mathscr{S})$ 弱准直系同源基因	$\Theta^w(\mathscr{S})$ 弱直系同源基因
μ(最近公共祖先$_T(x, y)) =$ 最近公共祖先$_S(\sigma$ $(x)\ \sigma(y))$	$\Psi^s(\mathscr{S})$ 严格准直系同源（严格）直系同源基因	$\Theta^s(\mathscr{S})$

弱准直系同源基因本质上是 Walter Fitch 最初提出的、纯粹基于事件的直系同源关系定义 [10]。在后来的研究中，Fitch [11] 强调了这样一个条件：“共同祖先位于获得这两个序列的分类单元的最近共同祖先中”，这对应于严格准直系同源基因的概念。其他直系同源关系定义明确排除了异源同源基因 [11, 12, 93], , 由此引出了弱直系同源基因和严格直系同源基因的概念。

如 [19], 所示，对于每个宽松的 σ-情景 $\mathscr{S}$，我们有

$$\Theta^s(\mathscr{S}) \subseteq \Psi^s(\mathscr{S}) \subseteq \Psi^w(\mathscr{S}) \text{ and } \Theta^s(\mathscr{S}) \subseteq \Theta^w(\mathscr{S}) \subseteq \Psi^w(\mathscr{S}), \qquad (7)$$

而 $\Psi_s(\mathscr{S})$ 与 $\Theta^w(\mathscr{S})$ 在子图关系下不可比较。此外，对于每个宽松的 $\mathscr{S}$，弱准直系同源图 $\Psi_w(\mathscr{S})$、弱直系同源图 $\Theta^w(\mathscr{S})$ 和严格直系同源图 $\Theta^s(\mathscr{S})$ 都是余图。一般而言，严格准直系同源图 $\Psi^s(\mathscr{S})$ 则并非如此。LDT 和 EDT 图与弱、严格直系同源基因之间存在密切联系:

命题 2。 *[19,* 命题 *5]* 给定两个图 G_1 和 G_2，当且仅当存在一个宽松的 σ-情景 $\mathscr{S}$，使得 $G_1 = \mathfrak{L}$ （ $\mathscr{S}$ ）和 $G_2 = \mathfrak{E}$ （ $\mathscr{S}$ ）; 同时，存在一个宽松的 σ- 情景 $\mathscr{S}'$，使得 $G_1 = \mathfrak{L}$（$\mathscr{S}'$）和 $G_2 = \mathfrak{E}$（$\mathscr{S}'$），并且还满足 Θ^w（$\mathscr{S}'$）$=\Theta^s$（$\mathscr{S}'$）。

也就是说，一对 LDT 图和 EDT图可以由一个共同的宽松情景加以解释，当且仅当这两个图可以由某个宽松情景加以解释，并且在该情景中弱直系同源关系与严格直系同源关系一致。

在一般情形下，彼此无关的进化事件发生在不同的时间点:

定义 14。 如果对于x、$y \in L(T)$以及$u \in V^0(S)$， $\tau_T(\mathrm{lca}_T(x$、$y)) = \tau_S(u)$，则宽松情景是一般的，且 $\mu(\mathrm{lca}_T(x$、$y)) = u$。

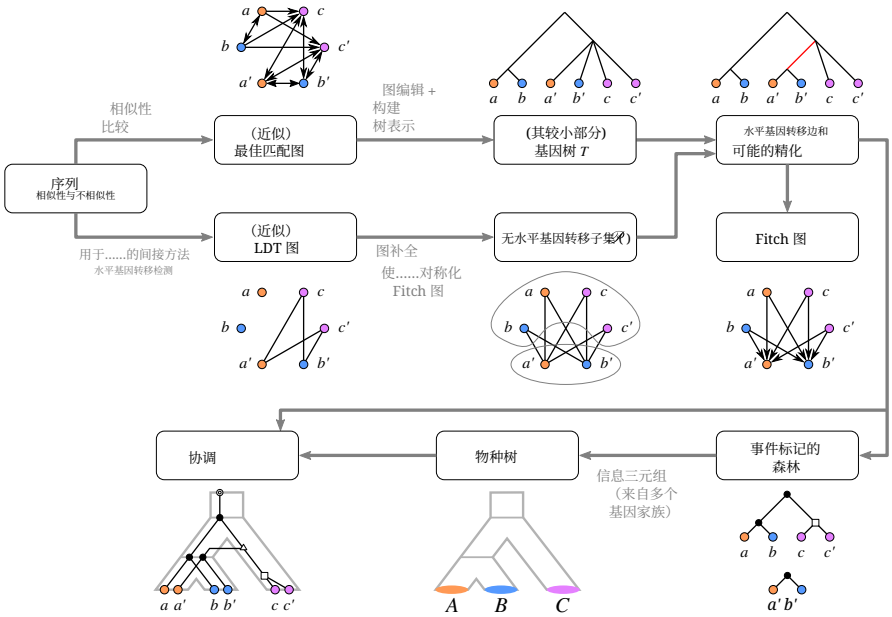


图 6 基于无树方法的直系同源关系检测以及事件标记基因树 T、物种树 S 和 σ-协调（T,S μ,σ）重建的概念性工作流程，假设发生了水平基因转移。在此，我们有 σ(x) = σ(x′) =X，对于 (x,x′,X) ∈ {(a,a′,A),(b,b′,B),(c,c′,C)}。利用序列相似性数据获得最佳匹配图的初始估计。随后，将其校正为数学上严谨的最佳匹配图和 LDT 图。通过多项式时间算法，从这些结果中获得通常未完全解析的基因树和无水平基因转移子集。下一步，可将图 5 所概述的工作流程应用于基因的无水平基因转移子集，从而得到事件标记基因树，其中也包含有关物种树的信息。整合多个基因家族中的后者信息，可以可靠地估计物种树。结合事件标记基因树，这意味着可以进行协调，从而获得完整的基因家族历史。

在一般情景中，EDT图与直系同源关系的不同概念连接如下：

$$\Theta^s(\mathscr{S}) \subseteq \Psi^s(\mathscr{S}) = \mathrm{EDT} \subseteq \Psi^w(\mathscr{S}) \tag{8}$$

在无水平基因转移情形下，如果 μ(最近公共祖先$_T$(x、 y)) ∈V^0(S) 意味着μ(最近公共祖先$_T$(x、 y)) =最近公共祖先$_S$(σ(x) σ(y)，那么所有这些图都相同。然而，对于存在水平基因转移时直系同源关系及其变体的完整形式化理解仍然缺失。

5 讨论与开放问题

本章概述了当前对基因家族历史的形式化、数学理解，重点介绍“无树”方法。我们有意未在讨论中涉及完整的概率模型。这些模型同样从演化情景的概念出发，并使用序列演化随机模型

来分配概率，随后利用最大似然或贝叶斯框架重建演化情景 [50, 94, 95]。尽管这些精细模型有望为小型和中型数据集提供准确结果，但它们能否扩展到超大规模问题仍不清楚。另一方面，此处概述的组合方法有望处理全球尺度的基因家族历史。

总而言之，图5和图6勾勒出一种通用的概念工作流程，可在“无树”框架下全面分析基因家族历史。我们不再使用基因树T 和物种树S，而是根据数据估计最佳匹配图和 LDT/EDT图。这两种图共同捕捉了关键信息，正如前文各节的形式化结果所概述的那样。与直接重建树相比，这种方法的主要优势在于，它能够利用数据中固有的冗余。在独立估计最佳匹配图和LDT/EDT图的边之后，可以利用这样一个事实来校正数据：结构良好的最佳匹配图和 LDT 图属于高度受限的特殊图类，因此可以将初始估计编辑为各自类型中最接近的结构良好的图。遗憾的是，这些图编辑问题是 NP完全的。不过，已有可行的启发式算法可用于解决这一问题。数据校正任务的困难也不足以构成反对图6所提工作流程的概念性理由；毕竟，多序列比对问题（见 [96]及其中的参考文献），以及通过最大简约法或最大似然重建系统发育树，同样是 NP完全的 [97–99]。随后，我们可以从最佳匹配图和LDT/EDT图中获得一棵不一定完全解析的基因树、关于物种的充分信息，以及相应的事件标记，以定位所有相关的进化事件。最后，可以在组合推断得到的树T 和S 作为约束的条件下，利用传统系统发育方法进一步解析基因树和物种树。尽管本章撰写时尚无成熟的完整实现，但已使用模拟数据对各个步骤进行了基准测试，并证明其在现实情境下可行 [100]。

协调这一概念的适用范围超出了本章的范畴。本章聚焦于基因组中的基因，或者说宿主-寄生物系统[45, 101]。研究蛋白质中结构域的演化时，也会出现相同的结构。这自然引出了多层次的协调版本[51, 102]。除了物种形成、基因复制、基因丢失和水平转移之外，研究其他类型的事件也很有意义。尤其是不完全谱系排序或深度共祖，以及杂交，都受到了广泛关注 [40, 41, 43, 103, 104]。有趣的是，在这种情况下，最大简约法协调变成了 NP 困难 [105]。系统发育网络是系统发育树的自然推广 [106, 107]。协调的概念在这类模型中依然适用。例如， [108, 109], 研究了基因树与某些受限类别的物种网络之间的协调。在[110], 中，研究了将基因网络（受重组启发）与物种树协调的问题，并提出了一种最近共同祖先协调的类比方法，为所谓的树子网络提供了解法。据我们所知，这些扩展模型尚未

迄今为止，这一问题都是从“无树/网络”视角进行研究的，而这一视角可扩展本章综述的结果。

自本书第一版问世以来，尽管已经取得了长足进展， [32], 仍有许多关键问题尚待解决，尤其是水平基因转移事件的检测与定位问题。最重要的是，可通过经验数据获取的关系（最佳匹配、较低／相等的分化时间）为协调（T，S μ）提供了独立约束，而推断出的直系同源关系和异源同源关系也可能施加额外条件。在大多数情况下，解释性情景存在所需条件中至少有一部分，是用三元组集合的一致性来表示的。这意味着应考察这些集合的并集是否一致。一般而言，这会引出一个 NP 完全问题：确定三元组集合的最大一致子集（或多个最大一致子集），或其变体 [111]。不过，更深入地理解冲突约束是如何产生的，或许能带来更准确且可能计算效率更高的方法，以处理数据中的内部冲突。另一方面，研究如何利用可信的物种树改进最佳匹配图的初始估计，也将颇有意义。在最大似然框架中，已有研究证明此类信息确实能发挥作用 [112]。这一主题的一个有趣变体，是从一组基因树推断物种树和协调映射的方法 [113]。对于给定的 T、S 和 σ，是否存在一项具有指定协调边集合 $H \subseteq E(T)$ 的协调，目前也仍不明确。我们猜想，与无水平基因转移情形类似，答案是肯定的。更一般地说，对于宽松情景，事件标记及其后果尚未得到充分理解。虽然事件标记作为协调映射的一部分被隐式定义，例如在 [26, 42, 43]，中，但这些定义也限制了其适用的情景，另见 [52, 53]。

从实用角度来看，最紧迫的问题是，为在现实应用无树方法时自然产生的图编辑问题开发更高效的启发式算法，即图5和图6所概述流程的中间步骤。对于余图 [72–74] 和最佳匹配图 [67]，近期已经取得了显著进展。面向大规模应用的完整生物信息学流程主要关注簇正交基因，而不是对基因家族历史进行更细粒度的呈现，因此对数学系统发育学方面的进展利用有限。结合大幅改进的全基因组比对方法 [114],，改进的基因家族历史也可能成为全面理解基因组演化的重要一步。

参考文献

[1] Michael Lynch 和 John S. Conery。重复基因的进化归宿及其后果。《科学》，290:1151–1155，2000年。[2] Asher D. Cutter 和 Richard Jovelin。自然选择何时对基因功能视而不见。《生物学论文集》，37(11):1169–1173，2015年。

[3] H Innan 和 F Kondrashov。基因复制的进化：对模型进行分类与区分。Nat Rev Genet，11:97–108，2010年。[4]Adrian M. Altenhoff、Romain A. Studer、Marc Robinson-Rechavi 和 Christophe Dessimoz。解决直系同源关系猜想：与旁系同源基因相比，直系同源基因的功能相似性往往较弱，但差异显著。《PLOS计算生物学》，8(e1002514)，2012年。[5] R. A. Studer 和 M. Robinson-Rechavi。我们有多大把握认为直系同源基因相似，而旁系同源基因有所不同？Trends in Genetics，25(210-216)，2009年。[6] T. Gabald´ón 和 E. V. Koonin。基因直系同源关系的功能与进化意义。Nat. Rev. Genet.，14(5):360–366，2013年。[7] N. L. Nehrt、W. T. Clark、P. Radivojac 和 M. W. Hahn。利用哺乳动物比较功能基因组数据检验直系同源关系猜想。《PLOS计算生物学》，7(e1002073)，2011年。[8] Erik Sonnhammer、Toni Gabald´ón、Alan Wilter Sousa da Silva、Maria Mar- tin、Marc Robinson-Rechavi、Brigitte Boeckmann、Paul Tho mas、Christophe Dessimoz 和直系同源基因研究联盟。大数据以及寻找直系同源基因面临的其他挑战。Bioinformatics，30(21):2993–2998，2014年。[9] Jes´us A. Ballesteros 和 Gustavo Hormiga。一种用于系统基因组数据的新型直系同源关系评估方法：无根系统发育直系同源关系。分子生物学与进化，33(8):2117–2134，2016年。[10] Walter M Fitch。区分同源蛋白质与类似蛋白质。Syst Zool，19:99–113，1970年。[11] Walter M. Fitch。同源关系：关于若干问题的个人看法。Trends Genet.，16:227–231，2000年。[12] Charlotte A Darby、Maureen Stolzer、Patrick J Ropp、Daniel Barker 和 Dannie Durand。异源基因分类。Bioinformatics，33:640–649，2017年。[13] Manuela Geiß、Marcos E. Gonz´ález Laffitte、Alitzel L´opez S´anchez、Dulce I. Valdivia、Marc Hellmuth、Maribel Hern´andez Rosales 和 Peter F. Stadler。最佳匹配图以及基因树与物种树的协调。《数学生物学杂志》，80:1459–1495，2020年。[14] Manuela Geiß、Peter F. Stadler 和 Marc Hellmuth。互为最佳匹配图。《数学生物学杂志》，80:865–953，2020年。[15] Peter F. Stadler、Manuela Geiß、David Schaller、Alitzel L´opez S´anchez、Marcos Gonz´alez Laffitte、Dulce I. Valdivia、Marc Hellmuth 和 Maribel Hern´andez Rosales。从最相似序列对到系统发育最佳匹配。Algorithms for Molecular Biology，15:5，2020年。[16]Manuela Geiß、Edgar Ch´avez、Marcos Gonz´alez Laffitte、Alitzel L´opez S´anchez、B¨arbel M. R. Stadler、Dulce I. Valdivia、Marc Hell- muth、Maribel Hern´andez Rosales 和 Peter F. Stadler。最佳匹配图。《数学生物学杂志》，78(7):2015–2057，2019年。[17] Pavel S Novichkov、Marina V Omelchenko、S Gelfand Mikhail、Andrei A Mironov、Yuri I Wolf 和 Eugene V Koonin。细菌进化中的全基因组分子钟与水平基因转移。J. Bacteriol.，186:6575–6585，2004年。

[18] David Schaller、Manuel Lafond、Peter F. Stadler、Nicolas Wiesecke 和 Marc Hellmuth。水平基因转移的间接识别。《数学生物学杂志》，83:10，2021年。arXiv 2012.08897。[19] David Schaller、Tom Hartmann、Manuel Lafond、Nicolas Wieseke、Marc Hellmuth 和 Peter F. Stadler。进化情景中的相对时间信息与直系同源关系。技术报告 2212.02201，arXiv，2023年。[20] V P Doyle、R E Young、G J Naylor 和 J M Brown。我们能否识别系统发育可靠性更高的基因？《系统生物学》，64(5):824–837，2015年。[21] B Boussau、G J Sz¨oll˝osi、L Duret、M Gouy、E Tannier 和 V Daubin。物种树与基因树的基因组尺度联合估计。《基因组研究》，23(2):323–330，2013年。[22] Gergely J. Sz¨oll˝osi、Eric Tannier、Vincent Daubin 和 Bastien Boussau。利用物种树推断基因树。*Syst. Biol.*，64:e42–e62，2015年。[23] Adrian M. Altenhoff 和 Christophe Dessimoz。直系同源基因推断项目与方法的系统发育和功能评估。《*PLOS*计算生物学》，5:e1000262，2009年。[24] A. M. Altenhoff、N. M. Glover 和 C. Dessimoz。直系同源关系与旁系同源关系推断。载于《进化基因组学》，第1910卷《分子生物学方法》，第149–175页。Humana，纽约，纽约州，2019年。ˇ[25] Matt Ravenhall、Nives Skunca、Florent Lassalle 和 Christophe Dessimoz。水平基因转移推断。《*PLOS*计算生物学》，11:e1004095，2015年。[26] Ali Tofigh、Michael Hallett 和 Jens Lagergren。复制与侧向基因转移的同时识别。*IEEE/ACM Trans.Comp.Biol.Bioinf.*，8(2):517–535，2011年。[27] Zhi-Zhong Chen、Fei Deng 和 Lusheng Wang。复制、丢失和侧向基因转移的同时识别。*IEEE/ACM Trans. Comp.Biol.Bioinf.*，9，2012年。[28] W Ma、D Smirnov、J Forman、A Schweickart、C. Slocum、S. Srinivasan 和 R. Libeskind-Hadas。DTL-RnB：用于概括DTL协调空间的算法与工具。*IEEE/ACM Trans. Comp.Biol. Bioinf.*，15:411–421，2018年。[29] Christine Dufraigne、Bernard Fertil、Sylvain Lespinats、Alain Giron 和 Patrick Deschavanne。利用基因组特征检测和表征原核生物中的水平转移。《核酸研究》，33:e6，2005年。[30] Jennifer Becq、C´ecile Churlaud 和 Patrick Deschavanne。用于检测水平转移的参数方法基准测试。《*PLOS ONE*》，5:e9989，2010年。[31] Aditi Kanhere 和 Martin Vingron。原核生物中的水平基因转移对代谢基因和翻译基因表现出不同的偏好。《*BMC*进化生物学》，9:9，2009年。[32] Jo˜ao C. Setubal 和 Peter F. Stadler。基因系统发育关系与直系同源群。载于Jo˜ao C. Setubal、Peter F. Stadler 和 Jens Stoye 编，《比较基因组学》，第1704卷，第1–28页。Springer，海德堡，2018年。

[33] A. V. Aho、Y. Sagiv、T. G. Szymanski 和 J. D. Ullman。利用最近共同祖先推断一棵树，并将其应用于关系表达式的优化。*SIAM J. Comput.*，10:405–421，1981年。[34] Ying-Jun He、Trinh N D Huynh、Jesper Jansson 和 Wing-Kin Sung。通过避开禁止的有根三元组推断系统发育关系。*J.Bioinf.Comp. Biol.*，4:59–74，2006年。[35] Frank Rutschmann。系统发育树的分子定年：当前分化时间估计方法简述。*Diversity and Distributions*，12:35–48，2006年。[36] Herv´e Sauquet。分子定年的实用指南。*Comptes RendusPalevol*，12:355–367，2013年。[37] Daniel Ford、Frederick A. Matsen 和 Tanja Stadler。一种研究系统发育树相对时间信息的方法。*SystematicBiology*，58:167–183，2009年。[38] Gergely Sz¨oll˝osi、Sebastian H¨ohna、Tom A Williams、Dominik Schrempf、Vincent Daubin 和 Bastien Boussau。相对时间约束可提高分子定年的准确性。*Systematic Biology*，71:797–809，2022年。[39] David A Marques、Joana I Meier 和 Ole Seehausen。从组合角度看物种形成与适应性辐射。*Trends Ecol Evol*，34:531–544，2019年。[40] Y Zheng 和 L. Zhang。不完全谱系排序对基于树协调的基因复制推断的影响。*IEEE/ACM Trans Com-put Biol Bioinform*，11:477–485，2014年。[41] Y B Chan、V Ranwez 和 C. J. Scornavacca。通过协调推断不完全谱系排序、复制、转移和丢失。*Theor.Biol.*，432:1–13，2017年。[42] Mukul S. Bansal、Eric J. Alm 和 Manolis Kellis。针对基因复制、水平转移和丢失协调问题的高效算法。*Bioinformatics*，28:i283–i291，2012年。[43] Maureen Stolzer、Han Lai、Minli Xu、Deepa Sathaye、Benjamin Vernot 和 Dannie Durand。利用非二叉物种树推断复制、丢失、转移和不完全谱系排序。*Bioinformatics*，28:i409–i415，2012年。[44] Roderic Guig´o、I Muchnik 和 T F Smith。古分子系统发育的重建。*Mol Phylogenet Evol*，6:189–213，1996年。[45] Roderic D. M. Page 和 Michael A. Charleston。从基因到生物体系统发育：协调树与基因树/物种树问题。*Mol.Phylog. Evol.*，7:231–240，1997年。[46]Louxin Zhang。关于比较分子系统发育的Mirkin-Muchnik-Smith猜想。*J. Comp.Biol.*，4:177–187，1997年。[47] K. Chen、D Durand 和 M Farach-Colton。NOTUNG：用于基因复制定年和优化基因家族树的程序。*J. Comput. Biol.*，7:e429–e447，2000年。[48] C M Zmasek 和 S R Eddy。一种在基因树上推断基因复制和物种形成事件的简单算法。*Bioinformatics*，17:821–828，2001年。

[49] P G´orecki、G J Burleigh 和 O Eulenstein。包含复制和丢失的基因树演化的最大似然模型和算法。《BMC基因组学》，12:S15，2011年。[50] L. Arvestad。使用马尔可夫链蒙特卡罗进行贝叶斯基因树/物种树协调和直系同源分析。Bioinformatics，19:i7–i15，2003年。[51] Hugo Menet、Vincent Daubin 和 Eric Tannier。系统发育协调。《PLOS计算生物学》，18:e1010621，2022年。[52] Nikolai Nøjgaard、Manuela Geiß、Daniel Merkle、Peter F. Stadler、Nicolas Wieseke 和 Marc Hellmuth。时间一致的协调映射与被禁止的时间旅行。Alg. Mol. Biol.，13:2，2018年。[53] Manuel Lafond 和 Marc Hellmuth。时间一致物种树的重建。Alg. Mol. Biol.，15:16，2020年。[54] Paweł G´orecki 和 Jerzy Tiuryn。DLS树：一种演化情景模型。Theor. Comp. Sci.，359:378–399，2006年。[55] B Vernot、M Stolzer、A Goldman 和 D Durand。与非二叉物种树的协调。J Comput Biol.，15:981–1006，2008年。[56] J-P Doyon、V Ranwez、V Daubin 和 V Berry。系统发育协调的模型、算法和程序。Brief Bioinform.，12:392–400，2011年。[57] L. Y. Rusin、E. Lyubetskaya、K. Y. Gorbunov 和 V. Lyubetsky。基因树和物种树的协调。BioMed Res Int.，2014:642089，2014年。[58] M. Hellmuth。生物学上可行的基因树、协调映射和信息性三元组。Alg. Mol. Biol.，12:23，2017年。[59] Marcus Lechner、Sven Findeiß、Lydia Steiner、Manja Marz、Peter F. Stadler 和 Sonja J. Prohaska。Proteinortho：大规模分析中的（共）直系同源基因检测。《BMC基因组学》，12:124，2011年。[60] David Schaller、Manuela Geiß、Edgar Ch´avez、Marcos Gonz´alez Laffitte、Al- itzel L´opez S´anchez、B¨arbel M. R. Stadler、Dulce I. Valdivia、Marc Hellmuth、Maribel Hern´andez Rosales 和 Peter F. Stadler。关于“最佳匹配图”的勘误。J. Math. Biol.，82:47，2021年。[61] David Schaller、Peter F Stadler 和 Marc Hellmuth。最佳匹配图修改问题的复杂性。Theor. Comp. Sci.，865:63–84，2021年。[62] W. Maddison。在多分叉支序图上重建性状演化。Cladistics，5:365–377，1989年。[63] R. DeSalle、R. Absher 和 G. Amato。物种形成与系统发育分辨率。Trends Ecol. Evol.，9:297–298，1994年。[64] H E Walsh、M G Kidd、T Moum 和 V L Friesen。多分叉与系统发育推断的能力。Evolution，53:932–937，1999年。[65] David Schaller、Manuela Geiß、Peter F. Stadler 和 Marc Hellmuth。最佳匹配图中错误直系同源关系分配的完整刻画。J. Math. Biol.，82:20，2021年。arXiv：2006.02249。[66] David Schaller、Manuela Geiß、Marc Hellmuth 和 Peter F. Stadler。具有二叉树的最佳匹配图。载于 Carlos Mart´ın-Vide、Miguel A. Vega-Rodr´ıguez 和 Travis Wheeler（编），《计算生物学算法》，第8届 AlCoB，第12715卷，Lect. Notes Comp. Sci.，第82–93页，2021年。

[67] David Schaller、Manuela Geiß、Marc Hellmuth 和 Peter F. Stadler。最佳匹配图编辑的启发式算法。*Alg.Mol. Biol.*，16:19，2021年。[68] Marc Hellmuth、Maribel Hernandez-Rosales、Katharina T. Huber、Vincent Moulton、Peter F. Stadler 和 Nicolas Wieseke。直系同源关系、符号超度量与余图。*J. Math.Biol.*，66:399–420，2013年。[69] D. G. Corneil、H. Lerchs 和 L. Stewart Burlingham。补可约图。《离散应用数学》，3(3):163–174，1981年。[70] Marc Hellmuth、Manuela Geiß 和 Peter F. Stadler。互为最佳匹配图修改问题的复杂性。《理论计算机科学》，809:384–393，2020年。[71] Yunlong Liu、Jianxin Wang、Jiong Guo 和 Jianer Chen。余图编辑的复杂性和参数化算法。*Theor. Comp.Sci.*，461:45–54，2012年。[72] W. Timothy J. White、Marcus Ludwig 和 Sebastian B¨ocker。余图编辑的精确算法和启发式算法。技术报告 1711.05839 v3，arXiv，2018年。[73] Marc Hellmuth、Adrian Fritz、Nicolas Wieseke 和 Peter F. Stadler。余图编辑问题的技术：模块合并等价于编辑P_4's。《艺术、离散与应用数学》，3:#P2.01，2020年。[74] Christophe Crespelle。线性时间最小余图编辑。载于 E. Bampis 和 A. Pagourtzis（编），计算理论基础（*FCT 2021*），第12867卷，*Lect. Notes Comp. Sci.*，第176–189页，Cham，CH，2021年。Springer。[75] Maribel Hernandez-Rosales、Marc Hellmuth、Nick Wieseke、Katharina T. Huber、Vincent Moulton 和 Peter F. Stadler。从事件标记基因树到物种树。《*BMC*基因组学》，13（增刊19）：S6，2012年。[76] Manuel Lafond、Riccardo Dondi 和 Nadia El-Mabrouk。直系同源关系与基因树之间的联系：修正视角。*Alg.Mol.Biol.*，11:4，2016年。[77] Manuel Lafond 和 Nadia El-Mabrouk。直系同源和旁系同源约束：可满足性与一致性。《*BMC*基因组学》，15 S6:S12，2014年。[78] Nikolai Nøjgaard、Nadia El-Mabrouk、Daniel Merkle、Nicolas Wieseke 和 Marc Hellmuth。部分同源关系——基于二余图的可满足性。载于 Lusheng Wang 和 Daming Zhu（编），计算与组合学 *COCOON'18*，第10976卷，*Lect. Notes Comp. Sci.*，第403–415页，Cham，2018年。施普林格国际出版。[79] Marc Hellmuth、Nicolas Wieseke、Marcus Lechner、Hans-Peter Lenhof、Mar- tin Middendorf 和 Peter F. Stadler。从旁系同源基因进行系统发育分析。《美国国家科学院院刊》，112:2058–2063，2015年。[80] R L Tatusov、E V Koonin 和 D J Lipman。蛋白质家族的基因组学视角。《科学》，278:631–637，1997年。[81] Alexander C J Roth、Gaston H Gonnet 和 Christophe Dessimoz。用于大规模直系同源推断的 OMA 算法。《*BMC*基因组学》，9:518，2008年。

[82] Sven Rahmann、Tobias Wittkop、Jan Baumbach、Marcel Martin、Anke Truß 和 Sebastian B¨ocker。加权簇编辑的精确算法与启发式算法。载于 第六届计算系统生物信息学 *LSS* 会议论文集（*CSB2007*），第 391–401 页。生命科学学会，2007年。[83] Craig Falls、Bradford Powell 和 Jack Snœyink。使用 Tur´an 型图计算高严格性簇正交基因。技术报告，北卡罗来纳大学，2008年。[84] Olivier Tremblay-Savard 和 Krister M Swenson。用于内旁系同源基因检测的图论方法。*BMCBioinformatics*，13:S16，2012年。[85] Manuela Geiß、John Anders、Peter F. Stadler、Nicolas Wieseke 和 Marc Hellmuth。根据 Fitch 的异源同源关系重建基因树。《数学生物学杂志》，77(5):1459–1491，2018年。[86] Marc Hellmuth 和 Carsten R. Seemann。Fitch 异源同源关系的替代刻画。《数学生物学杂志》，79(3):969–986，2019年。[87] Marc Hellmuth、Yangjing Long、Manuela Geiß 和 Peter F. Stadler。关于无向 Fitch 图的简短说明。《艺术、离散与应用数学》，1(1):#P1.08，2018年。[88] Marc Hellmuth、Mira Michel、Nikolai N. Nøjgaard、David Schaller 和 Pe- ter F. Stadler。在共同的系统发育树中结合直系同源关系与异源同源关系数据。载于 Peter F. Stadler Stadler、Maria Em´ılia M. T. Walter、Maribel Hern´andez-Rosales 和 Marcelo M. Brigido 编，生物信息学与计算生物学进展，第 *14* 届 *BSB*，*Lect. NotesBioinf*第 13063 卷，第 53–64 页，Cham，CH，2021年。Springer Nature。[89] Ron Shamir、Roded Sharan 和 Dekel Tsur。簇图修改问题。《离散应用数学》，144(1-2):173–182，2004年。[90] Yong Gao、Donovan R. Hare 和 James Nastos。余图的簇删除问题。离散数学，313(23):2763–2771，2013年。[91] Marc Hellmuth、David Schaller 和 Peter F. Stadler。分区与树、层次结构及分裂系统的相容性。《离散应用数学》，314:265–283，2022年。[92] David Schaller、Marc Hellmuth 和 Peter F. Stadler。Fitch 图的定向与基因树中水平基因转移的检测。*CoRR*，abs/2112.00403，2023年。[93] G. S. Gray 和 W. M. Fitch。抗生素抗性基因的进化：来自金黄色葡萄球菌的卡那霉素抗性基因 DNA 序列。分子生物学与进化，1:57–66，1983年。[94] ¨Orjan ˚Akerborg、Bengt Sennblad、Lars Arvestad 和 Jens Lagergren。同时进行贝叶斯基因树重建与协调分析。《美国国家科学院院刊》，106(14):5714–5719，2009年。[95] Bret R. Larget、Satish K. Kotha、Colin N. Dewey 和 C´ecile An´e。BUCKy：通过贝叶斯一致性分析进行基因树/物种树协调。*Bioinformatics*，26(22):2910–2911，2010年。[96] Isaac Elias。解决多重比对的棘手性。*J. Comp. Biol.*，13:1323–1339，2006年。[97] William H. E. Day。系统发育系统学中计算上困难的简约性问题。*J. Theor.Biol.*，103:429–438，1983年。

[98] William H. E. Day。从差异矩阵推断系统发育树的计算复杂性。*Bull. Math. Biol.*，49:461–467，1987年。[99] S Roch。系统发育树的最大似然重建是困难问题的一个简短证明。*IEEE/ACM Trans. Comp. Biol.Bioinf.*，3:92–94，2006年。[100] David Schaller、Marc Hellmuth 和 Peter F. Stadler。AsymmeTree：用于模拟复杂基因家族历史的灵活 Python 软件包。软件，1:276–298，2022年。[101] D Merkle、M Middendorf 和 N Wieseke。用于推断共系统发育关系的参数自适应动态规划方法。*BMCBioinformatics*，11(Suppl 1):S60，2010年。[102] Simon Penel、Hugo Menet、Th´eo Tricou、Vincent Daubin 和 Eric Tannier。Thirdkind：展示超越两级协调的系统发育交汇。*Bioinformatics*，38(8):2350–2352，2022年。[103] Peng Du 和 Luay Nakhleh。复制-丢失-共祖模型下的物种树与协调估计。载于 *2018 年 ACM 生物信息学、计算生物学与健康信息学国际会议论文集*，第 376–385 页，2018年。[104] J Ansarifar、A Markin、P G´orecki 和 O. Eulenstein。统一复制-丢失-共祖模型的整数线性规划表述。载于 Z. Cai、I. Mandoiu、G. Narasimhan、P. Skums 和 X. Guo 编，*生物信息学研究与应用（ISBRA 2020）*，*Lect.Notes Comp. Sci.*第 12304 卷，第 229–242 页，Cham，2020年。Springer。[105] Matthew LeMay、Yi-Chieh Wu 和 Ran Libeskind-Hadas。不完全谱系排序与杂交存在时的最简约协调问题是 NP 困难的。载于 Alessandra Carbone 和 Mohammed El-Kebir 编，*第 21 届生物信息学算法国际研讨会（WABI 2021）*，*Leibniz International Proceedings in Informatics*（*LIPIcs*）第 201 卷，第 1 页，德国达格斯图尔，2021年。Schloss Dagstuhl – Leibniz-Zentrum f¨ur Informatik。[106] Marc Hellmuth、Katharina T. Huber 和 Vincent Moulton。用多树与物种网络协调事件标记基因树。《数学生物学杂志》，79(5):1885–1925，2019年。[107] Daniel H. Huson、Regula Rupp 和 Celine Scornavacca。系统发育网络：概念、算法与应用。剑桥大学出版社，英国剑桥，2010年。[108] Thu-Hien To 和 Celine Scornavacca。通过复制与丢失事件协调基因树和物种网络的高效算法。*BMC Genomics*，16(Suppl 10):S6，2015年。[109] Celine Scornavacca、Joan Carles Pons Mayol 和 Gabriel Cardona。基因树与 LGT 网络协调的快速算法。*J.Theor.Biol.*，418:129–137，2017年。[110] Yao-Ban Chan 和 Charles Robin。基因网络与物种树的协调。*J Theor Biol*，472:54–66，2019年。[111] Jaroslaw Byrka、Sylvain Guillemot 和 Jesper Jansson。优化有根三元组一致性的新结果。《离散应用数学》，158:1136–1147，2010年。

[112] Benoit Morel、Alexey M Kozlov、Alexandros Stamatakis 和 Gergely J. Sz¨oll˝osi。GeneRax：一种在基因复制、转移和丢失情况下，进行物种树感知的基于最大似然的基因家族树推断工具。分子生物学与进化，37(9):2763–2774，2020年。[113] Benoit Morel、Paul Schade、Sarah Lutteropp、Tom A Williams、Gergely J. Sz¨oll˝osi 和 Alexandros Stamatakis。SpeciesRax：一种根据基因家族树推断物种树最大似然的方法，可处理复制、转移和丢失。分子生物学与进化，39(2):msab365，2022年。[114] J Armstrong、G Hickey、M Diekhans、I T Fiddes、A M Novak、A Deran、Q Fang、D Xie、S Feng、J Stiller、D Genereux、J Johnson、V D Marinescu、J Alf¨oldi、R S Harris、K Lindblad-Toh、D Haussler、E Karlsson、E D Jarvis、G Zhang 和 B Paten。Progressive Cactus 是一款面向千基因组时代的多基因组比对器。《自然》，587:246–251，2020年。