AsymmeTree

User Manual

David Schaller

This manual is work in progress!

Contents

Contents 1

1 Introduction 2

2 Manual 3
- 2.1 Installation . . . 3
  - 2.1.1 Easy Installation with Pip . . . 3
  - 2.1.2 Dependencies . . . 3
- 2.2 Overview . . . 4
  - 2.2.1 Tree Data Structures . . . 4
  - 2.2.2 Simulator for Species and Gene Trees . . . 4
  - 2.2.3 Best Match Inference . . . 4
  - 2.2.4 Supertree Computation . . . 4
  - 2.2.5 Cograph Editing and ParaPhylo . . . 5

3 Detailed Description 6
- 3.1 Mathematical Preliminaries . . . 6
  - 3.1.1 Graph and Tree Notation . . . 6
  - 3.1.2 Phylogenetic Trees . . . 7
  - 3.1.3 Metrics and Ultrametrics . . . 11
  - 3.1.4 Homology and Best Matches . . . 14

Bibliography 18

Notation 21

Chapter 1

Introduction

AsymmeTree is Python library for the simulation and analysis of phylogenetic scenarios. It includes a simulator for species and gene tree scenarios with asymmetric evolution rates, tools for the inference and analysis of best matches (resp. best hits) and orthology, as well as an algorithm to compute supertrees.

Chapter 2

Manual

2.1 Installation

2.1.1 Easy Installation with Pip

The `asymmetree` package is available on PyPI:

pip install asymmetree

For details about how to install Python packages see here.

2.1.2 Dependencies

AssymmeTree has several dependencies (which are installed automatically when using `pip`):

- NetworkX
- SciPy and NumPy
- Matplotlib

Furthermore, to use functions involving sequence simulation and alignment, the following packages must be installed (i.e., they are not installed automatically!):

- Biopython
- Pyvolve

To use the tree reconstruction method for best match inference and the C++ implementation of the quartet method, resp., the following software must be installed (I recommend that you compile these tools on your machine, place the binaries into a persistent location and add this location to your PATH environment variable):

- RapidNJ
- qinfer

2.2 Overview

2.2.1 Tree Data Structures

The two classes `Tree` (in `asymmetree.tools.Tree`) and `PhyloTree` (in `asymmetree.tools.PhyloTre` inherits from `Tree`) implement tree data structures which are essential for most of the modules in the package. The latter contains converters and parsers for the Newick format and a NetworkX graph format.

2.2.2 Simulator for Species and Gene Trees

The following steps are implemented in the Python package `asymmetree.simulator`:

- species tree simulation ('innovation model')
- gene tree simulation (Gillespie algorithm)
- gene tree imbalancing (asymmetric evolution rates of paralogous genes)
- computation of a (noisy) distance matrix from the gene tree

2.2.3 Best Match Inference

Inference of the best match relation either directly from the gene tree or from a distance matrix (several methods).

- `asymmetree.best_matches`

References:

- Geiß et al. (2019)
- Geiß et al. (2020b)
- Geiß et al. (2020a)
- Stadler et al. (2020)

2.2.4 Supertree Computation

Implementation of the BuildST algorithm described by Deng and Fernández-Baca (2016) to compute a supertree from a given list of tree based on the leaf labels. The algorithm uses the dynamic graph data structure described by Henzinger and King (1995) and Holm et al. (2001).

- `asymmetree.tools.BuildST`
- `asymmetree.tools.hdtgraph.DynamicGraph`

2.2.5 Cograph Editing and ParaPhylo

The subpackages `asymmetree.cograph` and `asymmetree.proteinortho` contain heuristics for cograph editing and a method to compute rooted species tree from orthology/paralogy relations. The latter is a reimplementation of ParaPhylo (Hellmuth et al. 2015) which uses heuristics for the NP-hard steps instead of exact ILP solutions.

Chapter 3

Detailed Description

3.1 Mathematical Preliminaries

3.1.1 Graph and Tree Notation

In this section, a very short introduction into the field of graph theory is given with a focus on the concepts that will be relevant in the subsequent sections. For standard textbooks see e.g. Diestel (2017) and Semple and Steel (2003) from which definitions in the following were taken. The notation concerning trees follows the one used by Geiß et al. (2019).

Definition 1 ((Di-)Graph)**.** *A* graph *is a pair $G = (V, E)$ of sets such that $E \subseteq [V]^2$, i.e., the elements of E are 2-element subsets of V. The elements in V are called vertices or nodes and the elements in E are called edges. Here, an (undirected) edge e between vertices x and y is written as xy.*

A directed graph *(or* digraph*) is s graph that assigns to every edge an initial and terminal vertex. In this case, an edge e is written as (x, y) if x is the initial, and y is the terminal vertex of edge e.*

As usual, it is often written $V(G) := V$ and $E(G) := E$ for the sets of vertices and edges of a graph $G = (V, E)$ in order to express affiliation. Two vertices x and y are *adjacent* if $xy \in E(G)$ [or $(x, y) \in E(G)$, respectively]. The *degree* of a vertex x is the number of adjacent vertices and denoted by $\mathsf{deg}(x)$.

The notation $G[x_1, ..., x_k]$ refers to an induced subgraph G' of a graph G on the set $V(G') = \{x_1, ..., x_k\} \subseteq V(G)$. Induced subgraphs are defined in the usual sense, i.e., $xy \in E(G')$ [or $(x, y) \in E(G')$ if G is directed] if and only if $xy \in E(G)$ [$(x, y) \in E(G)$] and $x, y \in V(G')$.

An undirected graph $G = (V, E)$ is called *connected* if any two vertices $x, y \in V(G)$ are linked by a path in G, i.e., there is a path $x = x_1 - x_2 - ... - x_k = y$ such that $x_i x_{i+1} \in E(G)$ for all $1 \leq i < k$. A maximal connected subgraph of G is called a *connected component* of G. In case of a directed graph $\vec{G}$, a subgraph $\vec{G}^*$ of $\vec{G}$ is called a *strongly connected component* if for any two vertices x and y there is directed path $x = x_1 \to x_2 \to ... \to x_k = y$ such that $(x_i, x_{i+1}) \in E(\vec{G})$ for all $1 \leq i < k$ and vice versa. Furthermore, $\vec{G}^*$ must be maximal in that sense.

Given a non-empty set of colors $\mathcal{C}$, a *proper vertex coloring* of a (di-)graph G is a map $\sigma\colon V(T) \to \mathcal{C}$ such that $xy \in E(G) \implies \sigma(x) \neq \sigma(y)$. In other words, two adjacent vertices must not have the same color.

Definition 2 (Rooted tree)**.** *A graph $T = (V, E)$ is a* tree *if it is connected and acyclic. A* rooted tree *is a tree T with a special node ρ called the* root*. As a consequence of this, a partial order $\preceq_T$ on the vertex set $V(T)$ can be defined where $v \preceq_T u$ if u lies on the (unique) path from ρ to v.*

A vertex u is called an *ancestor* of v in T if $v \preceq_T u$, whereas v is a descendant of u in this case. If furthermore $uv \in T$, then u is the parent of v, and v is a child of u. The set of all children of a vertex u is denoted by $\mathsf{child}(u)$. Likewise, $\mathsf{par}(v)$ refers to the unique parent of a vertex v. It is set $\mathsf{par}(v) = \emptyset$ if v has no parent, i.e., v is the root.

The set of all outer vertices $v \in V(T)$ (i.e., for which $\deg(v) = 1$) is called the *leaf set* of T and denoted by $L(T)$. Hence, the leafs $L(T)$ are the minima and ρ is the unique maximum w.r.t. $\preceq_T$. The set of inner vertices $V(T) \setminus L(T)$ is denoted by $V^0(T)$, and a subtree of T which is rooted at a vertex $u \in V^0(T)$ and contains all nodes and edges below u by $T(u)$. Likewise, $L(T(u))$ is the set of leaves under u.

A *leaf coloring* is a surjective map $\sigma\colon L(T) \to \mathcal{C}$ that assigns a color to the set of leaves of a tree. Given such a map, $L[r]$ refers to the set of leaves with color $r \in \mathcal{C}$; more formally $L[r] \coloneqq \{x \in L(T) \mid \sigma(x) = r\}$.

The last common ancestor $\operatorname{lca}(A)$ of a subset $A \subseteq V(T)$ is the smallest vertex $v \in V(T)$ (w.r.t. $\preceq_T$) such that $x \preceq_T v$ for all $x \in A$. For easier notation, let $\operatorname{lca}(x, y) \coloneqq \operatorname{lca}(\{x, y\})$.

A tree T *displays* another tree T' if T' can be obtained from a subtree of T by contraction of edges (in the usual sense). If not stated otherwise, the displayed tree for a restricted set of leaves L' of T means the tree that is obtained by subsequently removing all $v \in V(T) \setminus L'$ with $|\mathsf{child}(v)| = 0$ (including their incident edges) and contracting vertices $v \in V(T)$ with $|\mathsf{child}(v)| = 1$ and their two incident edges into a single edge. *Rooted triples* are a special kind of displayed trees. A rooted triple $xy|z$ is a rooted tree on three leaves x, y, z such that the path from x to y and the path from z to the root do not intersect. A rooted triple is called consistent with a tree T if $x, y, z \in L(T)$ and $\operatorname{lca}(x, y) \prec_T \operatorname{lca}(x, z) = \operatorname{lca}(y, z)$.

The term rooted triple is closely related to the definition of outgroups:

Definition 3 (Outgroup)**.** *A leaf $z \in L(T)$ is called an* outgroup *w.r.t. a set $X \subset L(T)$ if $\operatorname{lca}(X) \prec \operatorname{lca}(X \cup \{z\})$.*

In particular, T displays the rooted triple $x'x''|z$ for all distinct $x', x'' \in X$ and $z \in L(T) \setminus L(T(\operatorname{lca}(X)))$.

Finally, a tree T (and a graph in general) can be endowed with a map that assigns weights to its edges. In the case of positive weights, these can be interpreted as the lengths of the edges and represented by a weight function $\ell\colon E(T) \to \mathbb{R}^+$.

3.1.2 Phylogenetic Trees

This section describes the properties and the associated maps of phylogenetic trees, i.e., graph models of evolutionary histories of either species or genes.

Definition 4 (Phylogenetic tree (cf. Hernandez-Rosales et al. 2012; Geiß et al. 2020a))**.** *An unrooted tree $\overline{T}$ is a* phylogenetic tree *if every inner vertex $v \in V^0(\overline{T})$ has a degree of at least 3. A rooted tree T is a* phylogenetic tree *if every $v \in V^0(T)$*

has at least 2 children. $\overline{T}$ *and* T *are called* fully-resolved *if the respective equalities hold, i.e.,* $\mathsf{deg}(v) = 3$ *for every* $v \in V^0(\overline{T})$ *or* $|\mathsf{child}(v)| = 2$ *for every* $v \in V^0(T)$*, respectively.*
A planted *phylogenetic tree* T *is a rooted phylogenetic tree with a special vertex* 0_T*, called the* planted root*, that has a single child* ρ_T *such that* $T(\rho_T)$ *is a phylogenetic tree.*

The purpose of this definition is to avoid vertices of degree 2 (possibly with exception of the root). Such vertices would lack a justification by an evolutionary event that could be reconstructed from observable data with suitable methods. In contrast, a vertex that is not fully-resolved can be interpreted as missing information about the exact local topology in most cases. The planted edge $0_T\rho_T$ is useful in the wake of modeling events that predate the first branching event. This becomes especially relevant for the reconciliation of gene trees with the underlying history of the corresponding species, since, e.g., duplication events can occur before the first speciation.

Exact methods for phylogenetic reconstruction cannot per se determine the location of the root. Hence, it is often necessary to consider unrooted trees. An unrooted version $\overline{T}$ of a rooted tree T with distance function ℓ can be obtained by the following two operations (Stadler et al. 2020):

(i) Omit the planted root 0_T and its incident edge.

(ii) In case the root ρ_T has exactly two children u_1 and u_2, replace the path $u_1 - \rho_T - u_2$ by a single edge u_1u_2 with length $\ell(u_1u_2) := \ell(\rho_T u_1) + \ell(\rho_T u_2)$.

The weight function ℓ is the same for all other edges. However, note that the ancestor order $\preceq_T$ must be dropped.

As already mentioned, phylogenetic trees can either be species trees, that represent the relationship and branching history of different taxa, or gene trees, which constitute the history of a gene family. In the latter case, duplication and HGT events cause additional branching, whereas losses terminate existing branches. Since all members of a gene family reside in some species, there exists an embedding of a gene tree T into the corresponding species tree S.

In case of a gene tree T on a set of extant genes $L(T)$, the knowledge about which genes reside in which species is represented by a leaf coloring $\sigma\colon L(T) \to L(S)$. Therefore, the color and species of a leaf $v \in L(T)$ will be used as synonymous terms. The rest of the embedding can be formalized by a reconciliation map:

Definition 5 (Reconciliation Map (cf. Geiß et al. 2020a; Stadler et al. 2020))**.** *Let* $S = (W, F)$ *and* $T = (V, E)$ *be two planted phylogenetic trees and let* $\sigma\colon L(T) \to L(S)$ *be a surjective map. A reconciliation from* (T, σ) *to* S *is a map* $\mu\colon V \to W \cup F$ *satisfying*

(R0) Root Constraint. $\mu(x) = 0_S$ *if and only if* $x = 0_T$.

(R1) Leaf Constraint. *If* $x \in L(T)$*, then* $\mu(x) = \sigma(x)$.

(R2) Ancestor Preservation. *If* $x \prec_T y$*, then* $\mu(x) \preceq_S \mu(y)$.

(R3) Speciation Constraints. *Suppose* $\mu(x) \in W^0$.

(i) $\mu(x) = \mathrm{lca}_S(\mu(v'), \mu(v''))$ *for at least two distinct children* v', v'' *of* x *in* T*.*

(ii) $\mu(v')$ *and* $\mu(v'')$ *are incomparable in* S *for any two distinct children* v' *and* v'' *of* x *in* T*.*

(R4) Speciation Constraint II. *If* $\mu(\mathrm{lca}_T(x, y)) = \mu(\mathrm{lca}_T(x, z)) \in V^0(S)$*, then* $\mathrm{lca}_S(\sigma(x), \sigma(y)) = \mathrm{lca}_S(\sigma(x), \sigma(z))$ *for all distinct leaves* $x, y, z \in L(T)$*.*

Axioms (R0) to (R4) hold for gene family histories that do not include horizontal gene transfer events. The first two axioms ensure that the root and leaves of T also map to the root and leaves in the species tree S, respectively. Note that the *Leaf Constraint* thereby implicitly prohibits loss leaves since they could not be mapped to an extant species. Thus, both σ and μ would be undefined for such kind of leaves. The *Ancestor Preservation* constraint forbids that a descendant x of some vertex y in T is mapped above $\mu(y)$ in S which avoids time traveling of a gene into an ancestor species. The two constraints in (R3) refer to restrictions for (observable) speciation events in the gene tree. In particular, (R3 ii) states that any two children v' and v'' of such a vertex are not only incomparable in T but also in S, i.e., $\mu(v')$ and $\mu(v'')$ are incomparable w.r.t. $\preceq_S$. This corresponds to a separated evolution of the gene family in the descending branches of a speciation event. Finally, axiom (R4) has been introduced only recently by Stadler et al. (2020). It forbids to map two vertices $v_1, v_2 \in V^0(T)$ that represent distinct speciation events to the same vertex in S. This avoids ambiguous interpretations of single vertices as multiple events.

In case of the occurrence of HGT events, the axiom system has to be modified (Geiß et al. 2020a). First, a weaker version of axiom (R2) is satisfied:

(R2*) Weak Ancestor Preservation. *If* $x \prec_T y$*, then either* $\mu(x) \preceq_S \mu(y)$ *or* $\mu(x)$ *and* $\mu(y)$ *are incomparable w.r.t.* $\preceq_S$*.*

Moreover, two additional axioms are necessary:

(R3) *(iii)* (Addition to the Speciation Constraints.) *Suppose* $\mu(x) \in W^0$*. If* $\mu(x) \in W^0$*, then* $\mu(v) \preceq_S \mu(x)$ *for all* $v \in \mathsf{child}(x)$*.*

(R5) HGT Constraint. *If* x *has a child* y *such that* $\mu(x)$ *and* $\mu(y)$ *are incomparable, then* x *also has a child* y' *with* $\mu(y') \preceq_S \mu(x)$*.*

This extended axiom system is a valid generalization in most cases. However, it fails in some scenarios, e.g., (R3 ii) may not hold if an HGT event has no surviving members in the non-transferred branch (cf. Geiß et al. 2020a, Section 7.3).

An event labeling that is based on the reconciliation map is defined as follows:

Definition 6 (Event Labeling (Geiß et al. 2020a, Def. 3 with HGT))**.** *Given a reconciliation map* μ *from* (T, σ) *to* S*, the* event labeling on T (determined by μ) *is the map* $t\colon V(T) \to \{\circledcirc, \odot, \bullet, \square, \triangle\}$ *given by:*

$$t(u) = \begin{cases} \circledcirc & \textit{if } u = 0_T,\ \textit{i.e., } \mu(u) = 0_S \textit{ (root)} \\ \odot & \textit{if } u \in L(T),\ \textit{i.e., } \mu(u) \in L(S) \textit{ (leaf)} \\ \bullet & \textit{if } \mu(u) \in V^0(S) \textit{ (speciation)} \\ \square & \textit{if } \mu(u) \in E(S)\textit{, } u \textit{ and } v \textit{ are comparable for all } v \in \mathsf{child}(u) \textit{ (dupl.)} \\ \triangle & \textit{if } u \textit{ has a child } v \textit{ such that } \mu(u) \textit{ and } \mu(v) \textit{ are incomparable (HGT)} \end{cases}$$

Note that Definition 6 does not capture loss events. As they have shown to be an important cause of problems for inference methods, trees that additionally contain all branches leading to loss events will often be considered. Such complete gene family histories will be referred to as *extended gene trees*. In the examples, as well as in the simulated trees, loss events will be indicated by an asterisk ($*$), whilst the symbols of the event labeling are used for all other event types. In this context, phylogenetic trees in the usual sense correspond to the observable part of the complete gene trees. Observable gene trees are easily constructed from the latter:

(i) Remove all branches leading to loss events only.

(ii) Subsequently, contract all nodes of degree 2 (except the root) and the adjacent edges into a single edge.

In ambiguous cases, the observable gene tree will sometimes be denoted by T_{obs}. An example of an (extended) gene tree T embedding is shown in Figure 3.1. It contains the different event types including losses.

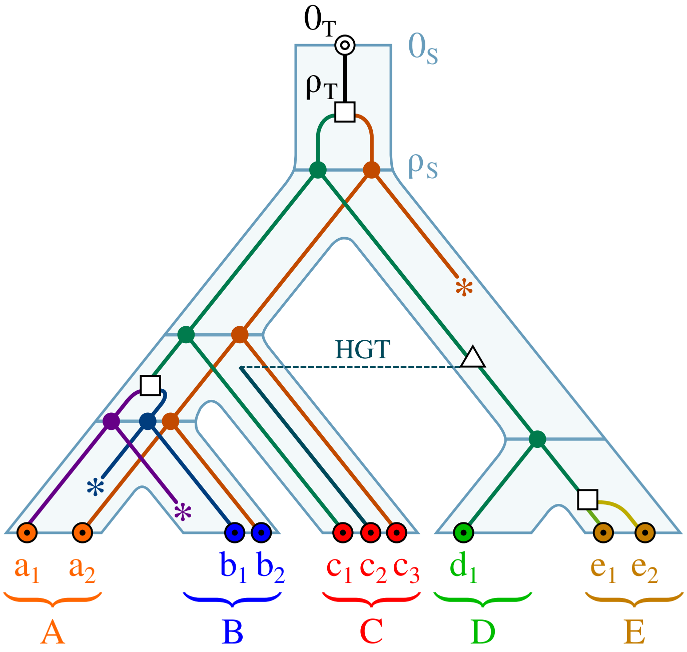

Figure 3.1: Embedding of an extended gene tree into the corresponding species tree. The species tree S on the set of extant species $\{A, B, C, D, E\}$ is shown in light blue. The branches of the gene tree T with planted root 0_T ($\circledcirc$) have different colors that change after duplication ($\square$) and HGT ($\triangle$) events to indicate their lineage. Moreover, the gene family history includes speciation ($\bullet$) and loss ($*$) events. The colors of the non-loss leaves ($\odot$) constitute known information about the extant genes, i.e., the species in which they reside.

Note that the root ρ_T is not a speciation event but a duplication in the example. Since it will be useful later, the following special type of duplication events is introduced:

Definition 7. *A duplication event $v \in V^0(T)$ is called* ancient *if v is mapped to the edge $0_S\rho_S$ under the reconciliation map μ.*

It will sometimes be necessary to consider duplication events that predate the last common ancestor of a subset $L' \subseteq L(T)$. Ancient duplications w.r.t. $u = \mathrm{lca}_T(L')$ are defined by applying Definition 7 to the subtree $T(u)$ where the unique path between 0_T and u is contracted into a planted edge for $T(u)$.

Since all vertices in phylogenetic trees represent evolutionary events, it is useful to have a dating function τ that assigns a time point to every vertex. Following the conventions e.g. used by Böcker and Dress (1998), these time points will be normalized such that $\tau(0_S) = 1$ and $\tau(x) = 0$ for all $x \in L(S)$. In case the tree is not planted, $\tau(\rho_T)$ is set to 1. Hence, the dating function maps to the unit interval: $\tau\colon V(T) \to [0, 1]$.

3.1.3 Metrics and Ultrametrics

Both the dating function $\tau\colon V(T) \to [0, 1]$ and the weight function $\ell\colon E(T) \to \mathbb{R}^+$ can be used to define a distance function on the set of vertices of a tree. Since the dating function assigns time points to the vertices of T, the time difference between two vertices x and y (that are comparable in T) can be interpreted as the time of divergence that lies between them. In general, the *divergence time* d_τ between two arbitrary nodes $x, y \in V(T)$ is given by

$$d_\tau(x, y) = \sum_{uv \in P} |\tau(u) - \tau(v)| \tag{3.1}$$

where P is the unique path between x and y. In contrast, the weighting function ℓ can be used to model different aspects of the evolutionary history. In particular, the weights can represent the dissimilarity or evolutionary distance of adjacent vertices in a gene tree T. Thus, the distance that is given by the sum of all edge weights on the unique path P between two vertices $x, y \in V(T)$

$$d(x, y) = \sum_{e \in P} \ell(e) \tag{3.2}$$

corresponds to the *evolutionary distance* between x and y.

The set of vertices $V(T)$ together with one of the two distance functions d_τ or d forms a *metric space.* A metric is defined as follows:

Definition 8 (Metric & Ultrametric)**.** *A* metric *(also called* distance function*) on a set X is a map $d\colon X \times X \to \mathbb{R}_0^+$ such that for all $x, y, z \in X$, the following conditions are satisfied:*

(i) Non-negativity: $d(x, y) \geq 0$.

(ii) Identity of indiscernibles: $d(x, y) = 0 \iff x = y$.

(iii) Symmetry: $d(x, y) = d(y, x)$.

(iv) Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

An ultrametric *is a metric that satisfies a stronger version of the triangle inequality:*

(iv)* $d(x,z) \leq \max(d(x,y), d(y,z))$.

Since they represent the extant members of a species or gene family, the distances between the leaves of a tree are of special relevance. In particular, evolutionary distances between extant genes often constitute the available information as discussed in later sections.

For a normalized dating function $\tau : V(T) \to [0,1]$ and two extant genes $x, y \in L(T)$ the divergence time simplifies to the map

$$d_\tau(x,y)\colon L(T) \times L(T) \to \mathbb{R}_0^+ \colon \quad (x,y) \mapsto 2\tau(\mathrm{lca}_T(x,y)) \tag{3.3}$$

which has the well-known property of representing a 1-to-1 correspondence between dated, rooted trees and ultrametrics (cf. Gordon 1987; Böcker and Dress 1998). An *ultrametric tree*, i.e., a rooted tree with an arbitrary distance function having the property of ultrametricity, is most naturally shown visually by positioning all its leaves on the same level such that they have the same distance to the root (see Figure 3.2, left).

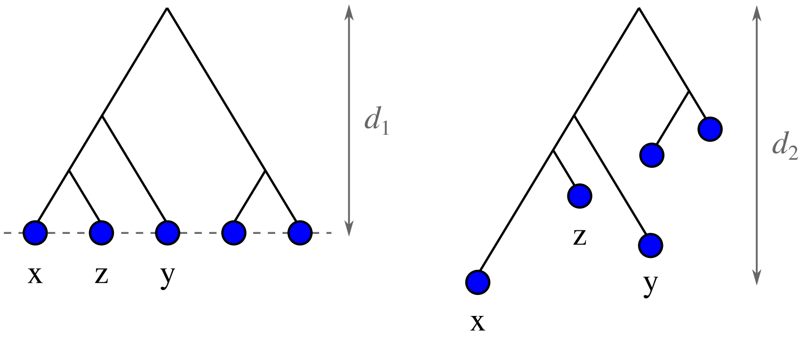

Figure 3.2: Ultrametric and non-ultrametric tree. The vertical components of the edges represent the distances. In the left tree, they induce an ultrametric on the set of its leaves, whereas they do not in the tree on the right side. This can, e.g., be seen by the violation of condition *(iv*)* in Definition 8 for the leaves x, y, z: $d_2(x,y) > \max(d_2(x,z), d_2(z,y))$.

In contrast, the distance function d on the set of leaves $L(T)$ is not an ultrametric on the set of leaves in general. This is especially the case for evolutionary distances in the presence of asymmetric divergence as it will be discussed later. However, the distance function d satisfies a weaker property by construction: A metric d on set X is *additive* if a (not necessarily rooted) weighted tree (T, ℓ) exists such that $L(T) = X$ and $d = d_T$ where d_T is the distance function of (T, ℓ) as defined by Equation 3.2. A metric on a set X can be tested for additivity with the four-point-condition: For any *quartet*, i.e., a set of four nodes $x, y, u, v \in X$, it must hold that out of the three distance sums

(i) $d(x,y) + d(u,v)$,

(ii) $d(x,u) + d(y,v)$,

(iii) $d(x,v) + d(y,u)$

two sums are equal and not smaller than the third (cf. Simões Pereira 1969; Buneman 1974). The concept of quartets will be essential throughout this work:

Definition 9 (Quartet Relation (cf. Stadler et al. 2020))**.** *Consider an unrooted tree $\overline{T}$ with leaf set $L(\overline{T})$. For any four distinct leaves $x, y, u, v \in L(\overline{T})$ denote by $\overline{T}[x, y, u, v]$ the unrooted tree obtained by suppressing all vertices of degree 2 in the union of the paths in $\overline{T}$ that connect x, y, u, v. The* quartet relation *for $\overline{T}$ is*

(i) $\overline{T}[x, y, u, v] = (xy|uv)$*,*

(ii) $\overline{T}[x, y, u, v] = (xu|yv)$ *or*

(iii) $\overline{T}[x, y, u, v] = (xv|yu)$

if there is an edge $e \in E(\overline{T})$ such that the respective pairs (that are separated by the vertical bar) are in different connected components of $\overline{T}$ after the removal of e. If there is no such edge write $\overline{T}[x, y, u, v] = \times$.

The four-point-condition is directly related to this notion of quartets: For an additive metric d, the smallest of the three distance sums induced by four distinct points x, y, u and v determines their topology in a corresponding tree $\overline{T}$ that explains d. The reason for this is that the edge separating two pairs in $\overline{T}[x, y, u, v]$ only contributes to the two larger sums. The case in which all three sums are equal corresponds to the absence of such a separating edge in $\overline{T}$ (and hence also in $\overline{T}[x, y, u, v]$) and is referred to as *star tree* or *star topology*. The possible cases are visualized in Figure 3.3.

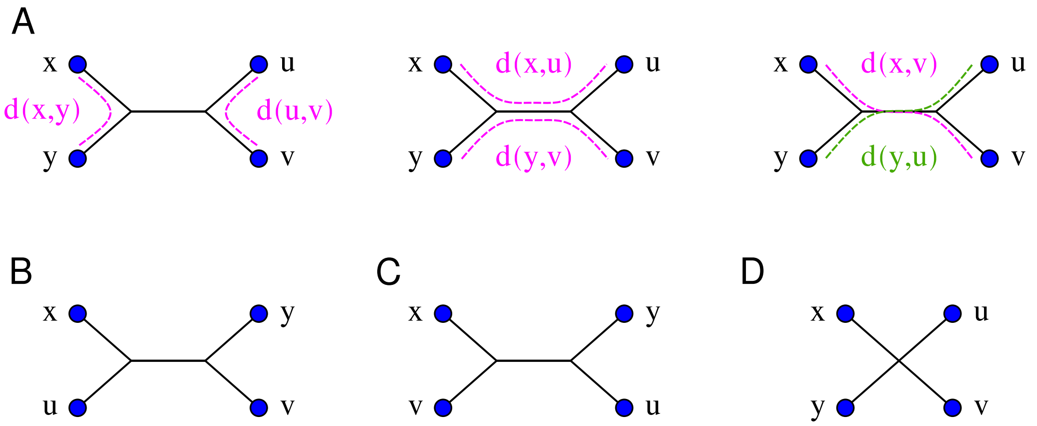

Figure 3.3: The four possible (unrooted) quartets. For case (A), the six distances that form the three distance sums are indicated. In formal, the cases are:

(A) $\overline{T}[x, y, u, v] = (xy|uv)$
$\iff d(x, y) + d(u, v) < d(x, u) + d(y, v) = d(x, v) + d(y, u)$

(B) $\overline{T}[x, y, u, v] = (xu|yv)$
$\iff d(x, u) + d(y, v) < d(x, y) + d(u, v) = d(x, v) + d(y, u)$

(C) $\overline{T}[x, y, u, v] = (xv|yu)$
$\iff d(x, v) + d(y, u) < d(x, y) + d(u, v) = d(x, u) + d(y, v)$

(D) $\overline{T}[x, y, u, v] = \times$
$\iff d(x, y) + d(u, v) = d(x, u) + d(y, v) = d(x, v) + d(y, u)$

Given a rooted or unrooted tree T and a unique numbering of the leaves $L(T)$, the distance function on the leaves can be represented by a symmetric square matrix which will be denoted by $\mathbf{D}$ in the following, where $\mathbf{D}(x, y)$ is the entry in the row and column corresponding to x and y, respectively.

3.1.4 Homology and Best Matches

As already outlined, the type of relationship between pairs of genes is of interest for the inference of gene functions. In particular, orthologous genes are considered to perform similar functions. Mathematically, both orthology and paralogy are binary relations on the set of (extant) genes in a gene tree T. They are defined w.r.t. the event type of the last common ancestor of two genes x and y:

Definition 10 (Orthology and Paralogy (Fitch 2000))**.** *Let (T, μ, t) be an event-labeled, rooted gene tree with reconciliation map μ. Two distinct leaves $x, y \in L(T)$ are* orthologs w.r.t. μ *if* $t(\mathrm{lca}_T(x, y)) = \bullet$*, and* paralogs *if* $t(\mathrm{lca}_T(x, y)) = \square$.

Of course, the true reconciliation map and event-labeling are not known for real-life data. Therefore, methods exist that aim to explicitly reconstruct the gene tree and the reconciliation map, but also others that attempt to directly infer orthology based on gene similarity. In both cases, a well-founded mathematical investigation of the orthology and paralogy relation is helpful.

Firstly, Defintion 10 unambiguously defines two distinct members of a gene family as either orthologs or paralogs in the absence of horizontal gene transfer. Otherwise, there may be pairs of genes $x, y \in L(T)$ having a horizontal gene transfer as last common ancestor, and, thus, $t(\mathrm{lca}_T(x, y)) = \triangle$. Clearly, x and y are neither orthologs nor paralogs given the definition above. Therefore, *xenology* was introduced as a third variant of homology. There are several definitions of xenologs. Fitch (2000) calls two genes xenologs if there is at least one HGT event on the unique path in T connecting them. Thus, it is not necessary that their last common ancestor was an HGT event. As a consequence, two genes can be both xenologs and ortho-/paralogs. They are commonly termed as *xeno-orthologs* and *xeno-paralogs* in this case. An alternative definition by Hellmuth and Wieseke (2016) avoids this ambiguity by calling two $x, y \in L(T)$ *(lca-)xenologs* if and only if $t(\mathrm{lca}_T(x, y)) = \triangle$. Analogously, *(lca-)orthologs* and *(lca-)paralogs* are defined. If not declared otherwise, all three terms will refer to this definition in the following.

Both orthology and paralogy are irreflexive and symmetric binary relations, since $t(\mathrm{lca}_T(x, x)) = t(x) = \odot$ for any $x \in L(T)$ and $t(\mathrm{lca}_T(x, y)) = t(\mathrm{lca}_T(y, x))$ for any two genes $x, y \in L(T)$. Hence, they can be represented by undirected graphs. However, they are not transitive, as the example in Figure 3.4 shows (consider, e.g., the three genes a_1, b_1, c_1 in the orthology graph Θ).

Definition 11 (Orthology Graph (Geiß et al. 2020a, Def. 5 mod.))**.** *Let T be a gene tree with event labeling t. Let Θ be the undirected graph on $L(T)$ with $xy \in E(\Theta)$ if and only if* $\mathrm{lca}_T(x, y) = \bullet$*. Then Θ is called an* orthology graph *that is explained by the orthology relation of (T, t).*

Similarly, the paralogy graph $\overline{\Theta}$ is defined. In the absence of HGT events, it is simply the complement of Θ.

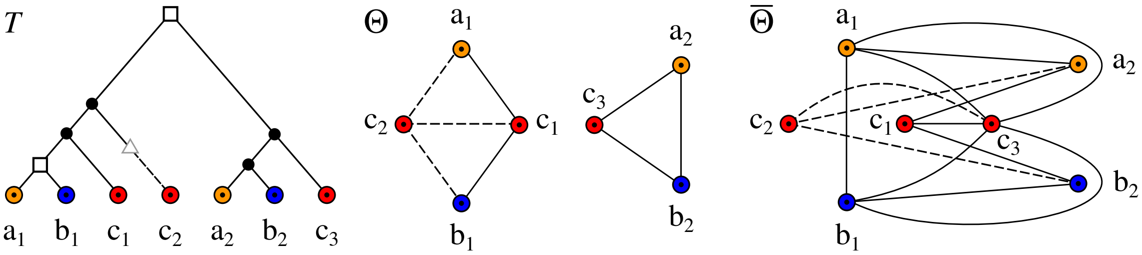

Figure 3.4: Corresponding orthology graph Θ and paralogy graph $\overline{\Theta}$ for the example gene tree in Figure 3.1 restricted to the set of species $\{A, B, C\}$ (see tree T on the left). Xeno-orthologous and -paralogous relations, respectively, are indicated by dashed lines.

Neither the orthology nor the paralogy relation can be observed directly. Therefore, heuristics have been developed that make use of the fact that orthologs are often the most closely related genes in two species. This approach has been used widely in phylogenetic reconstruction methods and requires some definition of the relationship between genes. To this end, different terms and variants have been introduced: Symmetric best matches (e.g. used by Tatusov 1997), bidirectional best hits (BBH) (e.g. Overbeek et al. 1999; Lafond et al. 2018), reciprocal best hits (RBH) (e.g. Bork et al. 1998), reciprocal best alignment heuristic (RBAH) (e.g. Lechner et al. 2011) and some others.

To understand the aim of this work, it is necessary to strictly distinguish between *best hits* and *best matches*:

Definition 12 ((Reciprocal) Best hit)**.** *Consider a gene tree T with leaf set $L(T)$, a surjective map $\sigma\colon L(T) \to L(S)$ (where $L(S)$ is the corresponding set of species) and a distance function $d : L(T) \times L(T) \to \mathbb{R}_0^+$. Then $y \in L(T)$ is a* best hit *of $x \in L(T)$ if and only if $d(x, y) \leq d(x, y')$ holds for all leaves y' from species $\sigma(y') = \sigma(y)$.*
If x is also a best hit of y, x and y are called reciprocal best hits.

Thus, *best hits* are defined in the context of the smallest evolutionary distance. On the other hand, *best matches* refer to the closest relatives of a gene w.r.t. the point in time when they were separated:

Definition 13 ((Reciprocal) Best match (Geiß et al. 2019))**.** *Consider a gene tree T with leaf set $L(T)$ and a surjective map $\sigma\colon L(T) \to L(S)$ (where $L(S)$ is the corresponding set of species). Then $y \in L(T)$ is a* best match *of $x \in L(T)$, in symbols $x \to y$, if and only if $\operatorname{lca}(x, y) \preceq \operatorname{lca}(x, y')$ holds for all leaves y' from species $\sigma(y') = \sigma(y)$.*
If x is also a best match of y in color $\sigma(x)$, i.e., $y \to x$, x and y are called reciprocal best matches.

A comprehensive mathematical theory on best matches was developed only recently by Geiß et al. (2019). Therein, the authors point out that best hits and best matches are the same if the mutation rate of the gene family is constant, i.e., the *Molecular Clock* holds.

The best match relation can be represented by a directed, vertex-colored graph on the set of leaves of a gene tree T as follows:

Definition 14 (cBMG and cRBMG (Geiß et al. 2019))**.** *Given a gene tree T and a map $\sigma : L(T) \to L(S)$, the* colored Best Match Graph (cBMG) *$\vec{G}(T,\sigma)$ has vertex set $L(T)$ and arcs $(x,y) \in E(\vec{G})$ if $x \neq y$ and $x \to y$. Each vertex $x \in L(T)$ obtains the color $\sigma(x)$.*
The rooted tree T explains *the vertex-colored graph $(\vec{G},\sigma)$ if $(\vec{G},\sigma)$ is isomorphic (in the usual sense, with preservation of colors) to the cBMG $\vec{G}(T,\sigma)$.*
The vertex-colored undirected graph $G(T,\sigma)$ that has vertex set $L(T)$ and edges $xy \in E(G)$ if $x \neq y$ and $x \to y$ as well as $y \to x$ is called the colored Reciprocal Best Match Graph (cRBMG).

Geiß et al. (2019) present two polynomial-time algorithms to decide whether a given digraph $(\vec{G},\sigma)$ is a valid colored Best Match Graph and to determine the unique least resolved tree, i.e., the corresponding tree which explains $(\vec{G},\sigma)$ and is minimal in the sense that no edge can be contracted such that the tree still explains $(\vec{G},\sigma)$. The algorithm for computing the least resolved tree that will be relevant later in this contribution makes use of informative triples:

Definition 15 (Informative triples (Geiß et al. 2019, Def. 8))**.** *Let $(\vec{G},\sigma)$ be a two-colored digraph. We say that a triple $ab|c$ is* informative *(for $(\vec{G},\sigma)$) if the three distinct vertices $a,b,c \in L$ induce a colored subgraph $\vec{G}[a,b,c]$ isomorphic (in the usual sense, with preservation of colors) to the graph X_1, X_2, X_3 or X_4 shown in Figure 3.5. The set of informative triples is denoted by $\mathcal{R}(\vec{G},\sigma)$.*

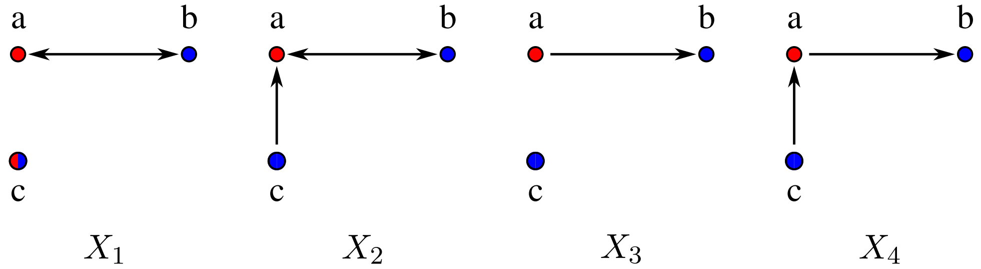

Figure 3.5: Each of the three-vertex induced subgraphs X_1, X_2, X_3 and X_4 gives a triple $ab|c$. If vertex c in the drawing has two colors, then the color $\sigma(c)$ does not matter (Geiß et al. 2019).

This definition can easily be extended for n-colored graphs. Geiß et al. (2019) shows that the well-known polynomial-time algorithm `BUILD` (Aho et al. 1981) can be used to compute the least resolved tree directly from the full set of informative triples $\mathcal{R}(\vec{G},\sigma)$ of an n-colored Best Match Graph. Moreover, a colored digraph $(\vec{G},\sigma)$ is a valid cBMG if and only if $(\vec{G},\sigma) = \vec{G}(\text{Aho}(\mathcal{R}(\vec{G},\sigma)),\sigma)$ where Aho($\mathcal{R}$) denotes the tree T that results from applying `BUILD` to a set of triples $\mathcal{R}$ (Geiß et al. 2019, cf. Theorems 6 and 9). In other words, the Best Match Graph of the resulting tree has to be equal to the original digraph. Therefore, the explicit construction of a least resolved tree can be used to decide whether a given digraph is a cBMG.

In a subsequent publication, the properties of *colored Reciprocal Best Match Graphs (cRBMG)* have been studied in detail (Geiß et al. 2020b). Two genes x and

y are reciprocal best matches if and only if both x is a best match of y and y is a best match of x. As mentioned above, the cRBMG (possibly with some corrections based on the corresponding cBMG) is a good heuristic for the true orthology relation. In fact, given a tree T with reconciliation map μ and a corresponding event-labeling t_μ, the orthology graph $\Theta(T, t_\mu)$ is a subgraph of the Reciprocal Best Match Graph $G(T, \sigma)$ (Geiß et al. 2020a, Theorem 2) in the absence of horizontal gene transfer. Thus, the edges of G cannot contain false positives w.r.t. the orthology relation.

Figure 3.6 shows the cBMG $\vec{G}$, the cRBMG G and the true orthology relation Θ that correspond to the previous example tree. The orthology graph is a subgraph of the cRBMG with exception of the xeno-orthologous relations. The simulations by Geiß et al. (2020a) show that reciprocal best matches are a useful heuristic for orthology also if HGT occurs (at moderate rates).

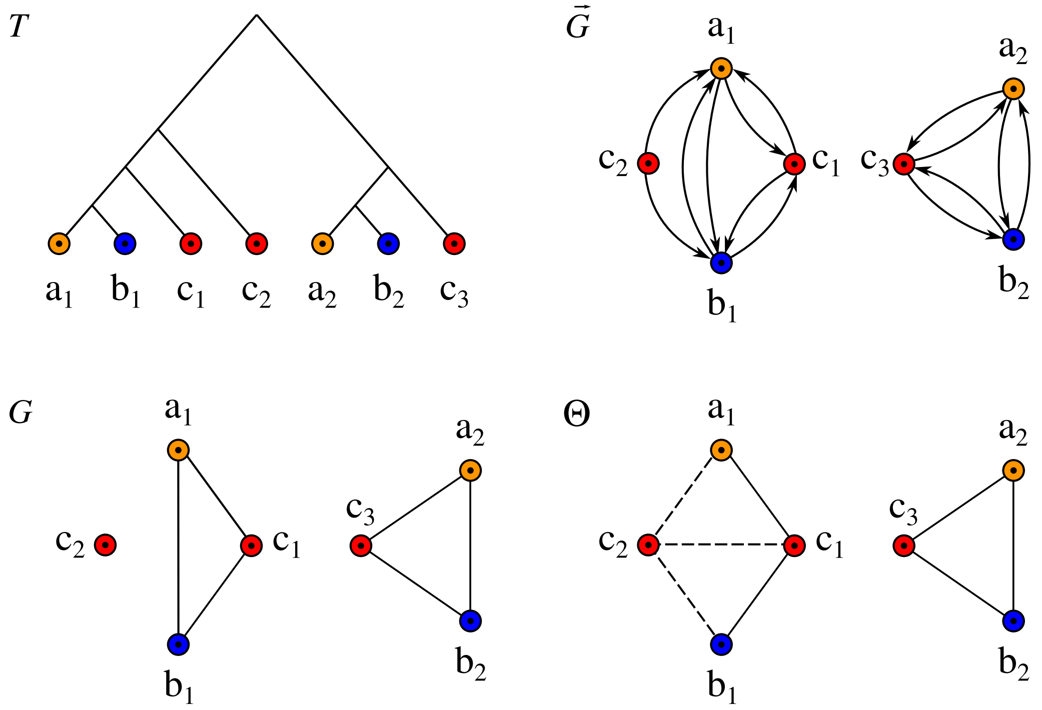

Figure 3.6: Gene tree with corresponding colored Best Match Graph $(\vec{G}, \sigma)$ and Reciprocal Best Match Graph (G, σ). The tree is a subtree of the gene tree in Figure 3.1 restricted to the set of species $\{A, B, C\}$. The true orthology graph Θ is again depicted for comparison.

However, especially in case of multiple losses, the true orthology relation is often hard to infer. The example includes the well-known problematic case of differential gene loss after a dupiclation event followed by speciation. The last common ancestor of genes a_1 and b_1 is a duplication (see Figure 3.1). The two loss events cause a false-positive edge when considering G as the estimated orthology relation.

Bibliography

Aho, A. V., Sagiv, Y., Szymanski, T. G., and Ullman, J. D. Inferring a Tree from Lowest Common Ancestors with an Application to the Optimization of Relational Expressions. *SIAM Journal on Computing*, 10(3):405–421, August 1981. ISSN 0097-5397, 1095-7111. doi: 10.1137/0210030.

Böcker, S. and Dress, A. W. Recovering Symbolically Dated, Rooted Trees from Symbolic Ultrametrics. *Advances in Mathematics*, 138(1):105–125, September 1998. ISSN 00018708. doi: 10.1006/aima.1998.1743.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. Predicting function: From genes to genomes and back 1 1Edited by P. E. Wright. *Journal of Molecular Biology*, 283(4):707–725, November 1998. ISSN 00222836. doi: 10.1006/jmbi.1998.2144.

Buneman, P. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50, August 1974. ISSN 00958956. doi: 10.1016/0095-8956(74)90047-1.

Deng, Y. and Fernández-Baca, D. Fast Compatibility Testing for Rooted Phylogenetic Trees. page 12 pages, 2016. doi: 10.4230/LIPICS.CPM.2016.12.

Diestel, R. *Graph Theory.* Springer Berlin Heidelberg, New York, NY, 2017. ISBN 978-3-662-53621-6.

Fitch, W. M. Homology. *Trends in Genetics*, 16(5):227–231, May 2000. ISSN 01689525. doi: 10.1016/S0168-9525(00)02005-9.

Geiß, M., Chávez, E., González Laffitte, M., López Sánchez, A., Stadler, B. M. R., Valdivia, D. I., Hellmuth, M., Hernández Rosales, M., and Stadler, P. F. Best match graphs. *Journal of Mathematical Biology*, 78(7):2015–2057, June 2019. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-019-01332-9.

Geiß, M., Laffitte, M. E. G., Sánchez, A. L., Valdivia, D. I., Hellmuth, M., Rosales, M. H., and Stadler, P. F. Best match graphs and reconciliation of gene trees with species trees. *Journal of Mathematical Biology*, January 2020a. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-020-01469-y.

Geiß, M., Stadler, P. F., and Hellmuth, M. Reciprocal best match graphs. *Journal of Mathematical Biology*, 80(3):865–953, February 2020b. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-019-01444-2.

Gordon, A. D. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General)*, 150(2):119, 1987. ISSN 00359238. doi: 10.2307/2981629.

Hellmuth, M. and Wieseke, N. From Sequence Data Including Orthologs, Paralogs, and Xenologs to Gene and Species Trees. In Pontarotti, P., editor, *Evolutionary Biology*, pages 373–392. Springer International Publishing, Cham, 2016. ISBN 978-3-319-41323-5 978-3-319-41324-2. doi: 10.1007/978-3-319-41324-2_21.

Hellmuth, M., Wieseke, N., Lechner, M., Lenhof, H.-P., Middendorf, M., and Stadler, P. F. Phylogenomics with paralogs. *Proceedings of the National Academy of Sciences*, 112(7):2058–2063, February 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1412770112.

Henzinger, M. R. and King, V. Randomized dynamic graph algorithms with polylogarithmic time per operation. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing - STOC '95*, pages 519–527, Las Vegas, Nevada, United States, 1995. ACM Press. ISBN 978-0-89791-718-6. doi: 10.1145/225058.225269.

Hernandez-Rosales, M., Hellmuth, M., Wieseke, N., Huber, K. T., Moulton, V., and Stadler, P. F. From event-labeled gene trees to species trees. *BMC Bioinformatics*, 13(S19), December 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-S19-S6.

Holm, J., de Lichtenberg, K., and Thorup, M. Poly-logarithmic deterministic fully-dynamic algorithms for connectivity, minimum spanning tree, 2-edge, and biconnectivity. *Journal of the ACM*, 48(4):723–760, July 2001. ISSN 00045411. doi: 10.1145/502090.502095.

Lafond, M., Meghdari Miardan, M., and Sankoff, D. Accurate prediction of orthologs in the presence of divergence after duplication. *Bioinformatics*, 34(13):i366–i375, July 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty242.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12(1), December 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-124.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, March 1999. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.96.6.2896.

Semple, C. and Steel, M. A. *Phylogenetics*. Number 24 in Oxford Lecture Series in Mathematics and Its Applications. Oxford University Press, Oxford ; New York, 2003. ISBN 978-0-19-850942-4.

Simões Pereira, J. A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, 6(3):303–310, April 1969. ISSN 00219800. doi: 10.1016/S0021-9800(69)80092-X.

Stadler, P. F., Geiß, M., Schaller, D., Sánchez, A. L., González, M. E., Valdivia, D. I., Hellmuth, M., and Rosales, M. H. From Best Hits to Best Matches. *arXiv:2001.00958 [q-bio]*, January 2020.

Tatusov, R. L. A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637, October 1997. ISSN 00368075, 10959203. doi: 10.1126/science.278.5338.631.

Notation

Graphs

$G = (V, E)$ – undirected graph
V – set of vertices
E – set of edges
$\vec{G}$ – directed graph
x, y – vertices
e, f – edges
xy – undirected edge between x and y
(x, y) – directed edge between x and y
P – path
$\mathsf{deg}(x)$ – degree of vertex x
$\mathcal{C}$ – set of connected components

(Phylogenetic) Trees

$T = (V, E)$ – rooted (gene) tree
T_{obs} – observable gene tree
$\overline{T}$ – unrooted (gene) tree
S – species tree
$0_T, 0_S$ – planted root
ρ_T, ρ_S – root (first branching event)
u, v, w – vertices
l, x, y, z – genes / leaves
r, s, s_1, s_2 – species
$\mathsf{par}(v)$ – parent of v
$\mathsf{child}(v)$ – set of children of v
$V^0(T)$ – inner vertices
$L(T)$ – leaves
$L[s]$ – leaves of color/species s
$T(v)$ – subtree rooted at v
$\preceq_T$ – ancestor relation
$\mathrm{lca}_T(A)$ – last common ancestor of set A
$xy|z$ – rooted triple
$(xy|uv)$ – quartet relation
$\sigma(x)$ – leaf coloring map
$\mu(v)$ – reconciliation map
$\tau(v)$ – dating function
$d_\tau(x, y)$ – divergence time
$\ell(e)$ – length of edge e
$d(x, y)$ – (evolutionary) distance

Event labeling

$t(v)$ – event labeling map
$\circledcirc$ – (planted) root
$\odot$ – leaf
● – speciation
$\square$ – duplication
$\triangle$ – horizontal gene transfer (HGT)
$*$ – loss

Orthology and Best Matches

Θ – (true) orthology graph
$\overline{\Theta}$ – (true) paralogy graph
$\vec{G}$ – colored Best Match Graph (cBMG)
G – colored Reciprocal Best Match Graph (cRBMG)
$x \to y$ – best match relation
$\langle a_1 b c a_2 \rangle$ – induced P_4

Inference methods

$\mathbf{D}$ – distance matrix
ϵ – relative tolerance threshold
η – discarding threshold for outgroup species
$\mathcal{R}$ – triple set
$\mathcal{L}$ – leaf set
$\mathcal{H}$ – auxiliary graph (`BUILD`)
L_e – leaf set (extant genes)
L_0 – leaf set (losses)
Z_v – set of outgroups for vertex $v \in V(S)$

Simulation

d, l, h – event rates
$g \in G(\tau)$ – extant gene at time τ
$q \in Q$ – event type
$\xi = (g, q)$ – 'reaction' in Gillespie algorithm
$r_\xi(\tau)$ – rate of reaction ξ
$R(\tau)$ – total rate
$\Delta\tau$ – waiting time
$\mathcal{P}$ – priority queue
K_A – non-synonymous substitution rate
K_S – synonymous substitution rate
ω – measure for selection pressure
R, R' – asymmetry measures
ϱ – evolution rate
k, θ – shape and scale (Gamma distribution)
$\mathfrak{L}_e$ – list of assigned rates to edge e
w_C, w_N, w_S – weights for duplication types
σ – standard deviation (normal distribution)
α – contribution of the disturbance matrix in a convex combination
$\mathbf{D}'$ – disturbed distance matrix
$\mathfrak{S}$ – standard simulated data set