# Tree reconciliations: beyond the LCA mapping

Taoyang Wu
Joint work with Zhang Louxin

Taoyang.Wu@gmail.com
Department of Mathematics,
National University of Singapore

The Isaac Newton Institute for Mathematical Sciences
June 2011

# Phylogenetic tree

A phylogenetic tree (in this talk) is a tree $T = (V, E)$ that is

- rooted: $\rho(T)$
- binary: each node has either two children (internal node) or no child (leaf node)
- leave-labelled: each leaf node is labelled by an element from a finite set $L(T)$

# Gene tree and species tree

A species tree $S$ is a phylogenetic tree such that

- its leaves are uniquely labeled
- a leaf represents a species, the label of it.

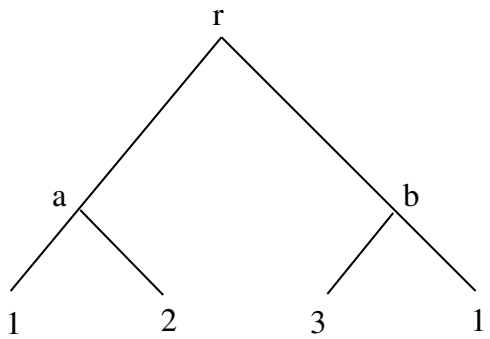
# Gene tree and species tree

A species tree $S$ is a phylogenetic tree such that

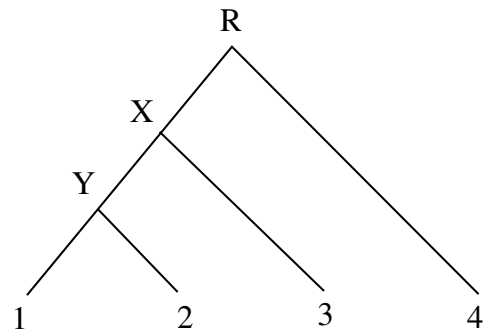
- its leaves are uniquely labeled
- a leaf represents a species, the label of it.

A gene tree $G$ is a phylogenetic tree such that

- MUL-trees, i.e., its leaves are not necessarily uniquely labeled;
- a leaf typically represents a gene found in a species, and is labelled by this species.



Gene tree G

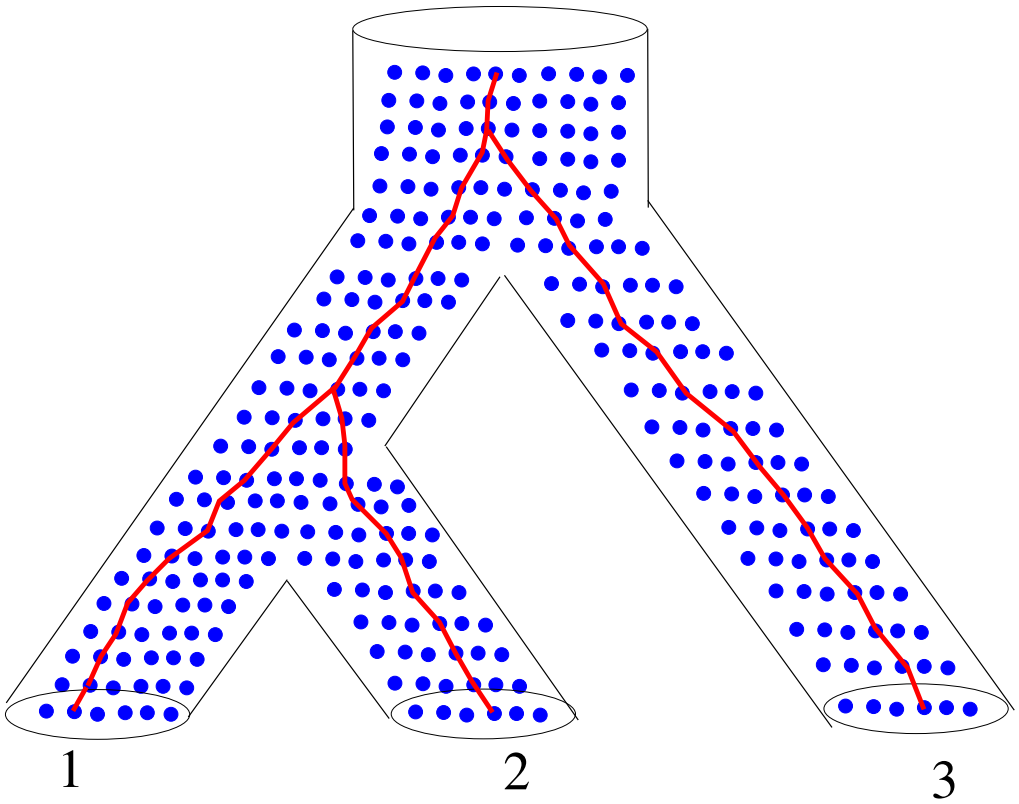


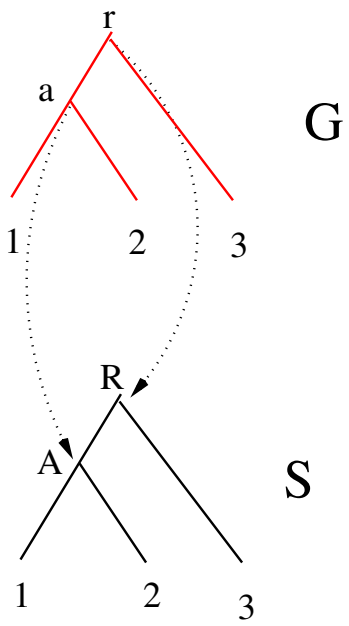
Species tree S

# Tree reconciliation

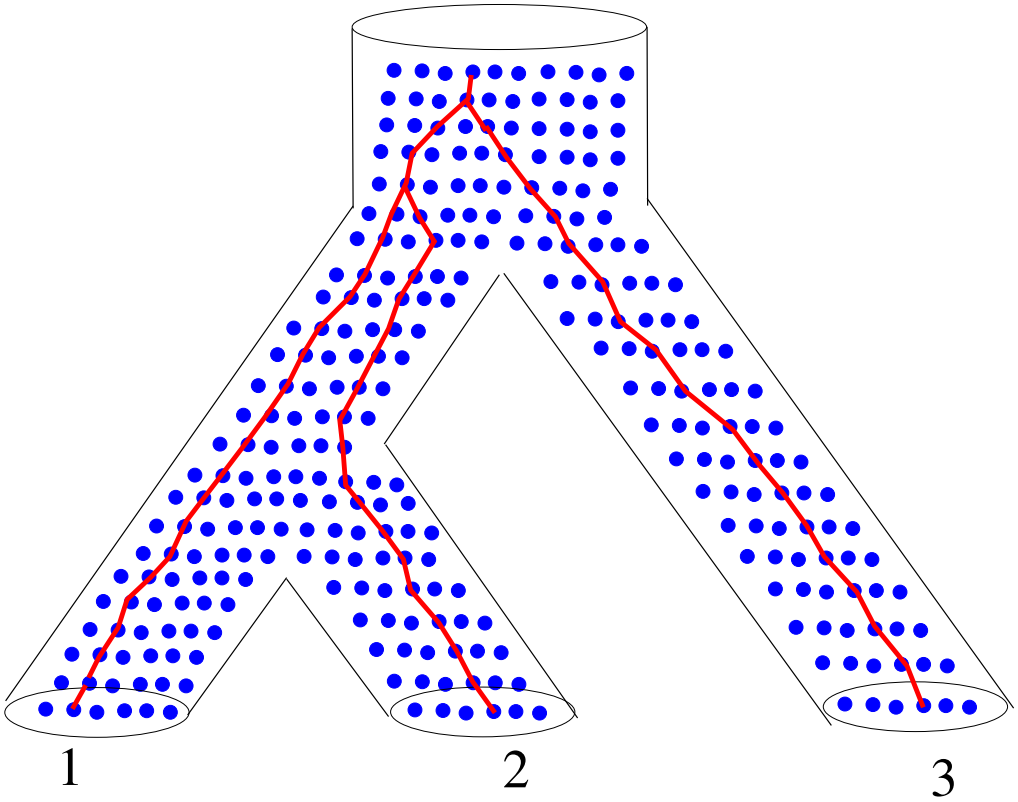
- Definition (informal): Embedding one tree into the other
- Motivation: discordance between gene tree and species tree
- A different approach: mappings between trees

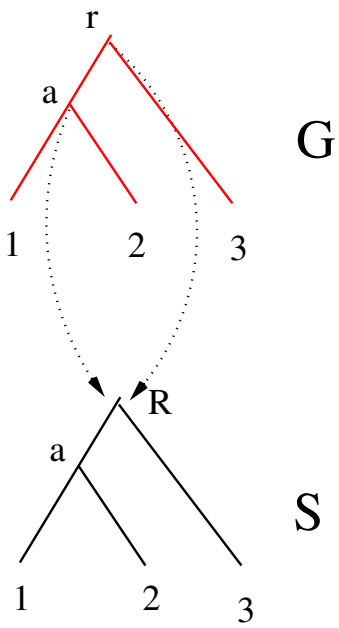
# Mappings from coalescence history: LCA



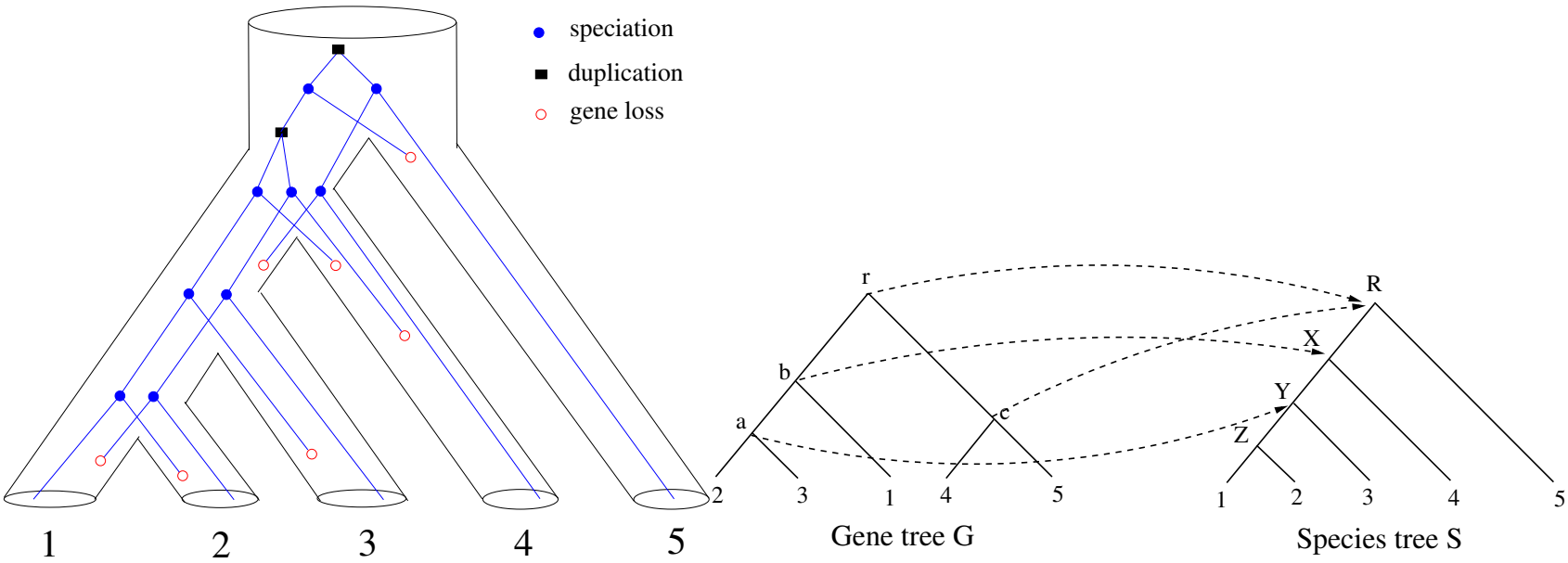



# Mappings from coalescence history: general



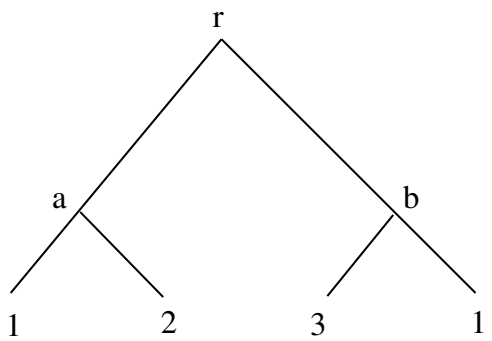
# Mappings from gene duplication and loss



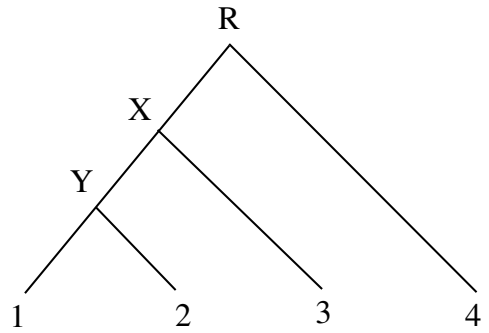
# Some notation

Given two vertices $u$ and $v$ in a phylogenetic tree $T$

- $v \leq u$: if $u$ is in $P(\rho(T), v)$, the unique path from $\rho(T)$ to $v$;
- $u$ is an ancestor of $v$, and $v$ is an descendant of $u$;
- $L(u)$: the cluster induced by $u$, i.e., the set of the labels of its leaf descendants;
- $u$ is a common ancestor of a set $A \subseteq V(T)$: $a \leq u$ for all $a \in A$.
- lca$(A)$: the *least common ancestor* of $A$



Gene tree G



Species tree S

# Reconciliation

A map $f$ from $V(G)$ to $V(S)$ is

- order-preserving: "$u \leq v$" $\Rightarrow$ "$f(u) \leq f(v)$", $\forall u, v \in V(G)$;
- leaf-preserving: for each leave $x$ in $G$, $f(x)$ and $x$ has the same label.

# Reconciliation

A map $f$ from $V(G)$ to $V(S)$ is

- order-preserving: "$u \leq v$" $\Rightarrow$ "$f(u) \leq f(v)$", $\forall u, v \in V(G)$;
- leaf-preserving: for each leave $x$ in $G$, $f(x)$ and $x$ has the same label.

A *reconciliation* between a gene tree $G$ and species tree $S$ is a leaf-preserving and order-preserving map from $V(G)$ to $V(S)$.

# Reconciliation

A map $f$ from $V(G)$ to $V(S)$ is

- order-preserving: "$u \leq v$" $\Rightarrow$ "$f(u) \leq f(v)$", $\forall u, v \in V(G)$;
- leaf-preserving: for each leave $x$ in $G$, $f(x)$ and $x$ has the same label.

A *reconciliation* between a gene tree $G$ and species tree $S$ is a leaf-preserving and order-preserving map from $V(G)$ to $V(S)$.

**Note:** the existence of reconciliations between $G$ and $S \Leftrightarrow$ $L(G) \subseteq L(S)$

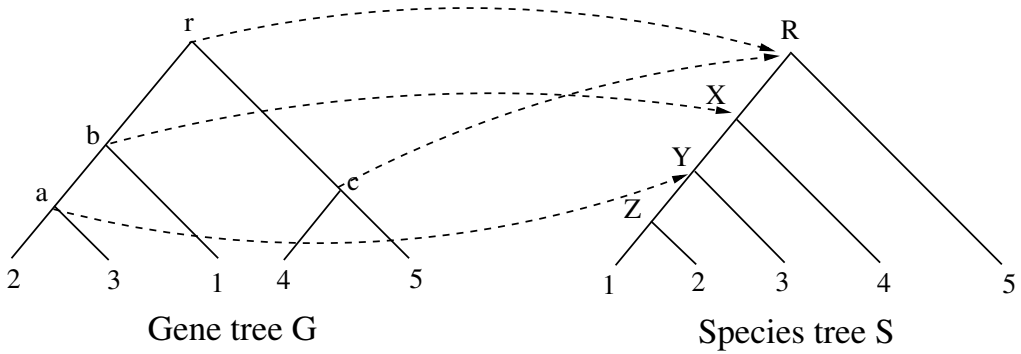
# LCA mapping

Define a mapping $M$ from $G$ to $S$ as

$$M(u) = \begin{cases} \text{The unique leaf with the same label} & \text{if } u \text{ is a leaf,} \\ \text{lca}\big((M(u_1), M(u_2)\big) & \text{if } u \text{ has children } u_1,\ u_2. \end{cases}$$

$M$ is called the *least common ancestor* (LCA) mapping from $G$ to $S$.

# Reconciliation: an example




The map $f$ defined as $f(a) = Y, f(b) = X, f(c) = f(r) = R$ and $f(i) = i$ for $1 \leq i \leq 5$, is a reconciliation between the gene tree $G$ and the species tree $S$. Note that we have

$$M(a) = M(b) = Y, \quad \text{and} \quad M(c) = M(r) = R.$$

# Deep coalescence cost

Given a reconciliation $f$ between $G$ and $S$, and a branch $e$ in $S$, we define

- $\tau_f(e) = k$ $(k > 0)$, if there exist $k+1$ distinct edges $(u_i, v_i)$ $(1 \leq i \leq k+1)$ in $G$ s.t. $e$ is on the path $P(f(u_i), f(v_i))$;
- $\tau_f(e) = 0$, otherwise.

# Deep coalescence cost

Given a reconciliation $f$ between $G$ and $S$, and a branch $e$ in $S$, we define

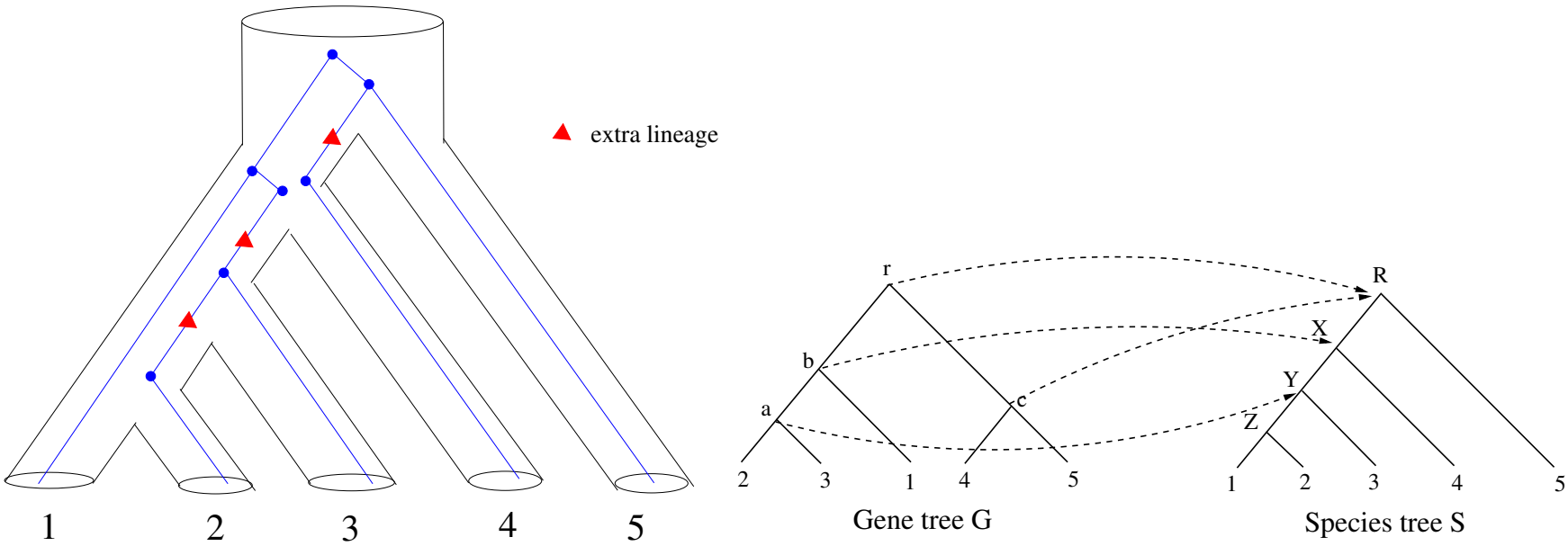
- $\tau_f(e) = k$ $(k > 0)$, if there exist $k+1$ distinct edges $(u_i, v_i)$ $(1 \leq i \leq k+1)$ in $G$ s.t. $e$ is on the path $P(f(u_i), f(v_i))$;
- $\tau_f(e) = 0$, otherwise.

The *deep coalescence* (DC) cost $\mathrm{dc}(f)$ of $f$ is defined by Maddison [Sys. Bio. 1997] as

$$\mathrm{dc}(f) := \sum_{e \in E(S)} \tau_f(e)$$

i.e., the total number of the extra lineages with respect to $f$ on all branches of $S$.

# DC cost: an example




For the reconciliation $f$ , we have $\mathrm{dc}(f) = 3$, because

$$\tau_f((R, X)) = \tau_f((X, Y)) = \tau_f((Y, Z)) = 1.$$

# Gene duplication: LCA mapping

A *duplication event* is associated with an internal vertex $u$ in $G$ ( w.r.t. the LCA reconciliation $M$) if

$$M(u) \in \{M(u_1), M(u_2)\} \tag{1}$$

where $u_1$ and $u_2$ are the children of $u$ [Page, Sys. Bio., 1994]. That is,

$$L(M(u_1)) \cap L(M(u_2)) \neq \emptyset. \tag{2}$$

# Gene duplication: LCA mapping

A *duplication event* is associated with an internal vertex $u$ in $G$ ( w.r.t. the LCA reconciliation $M$) if

$$M(u) \in \{M(u_1), M(u_2)\} \tag{1}$$

where $u_1$ and $u_2$ are the children of $u$ [Page, Sys. Bio., 1994]. That is,

$$L(M(u_1)) \cap L(M(u_2)) \neq \emptyset. \tag{2}$$

**Note:** A direct generalization of (1) does not work, but that of (2) works.

# Gene duplication: general case

For an internal vertex $u$, a *duplication event* is associated with $u$ (w.r.t. a reconciliation $f$) if and only if

$$L(f(u_1)) \cap L(f(u_2)) \neq \emptyset.$$

# Gene duplication: general case

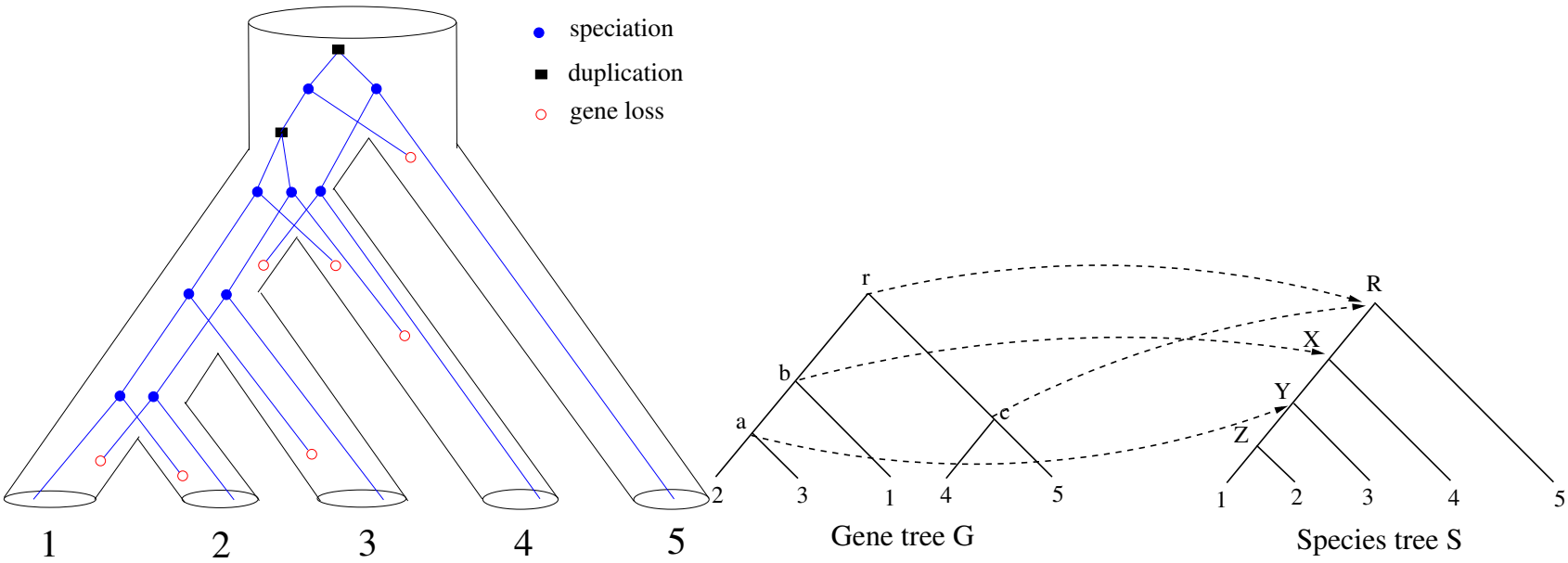
For an internal vertex $u$, a *duplication event* is associated with $u$ (w.r.t. a reconciliation $f$) if and only if

$$L(f(u_1)) \cap L(f(u_2)) \neq \emptyset.$$

Now let $\delta_f(u) = 1$ if $u$ is an internal vertex and there is a duplication event associated with it, and $\delta_f(u) = 0$ otherwise. Then the *gene duplication* (GD) cost $\mathrm{gd}(f)$ of $f$ is defined as

$$\mathrm{gd}(f) := \sum_{u \in V(G)} \delta_f(u).$$

# GD cost: an example




By definition, we have $\mathrm{gd}(f) = 2$ because $\delta_f(r) = \delta_f(b) = 1$.

# Gene loss cost

The *number of the losses* $l_f(u)$ associated to $u$ is defined as

$$l_f(u) = \begin{cases} d\big(f(u), f(u_1)\big) + d\big(f(u), f(u_2)\big) & \text{if } \delta_f(u) = 1, \\ d\big(f(u), f(u_1)\big) + d\big(f(u), f(u_2)\big) - d(f(u), p) - 2. & \text{otherwise,} \end{cases}$$

where $p := \mathrm{lca}(f(u_1), f(u_2))$.

# Gene loss cost

The *number of the losses* $l_f(u)$ associated to $u$ is defined as

$$l_f(u) = \begin{cases} d\big(f(u), f(u_1)\big) + d\big(f(u), f(u_2)\big) & \text{if } \delta_f(u) = 1, \\ d\big(f(u), f(u_1)\big) + d\big(f(u), f(u_2)\big) - d(f(u), p) - 2. & \text{otherwise,} \end{cases}$$

where $p := \mathrm{lca}(f(u_1), f(u_2))$.

If $f$ is the LCA reconciliation $M$, we have $f(u) = p$, and hence $d(f(u), p) = 0$ [Page, 94].

# Gene loss cost

The *number of the losses* $l_f(u)$ associated to $u$ is defined as

$$l_f(u) = \begin{cases} d\big(f(u), f(u_1)\big) + d\big(f(u), f(u_2)\big) & \text{if } \delta_f(u) = 1, \\ d\big(f(u), f(u_1)\big) + d\big(f(u), f(u_2)\big) - d(f(u), p) - 2. & \text{otherwise,} \end{cases}$$

where $p := \mathrm{lca}(f(u_1), f(u_2))$.

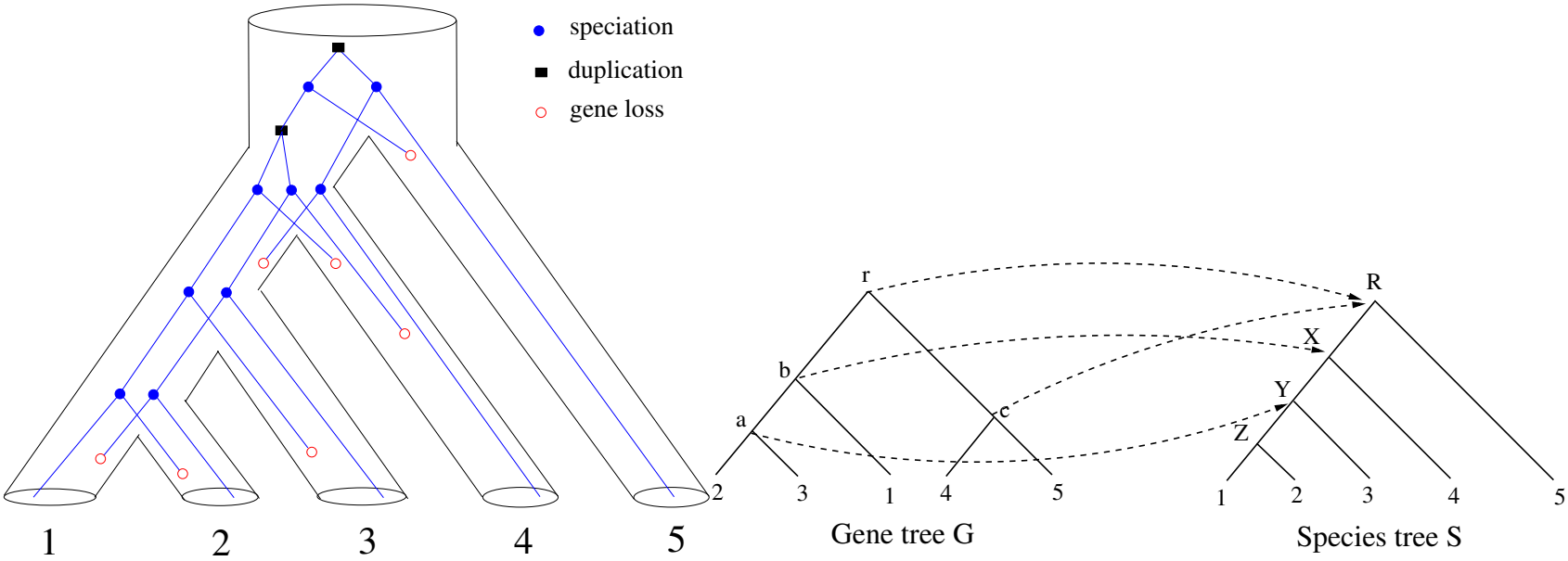
If $f$ is the LCA reconciliation $M$, we have $f(u) = p$, and hence $d(f(u), p) = 0$ [Page, 94].

The *gene loss* (GL) cost $\mathrm{gl}(f)$ of $f$ is defined as

$$\mathrm{gl}(f) := \sum_{u \in V(G)} l_f(u),$$

where we set $l_f(x) = 0$ for any leaf $x$ of $G$.

# GL cost: an example




By definition, we have

$$l_f(a) = l_f(c) = l_f(r) = 1 \quad \textit{and} \quad l_f(b) = 4,$$

and hence $\mathrm{gl}(f) = 7$.

# Reconciliation space

There is a canonical partial order $\preceq$ on the set of reconciliations between $G$ and $S$: for any $f'$ and $f$, $f' \preceq f$ if and only if $f'(v) \leq f(v)$ holds for every vertex $v$ in $G$.

# Reconciliation space

There is a canonical partial order $\preceq$ on the set of reconciliations between $G$ and $S$: for any $f'$ and $f$, $f' \preceq f$ if and only if $f'(v) \leq f(v)$ holds for every vertex $v$ in $G$.

In this space, there is a unique

- minimal element, i.e., the LCA reconciliation $M$;
- maximal element, which maps each internal node in $G$ to $\rho(S)$.

# The monotonicity result

## Theorem (W-Zhang, 2011)

*Let $f' \preceq f$ be two distinct reconciliations between a gene tree $G$ and a species tree $S$; then we have*

- $\mathrm{gd}(f') \leq \mathrm{gd}(f)$,
- $\mathrm{gl}(f') \leq \mathrm{gl}(f)$, *and*
- $\mathrm{dc}(f') < \mathrm{dc}(f)$.

# The monotonicity result

## Theorem (W-Zhang, 2011)

*Let $f' \preceq f$ be two distinct reconciliations between a gene tree $G$ and a species tree $S$; then we have*

- $\mathrm{gd}(f') \leq \mathrm{gd}(f)$,
- $\mathrm{gl}(f') \leq \mathrm{gl}(f)$, *and*
- $\mathrm{dc}(f') < \mathrm{dc}(f)$.

**Corollary 1:** LCA reconciliation is optimal, for

- the gene duplication cost [Gorecki and Tiuryn, TCS, 2006],
- the gene loss cost [Chauve et al. RECOMB 2009],
- the deep coalescence cost.

**Corollary 2:** $\mathrm{gd}(M) = \mathrm{gd}(f) \Leftrightarrow \delta_M(u) = \delta_f(u), \forall u \in V(G)$.

# An interpolation lemma

$D(f', f) := |\{v \in V(G) : f(v) \neq f'(v)\}|.$

**Lemma**

*Let $f' \preceq f$ be two distinct reconciliations between $G$ and $S$. Then there exists a reconciliation $f^*$ between $G$ and $S$ such that:*

- $f' \preceq f^* \preceq f$,
- $D(f', f^*) = D(f', f) - 1$, *and*
- $D(f^*, f) = 1$.

# Enumeration problems

Problem I:
**Input:** A gene tree $G$, species tree $S$ with $L(G) \subseteq L(S)$.
**Output:** All reconciliations with the minimum gene duplication cost.

Ans: a dynamic programming approach

# Enumeration problems

Problem I:
**Input:** A gene tree $G$, species tree $S$ with $L(G) \subseteq L(S)$.
**Output:** All reconciliations with the minimum gene duplication cost.

Ans: a dynamic programming approach

Problem II:
**Input:** A gene tree $G$, species tree $S$ with $L(G) \subseteq L(S)$, and $\varepsilon \geq 0$.
**Output:** The set of (nearly-optimal) reconciliations $f$ with $\mathrm{gd}(f) \leq \mathrm{gd}(M) + \varepsilon$.

Ans: a greedy approach

# Conclusions

- A theoretic-graph framework to study reconciliation is introduced: separating reconciliation concept from the cost models that are used to measure the tree discordance.
- We show that three cost measures are monotonic.

- Some algorithms to enumerate (nearly)-optimal reconciliations have been developed; they are implemented in c++.

# Thanks

- Thanks to the organizers: Tandy, Mike and Vincent
- Thanks to the INI institute
- Thanks for your attention

# Thanks

- Thanks to the organizers: Tandy, Mike and Vincent
- Thanks to the INI institute
- Thanks for your attention
- Thanks