Reconciliation Problems for Duplication, Loss and Horizontal Gene Transfer

Pawel Górecki

Presented by Connor Magill

November 20, 2008

Introduction

- Problem:

  Relationships between species cannot always be inferred from a single gene family

  Gene Duplication, gene loss, gene convergence, and horizontal gene transfer can cause dissimilarities between gene family trees

  How can these be explained?

Introduction

- Earlier Approaches:

  Goodman: *Mapping* and *reconciling trees* - inspired research on *duplication-loss model*

  D-L problem is NP-hard

  Page & Charleston - Transfer extension to D-L with host switching & parasite duplications.

  Charleston-"jungles" algorithm - finds an optimal set of transfers for given gene & species trees (complexity unknown)

Introduction

- Earlier Approaches:

Hallett and Lagergren - model based on the reconciliation,

*Gene duplications disallowed.

*Various formulations of the problem are NP-complete

*Polynomial complexity in these formulations of the problem

- Motivation:

Lack of a formal definition of the DLT model

Basic problem - “find the optimal reconciliation cost (or tree) for a given gene tree, a species tree and a set of (biologically well defined) gene transfers”

Duplication-Loss Model

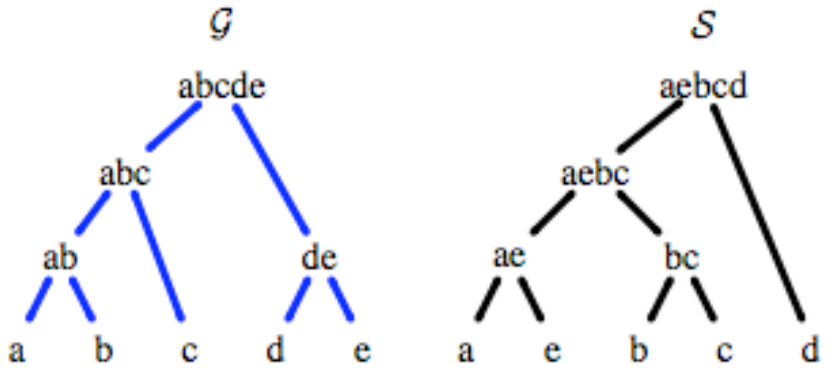


Figure 1: Gene tree and species tree

I = Set of Species

S and G are rooted binary directed trees
S has N leaves uniquely labeled by the elements from I

In the gene tree, a given gene x is related to the species x

L(T) is the cluster for the root of tree T
L(G) = L(S) is the usually considered variant of reconciliation (and leaves of the trees have unique labeling)

Duplication-Loss Model

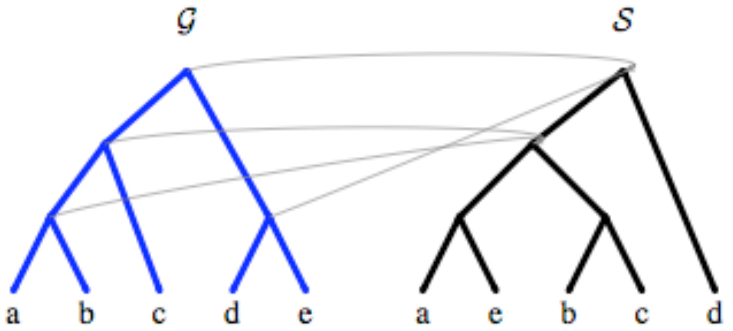


Figure 2: Mapping M for internal nodes of $\mathcal{G}$

Fig. 2 presents trees with two duplications (nodes in $\mathcal{G}$): the root and *abc*.

For each node g ∈ G, M(g) is the node from S which is the least common ancestor of 'g' in S

Duplication occurs in G if M(g) = M(h), where g and h are internal nodes of G, and h is a child of g

Duplication Cost - the total number of duplications in G, can be considered as a dissimilarity measure between G & S

Duplication-Loss Model

For a disjoint set L in tree S, let S^L be the smallest subtree in S containing L as its leaves

Contract all nodes of degree 2 of S^L (except for the root) to get the homomorphic tree $S|_L$ of S induced by L

Let M' be the mapping from $\mathcal{G}$ to $\mathcal{S}|_{L(\mathcal{G})}$ induced by the mapping M from $\mathcal{G}$ to $\mathcal{S}$. Let $\mathbf{loss}_g$ denote the number of gene losses (or losses) associated with the node g.

If g is a leaf node in $\mathcal{G}$ then the number of losses equals zero.

If g is an internal node in $\mathcal{G}$ then let a and b denote the two children of g. The number of losses is defined by

$$\mathbf{loss}_g = \begin{cases} d(M'(a), M'(g)) + 1 \\ \qquad \text{if } M'(a) \subsetneq M'(g) = M'(b), \\ d(M'(a), M'(g)) + d(M'(b), M'(g)) \\ \qquad \text{if } M'(a) \subsetneq M'(g) \supsetneq M'(b), \\ 0 \qquad \text{otherwise} \end{cases}$$

where $d(s, s') = |\{h \in \mathcal{S}|_{L(G)} | s \subsetneq h \subsetneq s'\}|$ for $s, s' \in \mathcal{S}|_{L(G)}$.

The number of *gene losses* is given by

$$\mathbf{loss}(\mathcal{G}, \mathcal{S}) = \sum_{g \in \mathcal{G}} \mathbf{loss}_g.$$

Reconciled Trees

- $T_r(G,S)$
  -Contains only clusters from S
  -G is a subtree
  -For different children a & b of node g, a & b are either disjoint or the same

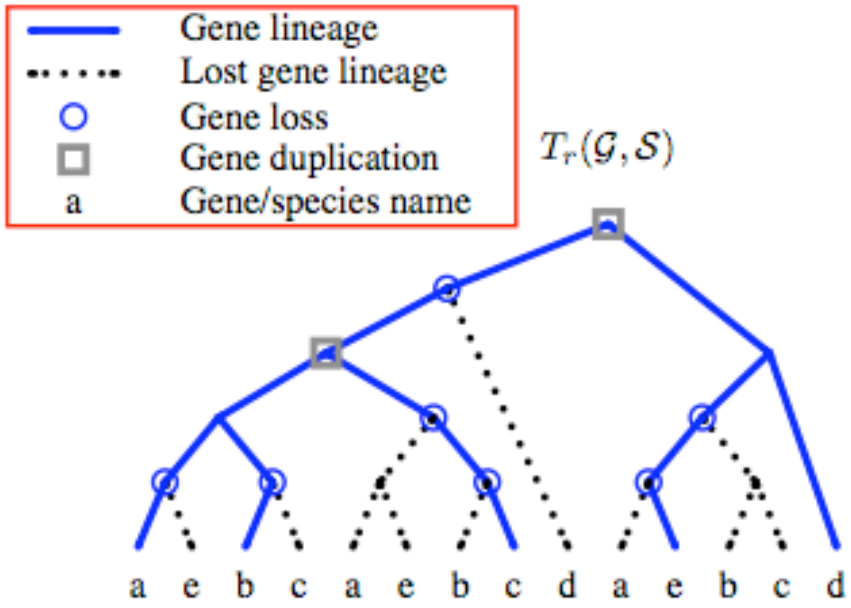


Figure 3: Reconciled tree and embedded gene tree (solid lines - tree $T_r^{L(\mathcal{G})}$)

Horizontal Transfer

Contract all nodes of degree 2 (except the root) of S, this can also be viewed as a directed graph <V,E>:

Let $H \subseteq V \times V$ satisfy the following conditions

(H1) $\langle v, v\rangle \notin H$, for each $v \in V$,

(H2) For all $\langle v, w\rangle \in H$ nodes v, w have exactly one child in S.

(H3) For all $h_1, h_2 \in H$ if $h_1 \neq h_2$ and $h_i = \langle v_i, w_i\rangle$, $i = 1, 2$ then $\{v_1, w_1\} \cap \{v_2, w_2\} = \emptyset$.

A relation $H \subseteq V \times V$ satisfying (H1)-(H3) is said to be *horizontal* for S if exists $\delta_H : V \to \mathcal{N}$ that (H4)-(H6) hold:

(H4) $\delta_H(v) = \delta_H(w)$, for each $\langle v, w\rangle \in H$,

(H5) $\delta_H(v) < \delta_H(w)$, for each $\langle v, w\rangle \in E$,

(H6) $\delta_H(v) \neq \delta_H(w)$, for each $\langle v, v'\rangle, \langle w, w'\rangle \in H$ satisfying $v \neq w$.

H’ (reverse direction of some edges from H) is horizontal for S

Horizontal Transfer

Horizontal transfers added to the directed graph cannot create a cycle in the graph

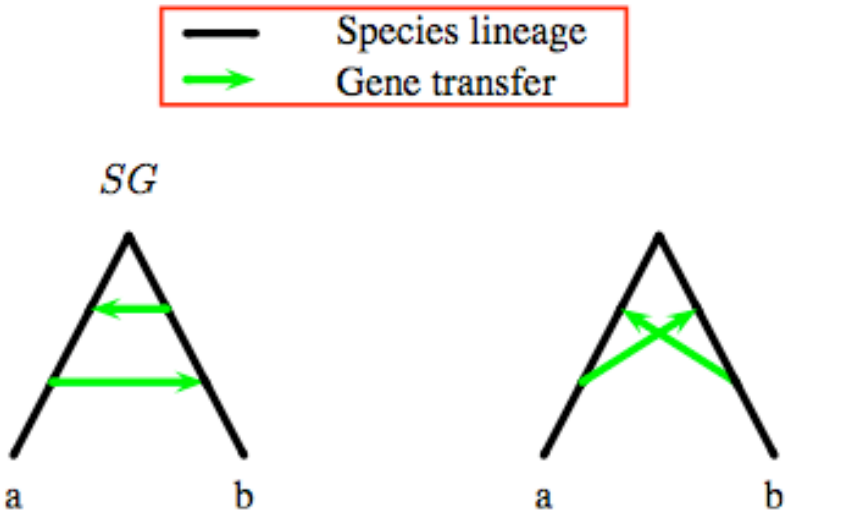


Figure 5: Only the graph on the left is a species graph

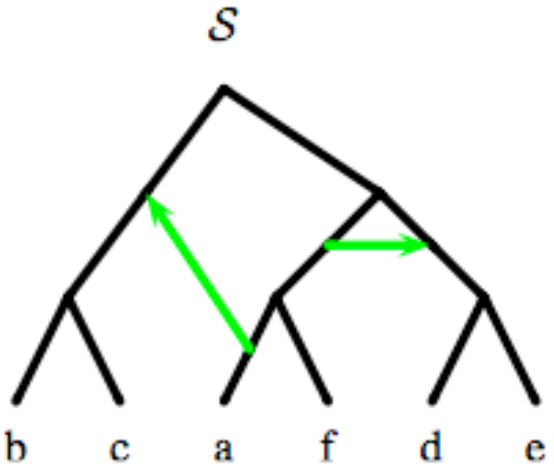


Figure 6: Species graph with two transfers

Reconciliation & Gene Transfer

Gene transfer gives additional ways of reconciliation between species and gene trees

Several gene lineages could be transferred using one transfer edge

- Problem:

  Which gene lineages should be transferred?

Reconciliation & Gene Transfer

3.4.1 *Homomorphic species graph*

The model of gene trees is the same as in the duplication-loss model.

Let us consider a species graph $\mathrm{SG} = \langle S, H \rangle$ and a gene tree $\mathcal{G}$ such that $L(\mathcal{G}) \subseteq L(S)$. We define a *homomorphic species graph* $\mathrm{SG}|_{L(\mathcal{G})} = \langle T, I \rangle$ where $I \subseteq H$ is a set of edges such that if $\langle v, w \rangle \in I$ then $v, w \in S^{L(\mathcal{G})}$; moreover T is a tree obtained from $S^{L(\mathcal{G})}$ by contracting all nodes of degree 2 except

- the root,
- the nodes of edges from I.

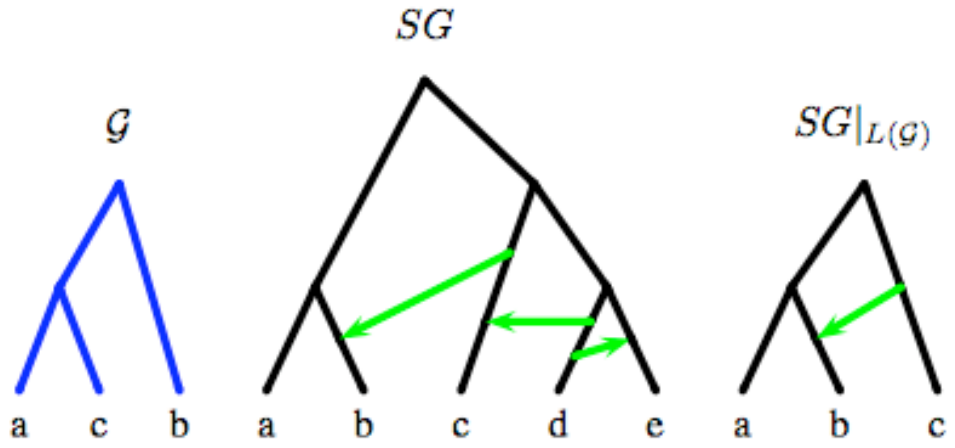


Figure 7: Homomorphic species graph example

Reconciliation & Gene Transfer

- Extended Species Graph

- Let $G_{k+1} = \vec{SG}$. Let $A_{k+1} = \emptyset$ be the set of horizontal edges for the graph G_{k+1}.
- Assume that $G_{i+1} = \langle V_{i+1}, E_{i+1} \rangle$ and A_{i+1} are constructed.
- We define the graph $G_i = \langle V_i, E_i \rangle$ and A_i. Let $h_i = \langle v, w \rangle$ then exists exactly one z such that $\langle w, z \rangle \in E_{i+1}$. Let $K = \langle V', E' \rangle$ be an isomorphic disjoint copy of $G_{i+1}(z)$. Let B be the set of the horizontal edges in K induced by this isomorphism.
  - $V_i = V_{i+1} \cup V'$,
  - $E_i = (E_{i+1} \setminus \{h_i\}) \cup E' \cup \{\langle v, r \rangle\}$, where r is the root of K,
  - $A_i = A_{i+1} \cup B \cup \{\langle v, r \rangle\}$,
- The labelling in G_i is induced by labelling in G_{i+1} and K.

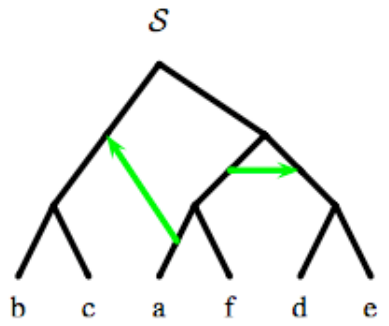


Figure 6: Species graph with two transfers

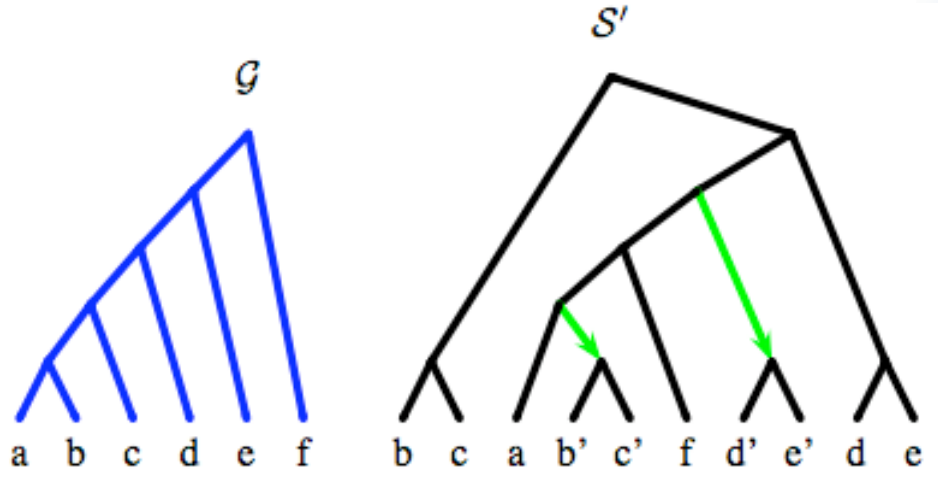


Figure 8: The gene tree and the extended species tree (SG - see Fig. 6)

Reconciliation & Gene Transfer

The number of horizontal edges could be exponential in the size of H

Table 1: Scenario θ - the example cont.

Label of g ($\mathcal{G}$ leaf node)	$\theta(g)$ ($\mathcal{S}'$ leaf node)	$\theta(g)$ label
a	a	a
b	b'	b
c	c'	c
d	d'	d
e	e	e
f	f	f

We write x and x' to distinguish between different leaves

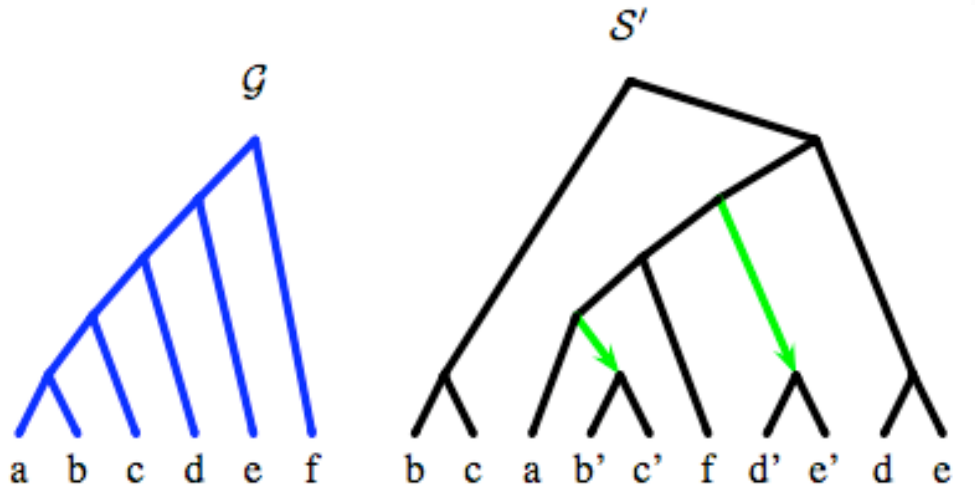


Figure 8: The gene tree and the extended species tree (SG - see Fig. 6)

Reconciliation & Gene Transfer

Reconciled tree and the evolutionary events depend on the transfer scenario

Let $\mathcal{G}$ be a gene tree. A *scenario* is a function from leaves of $\mathcal{G}$ into leaves of $\mathcal{S}'$ which preserves labels. A scenario ξ naturally extends to the least common ancestor mapping $M_\xi : \mathcal{G} \to \mathcal{S}'$.

For a directed graph G with horizontal edges let H_G denote the set of horizontal edges in G. Let $\mathrm{HS}_{G,\xi} = \{v | \langle v, w\rangle \in \mathrm{H}_G\}$ and $\mathrm{HE}_{G,\xi} = \{w | \langle v, w\rangle \in \mathrm{H}_G\}$.

The definition of the number of gene losses $\mathbf{loss}_{\xi,g}$ associated with the gene node g is given below. If g is an internal node and its children are a and b then

$$\mathbf{loss}_{\xi,g} = \begin{cases} d(M_\xi(a), M_\xi(g)) \\ \qquad \text{if } M_\xi(a) \subsetneq M_\xi(g) = M_\xi(b), g \in \mathrm{HS}_{S'}, \\ d(M_\xi(a), M_\xi(g)) + 1 \\ \qquad \text{if } M_\xi(a) \subsetneq M_\xi(g) = M_\xi(b), g \notin \mathrm{HS}_{S'}, \\ d(M_\xi(a), M_\xi(g)) + d(M_\xi(b), M_\xi(g)) \\ \qquad \text{if } M_\xi(a) \subsetneq M_\xi(g) \supsetneq M_\xi(b), \\ 0 \qquad \text{otherwise} \end{cases}$$

where $d(s, s') = |\{h \in S' | s \subsetneq h \subsetneq s' \text{ and if } h \in \mathrm{HS}_{S'} \text{ then } s' \subseteq c \text{ and } \langle h, c\rangle \in \mathrm{H}_{S'}\}|$ for $s, s' \in S'$.

The number of gene losses for the scenario ξ is given by

$$\mathbf{loss}_\xi(\mathcal{G}, \mathcal{S}) = \sum_{g \in \mathcal{G}} \mathbf{loss}_{\xi,g}.$$

Reconciliation & Gene Transfer

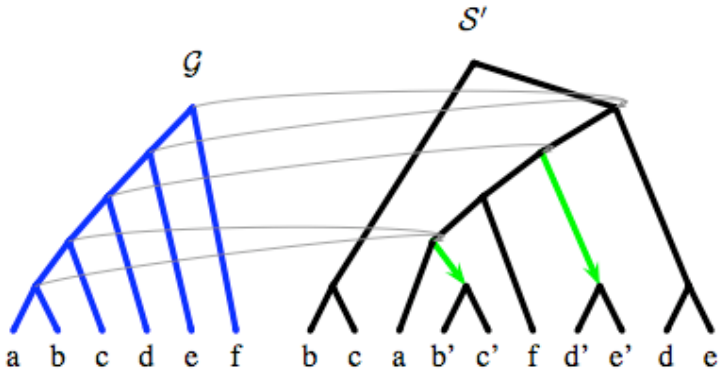
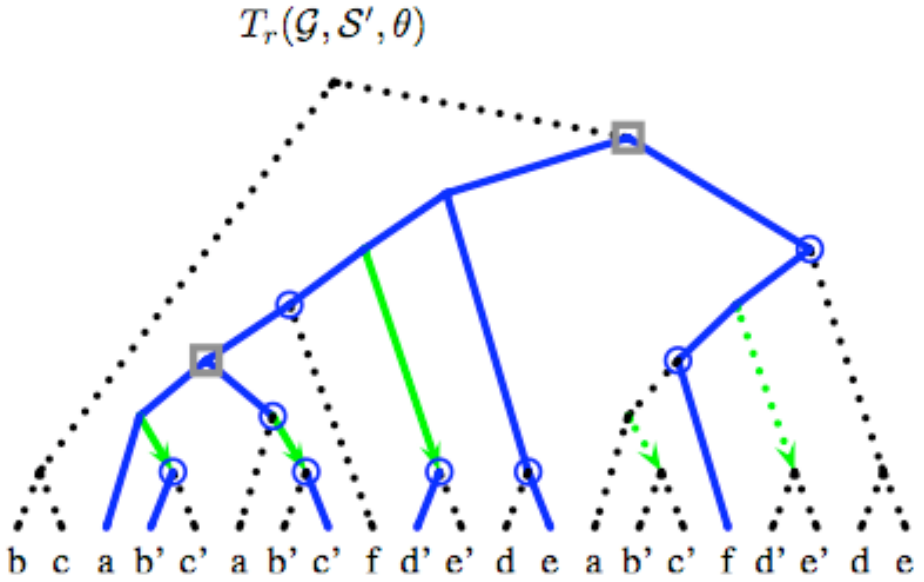


Figure 9: Mapping M_θ for internal nodes of $\mathcal{G}$

Reconciled Tree for the scenario:



Scenarios

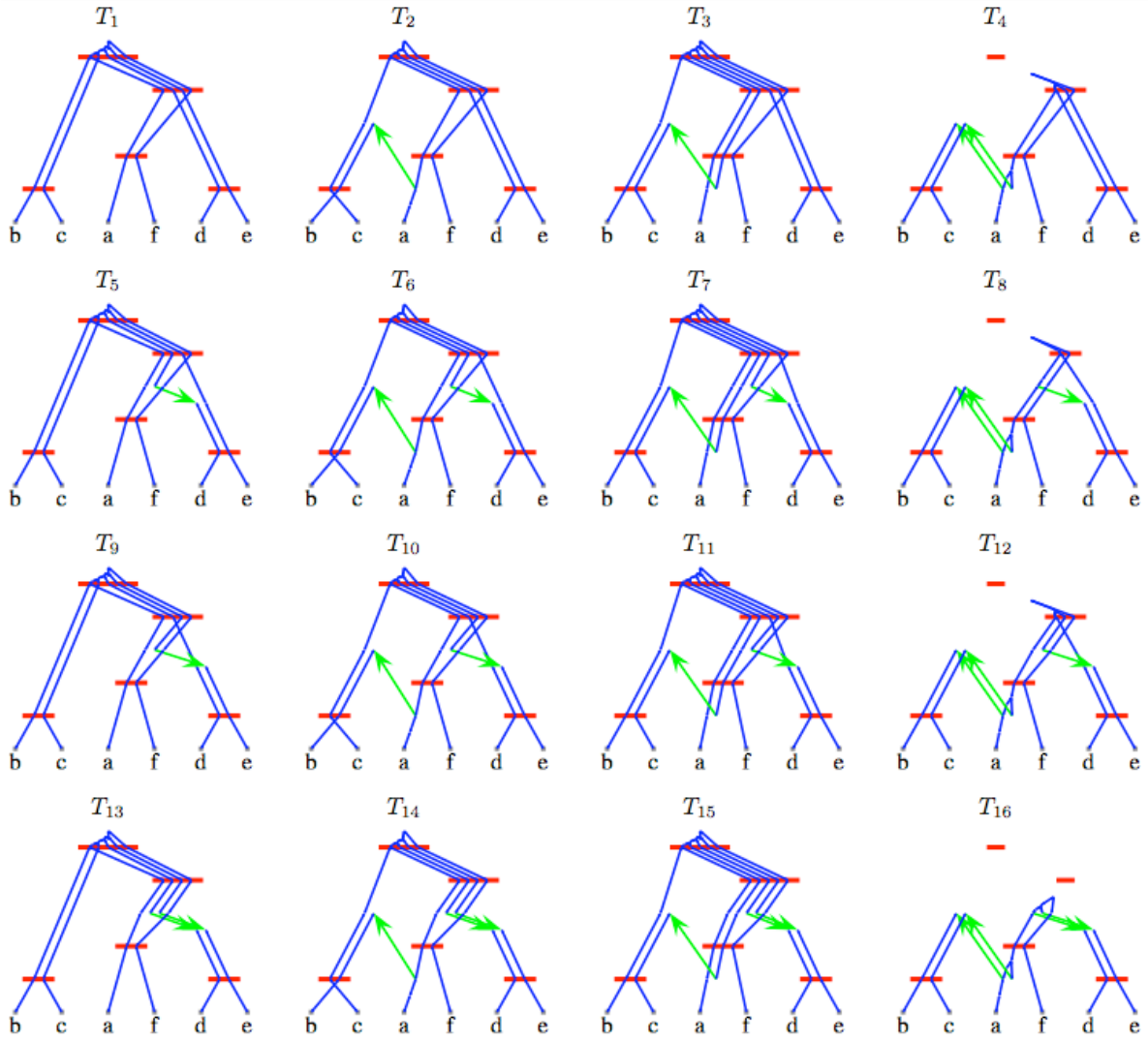


Table 2: Reconciliation costs summary

Tree no	**Total cost**	Dup	Loss	Hgt	Scen. - leaves x s.that $M^*(x) = x'$
T_1	**18**	4	14	0	
T_2	**17**	3	13	1	b
T_3	**22**	4	17	1	c
T_4	**14**	3	9	2	b c
T_5	**20**	4	15	1	d
T_6	**19**	3	14	2	b d
T_7	**24**	4	18	2	c d
$\mathbf{T_8}$	**13**	**2**	**8**	**3**	**b c d**
T_9	**20**	4	15	1	e
T_{10}	**19**	3	14	2	b e
T_{11}	**24**	4	18	2	c e
T_{12}	**16**	3	10	3	b c e
T_{13}	**22**	4	16	2	d e
T_{14}	**21**	3	15	3	b d e
T_{15}	**26**	4	19	3	c d e
T_{16}	**15**	3	8	4	b c d e

Optimal Scenario Algorithm

Let $a \geq 0$, $b \geq 0$ and $c \geq 0$. Let $\mathbf{rcost}(\mathcal{G}, \mathrm{SG}, \xi)$ denote the reconciliation cost for the scenario ξ. This cost equals

$$a * \mathbf{dup}_{\xi}(\mathcal{G}, \mathrm{SG}) + b * \mathbf{loss}_{\xi}(\mathcal{G}, \mathrm{SG}) + c * \mathbf{hgt}_{\xi}(\mathcal{G}, \mathrm{SG}).$$

An optimal scenario for a species graph and gene tree minimizes the reconciliation cost. (Could be more than one optimal scenario)

The input to the algorithm requires

- a gene tree $\mathcal{G}$,
- a species graph $\mathrm{SG} = \langle\langle V, E\rangle, H\rangle$ such that $L(\mathcal{G}) = L(\mathrm{SG})$,
- constants a, b and c.

Output: $\mathbf{rcost}(\mathcal{G}, \mathrm{SG}, \xi)$

Optimal Scenario Algorithm

Optimal path cost (pc): $(E \cup H) \times V \rightarrow R^{+} \cup \{\infty\}$

Cost judged by number of losses and transfers in the path

Let HS denote the set of starts of the transfers. Let HE denote the set of ends of the transfers. Let IN denote the set of all internal nodes except nodes from HS and HE.

$$pc(e,s) = \begin{cases} 0 & \text{if } w = s \text{ and } e \in \mathrm{E} \\ c & \text{if } w = s \text{ and } e \in \mathrm{H} \\ pc(\langle w, w.c\rangle, s) & \text{if } w \neq s,\ w \in \mathrm{HE} \text{ and } e \in \mathrm{E} \\ pc(\langle w, w.c\rangle, s) + c & \text{if } w \neq s,\ w \in \mathrm{HE} \text{ and } e \in \mathrm{H} \\ pc(\langle w, w.a\rangle, s) + b & \text{if } w \neq s,\ w \in \mathrm{IN} \text{ and } w.a \rightarrow s \nleftarrow w.b \\ pc(\langle w, w.b\rangle, s) + b & \text{if } w \neq s,\ w \in \mathrm{IN} \text{ and } w.a \nrightarrow s \leftarrow w.b \\ \min\{pc(\langle w, w.a\rangle, s),\ pc(\langle w, w.b\rangle, s)\} + b & \text{if } w \neq s,\ w \in \mathrm{IN} \text{ and } w.a \rightarrow s \leftarrow w.b \\ pc(\langle w, w.a\rangle, s) + b & \text{if } w \neq s,\ w \in \mathrm{HS} \text{ and } w.a \rightarrow s \nleftarrow w.b \\ pc(\langle w, w.b\rangle, s) & \text{if } w \neq s,\ w \in \mathrm{HS} \text{ and } w.a \nrightarrow s \leftarrow w.b \\ \min\{pc(\langle w, w.a\rangle, s) + b,\ pc(\langle w, w.b\rangle, s)\} & \text{if } w \neq s,\ w \in \mathrm{HS} \text{ and } w.a \rightarrow s \leftarrow w.b \\ +\infty & \text{otherwise} \end{cases}$$

Figure 12: Function pc - the optimal cost of a path for $e = \langle v, w\rangle$ and s

Optimal Scenario Algorithm

1: **function** optCostPairs $(\langle p_1, c_1\rangle,\ \langle p_2, c_2\rangle,\ g, s)$
2: **if** $s \in \text{HE}$ **then**
3: return optCostPairs$(\langle p_1, c_1\rangle,\ \langle p_2, c_2\rangle,\ g,\ s.c)$
4: **end if**
5:

$$opt := \begin{cases} a & \text{if } p_1 = s = p_2 \\ a + \min\{pc'(a, p_2) + b,\ pc'(b, p_2) + b\} & \text{if } p_1 = s \neq p_2,\ s \to p_2 \text{ and } s \in HS \\ a + \min\{pc'(a, p_2),\ pc'(b, p_2) + b\} & \text{if } p_1 = s \neq p_2,\ s \to p_2 \text{ and } s \notin HS \\ a + \min\{pc'(a, p_1) + b,\ pc'(b, p_1) + b\} & \text{if } p_2 = s \neq p_1,\ s \to p_1 \text{ and } s \in HS \\ a + \min\{pc'(a, p_1),\ pc'(b, p_1) + b\} & \text{if } p_2 = s \neq p_1,\ s \to p_1 \text{ and } s \notin HS \\ \min(pc'(a, p_1) + pc'(b, p_2),\ pc'(a, p_2) + pc'(b, p_1)) & \text{if } p_1 \neq s \neq p_2 \text{ and } p_1 \leftarrow s \to p_2 \\ +\infty & \text{otherwise} \end{cases}$$

where $pc'(x, p_i) = pc(\langle s, s.x\rangle, p_i)$ for $x \in \{a, b\}$.
6: $Q := \emptyset$
7: **if** $opt < +\infty$ **then** $Q := \{\langle s, c_1 + c_2 + opt\rangle\}$ **fi**
8: **if** g.labels $\subseteq s.a$.labels and $s.a \to p_1, p_2$ **then** $Q := Q \cup optCostPairs(\langle p_1, c_1\rangle,\ \langle p_2, c_2\rangle,\ g,\ s.a)$ **end if**
9: **if** g.labels $\subseteq s.b$.labels and $s.b \to p_1, p_2$ **then** $Q := Q \cup optCostPairs(\langle p_1, c_1\rangle,\ \langle p_2, c_2\rangle,\ g,\ s.b)$ **end if**
10: return Q
11: **BEGIN** - main algorithm
12: **for** each node g of $\mathcal{G}$ in postfix order **do**
13: **if** g is leaf node **then**
14: $P_g := \{\langle s, 0\rangle\}$, s is leaf node from the species graph with the label g.
15: **else**
16: $Q := \bigcup_{p \in P_{g.a}, p' \in P_{g.b}} optCostPairs(p,\ p',\ g,\ \text{root(SG)})$
17: $N := \{n | \langle n, c\rangle \in Q\}$
18: $P_g := \{\langle n, \min\{c | \langle n, c\rangle \in Q\}\rangle\}_{n \in N}$
19: **end if**
20: **end for**
21: return $\min\{c | \langle n, c\rangle \in P_{\text{root}(\mathcal{G})}\}$
22: **END** - main algorithm

Algorithm 1: Find optimal cost

Optimal Scenario Algorithm

Time complexity of Algorithm 1 is $O(|V_g||V|^3)$,
V_g is the set of nodes of G

Space complexity of Algorithm 1 is $O(|E||V|)$

Test with randomly generated trees and set transfers shows that this algorithm can compute optimal cost for trees with 1000 taxons and several hundred transfers hypothesis in a few seconds.

Optimal Scenario Algorithm

Sequences of isoleucyl-tRNA synthetase, ClustalW used for multiple alignment, gene tree computed using NJ & ML ((**Eukaryotes**, *Archea*), Bacteria)

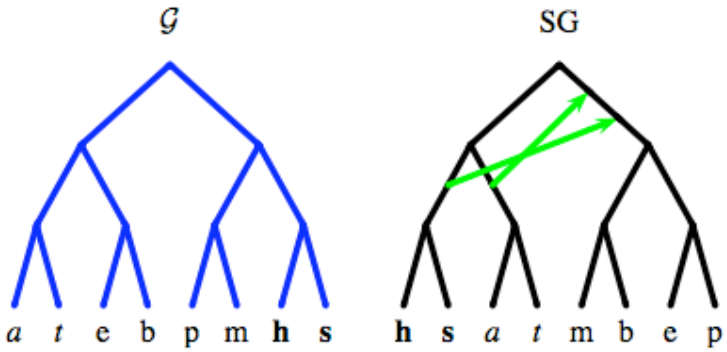


Figure 13: Gene tree and species graph

Abbreviations: **h** - **H.sapiens**, **s** - **Saccharomyces cerevisiae**, *t* - *Methanothermobacter. thermautotrophicus*, *a* - *Pyrococus abyssi*, m - Mycobacterium leprae, b - Bacillus subtilis, e - Esterichia coli, p - Ricketsia prowazekii.

Optimal solution has 2 transfers and 4 gene losses
DL-Solution has 1 gene duplication and 6 gene losses

Future

Extensions to Algorithm 1:

- Using gene trees without unique leaves labelling
- Reconciling gene trees and species graph based on non binary trees

Conclusion

If this really is the only model which brings together gene duplication, loss, and transfer well, then it is good, but it means that there are not other models to really compare it to.

Glossed over test results in the paper, just stating that "we can compute optimal cost for trees with 1000 taxons(leaves) and several hundred transfers hypothesis in a few seconds"

Not really that useful as of now, need to implement the first and second extensions to Algorithm 1 for the model to truly be useful