Comparing MUL-trees

Katharina Huber[1],
School of Computing Sciences,
University of East Anglia (UEA),
UK.

Phylogenetics: New Data, New Phylogenetic Challenges,
June 2011.

[1]This is joint work V.Moulton, R. Suchecki (both UEA) and A. Spillner (Greifswald University, Germany)

Gene trees encountered in polyploidy studies

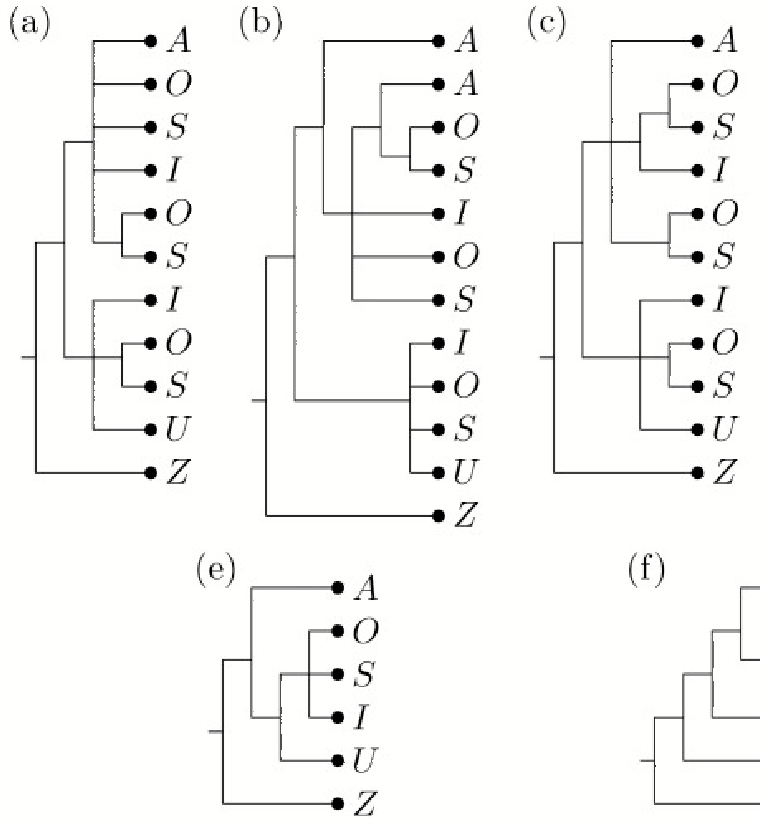
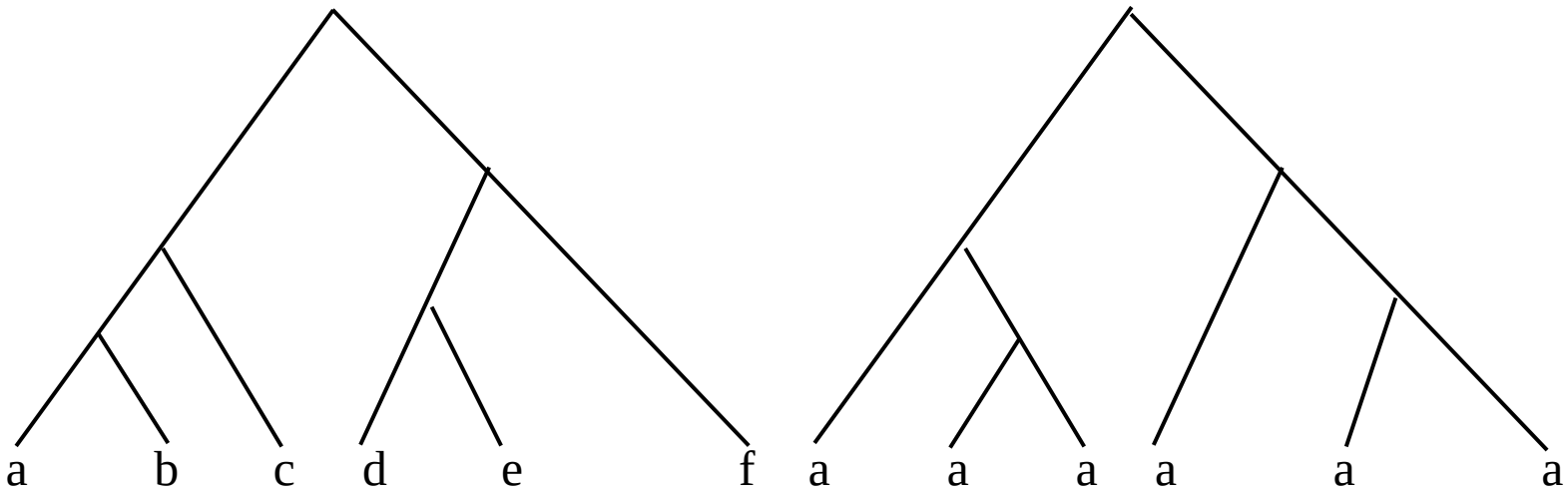


Figure: Diploids: Silene.ajanensis (A) and S. uralensis (U), tetraploid: S.involucrata (I), and hexaploids: S.sorensenis (S) and S.ostenfeldii (O). Root: S.zawadskii (Z).

Extreme cases



Left: Phylogenetic tree on $X = \{a, \ldots, f\}$. Right: "Tree shape" on $M = \{a, a, a, a, a\}$.

A crucial concept: Metrics

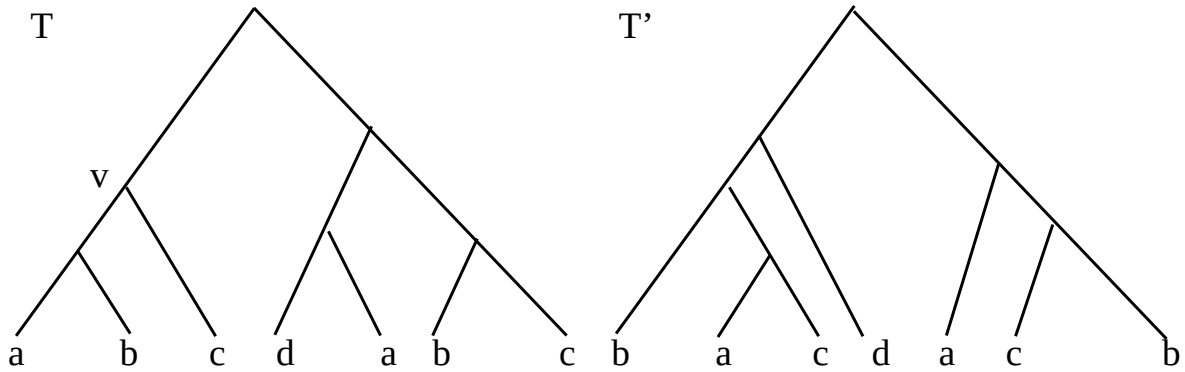
Notation: $\mathbf{T}(M)$ the class of all MUL-trees with leaf set M.

$D : \mathbf{T}(M) \times \mathbf{T}(M) \to \mathbb{R}$ is called a *metric* if, for all T_1, T_2, and T_3 in $T(M)$,

- $D(T_1, T_1) = 0$,
- $D(T_1, T_2) = D(T_2, T_1)$ (symmetric),
- $D(T_1, T_2) \leq D(T_1, T_3) + D(T_3, T_2)$ (triangle inequality).

A metric is called *proper* if $D(T_1, T_2) = 0$ implies that T_1 and T_2 are isomorphic.

Cluster metric D_c (not a proper metric (tree shapes!))



T_1, T_2 MUL-trees on M:

$$D_c(T_1, T_2) = |\mathcal{C}(T_1)\Delta\mathcal{C}(T_2)|/2$$

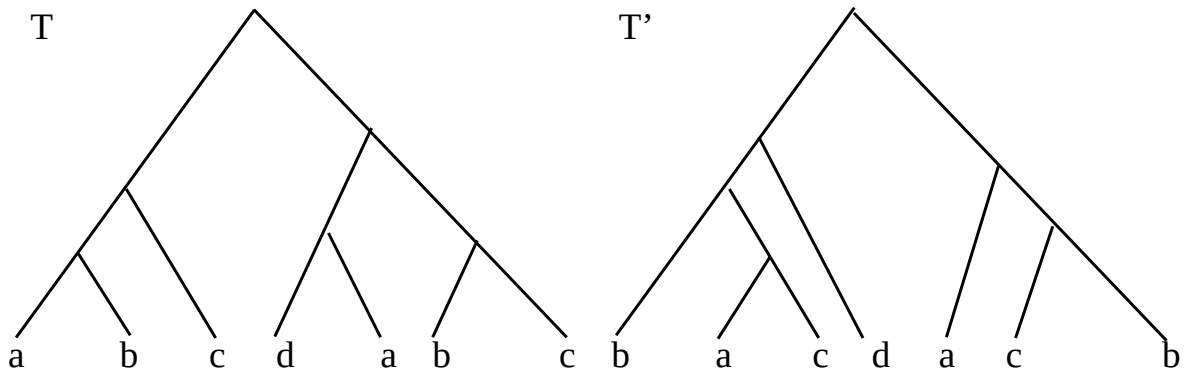
where for a MUL- tree R we have

$$\mathcal{C}(R) = \{C(v) : V \text{ vertex of } R\}$$

and $C(v)$ is the cluster induced a vertex of R.

Example: $C(v) = \{a, b, c\}$ and $D_c(T, T') = 1$.

Robinson-Foulds metric D_R (proper metric!)

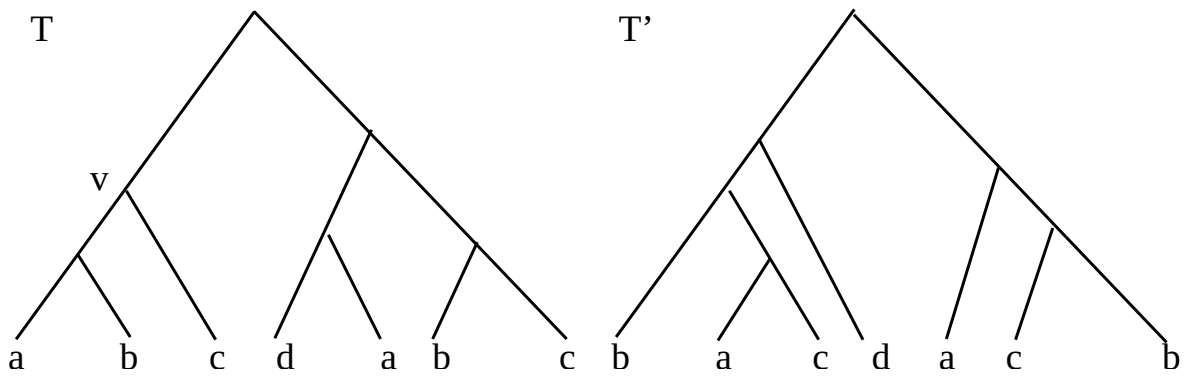


T_1, T_2 MUL-trees on M:

$$D_R(T_1, T_2) = \text{(min number of edge contractions and vertex expansions to transform } T_1 \text{ into } T_2)/2.$$

Example: $D_R(T, T') = 3$.

Nested label metric D_N (proper metric!)



T_1, T_2 MUL-trees on M:

$$D_N(T_1, T_2) = |\Gamma(T_1)\Delta\Gamma(T_2)|/2$$

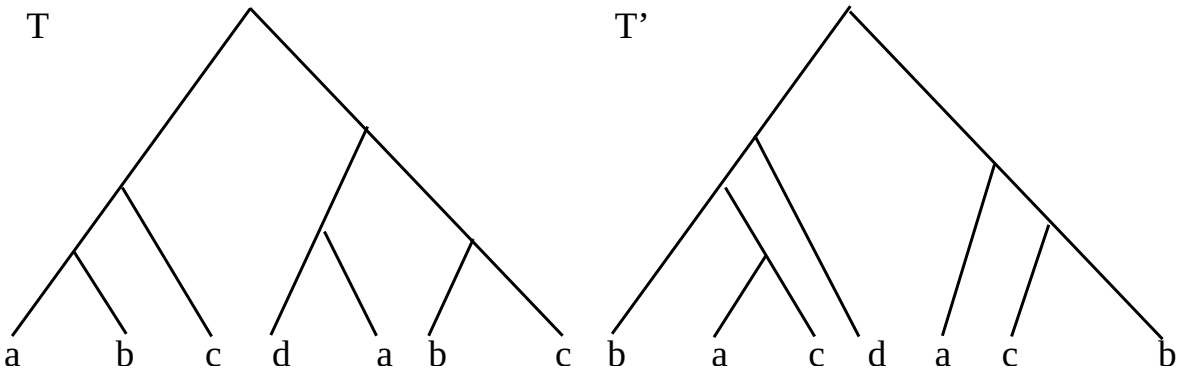
where for a MUL-tree R, we put

$$\Gamma(T) = \{\gamma(v) : v \in V(T)\}$$

and for all vertices v in T we have: $\gamma(v) = \{v\}$ if v is a leaf of R and $\gamma(v_1) \cup \ldots, \cup\gamma(v_l)$ else where $v_1, \ldots, v_l$ are the children of v.

Example: $\gamma(v) = \{\{\{a\}, \{b\}\}, \{c\}\}$ and $D_N(T, T') = 4$.

MAST-metric D_{MA} (proper metric!): Part I



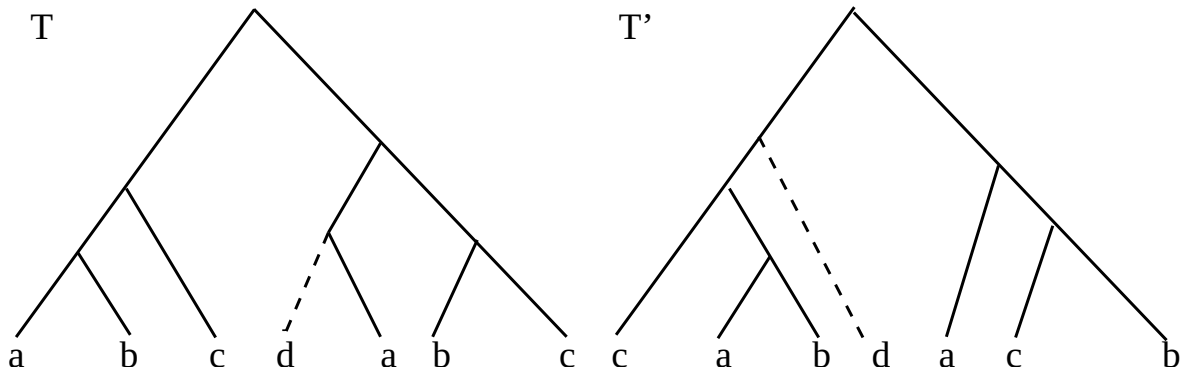
T_1, T_2 MUL-trees on M:

$D_{MA}(T_1, T_2) =$ (number of elements in M with multiplicities)$-|L|$.

L is a subset of the leaf set of T_1 of maximum size such that the subtree induced by T_1 on L is isomorphic to the subtree induced by T_2 on L.

MAST-metric D_{MA}: Part II

Example:



$$D_{MA}(T, T') = 7 - 6 = 1.$$

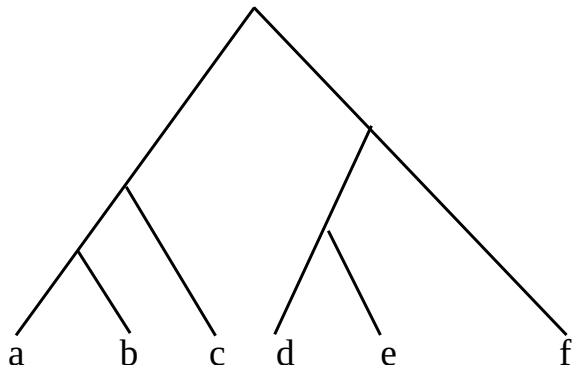
Triplet metric (proper metric!) D_{Tr}: Part I - phylogenetic trees

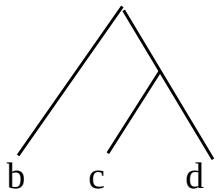
T_1 and T_2 phylogenetic trees on X:

$$\overline{D_{Tr}}(T_1, T_2) = |\mathcal{R}(T_1)\Delta\mathcal{R}(T_2)|/2$$

where for a phylogenetic tree T we put

$$\mathcal{R}(T) = \{\text{triplets on } X \text{ induced by } T\}.$$

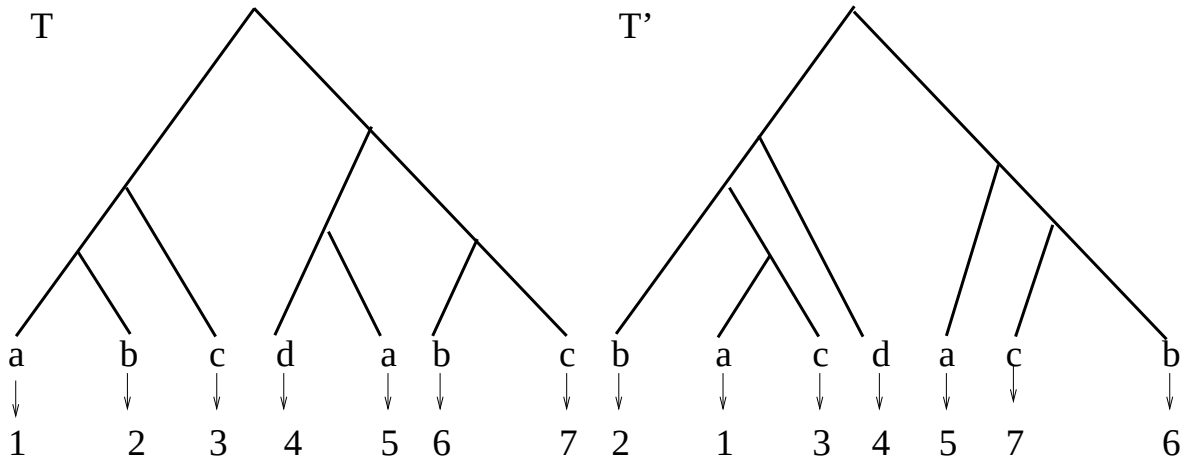




induced triplet

Triplet metric (proper metric!) D_{Tr}: Part II - turning MUL-trees into phylogenetic trees: consistent relabellings

Example:

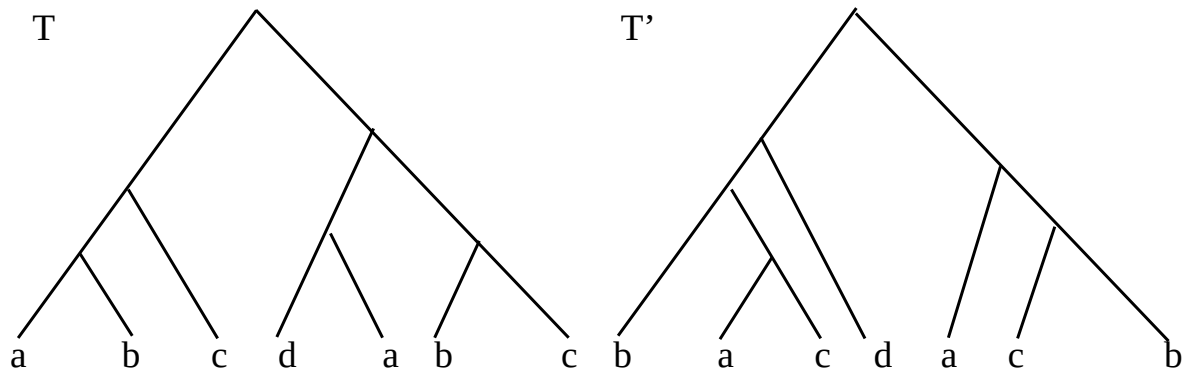


T_1, T_2 MUL-trees on M:

$$D_{Tr}(T_1, T_2) = \min\{\overline{D_{Tr}}(T_1^*, T_2^*) : \ T_1^* \text{ and } T_2^* \text{ consistent relabellings of } (T_1, T_2)\}.$$

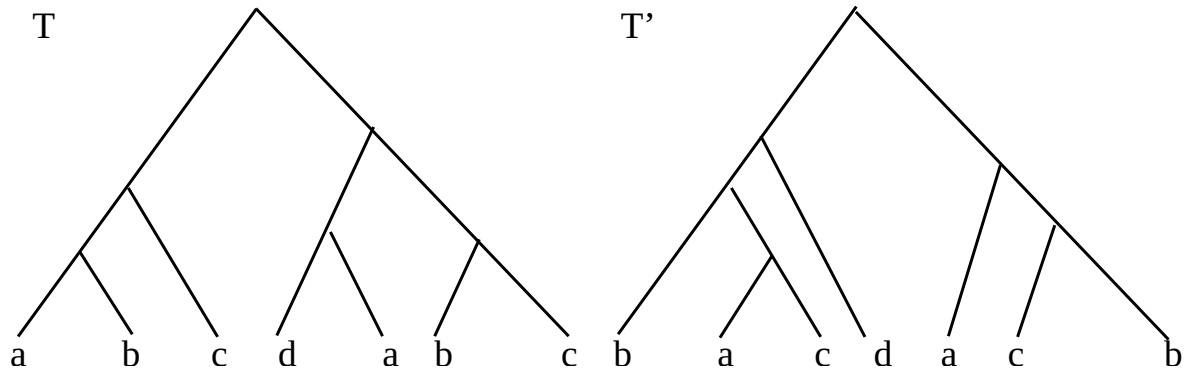
Triplet metric (proper metric!) D_{Tr}: Part III

Example:



$$
\begin{aligned}
D_{Tr}(T, T') &= \min\{\overline{D_{Tr}}(T^*, T'^*) : \ T^* \text{ and } T'^* \\
&\qquad \text{consistent relabellings of } (T, T')\} \\
&= 4.
\end{aligned}
$$

rooted Subtree Prune and Regraft metric D_{SPR} and rooted Nearest Neighbor Interchange metric D_{NNI} (proper metrics!)



T_1, T_2 MUL-trees on M:

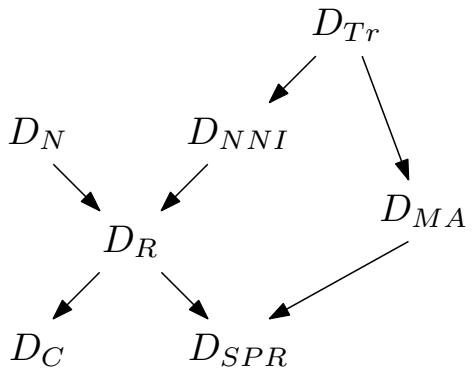
$D_{SPR/NNI}(T_1, T_2)$ = min number of rooted SPR/NNI operations to transform T_1 into T_2.

Example: $D_{SPR}(T, T') = 1$ and $D_{NNI}(T, T') = 3$.

Domination relationships

Theorem (Huber, Moulton Suchcki, Spillner, 2010)

For every multiset M the following diagram depicts the domination relationships for the class of binary MUL-trees on M:



Moreover, there exist multistets M and MUL-trees on M for which the above domination relationships are strict.

Diameter bounds

Theorem (Huber, Moulton Suchcki, Spillner, 2010)

For every multiset M with $|M| \geq 3$ the following hold.

(i) *For $D \in \{D_C, D_R\}$, if $\Delta_\infty(M) < |M|$, then $diam(\mathbf{T}_{bin}(M), D) = |M| - 2$ holds. Otherwise $diam(\mathbf{TS}_{bin}(|M|), D) = |M| - \lceil \log \frac{|M|+2}{3} \rceil - 2$ holds.*

(ii) $$diam(\mathbf{T}_{bin}(M), D_N) = \begin{cases} 0 & \text{if } \Delta_\infty(M) = |M| = 3, \\ |M| - 1 & \text{if } \Delta_\infty(M) < |M|, \\ |M| - 2 & \text{else.} \end{cases}$$

(iii) $$diam(\mathbf{T}_{bin}(M), D_{MA}) = |M| - \max\{2, \lceil \log \Delta_\infty(M) \rceil + 1\}.$$

where: • $\mathbf{T}_{bin}(M)$ is the class of binary MUL-trees on M,
• $\mathbf{TS}_{bin}(n)$ is the set of binary tree shapes with n leaves, and
• $\Delta_\infty(M)$ is the largest number of times an element occurs in M.
E.g. $M = \{a, a, b, c\}$ then $\Delta_\infty(M) = 2$.