Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Lehrstuhl für
INFORMATIK I
Effiziente Algorithmen und
wissensbasierte Systeme

Institut für Informatik

Advanced Algorithms

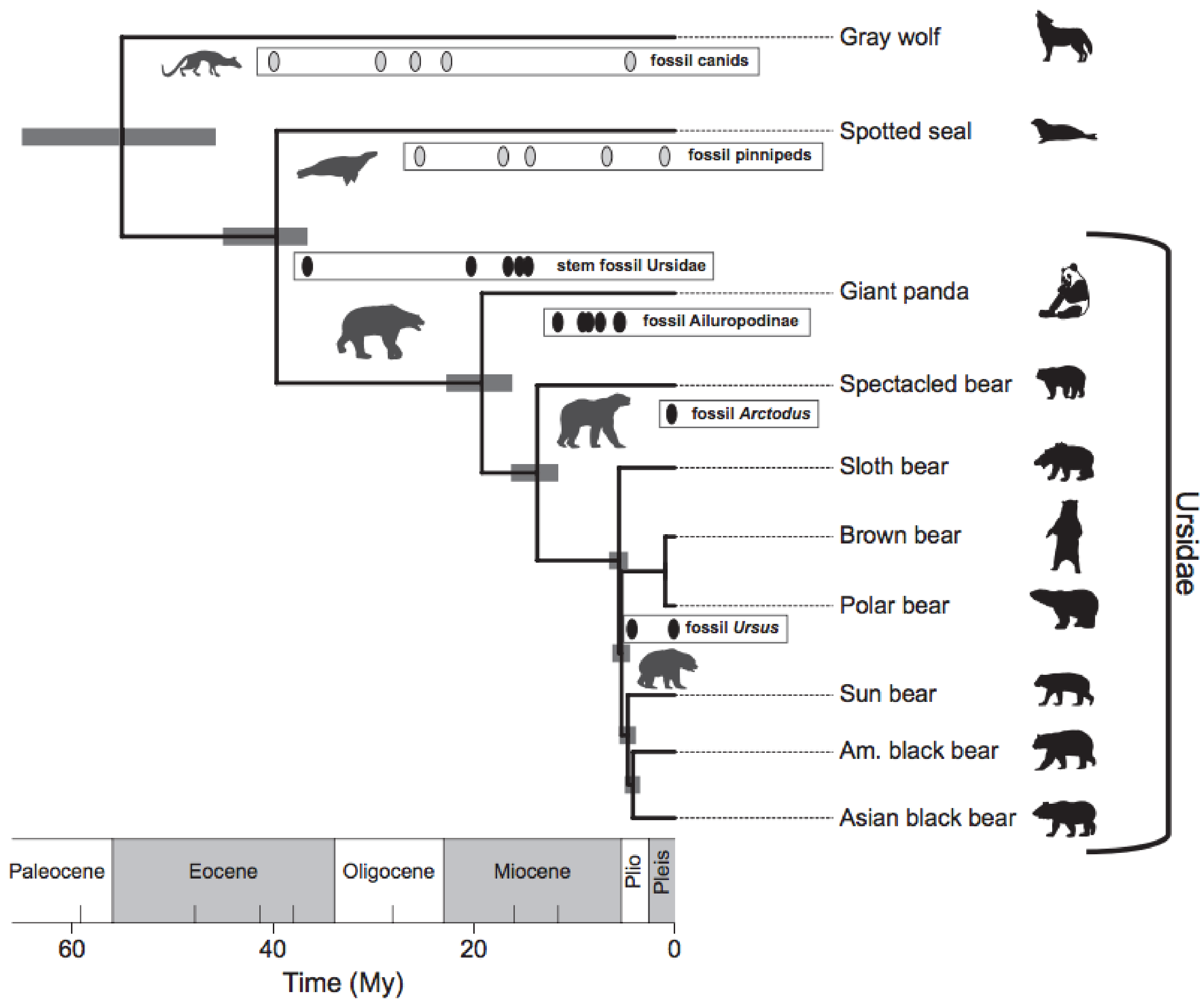
Winter term 2019/20

Lecture 12. Rearrangement distance of phylogenetic trees

Jonathan Klawitter

Chair for Computer Science I

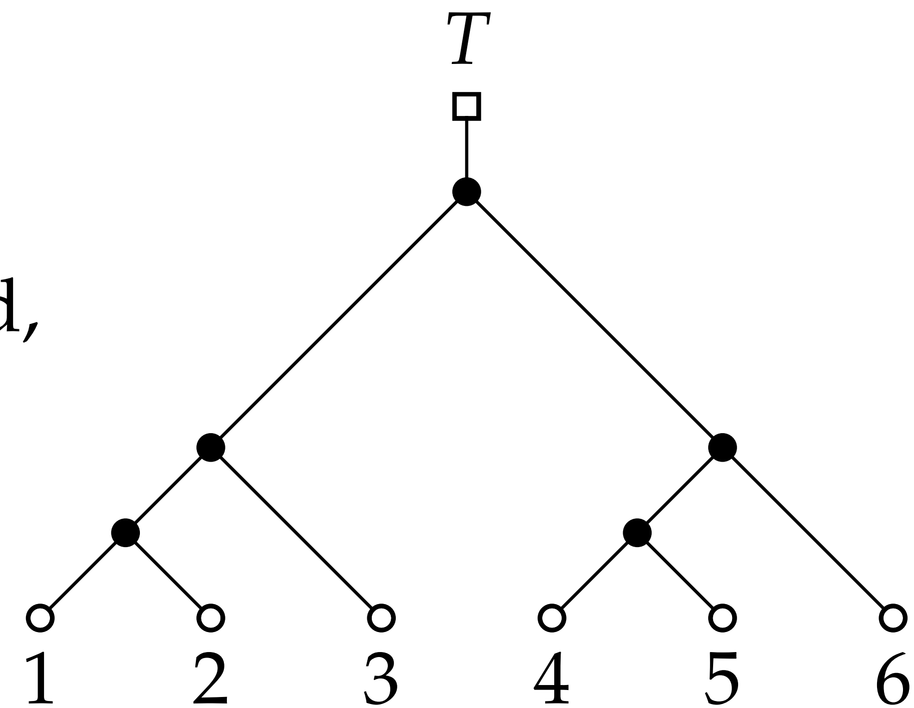
Phylogenetic trees



Phylogenetic trees

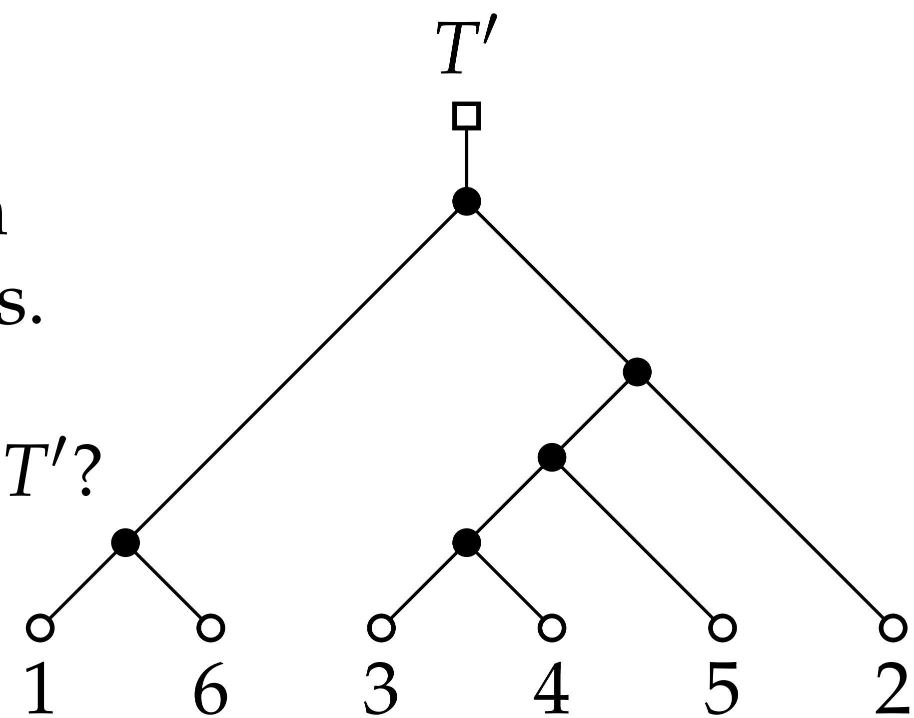
Let $X = \{1, 2, \ldots, n\}$.

A phylogenetic tree T is a rooted, binary tree where the leaves are bijectively labelled with X.



Inference methods compute a phylogenetic tree based on some model and data.

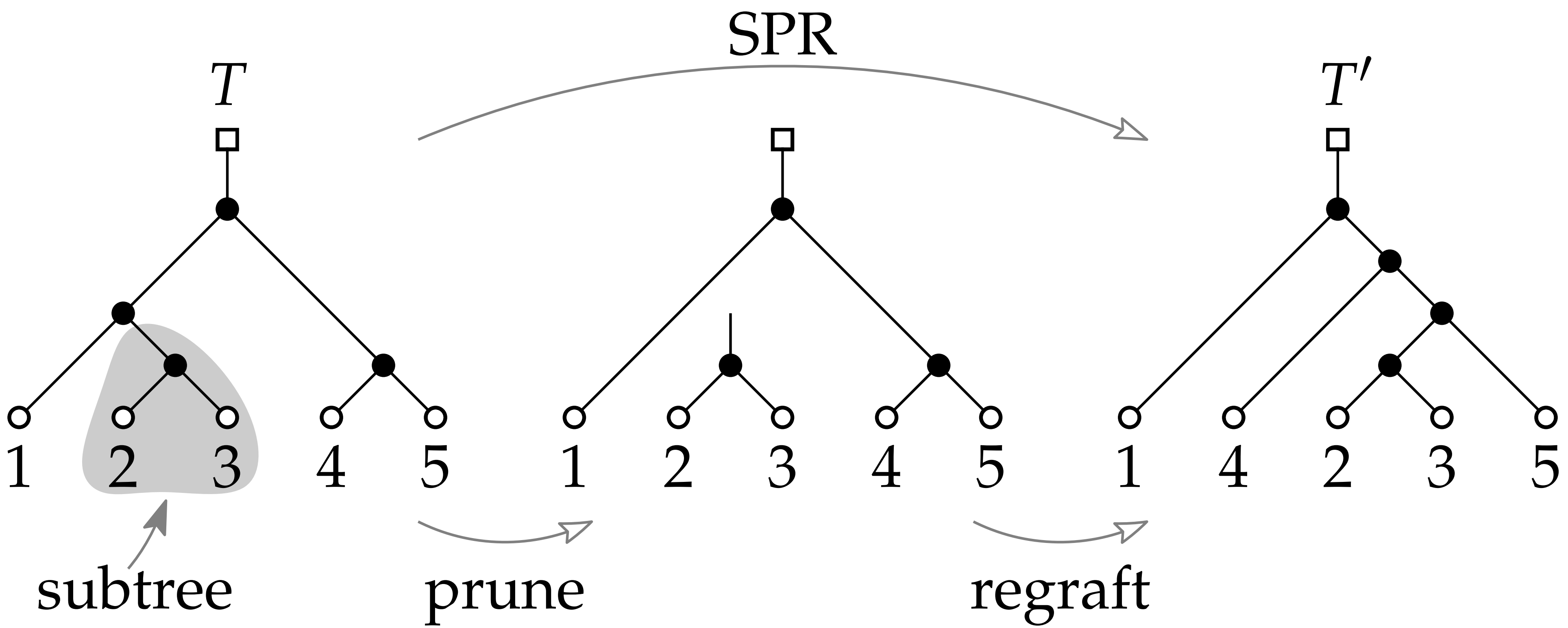
Different methods/models/data yield different phylogenetic trees.



- How can we compare T and T'?

→ We want a metric on phylogenetic trees.

Subtree Prune & Regraft (SPR)



Define SPR-rearrangement graph $G = (V, E)$ with

- $V = \{$ all phylogenetic trees on X $\}$ and
- $\{T, T'\} \in E$ if T can be transformed into T' with an SPR.

SPR-distance

Define the SPR-distance of T and T' as

$$\mathrm{d}_{\mathrm{SPR}}(T, T') = \text{ distance of } T \text{ and } T' \text{ in } G.$$

Lemma. The SPR-rearrangement graph G is connected.

Proof. See blackboard (or exercise).

Corollary. The SPR-distance is a metric.

Proof. G is connected and undirected.

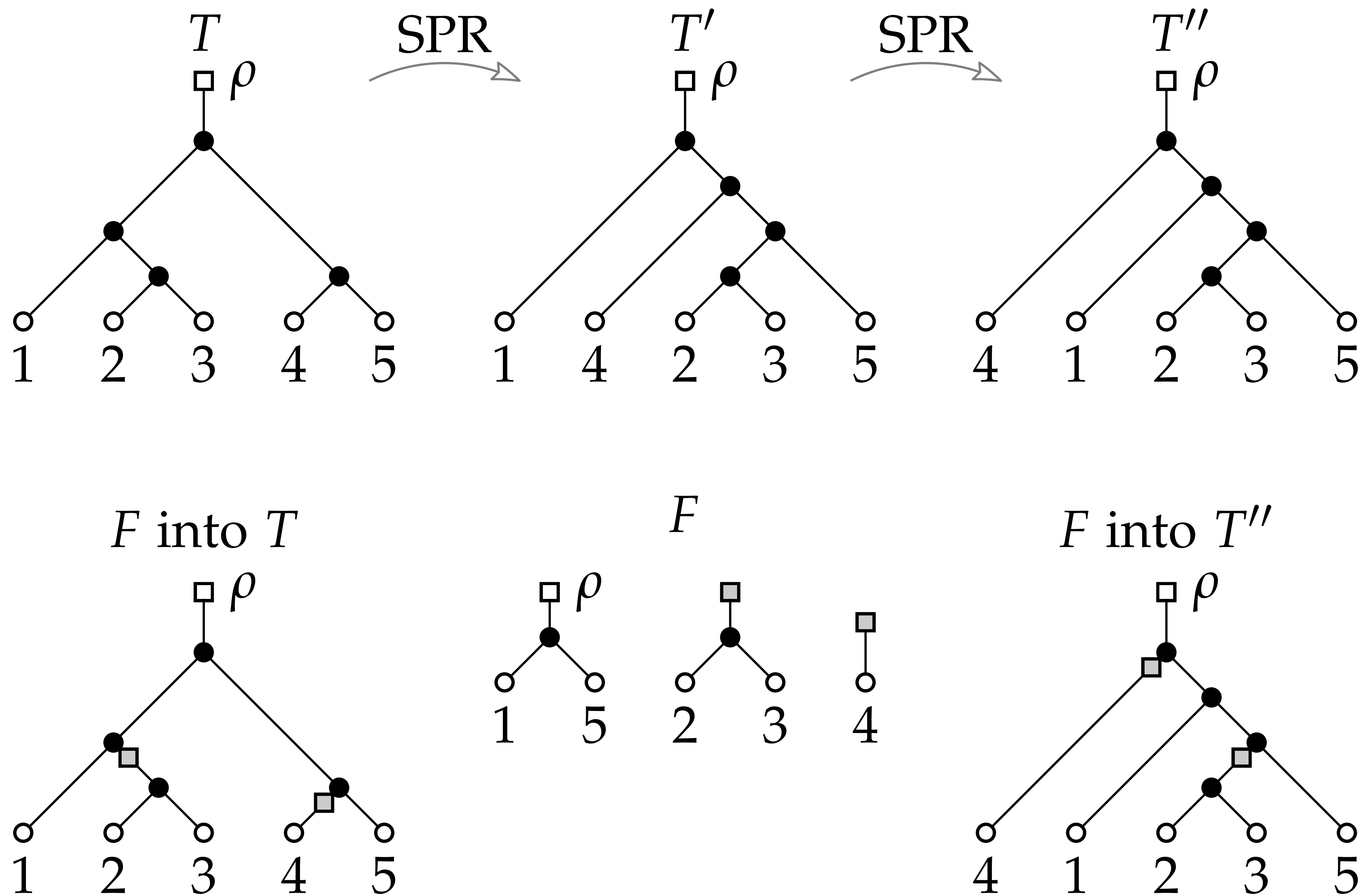
Goal. Compute the SPR-distance $\mathrm{d}_{\mathrm{SPR}}(T, T')$.

Problem. G is huge!

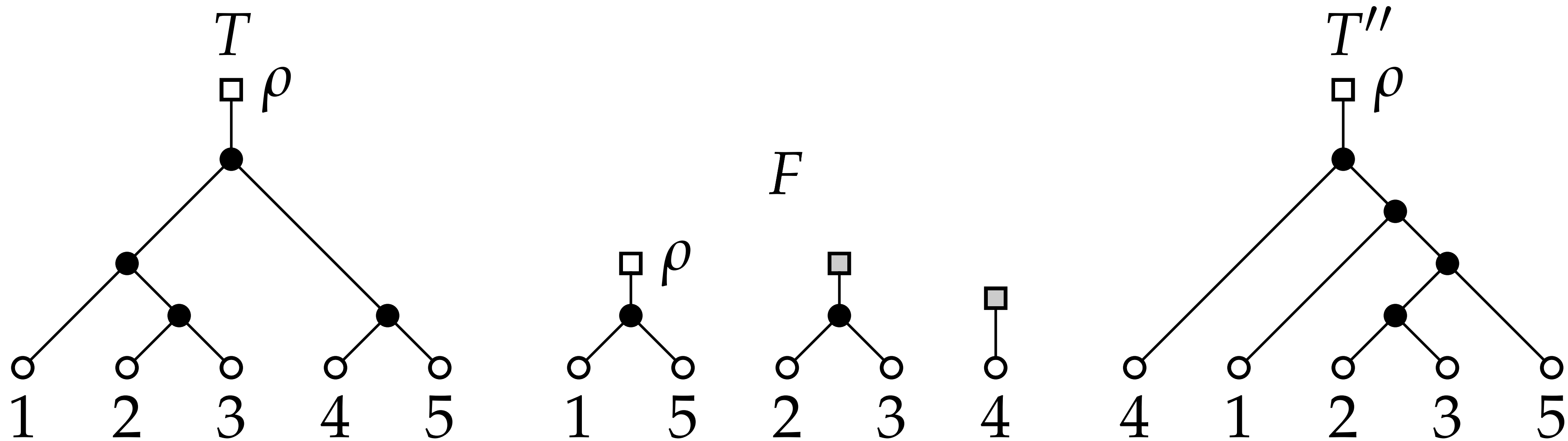
$|V(G)| = (2n-3)!! = (2n-3) \cdot (2n-5) \cdot \ldots \cdot 5 \cdot 3$

- Can we rephrase the problem?

Maximum agreement forests



Maximum agreement forests



An agreement forest F of T and T'' is a forest $\{T_\rho, T_1, T_2, \ldots, T_k\}$ such that

- label sets of the T_i partition $X \cup \{\rho\}$,
- ρ is in label set of T_ρ, and
- there exist edge-disjoint embeddings of subdivisions of the T_i's into T and T'' that cover all edges.

If k is minimal, F is a maximum agreement forest (MAF).

Characterisation

Let $F = \{T_\rho, T_1, T_2, \ldots, T_k\}$ be a MAF of T and T'.
Then define

$$m(T, T') = k.$$

Theorem. Let T and T' be two phylogenetic trees on X. Then

$$m(T, T') = d_{\mathrm{SPR}}(T, T').$$

Proof. See blackboard.

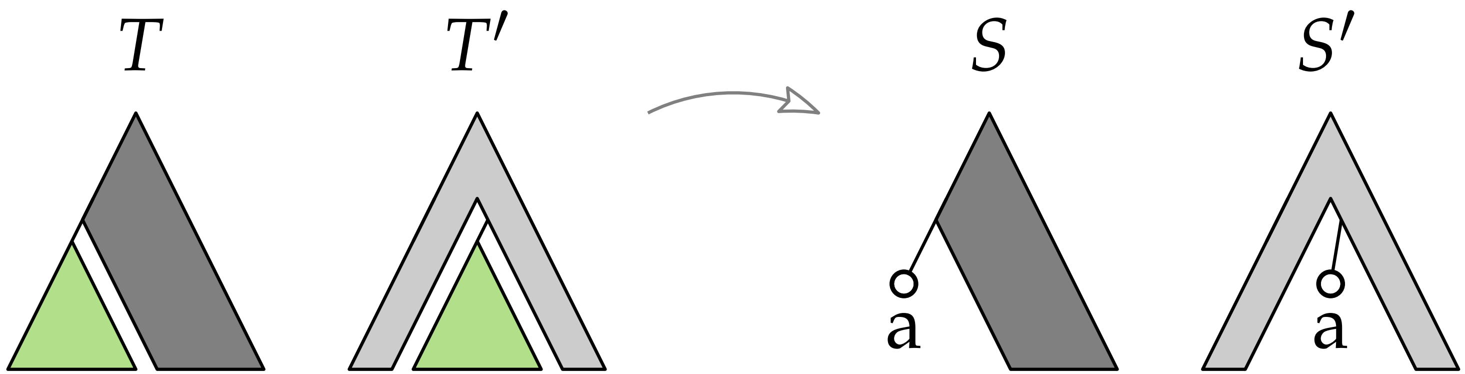
Theorem. Computing the SPR-distance of T and T' is NP-hard.

Proof is via reduction from Exact Cover by 3-Sets.

See Bordewich, Semple, "On the computational complexity of the rooted subtree prune and regraft distanc" and Hein et al., "On the complexity of comparing evolutionary trees" for details.

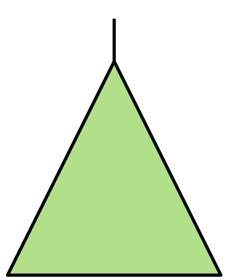
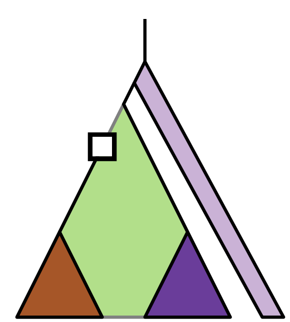
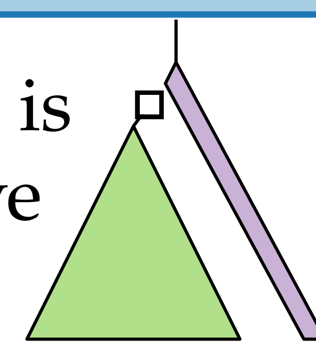
Kernelisation (1 of 2)

- **Common subtree reduction**:
  Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label.



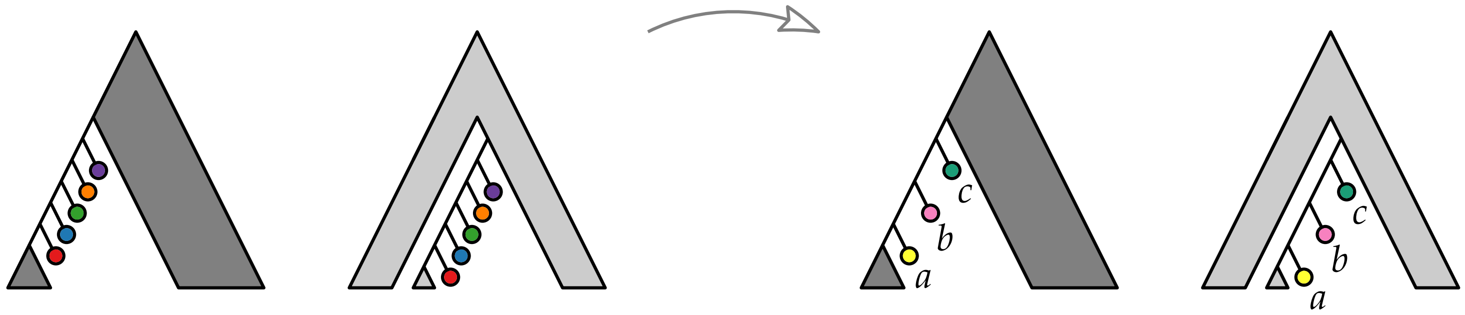
Lemma. Applying common subtree reduction is safe; i.e. $\mathrm{d}_{\mathrm{SPR}}(T, T') = \mathrm{d}_{\mathrm{SPR}}(S, S')$.

Proof.

Suppose  is covered by two trees of MAF  then there is alternative MAF 

Kernelisation (2 of 2)

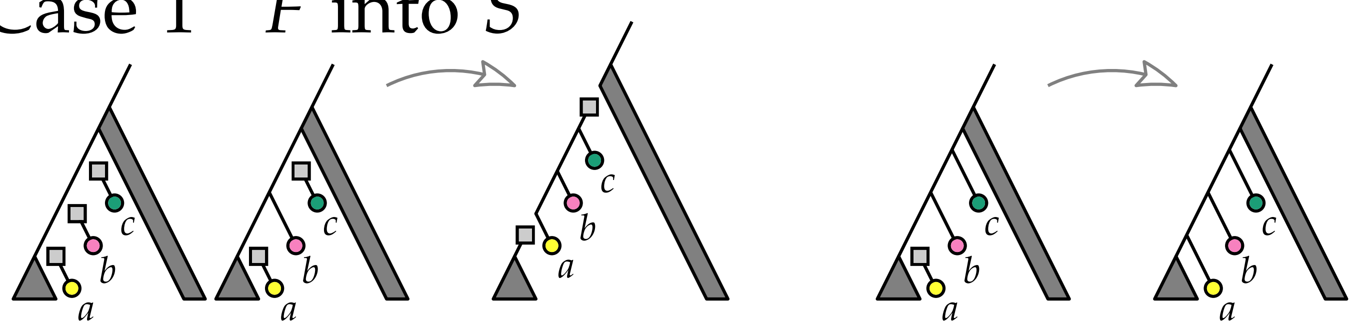
- Chain reduction:
  Replace any chain of leaves that occurs identically in both trees by three new leaves.



Lemma. Applying chain reduction is safe; i.e. $d_{\mathrm{SPR}}(T, T') = d_{\mathrm{SPR}}(S, S')$.

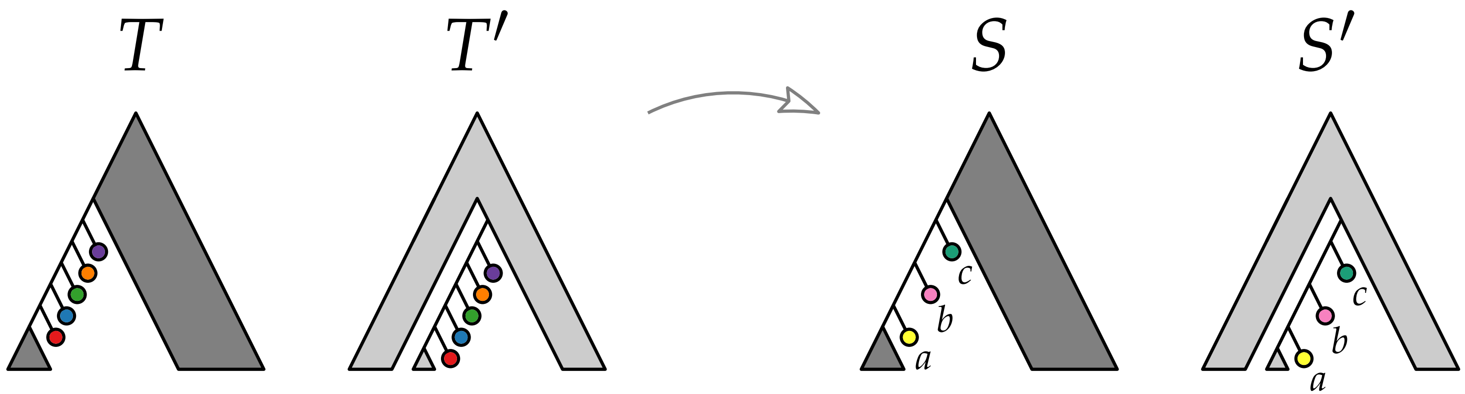
Proof.
Show there is a tree with abc-chain in a MAF.

Case 1 F into S



Kernelisation (2 of 2)

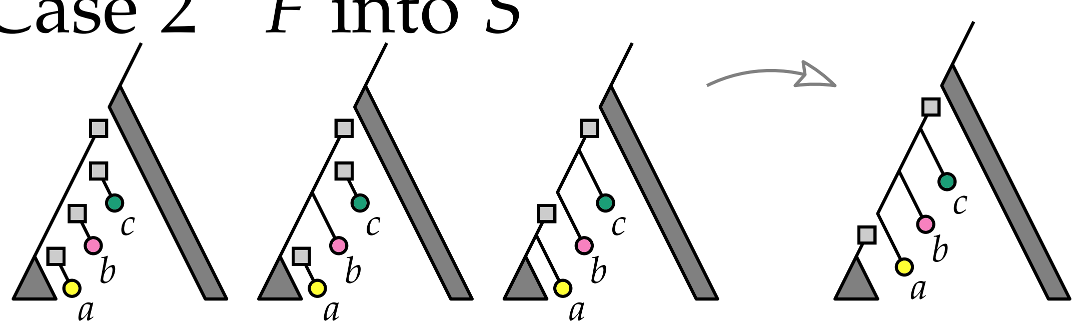
- **Chain reduction**:
  Replace any chain of leaves that occurs identically in both trees by three new leaves.



Lemma. Applying chain reduction is safe; i.e. $d_{SPR}(T, T') = d_{SPR}(S, S')$.

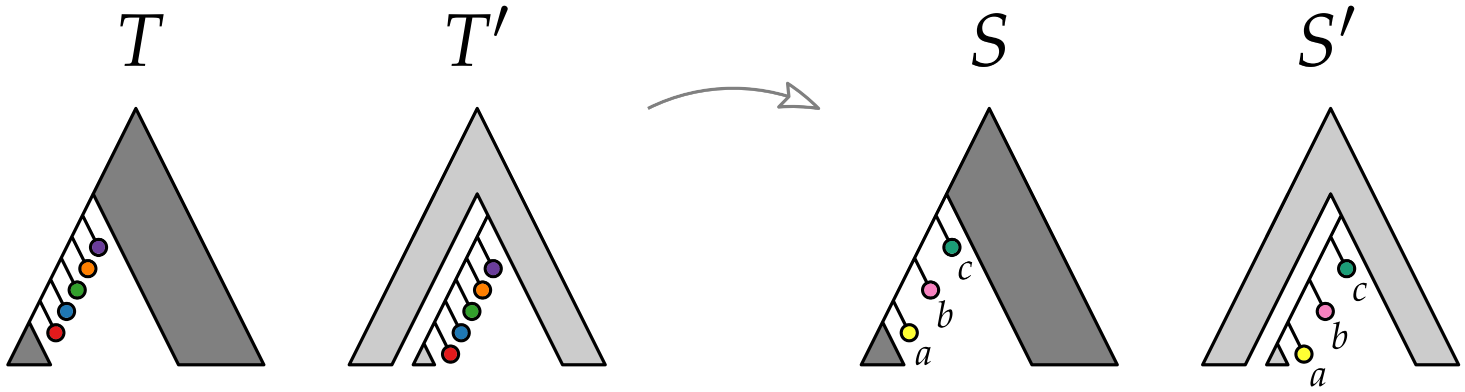
Proof.
Show there is a tree with abc-chain in a MAF.

Case 2 F into S



Kernelisation (2 of 2)

- Chain reduction:
  Replace any chain of leaves that occurs identically in both trees by three new leaves.



Lemma. Applying chain reduction is safe; i.e. $d_{\mathrm{SPR}}(T, T') = d_{\mathrm{SPR}}(S, S')$.

Proof.

Show there is a tree with abc-chain in a MAF.

Swap abc-chain with original chain for MAF of T and T'.

Kernelisation and fpt algorithm

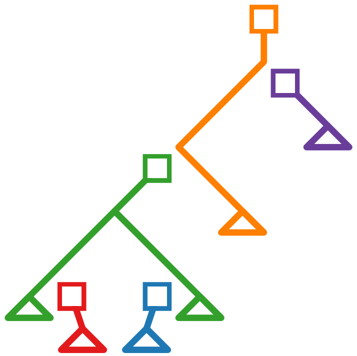
Theorem. Reduce T and T' to S and S' by exhaustively applying the reduction rules.
Let S and S' be on X'. Then

$$|X'| \leq 28\, \mathrm{d_{SPR}}(T, T').$$

Proof. Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for S and S'.

Let $n(T_i)$ be # T_j it overlaps with in embedding of F into T.

Claim 1. $\sum_{i=\rho}^{k}(n(T_i) + n'(T_i)) \leq 4k = 4\, \mathrm{d_{SPR}}(T, T')$.



Kernelisation and fpt algorithm

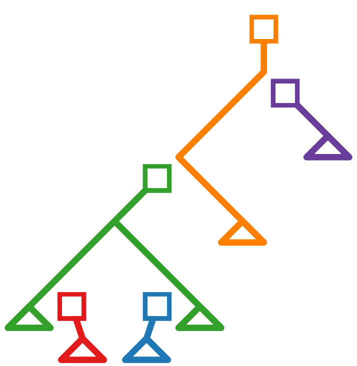
Theorem. Reduce T and T' to S and S' by exhaustively applying the reduction rules.
Let S and S' be on X'. Then

$$|X'| \leq 28\,\mathrm{d}_{\mathrm{SPR}}(T, T').$$

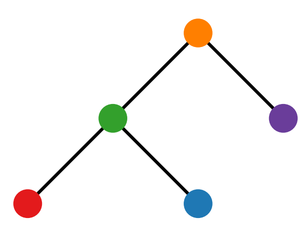
Proof. Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for S and S'.

Let $n(T_i)$ be # T_j it overlaps with in embedding of F into T.

Claim 1. $\sum_{i=\rho}^{k}(n(T_i) + n'(T_i)) \leq 4k = 4\,\mathrm{d}_{\mathrm{SPR}}(T, T')$.



H



$$|V(H)| = k + 1$$
$$= |E(H)| + 1$$

$$\sum n(T_i) = 2|E(H)| \leq 2k$$

Kernelisation and fpt algorithm

Theorem. Reduce T and T' to S and S' by exhaustively applying the reduction rules.
Let S and S' be on X'. Then

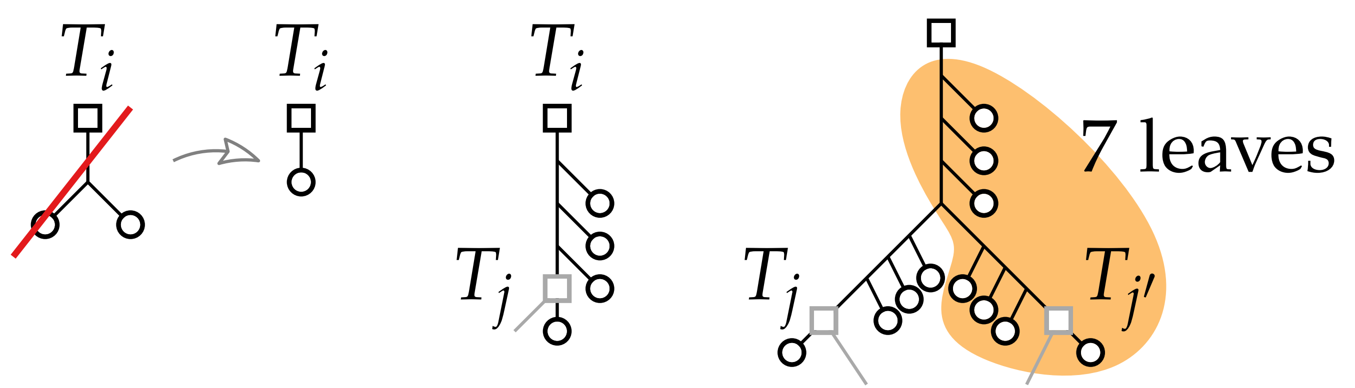
$$|X'| \leq 28\,\mathrm{d}_{\mathrm{SPR}}(T, T').$$

Proof. Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for S and S'.

Let $n(T_i)$ be # T_j it overlaps with in embedding of F into T.

Claim 1. $\sum_{i=\rho}^{k}(n(T_i) + n'(T_i)) \leq 4k = 4\,\mathrm{d}_{\mathrm{SPR}}(T, T')$.

Claim 2. # leaves of $T_i \leq 7(n(T_i) + n'(T_i))$.



Kernelisation and fpt algorithm

Theorem. Reduce T and T' to S and S' by exhaustively applying the reduction rules.
Let S and S' be on X'. Then

$$|X'| \leq 28\,\mathrm{d_{SPR}}(T, T').$$

Proof. Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for S and S'.

Let $n(T_i)$ be # T_j it overlaps with in embedding of F into T.

Claim 1. $\sum_{i=\rho}^{k}(n(T_i) + n'(T_i)) \leq 4k = 4\,\mathrm{d_{SPR}}(T, T')$.

Claim 2. # leaves of $T_i \leq 7(n(T_i) + n'(T_i))$.

$\sum_{i=\rho}^{k}$ # leaves of $T_i \leq$

Kernelisation and fpt algorithm

Theorem. Reduce T and T' to S and S' by exhaustively applying the reduction rules.
Let S and S' be on X'. Then

$$|X'| \leq 28\,\mathrm{d}_{\mathrm{SPR}}(T, T').$$

Proof. Let $F = \{T_\rho, T_1, \ldots, T_k\}$ be MAF for S and S'.

Let $n(T_i)$ be # T_j it overlaps with in embedding of F into T.

Claim 1. $\sum_{i=\rho}^{k}(n(T_i) + n'(T_i)) \leq 4k = 4\,\mathrm{d}_{\mathrm{SPR}}(T, T')$.

Claim 2. # leaves of $T_i \leq 7(n(T_i) + n'(T_i))$.

$\sum_{i=\rho}^{k}$ # leaves of $T_i \leq \sum_{i=\rho}^{k} 7(n(T_i) + n'(T_i)) \leq 28k$.

Kernelisation and fpt algorithm

Theorem. Reduce T and T' to S and S' by exhaustively applying the reduction rules.
Let S and S' be on X'. Then

$$|X'| \leq 28\, \mathrm{d_{SPR}}(T, T').$$

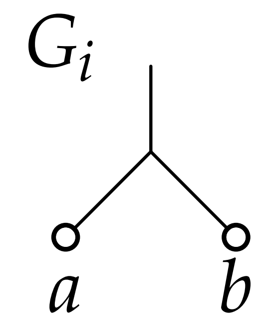
Corollary. Computing $\mathrm{d_{SPR}}(T, T')$ is fixed-parameter tractable when parameterized by $\mathrm{d_{SPR}}(T, T')$.

Proof.
- Reduce T and T' to S and S'. Let $k = \mathrm{d_{SPR}}(S, S')$.
- S has at most $4|X'|^2$ neighbours.
  - S has $\leq 2|X'|$ edges to cut and attach to.
- Length-k BFS from S visits at most $O((4|X'|^2)^k) = O((56k)^{2k})$ trees.

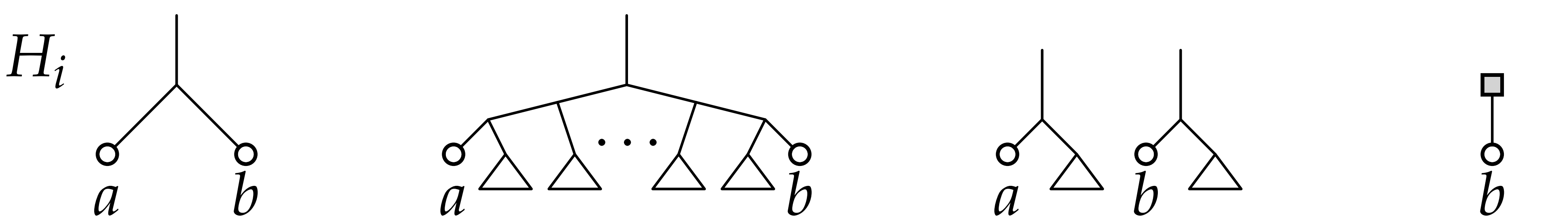
Approximation algorithm

Algorithm: dSPRApprox(T, T')

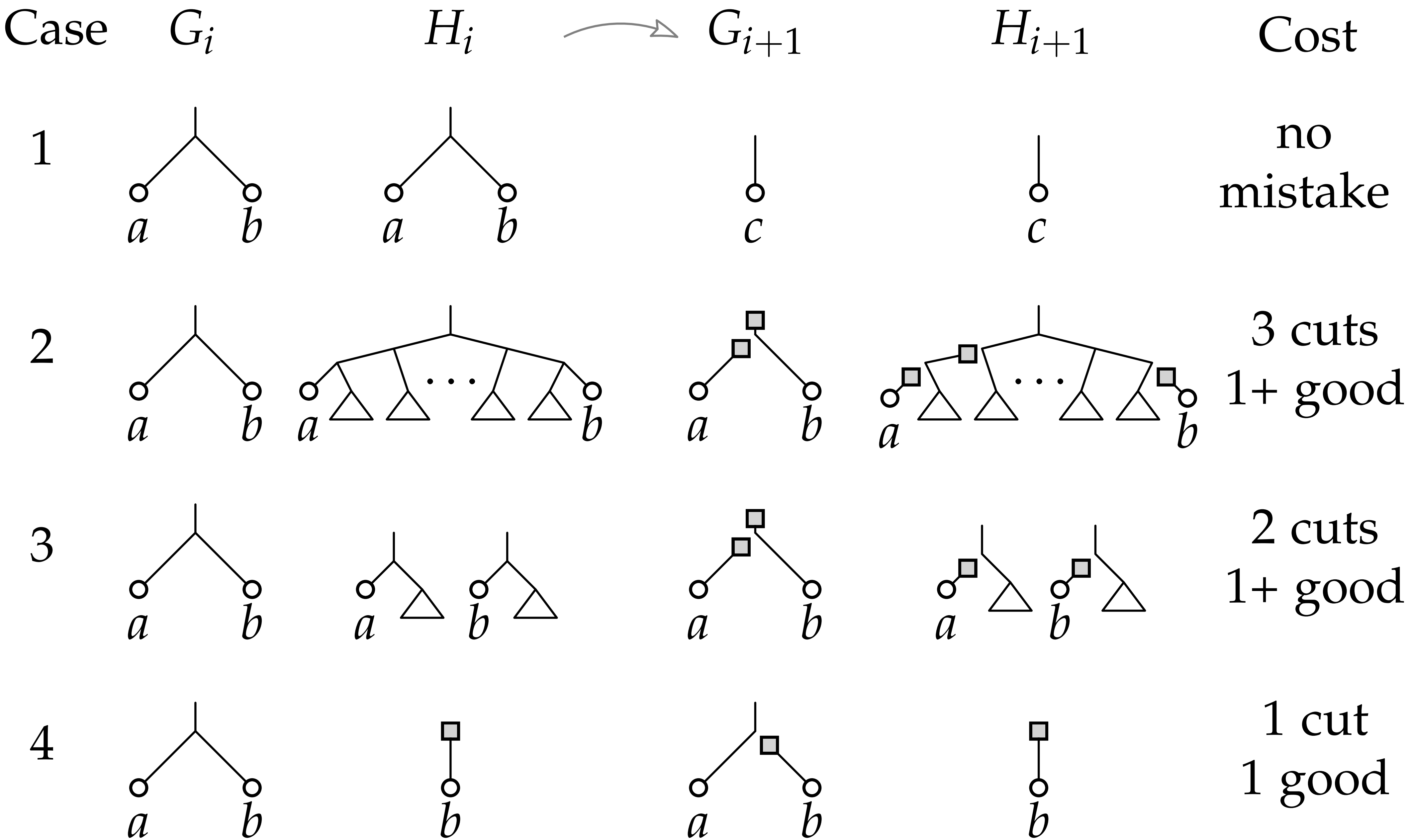
$i \leftarrow 1$
$G_i \leftarrow T$
$H_i \leftarrow T'$
while $\exists$ pair of sibling leaves a and b in G_i **do**
    find the case that applies to a and b in H_i
    apply the corresponding transaction
    to obtain G_{i+1} from G_i and H_{i+1} from H_i
    $i++$
return H_i



Case 1 | Case 2 | Case 3 | Case 4

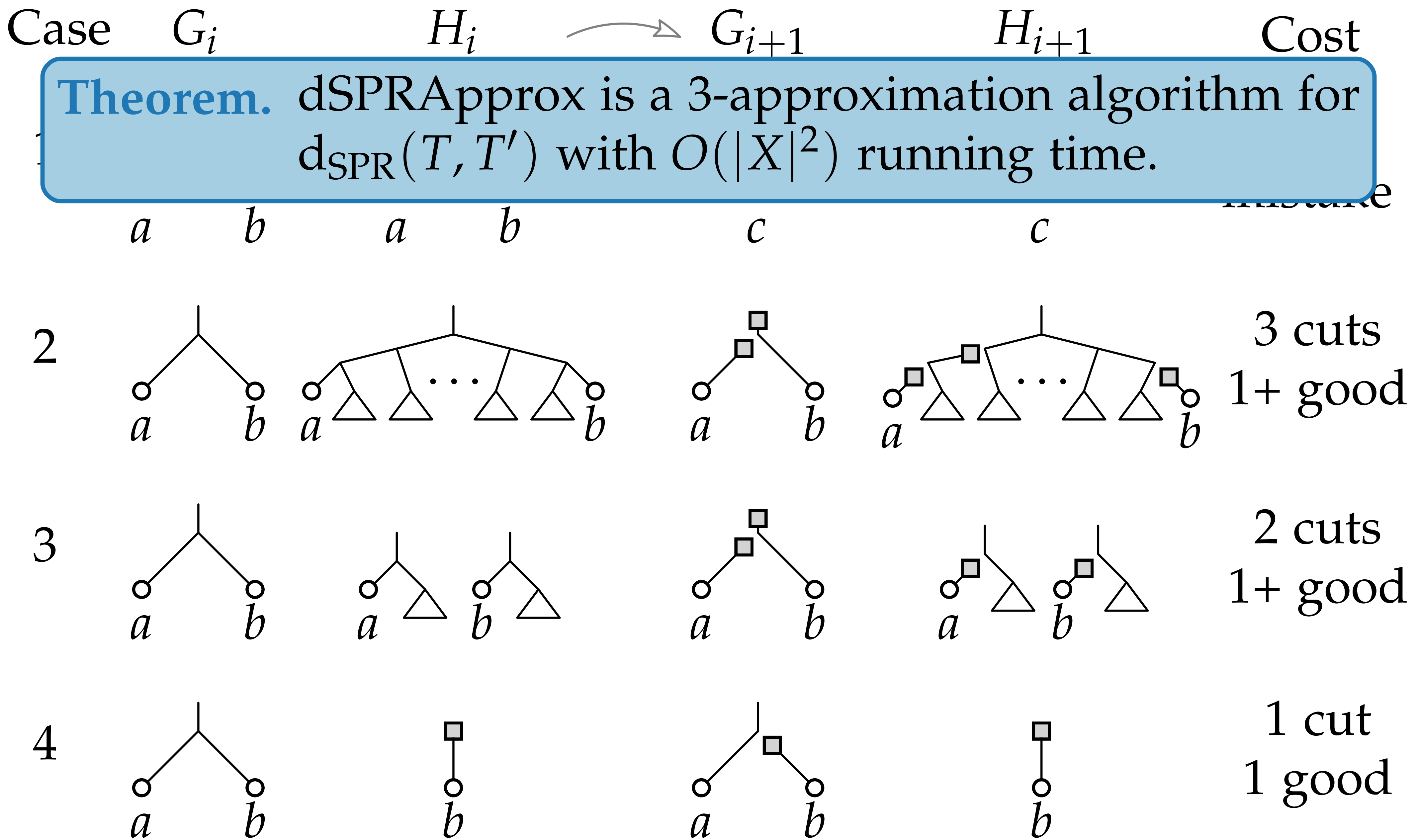


Approximation algorithm



Approximation algorithm

Theorem. dSPRApprox is a 3-approximation algorithm for $d_{\mathrm{SPR}}(T, T')$ with $O(|X|^2)$ running time.



References

- Bordewich, Semple, “On the computational complexity of the rooted subtree prune and regraft distance”, 2005
  for SPR, MAF, characterisation, fpt, divide & conquer

- Hein et al., “On the complexity of comparing evolutionary trees”, 1996
  for NP-hardness proof

- Rodrigues et al., “The maximum agreement forest problem: Approximation algorithms and computational experiments”, 2006