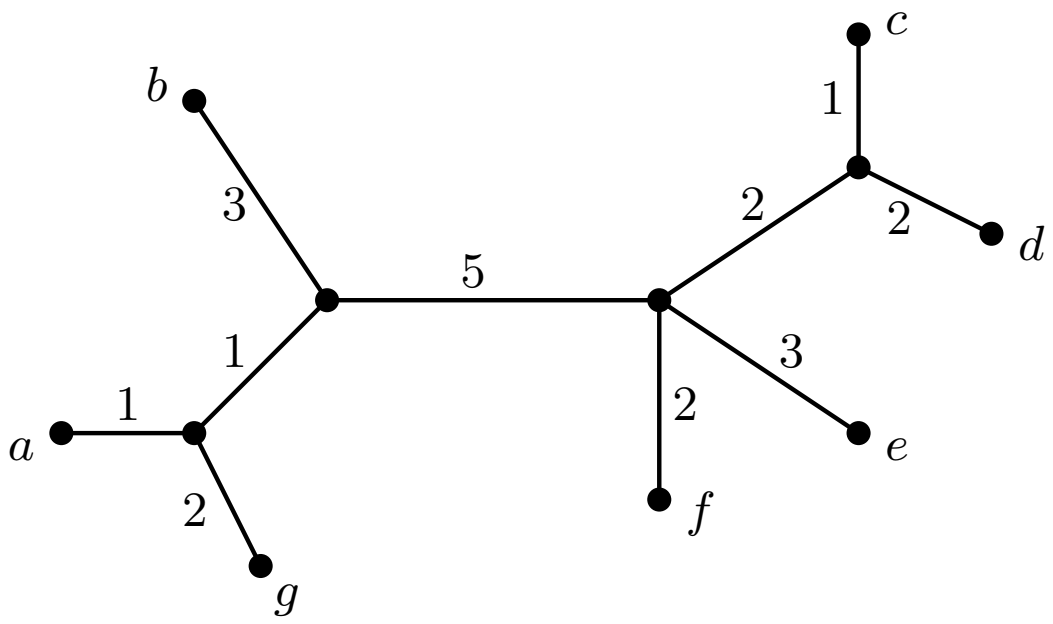
Consistency and Balanced Minimum Evolution

Newton Institute – Sept. '07

Magnus Bordewich

Joint work with O. Gascuel, K. Huber, V. Moulton

Reconstructing trees from a distance matrix



We are interested in the evolutionary history of a set of taxa X.

We are given an estimated distance matrix $\Delta = [\delta_{ij}]$.

Goal: to reconstruct the true phylogenetic tree $\mathcal{T}^*$ for the set X.

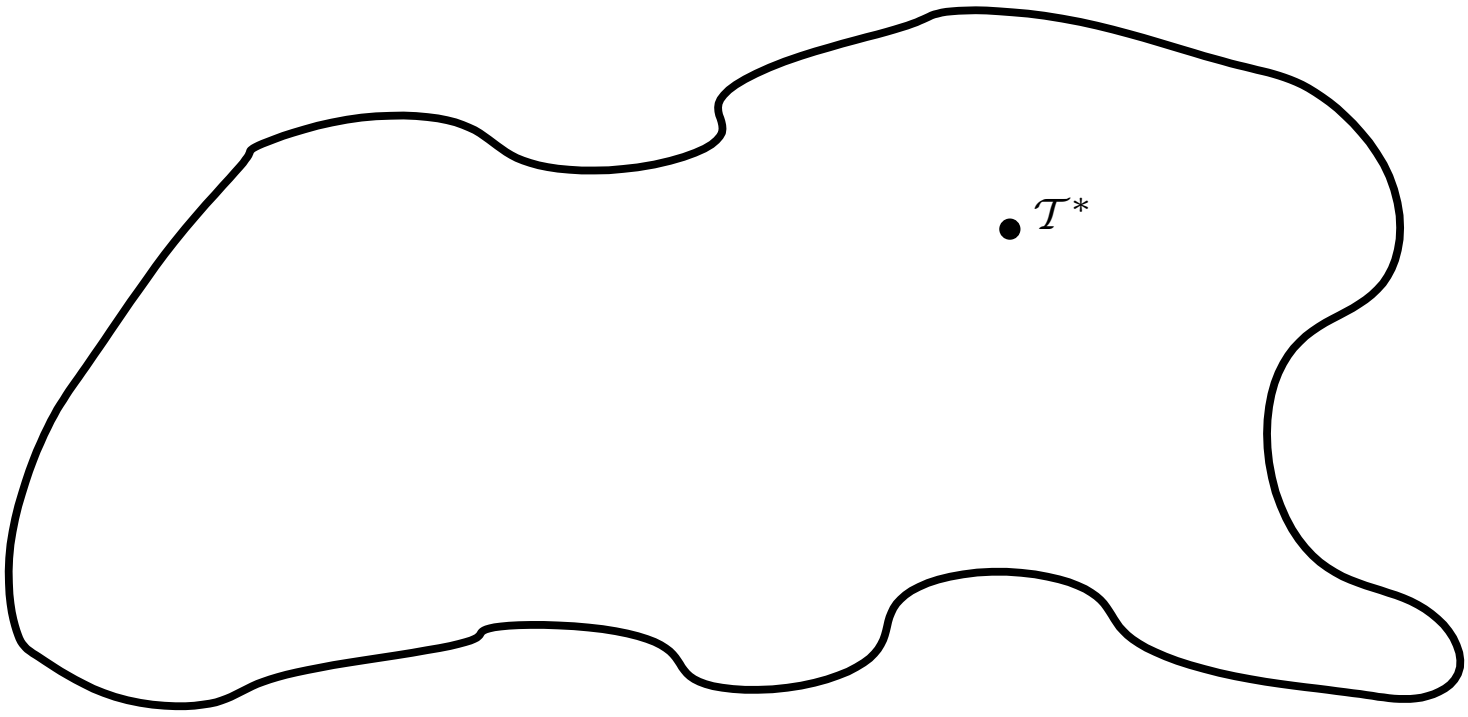
Balanced Minimum Evolution (BME) principle

BME introduced by Desper and Gascuel (2002), based on a tree length estimation scheme of Pauplin (2000).

- We are given the estimated matrix Δ.
- For a topology $\mathcal{T}$ we can estimate the branch lengths ($\hat{l}(e)$).
- We want the topology of smallest total length ($\sum_e \hat{l}(e)$).
- NJ greedily minimises this tree length.

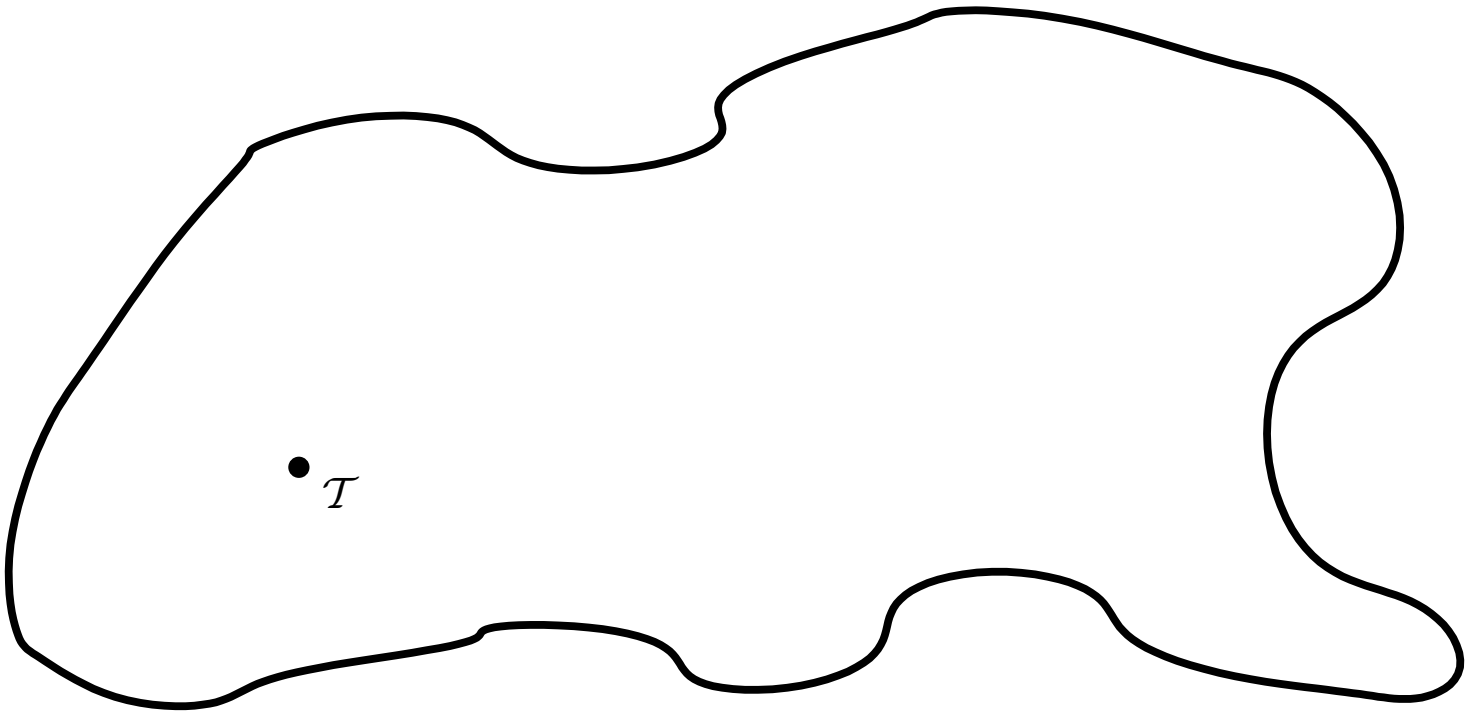
Idea: find this minimal tree using a local topology search.

Local topology search



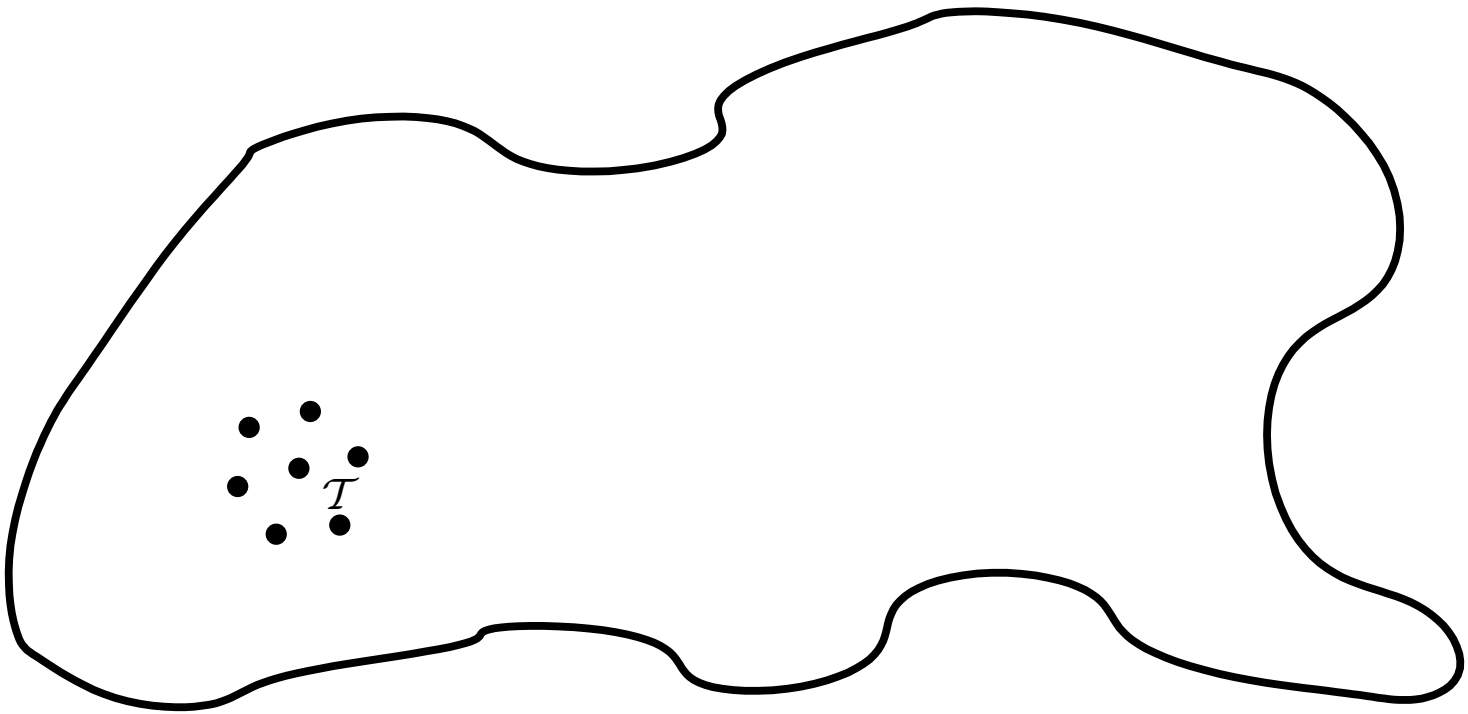
Local topology search

1. Compute a starting tree $\mathcal{T}$ (e.g. using a greedy method).



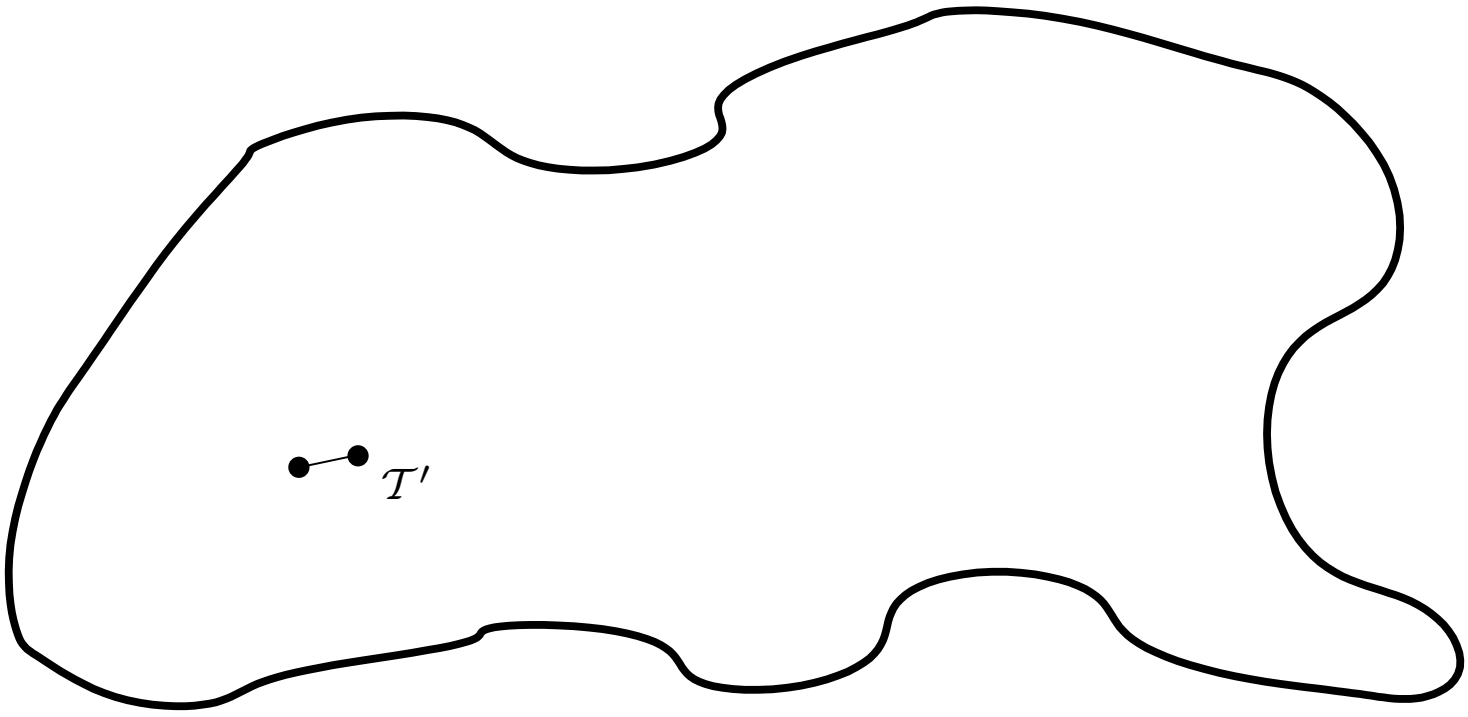
Local topology search

1. Compute a starting tree $\mathcal{T}$ (e.g. using a greedy method).
2. Then iteratively
   - check all topologies one NNI (or SPR) move away from $\mathcal{T}$



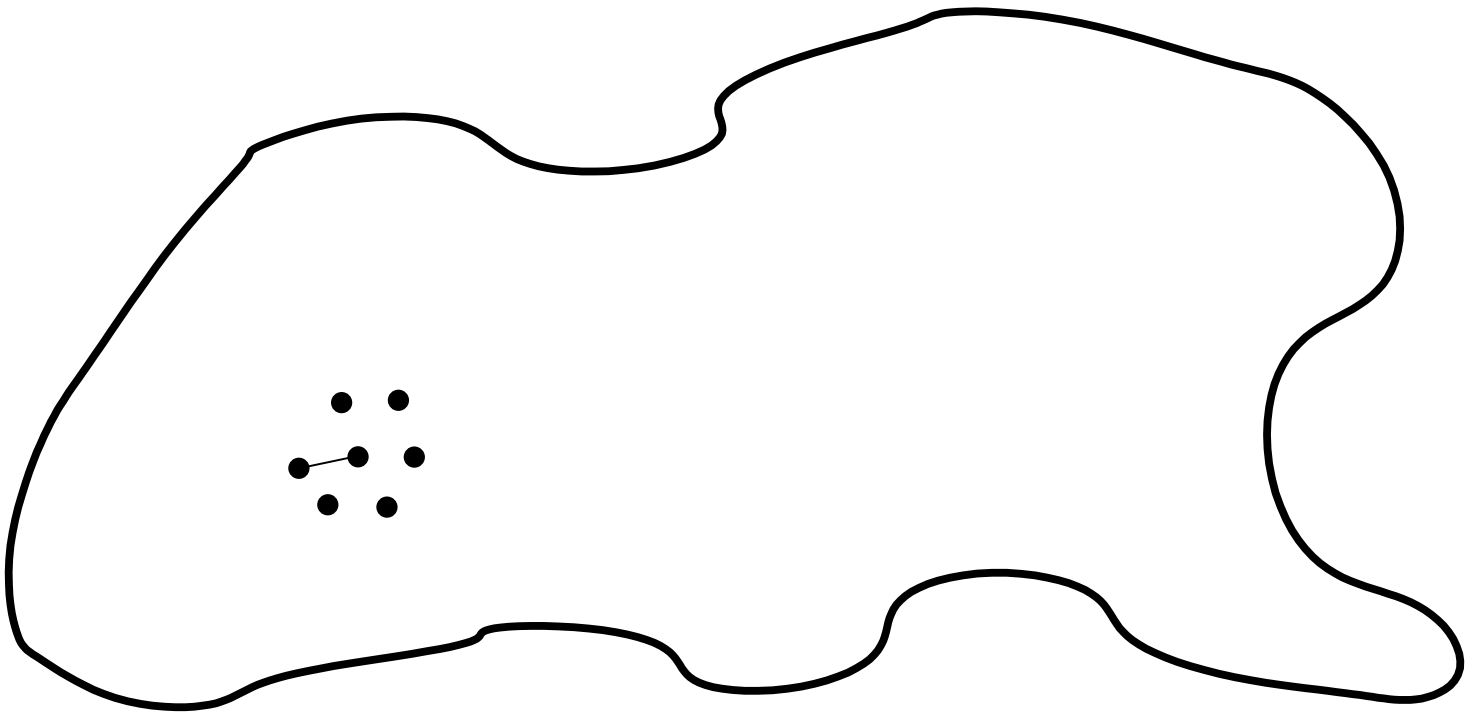
Local topology search

1. Compute a starting tree $\mathcal{T}$ (e.g. using a greedy method).
2. Then iteratively
   - check all topologies one NNI (or SPR) move away from $\mathcal{T}$
   - move to the topology $\mathcal{T}'$ that minimizes the tree length.



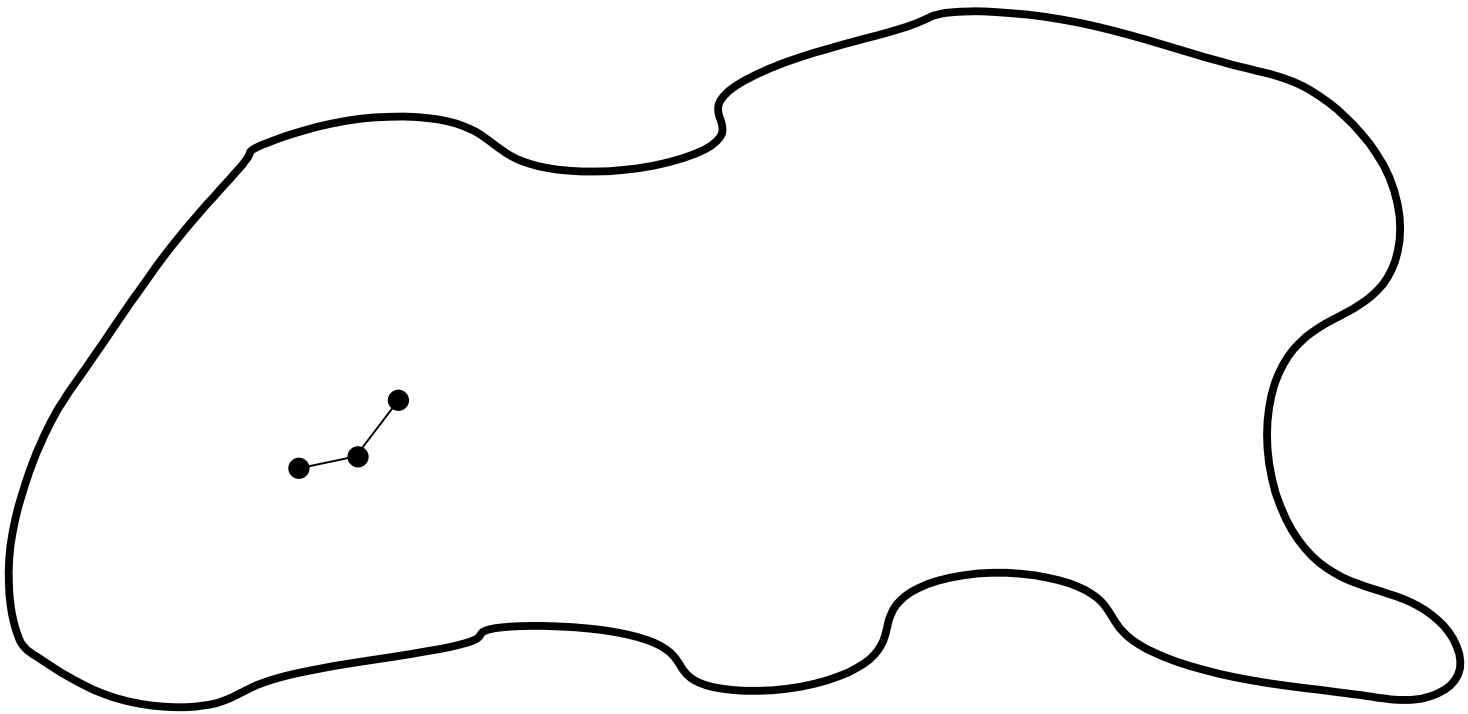
Local topology search

1. Compute a starting tree $\mathcal{T}$ (e.g. using a greedy method).
2. Then iteratively
   - check all topologies one NNI (or SPR) move away from $\mathcal{T}$
   - move to the topology $\mathcal{T}'$ that minimizes the tree length.



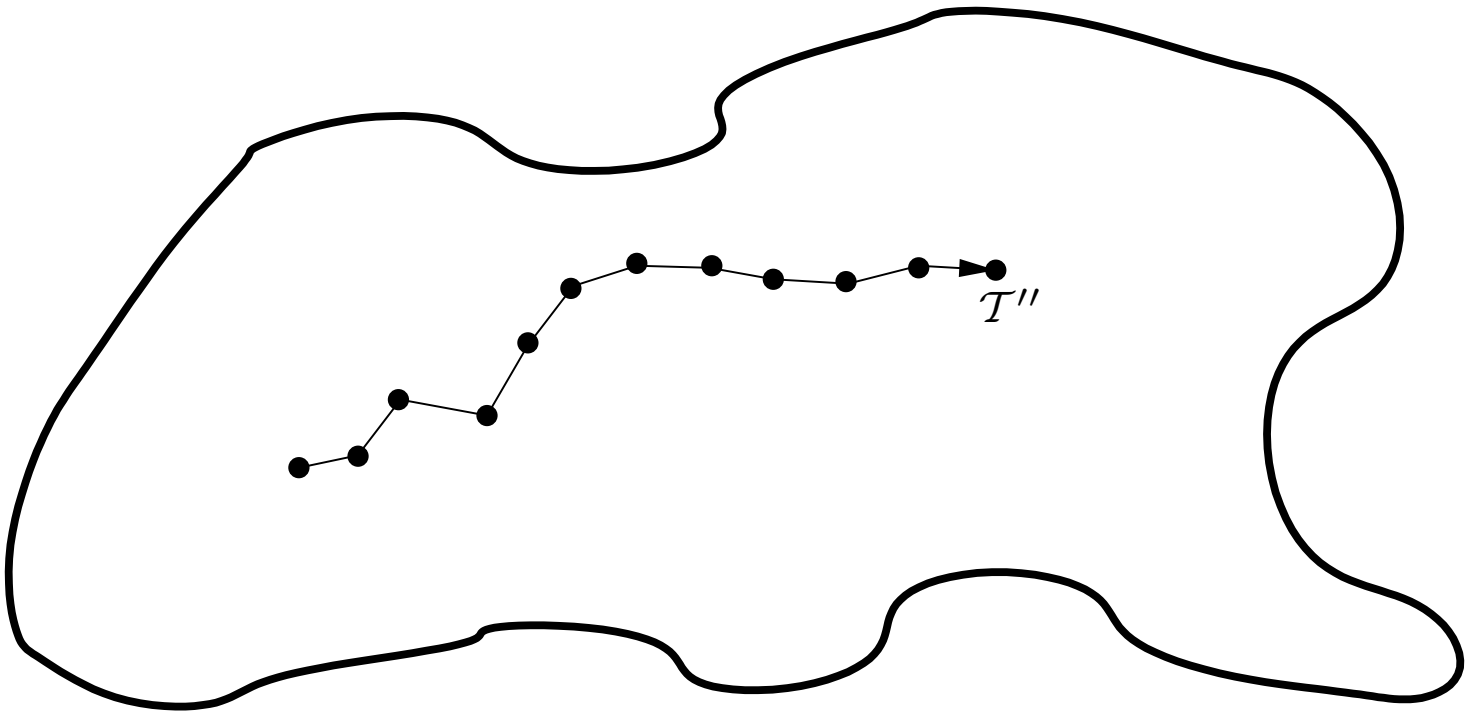
Local topology search

1. Compute a starting tree $\mathcal{T}$ (e.g. using a greedy method).
2. Then iteratively
   - check all topologies one NNI (or SPR) move away from $\mathcal{T}$
   - move to the topology $\mathcal{T}'$ that minimizes the tree length.



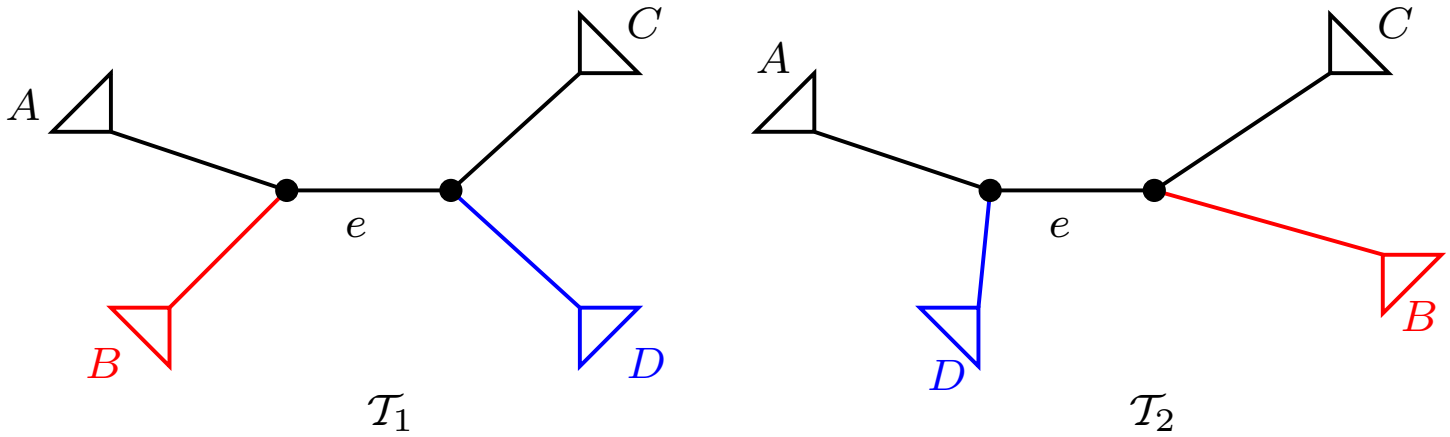
Local topology search

1. Compute a starting tree $\mathcal{T}$ (e.g. using a greedy method).
2. Then iteratively
   - check all topologies one NNI (or SPR) move away from $\mathcal{T}$
   - move to the topology $\mathcal{T}'$ that minimizes the tree length.
3. Output the final topology: a local minimum in tree length.

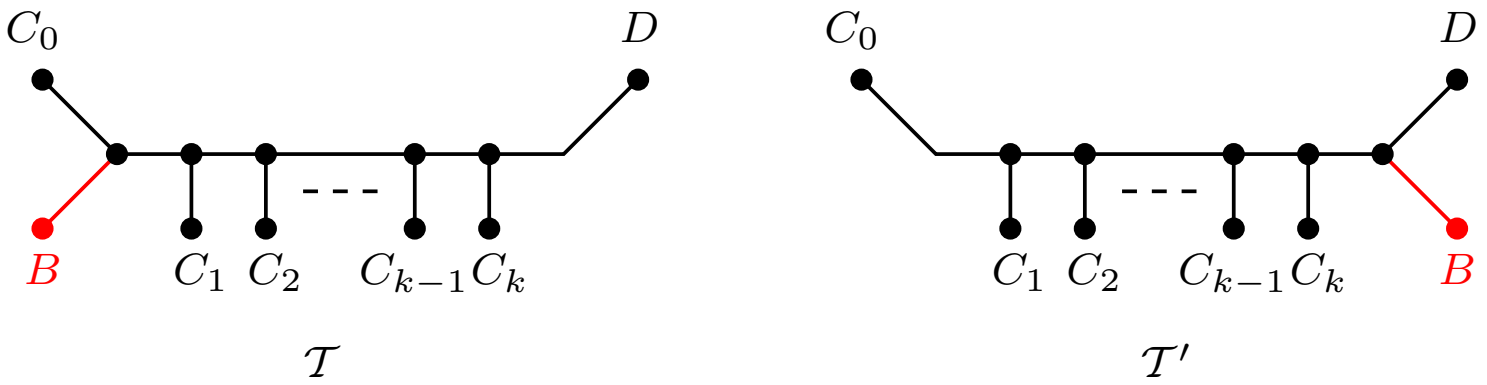


Subtree transfer operations

Nearest Neighbour Interchange (NNI):



Subtree Prune and Regraft (SPR):



FASTME

- This approach is has been programmed as FASTME, available on the web (Desper and Gasuel).
- It is extremely fast (SPR only slightly slower than NNI).
- Extensive experimental results suggest that it gives very accurate results (using either NNI or SPR).

FASTME

- This approach is has been programmed as FASTME, available on the web (Desper and Gasuel).
- It is extremely fast (SPR only slightly slower than NNI).
- Extensive experimental results suggest that it gives very accurate results (using either NNI or SPR).

But is it provably consistent? *I.e.* given good data do we get the correct answer?

Results

Theorem (B., Gascuel, Huber, Moulton '07)

From any starting tree, Balanced Minimum Evolution using an SPR local topology search is consistent.

Results

Theorem (B., Gascuel, Huber, Moulton '07)

From any starting tree, Balanced Minimum Evolution using an SPR local topology search is consistent.

With perfect data we do get to the true tree.

Results

Theorem (B., Gascuel, Huber, Moulton '07)

From any starting tree, Balanced Minimum Evolution using an SPR local topology search is consistent.

With perfect data we do get to the true tree.

Theorem (B., Gascuel, Huber, Moulton '07)

From any starting tree, Balanced Minimum Evolution using an SPR local topology search has a safety radius of 1/3.

Results

Theorem (B., Gascuel, Huber, Moulton '07)

From any starting tree, Balanced Minimum Evolution using an SPR local topology search is consistent.

With perfect data we do get to the true tree.

Theorem (B., Gascuel, Huber, Moulton '07)

From any starting tree, Balanced Minimum Evolution using an SPR local topology search has a safety radius of 1/3.

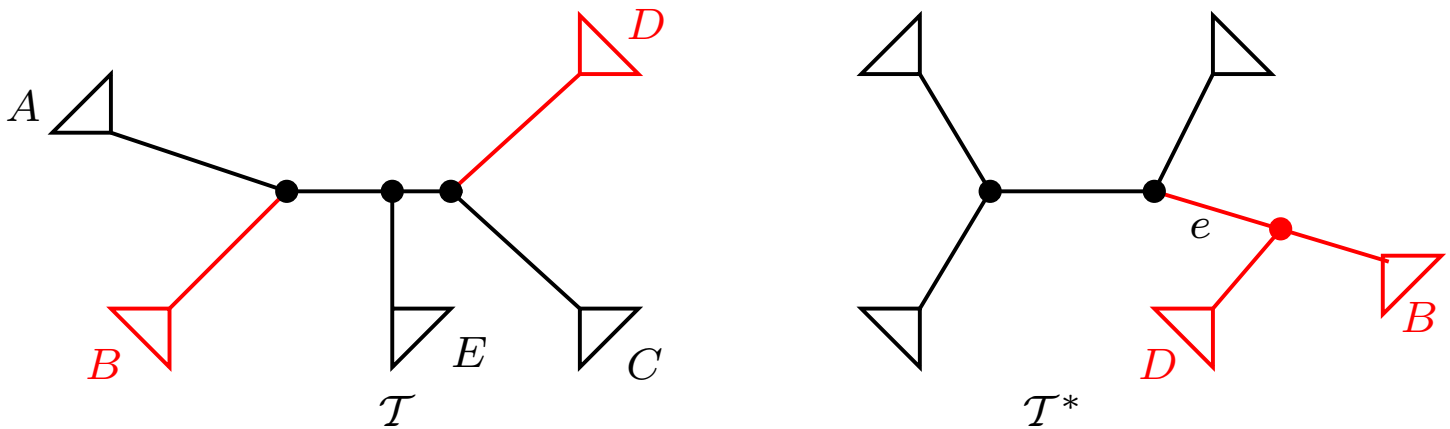
Even with small errors in the data we get to the true tree.

Robinson-Foulds distance

Lemma

For any tree $\mathcal{T} \neq \mathcal{T}^*$ there is a tree $\mathcal{T}'$ one SPR from $\mathcal{T}$ which reduces the Robinson-Foulds distance to $\mathcal{T}^*$.

We find subtrees B and D of $\mathcal{T}$ as follows

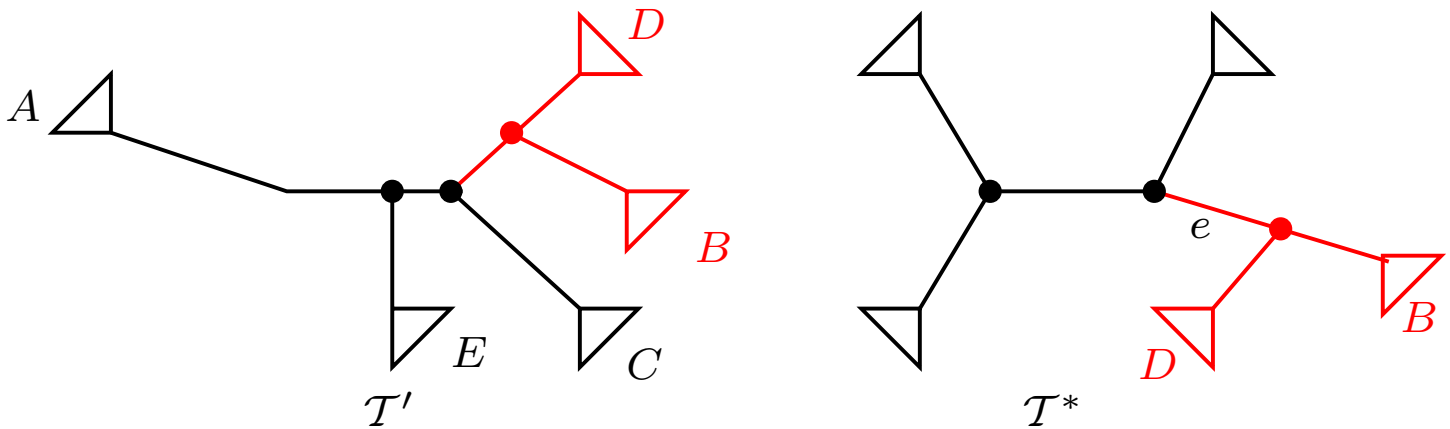


Robinson-Foulds distance

Lemma

For any tree $\mathcal{T} \neq \mathcal{T}^*$ there is a tree $\mathcal{T}'$ one SPR from $\mathcal{T}$ which reduces the Robinson-Foulds distance to $\mathcal{T}^*$.

We find subtrees B and D of $\mathcal{T}$ as follows



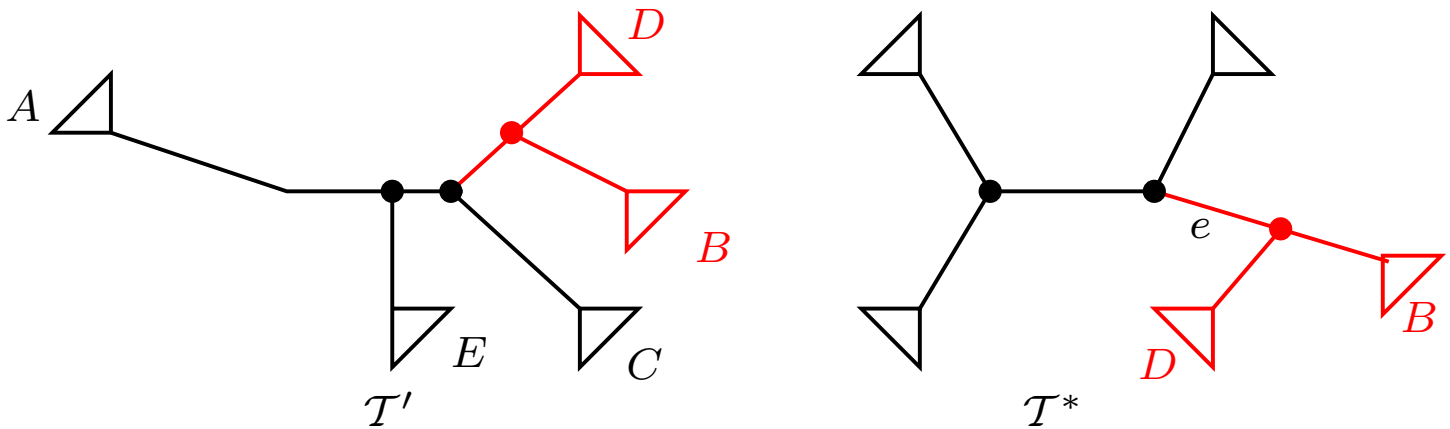
- We then form $\mathcal{T}'$ from $\mathcal{T}$ by uniting B and D.

Robinson-Foulds distance

Lemma

For any tree $\mathcal{T} \neq \mathcal{T}^*$ there is a tree $\mathcal{T}'$ one SPR from $\mathcal{T}$ which reduces the Robinson-Foulds distance to $\mathcal{T}^*$.

We find subtrees B and D of $\mathcal{T}$ as follows



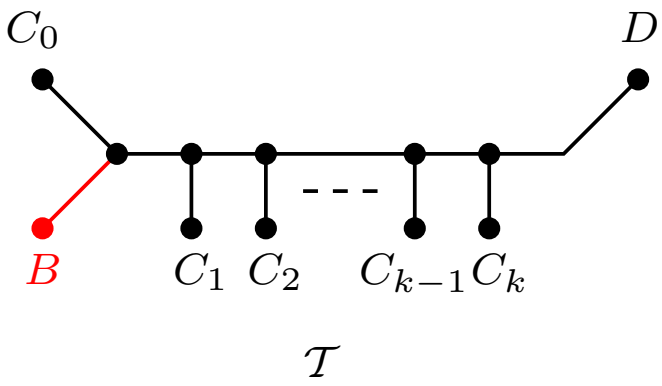
- We then form $\mathcal{T}'$ from $\mathcal{T}$ by uniting B and D.
- Split corresponding to e now appears in $\mathcal{T}'$ and $\mathcal{T}^*$.

Pauplin tree length

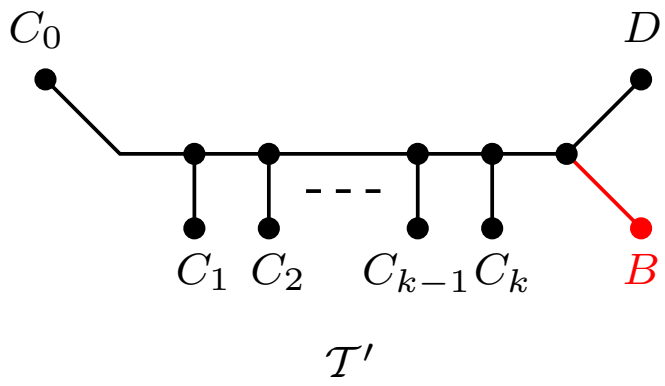
Lemma

Let ϵ be the maximum error of any entry in the distance matrix Δ.

Let k be the number of intermediate subtrees between B and D.



$\mathcal{T}$



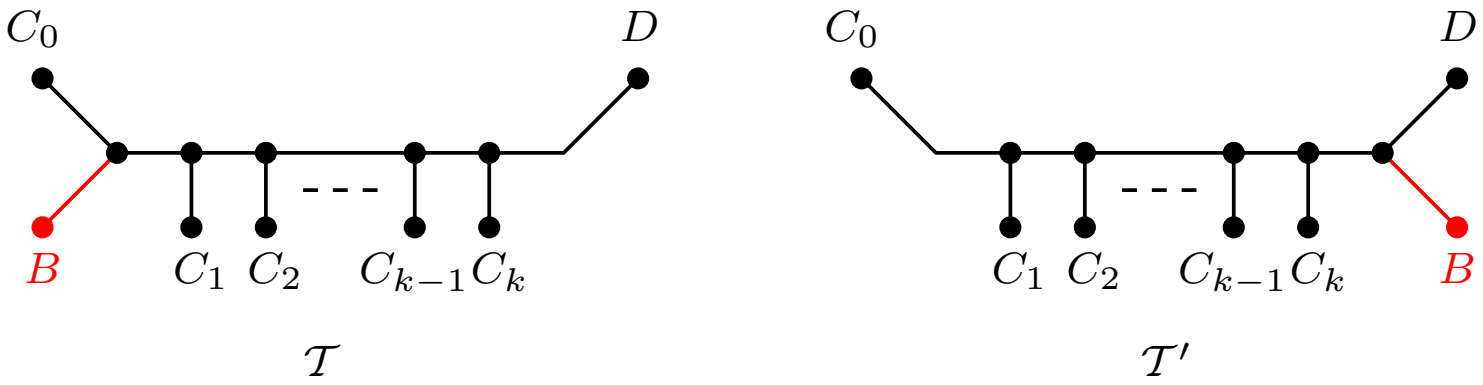
$\mathcal{T}'$

Pauplin tree length

Lemma

Let ϵ be the maximum error of any entry in the distance matrix Δ.

Let k be the number of intermediate subtrees between B and D.



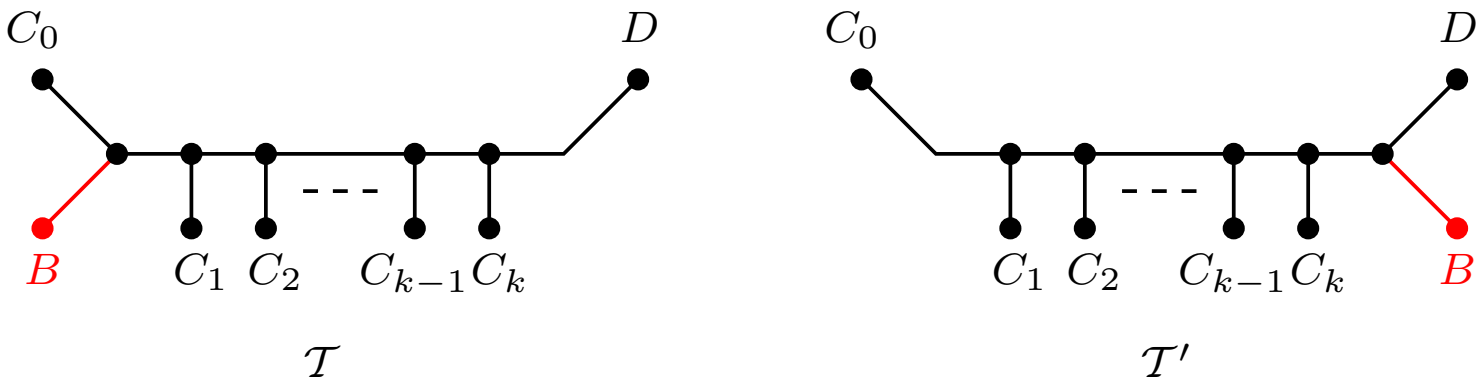
Then $\hat{l}(\mathcal{T}) - \hat{l}(\mathcal{T}') \geq (1 - 2^{-k})(l(e) - 3\epsilon)$.

Pauplin tree length

Lemma

Let ϵ be the maximum error of any entry in the distance matrix Δ.

Let k be the number of intermediate subtrees between B and D.



Then $\hat{l}(\mathcal{T}) - \hat{l}(\mathcal{T}') \geq (1 - 2^{-k})(l(e) - 3\epsilon)$.

Corollary

If $\epsilon < l_{\min}/3$, then there is no local minimum except $\mathcal{T}^*$.

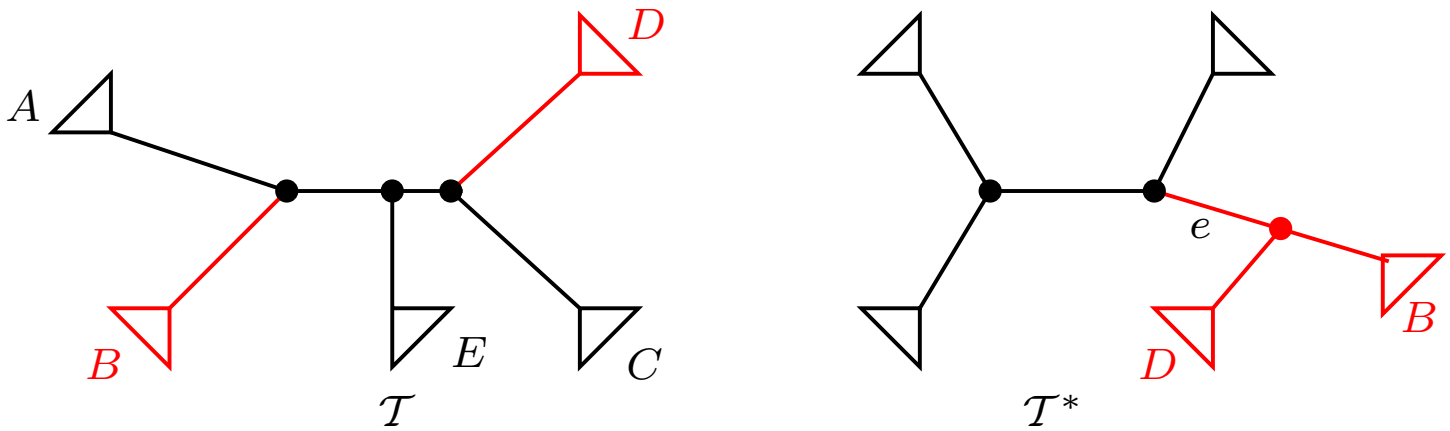
Quartet distance

If $\mathcal{T} \neq \mathcal{T}^*$ then can we find a single SPR move which reduces the quartet distance to $\mathcal{T}^*$?

Quartet distance

If $\mathcal{T} \neq \mathcal{T}^*$ then can we find a single SPR move which reduces the quartet distance to $\mathcal{T}^*$?

Recall we have found subtrees B and D of $\mathcal{T}$ as follows

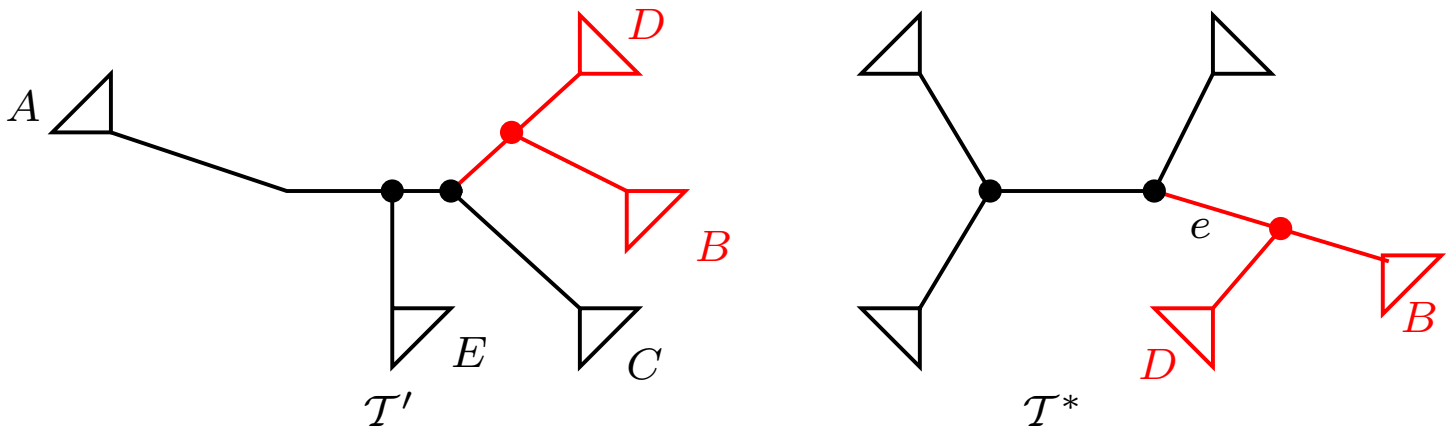


- Previously we arbitrarily moved B to D.

Quartet distance

If $\mathcal{T} \neq \mathcal{T}^*$ then can we find a single SPR move which reduces the quartet distance to $\mathcal{T}^*$?

Recall we have found subtrees B and D of $\mathcal{T}$ as follows

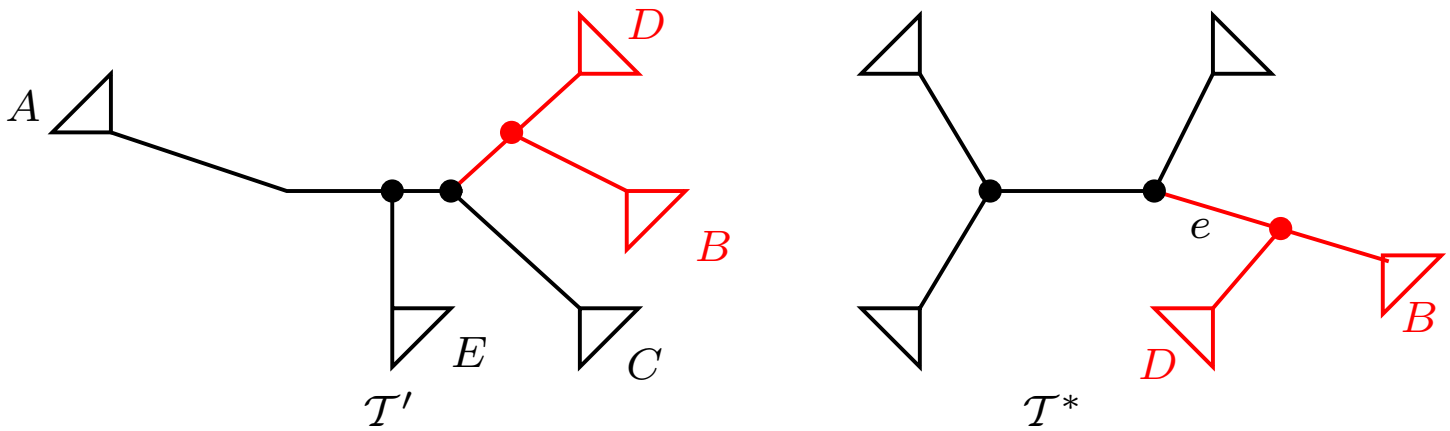


- Previously we arbitrarily moved B to D.

Quartet distance

If $\mathcal{T} \neq \mathcal{T}^*$ then can we find a single SPR move which reduces the quartet distance to $\mathcal{T}^*$?

Recall we have found subtrees B and D of $\mathcal{T}$ as follows



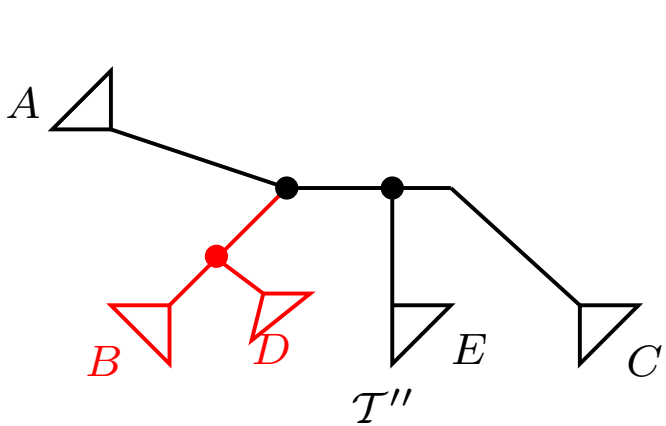
- Previously we arbitrarily moved B to D.
- We also show that one of the two moves
  - uniting B and D by moving B, or
  - uniting B and D by moving D

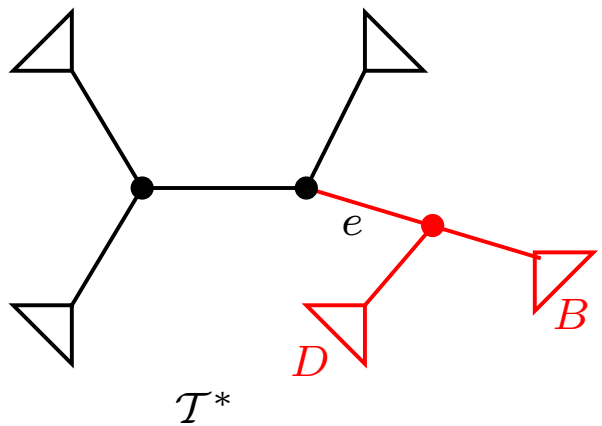
  must reduce the quartet distance to $\mathcal{T}^*$.

Quartet distance

If $\mathcal{T} \neq \mathcal{T}^*$ then can we find a single SPR move which reduces the quartet distance to $\mathcal{T}^*$?

Recall we have found subtrees B and D of $\mathcal{T}$ as follows





- Previously we arbitrarily moved B to D.
- We also show that one of the two moves
  - uniting B and D by moving B, or
  - uniting B and D by moving D

  must reduce the quartet distance to $\mathcal{T}^*$.

Open questions

- Is the safety radius larger than $1/3$?

  Note: there is a counter example to any distance based method having safety radius greater than $1/2$.

Open questions

- Is the safety radius larger than $1/3$?

  Note: there is a counter example to any distance based method having safety radius greater than $1/2$.

- Can we prove other forms of robustness?

Open questions

- Is the safety radius larger than $1/3$?

  Note: there is a counter example to any distance based method having safety radius greater than $1/2$.

- Can we prove other forms of robustness?

- Do any of the results hold for an NNI based local search?

  Note: NNI is more commonly used in practice, and empirically appears as good as SPR.