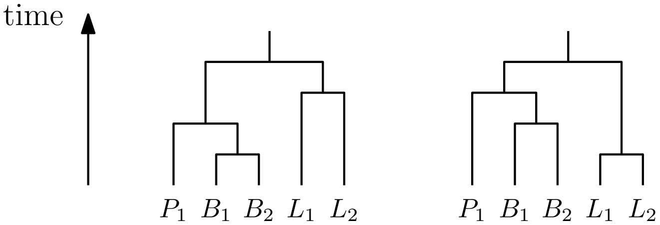
The Space of Discrete Coalescent Trees

Lena Collienne

Biological Data Science Lab
Department of Computer Science
University of Otago

30/03/2021

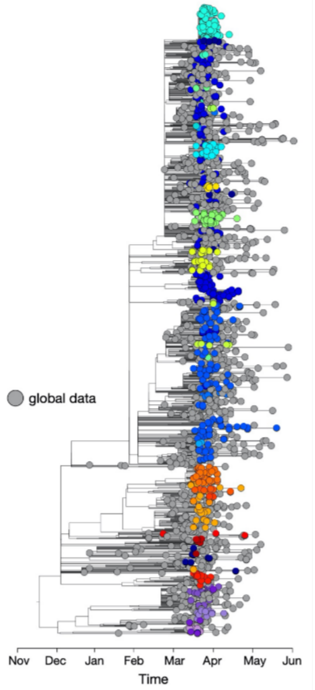
Time Trees
Cancer Phylogenies



P: pancreas (main tumor), L: lung, B: brain

Time Trees

SARS-CoV-2

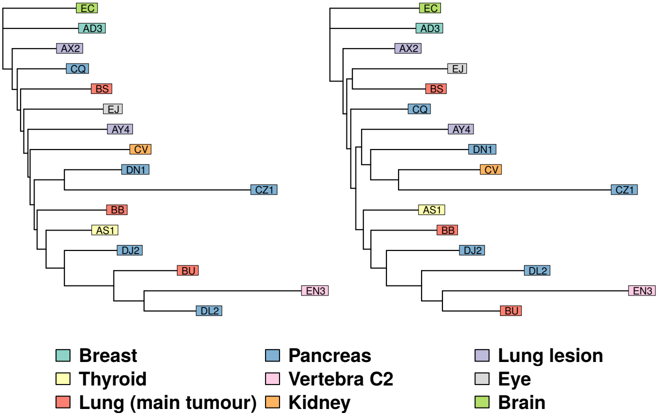


Why Distances?

Tree Inference

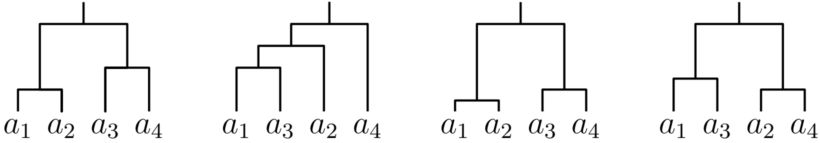
Why Distances?

Tree Inference



Why Distances?

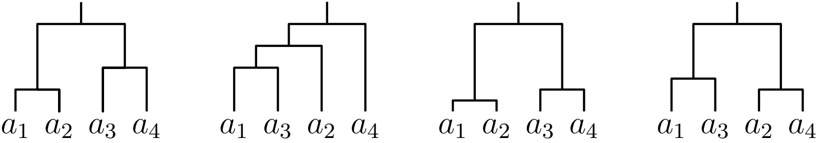
Summarising Trees



What is the mean tree?

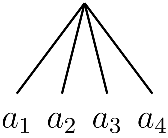
Why Distances?

Summarising Trees

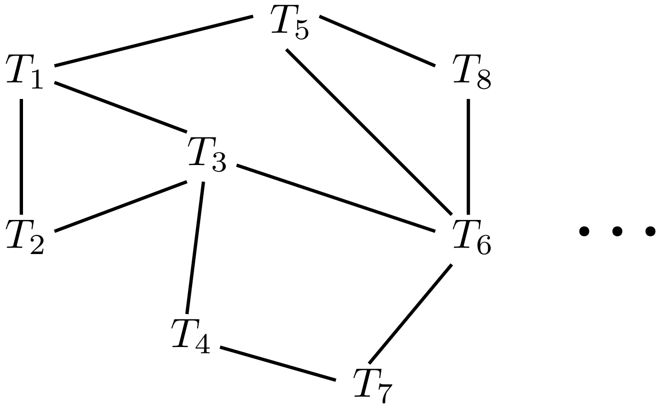


What is the mean tree?

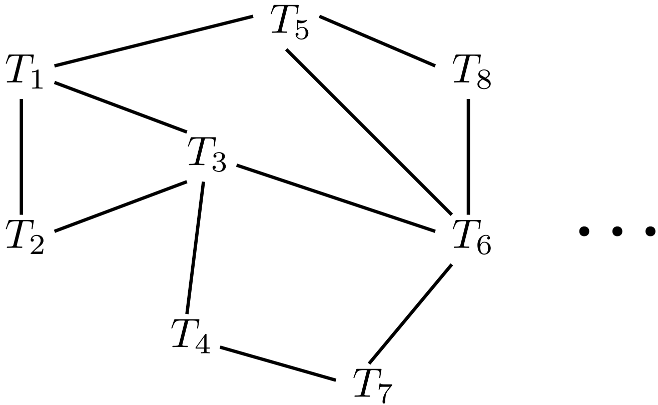
Problem: In most tree spaces the mean tree is a start tree:



Discrete Tree Spaces



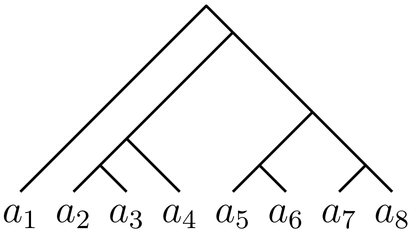
Discrete Tree Spaces



Tree Rearrangement operations: NNI, SPR, TBR

Phylogenetic Trees

Rooted, binary

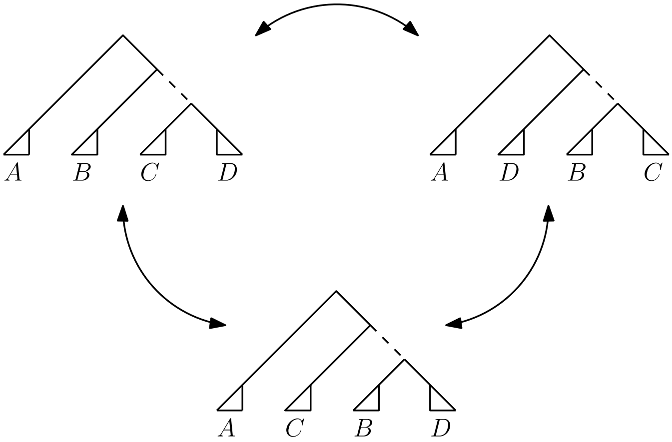


NNI – Nearest Neighbour Interchange

Definition 1

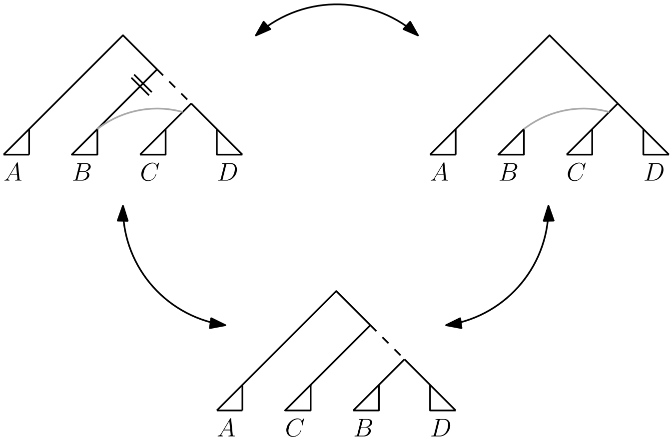
NNI – Nearest Neighbour Interchange

Definition 1



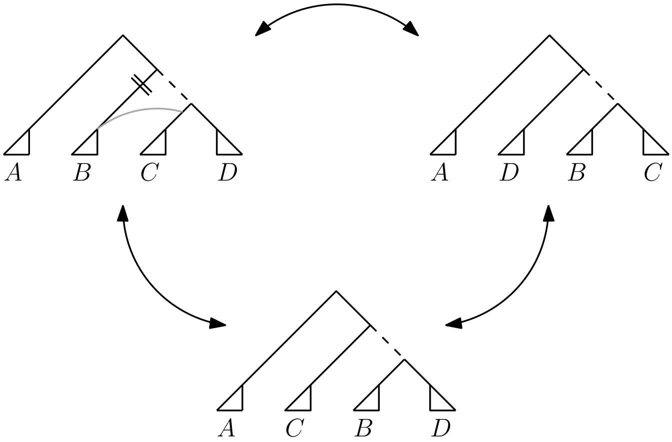
NNI – Nearest Neighbour Interchange

Definition 1



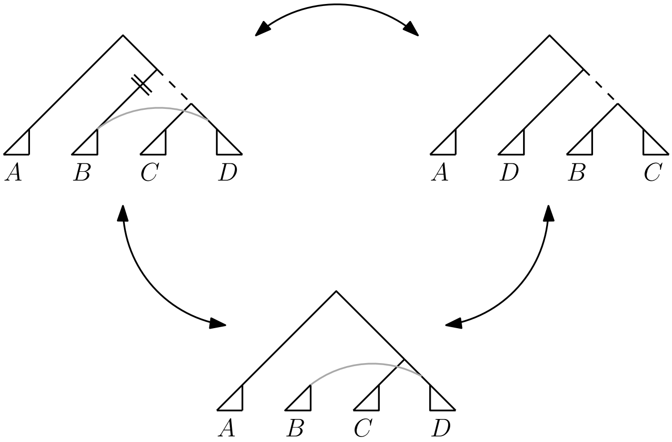
NNI – Nearest Neighbour Interchange

Definition 1



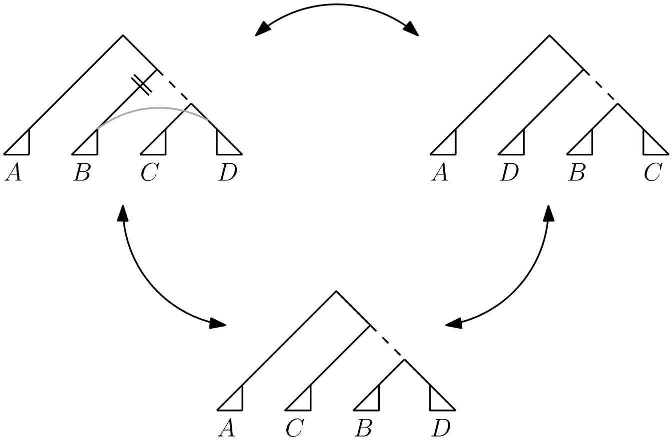
NNI – Nearest Neighbour Interchange

Definition 1



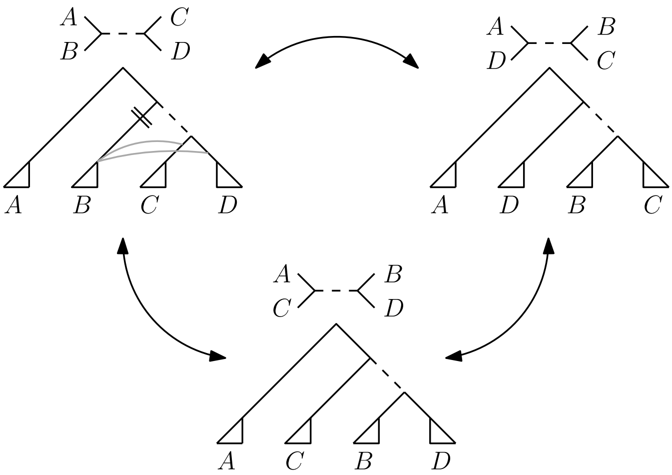
NNI – Nearest Neighbour Interchange

Definition 1



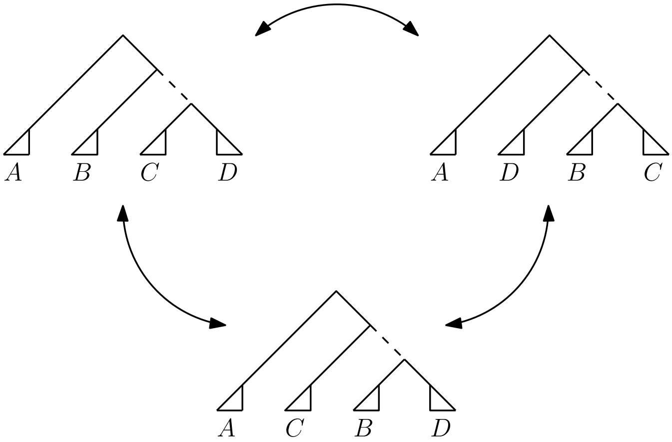
NNI – Nearest Neighbour Interchange

Definition 1



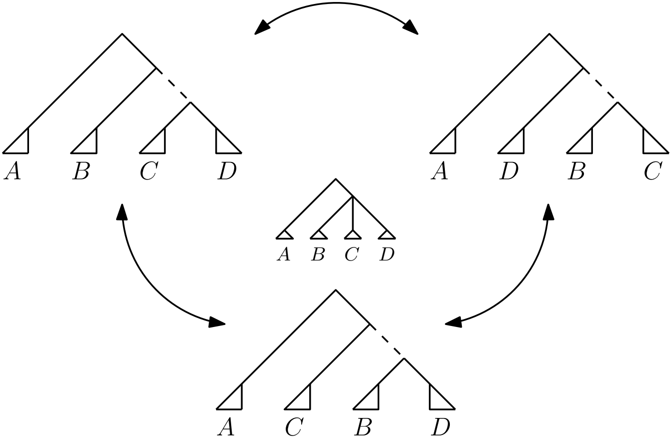
NNI – Nearest Neighbour Interchange

Definition 2



NNI – Nearest Neighbour Interchange

Definition 2



Computing Distances

NNI

NNI-DIST:

INSTANCE: A pair of trees T and R
FIND: Distance between T and R in NNI

Computing Distances

NNI

NNI-DIST:

INSTANCE: A pair of trees T and R
FIND: Distance between T and R in NNI

- $\mathcal{NP}$-hard

Computing Distances
NNI

NNI-DIST:

INSTANCE: A pair of trees T and R
FIND: Distance between T and R in NNI

- $\mathcal{NP}$-hard
- BUT: fixed-parameter tractable (FPT):

  distance computable in $\mathcal{O}(2^{\frac{21k}{2}} * n)$ where $d(T, R) \leq k$

Computing Distances

NNI-DIST:

INSTANCE: A pair of trees T and R
FIND: Distance between T and R in NNI

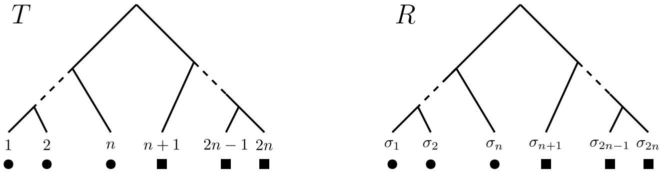
- $\mathcal{NP}$-hard
- BUT: fixed-parameter tractable (FPT):

  distance computable in $\mathcal{O}(2^{\frac{21k}{2}} * n)$ where $d(T, R) \leq k$

- Approximation algorithm: ratio $\mathcal{O}(log(n))$

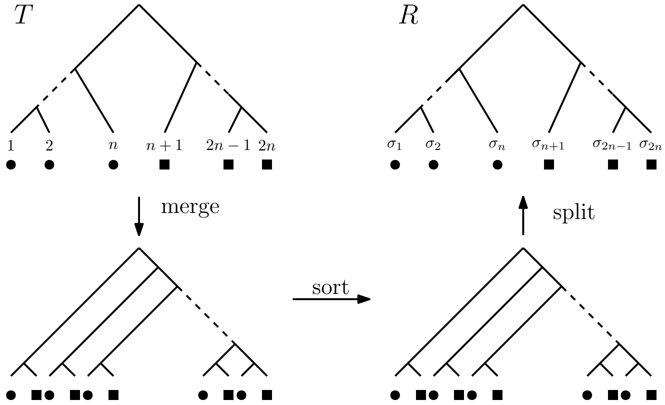
Biological Interpretability

Cluster Property



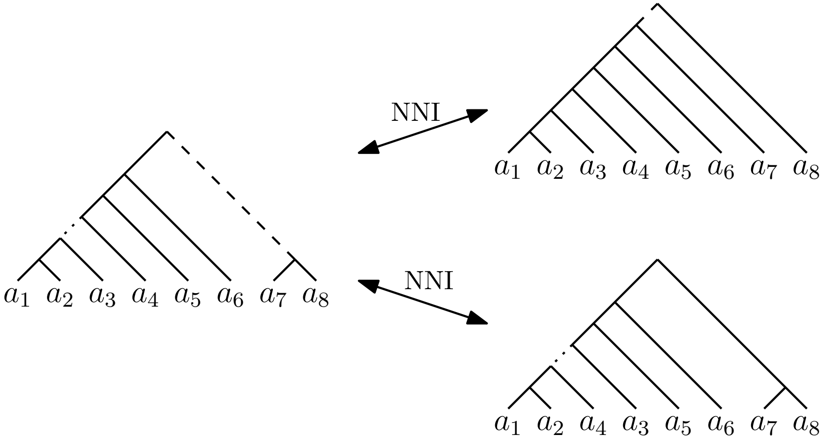
Biological Interpretability

Cluster Property

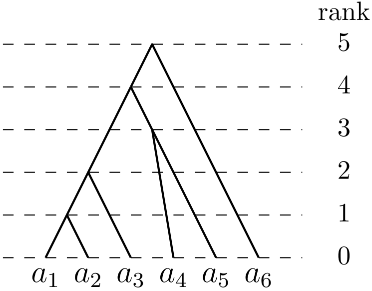


Biological Interpretability

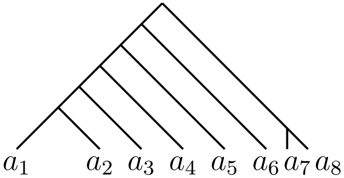
NNI move = NNI move?

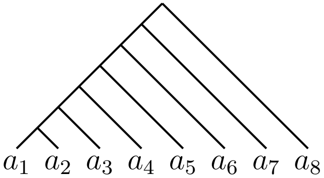


Ranked Trees

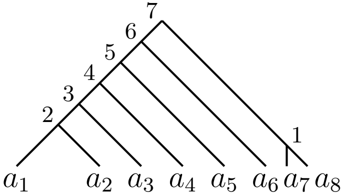


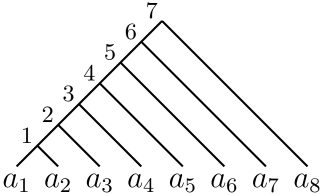
NNI on Ranked Trees



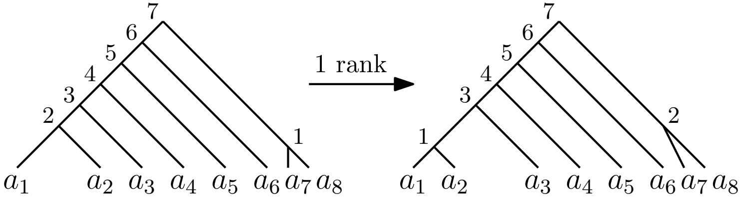


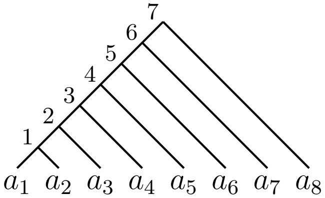
NNI on Ranked Trees



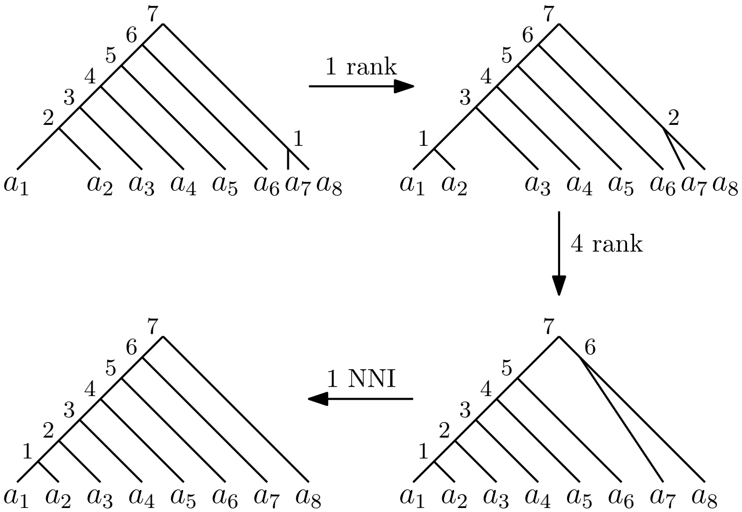


NNI on Ranked Trees

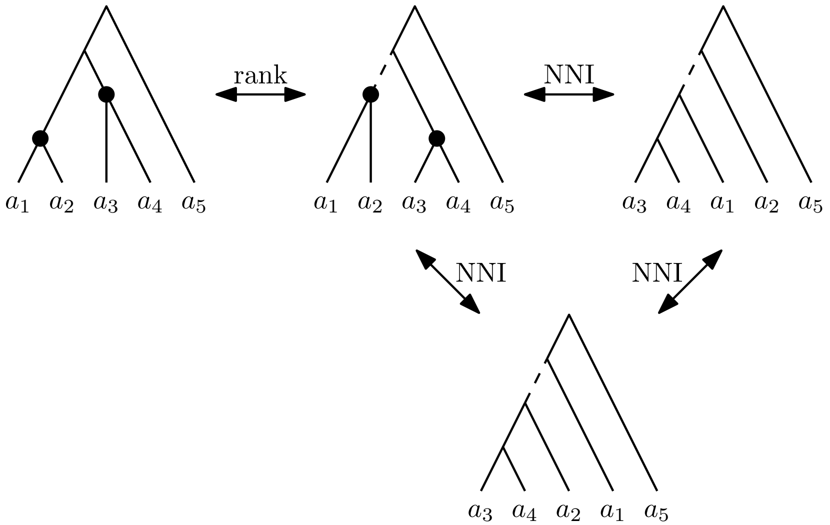




NNI on Ranked Trees



RNNI


rank
NNI
NNI
NNI
a_1 a_2 a_3 a_4 a_5
a_1 a_2 a_3 a_4 a_5
a_3 a_4 a_1 a_2 a_5
a_3 a_4 a_2 a_1 a_5

Shortest Path Problem

RNNI

RNNI-SP:

INSTANCE: A pair of ranked trees T and R
FIND: Shortest Path between T and R in RNNI

FindPath

- Greedy algorithm for approximating RNNI-SP

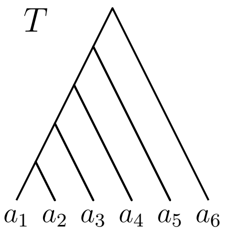
FindPath

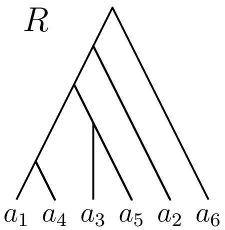
- Greedy algorithm for approximating RNNI-SP
- Running time $\mathcal{O}(n^2)$

FindPath

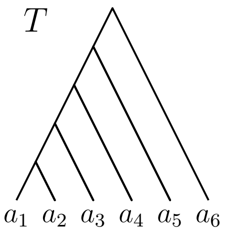
- Greedy algorithm for approximating RNNI-SP
- Running time $\mathcal{O}(n^2)$
- Shortest paths for up to 7 leaves

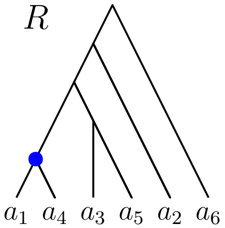
FindPath



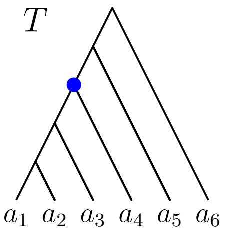


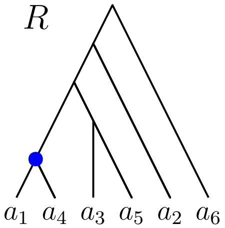
FindPath



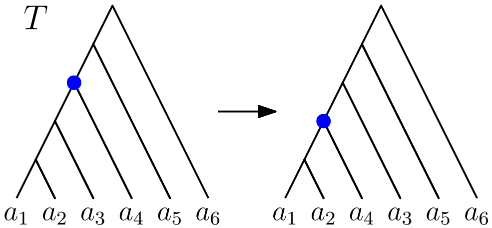


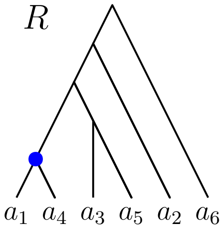
FindPath



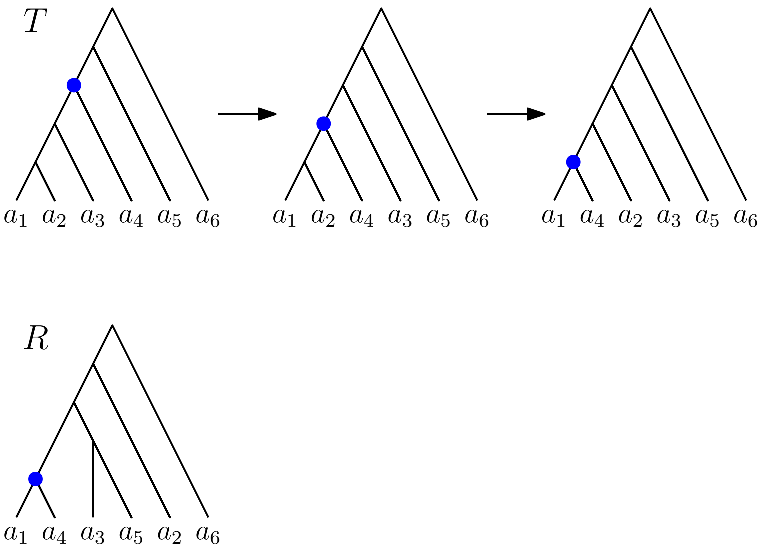


FindPath



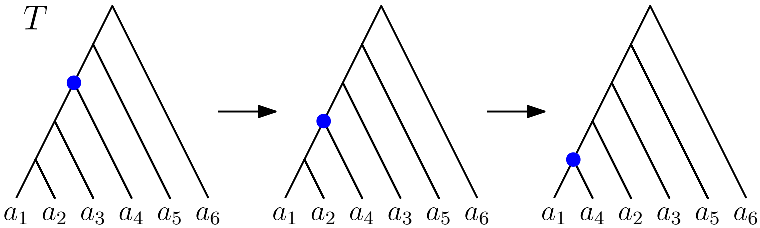


FindPath

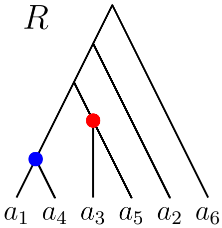


T
a_1 a_2 a_3 a_4 a_5 a_6
a_1 a_2 a_4 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6
R
a_1 a_4 a_3 a_5 a_2 a_6

FindPath

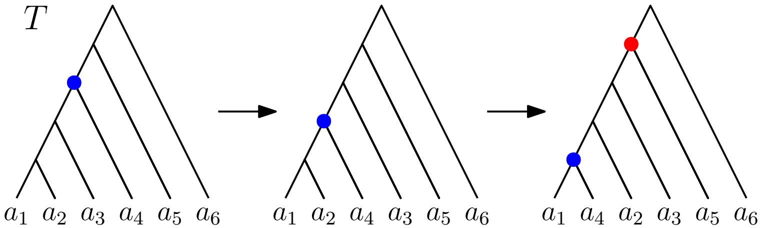


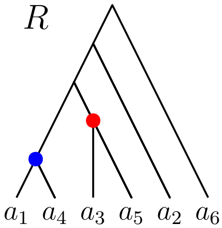
T
a_1 a_2 a_3 a_4 a_5 a_6
a_1 a_2 a_4 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6



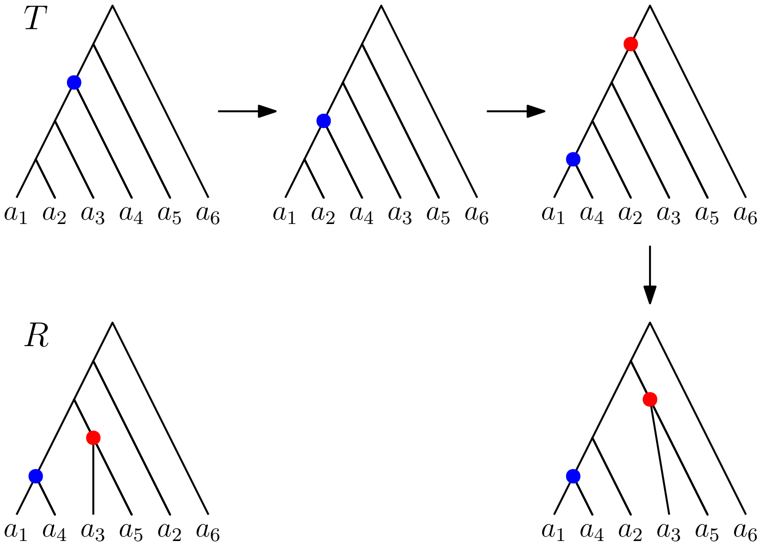
R
a_1 a_4 a_3 a_5 a_2 a_6

FindPath


T
a_1 a_2 a_3 a_4 a_5 a_6
a_1 a_2 a_4 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6

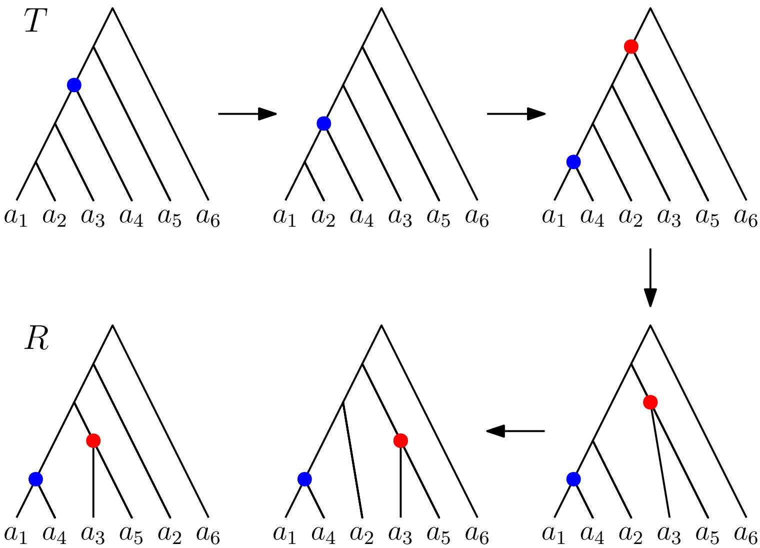

R
a_1 a_4 a_3 a_5 a_2 a_6

FindPath



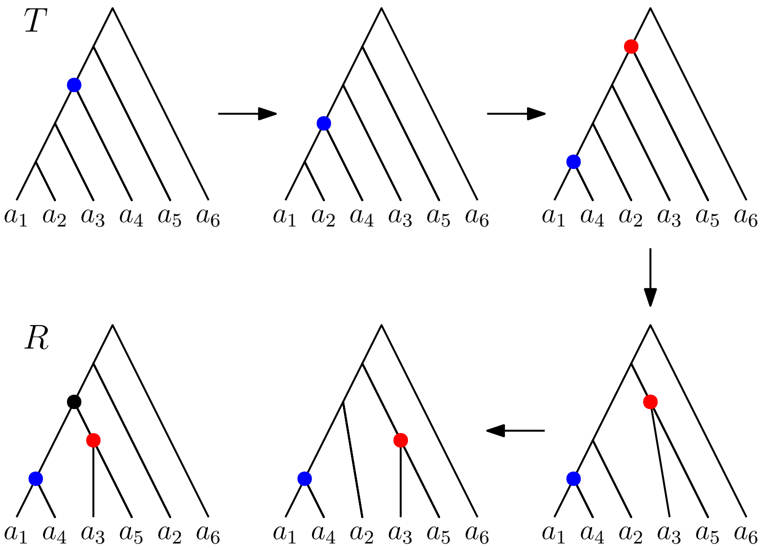
T
a1 a2 a3 a4 a5 a6
a1 a2 a4 a3 a5 a6
a1 a4 a2 a3 a5 a6
R
a1 a4 a3 a5 a2 a6
a1 a4 a2 a3 a5 a6

FindPath

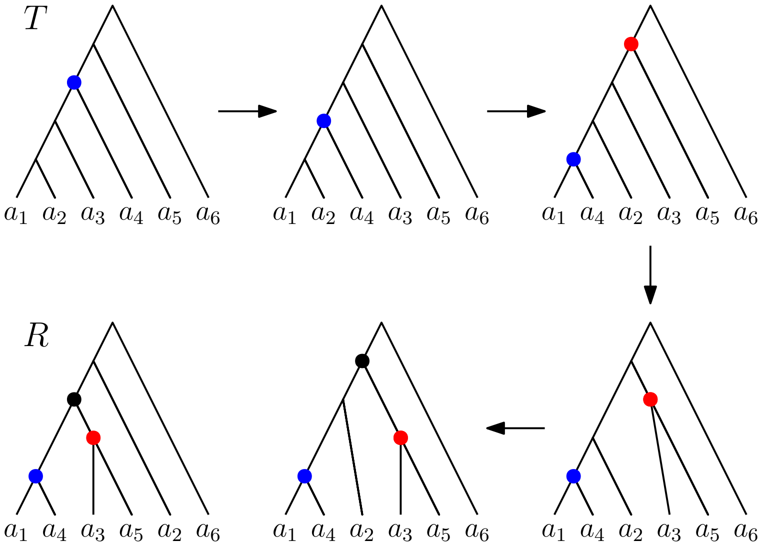


T
a_1 a_2 a_3 a_4 a_5 a_6
a_1 a_2 a_4 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6
R
a_1 a_4 a_3 a_5 a_2 a_6
a_1 a_4 a_2 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6

FindPath

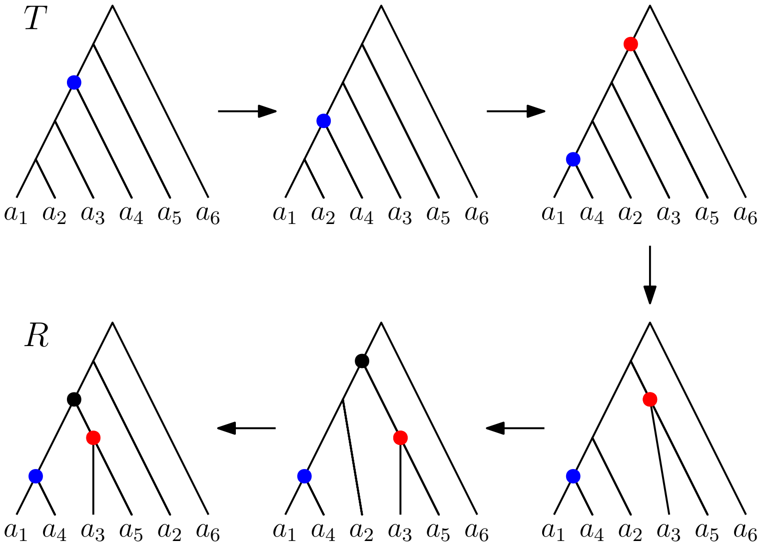

T
a_1 a_2 a_3 a_4 a_5 a_6
a_1 a_2 a_4 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6
R
a_1 a_4 a_3 a_5 a_2 a_6
a_1 a_4 a_2 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6

FindPath



T
a_1 a_2 a_3 a_4 a_5 a_6
a_1 a_2 a_4 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6
R
a_1 a_4 a_3 a_5 a_2 a_6
a_1 a_4 a_2 a_3 a_5 a_6
a_1 a_4 a_2 a_3 a_5 a_6

FindPath


T
a1 a2 a3 a4 a5 a6
a1 a2 a4 a3 a5 a6
a1 a4 a2 a3 a5 a6
R
a1 a4 a3 a5 a2 a6
a1 a4 a2 a3 a5 a6
a1 a4 a2 a3 a5 a6

FindPath

Theorem

FindPath *computes shortest paths in* RNNI*.*

FindPath

Theorem

FindPath *computes shortest paths in* RNNI.

Idea for proof

$\mathrm{FP}(T, R) :=$ *path between* T *and* R *computed by* FindPath

FindPath

Theorem

FindPath *computes shortest paths in* RNNI.

Idea for proof

$\mathrm{FP}(T, R) :=$ *path between* T *and* R *computed by* FindPath

Lemma

If for all trees T, R *and neighbour* T' *of* T *it is*

$$|\mathrm{FP}(T', R)| \geq |\mathrm{FP}(T, R)| - 1,$$

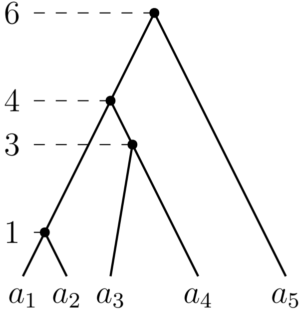
then

$$|\mathrm{FP}(T, R)| = d(T, R)$$

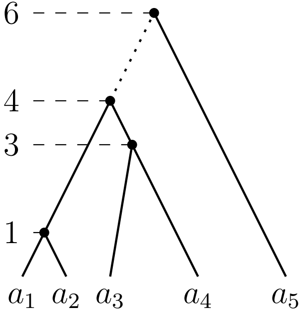
for all trees T *and* R

Discrete Coalescent Trees

Discrete Coalescent Trees

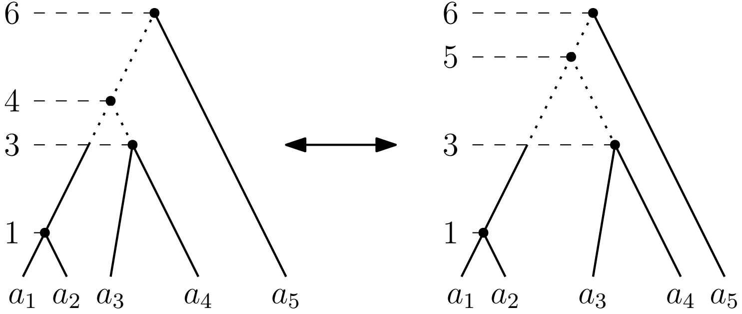


Discrete Coalescent Trees



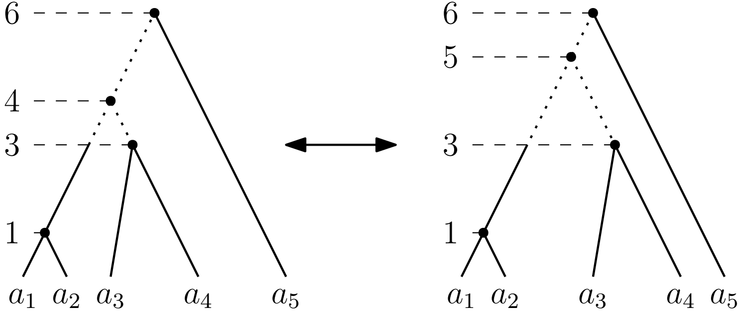
Discrete Coalescent Trees

Length move



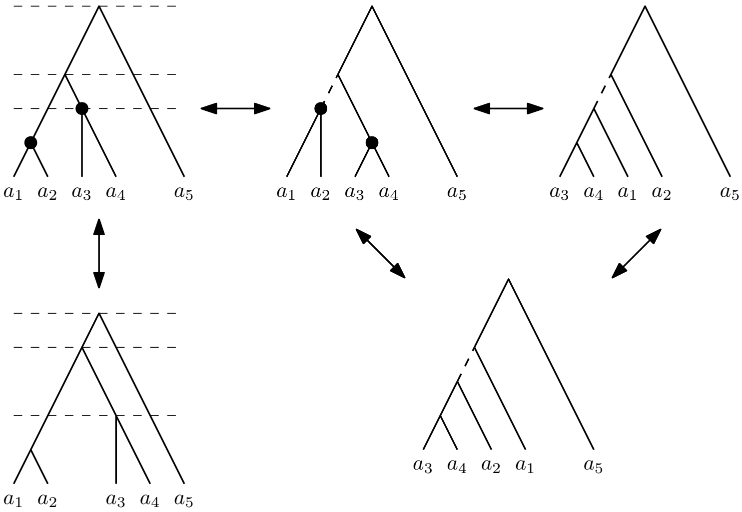
Discrete Coalescent Trees

Length move

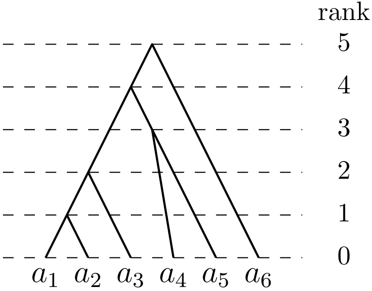


New parameter: m = max height of tree

DCT$_m$


a_1 a_2 a_3 a_4 a_5
a_1 a_2 a_3 a_4 a_5
a_3 a_4 a_1 a_2 a_5
a_1 a_2 a_3 a_4 a_5
a_3 a_4 a_2 a_1 a_5

$\mathrm{DCT}_{n-1} = \mathit{RNNI}$

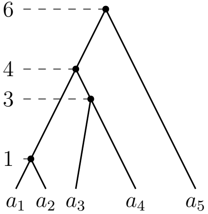


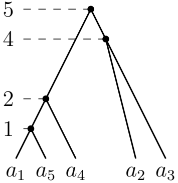
FindPath in DCT_m

$m = 6$

FINDPATH in DCT_m

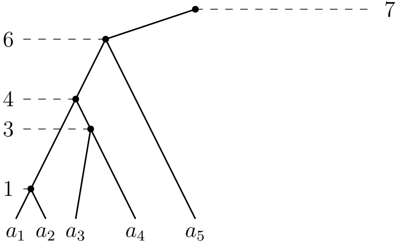
$m = 6$

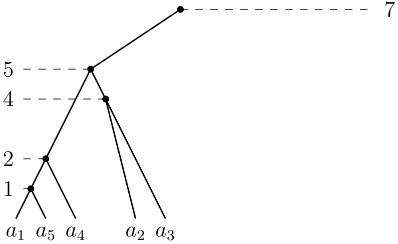




FindPath in DCT_m

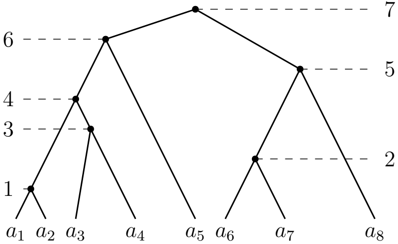
$m = 6$

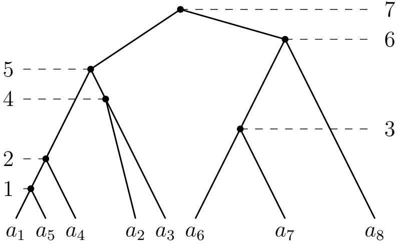




FindPath in DCT_m

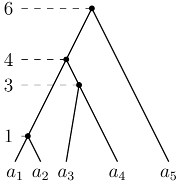
$m = 6$

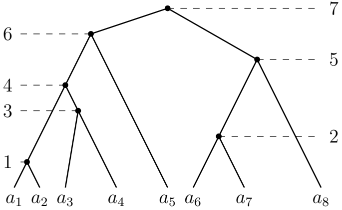




FindPath in DCT_m

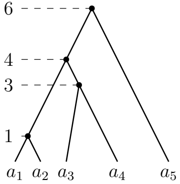
$m = 6$

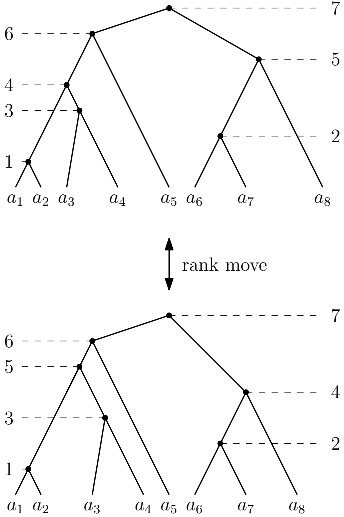




FindPath in DCT_m

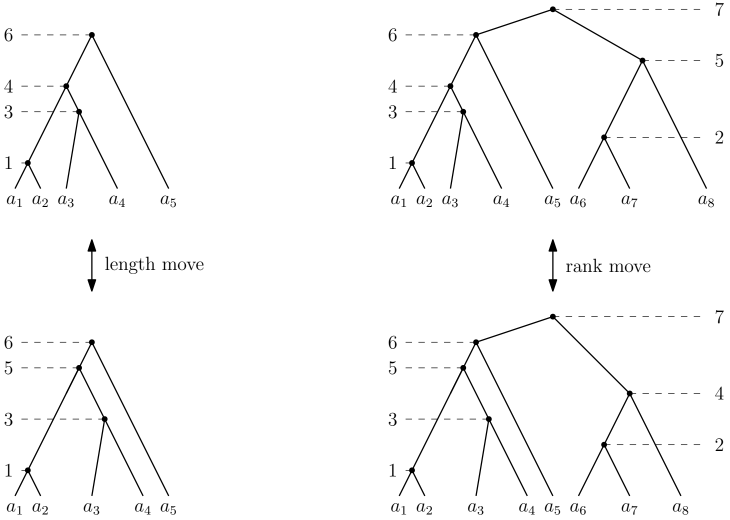
$m = 6$





FindPath in DCT_m

$m = 6$



FindPath in DCT_m

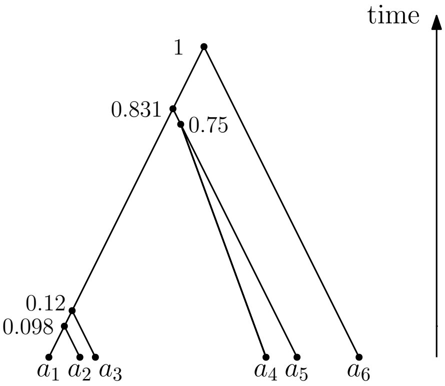
$m = 6$

Theorem

FindPath *computes shortest paths between discrete coalescent trees T and R in $\mathcal{O}(m^2)$.*

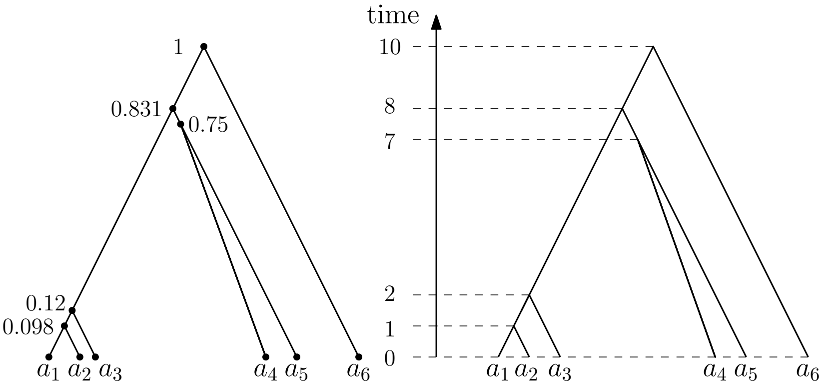
FindPath in DCT_m

Scalability



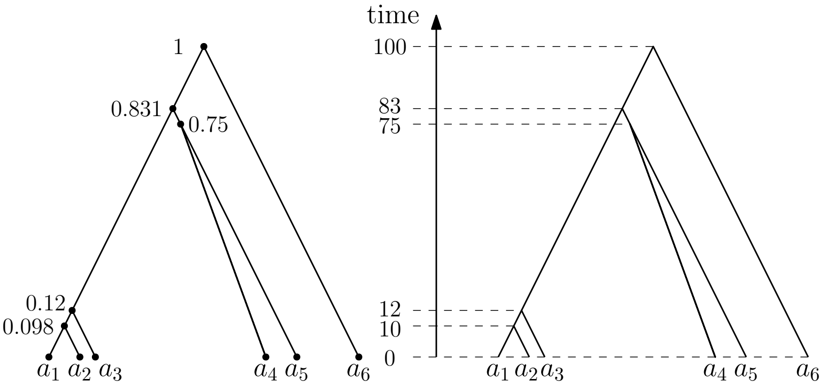
FindPath in DCT_m

Scalability



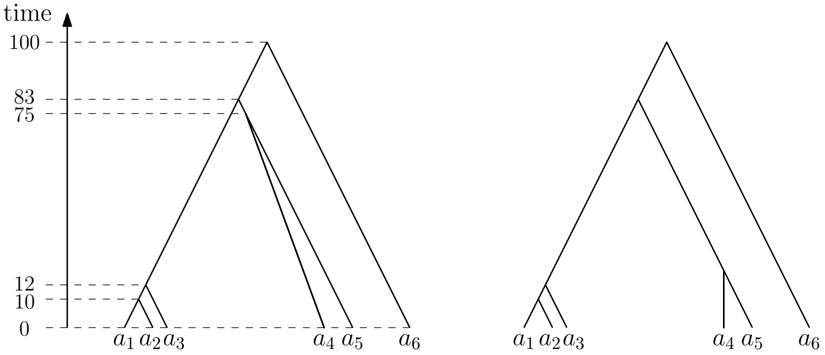
FindPath in DCT_m

Scalability



FindPath in DCT_m

Scalability



Properties of DCT_m

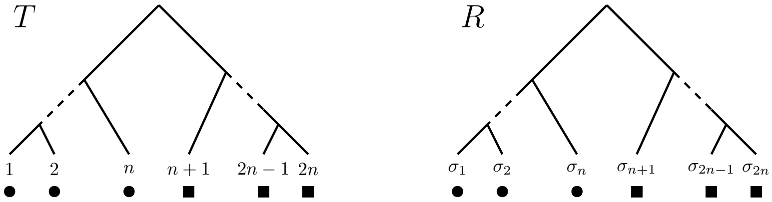
	# Trees	Diameter	Radius
RNNI	$\frac{n!(n-1)!}{2^{n-1}}$	$\binom{n-1}{2}$	$\binom{n-1}{2}$

Properties of DCT_m

	# Trees	Diameter	Radius
RNNI	$\frac{n!(n-1)!}{2^{n-1}}$	$\binom{n-1}{2}$	$\binom{n-1}{2}$
DCT_m	$\frac{n!(n-1)!}{2^{n-1}}\binom{m}{n-1}$	$\binom{n-1}{2} + (m-n+1)(n-1)$	?

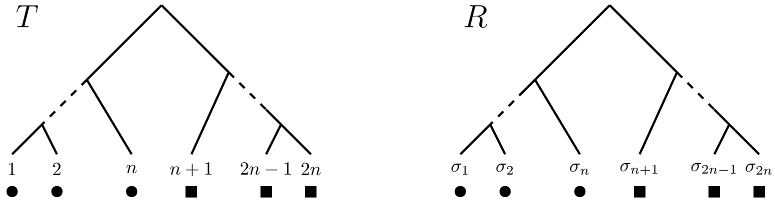
Properties of DCT_m

Cluster Property



Properties of DCT_m

Cluster Property

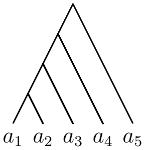


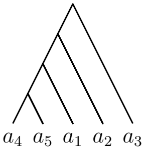
Theorem

DCT_m *has the cluster property.*

Properties of DCT_m

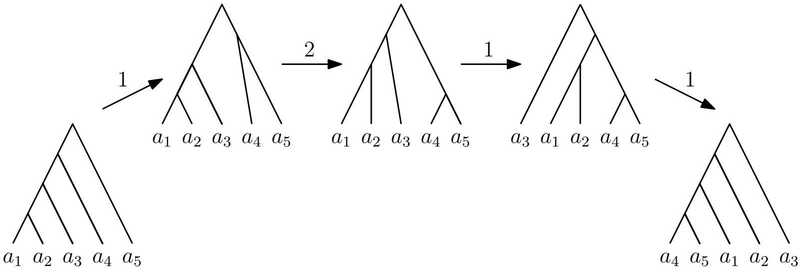
Caterpillar Trees





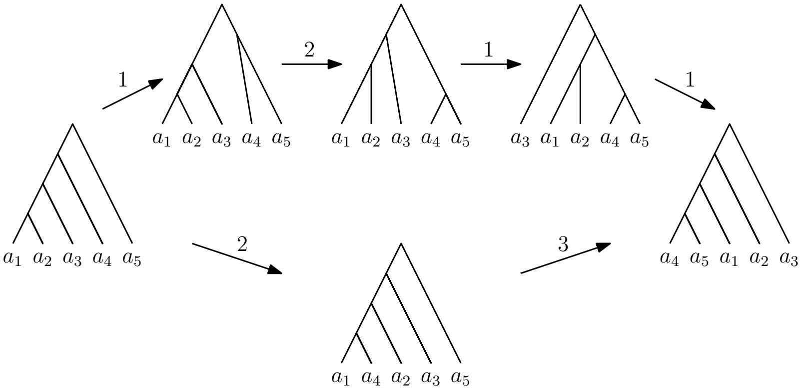
Properties of DCT_m

Caterpillar Trees



Properties of DCT_m

Caterpillar Trees



Properties of DCT_m

Caterpillar Trees

Theorem

The set of caterpillar trees is convex.

Properties of DCT_m

Caterpillar Trees

Theorem

The set of caterpillar trees is convex.

Corollary

The distance between caterpillar trees can be computed in $\mathcal{O}(n\sqrt{\log(n)})$.

Conclusion

Solved problems:

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!
- We know diameter, radius, cluster property, convexity of set of caterpillar trees

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!
- We know diameter, radius, cluster property, convexity of set of caterpillar trees

Open Problems:

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!
- We know diameter, radius, cluster property, convexity of set of caterpillar trees

Open Problems:

- Can we compute distances more efficiently?

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!
- We know diameter, radius, cluster property, convexity of set of caterpillar trees

Open Problems:

- Can we compute distances more efficiently?
- How can we summarise trees?

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!
- We know diameter, radius, cluster property, convexity of set of caterpillar trees

Open Problems:

- Can we compute distances more efficiently?
- How can we summarise trees?
- Does this help us doing statistics in tree space? Confidence intervals?

Conclusion

Solved problems:

- RNNI and DCT_m: shortest paths in $\mathcal{O}(n^2)$!
- We know diameter, radius, cluster property, convexity of set of caterpillar trees

Open Problems:

- Can we compute distances more efficiently?
- How can we summarise trees?
- Does this help us doing statistics in tree space? Confidence intervals?
- ...

Thank you

- Alex Gavryushkin (University of Otago)
- David Bryant (University of Otago)
- BioDS Lab: