A forest: deep kernelization for Tree Bisection and Reconnnect (TBR) distance

Steven Kelk

Department of Advanced Computing Sciences
Maastricht University
The Netherlands

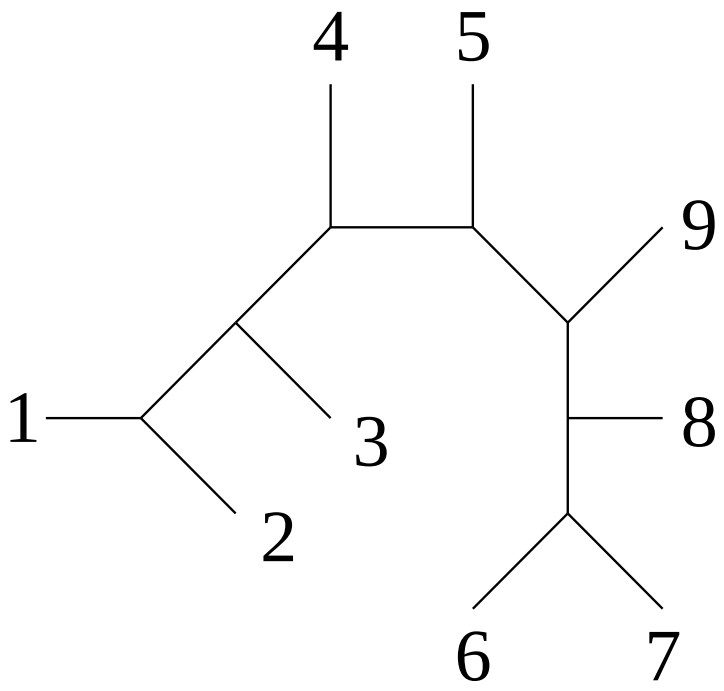
Email: steven.kelk@maastrichtuniversity.nl

Joint work with Simone Linz (Auckland) and Ruben Meuwese (Maastricht)

Emerging Mathematical Frontiers in Molecular Evolution
1st August 2022

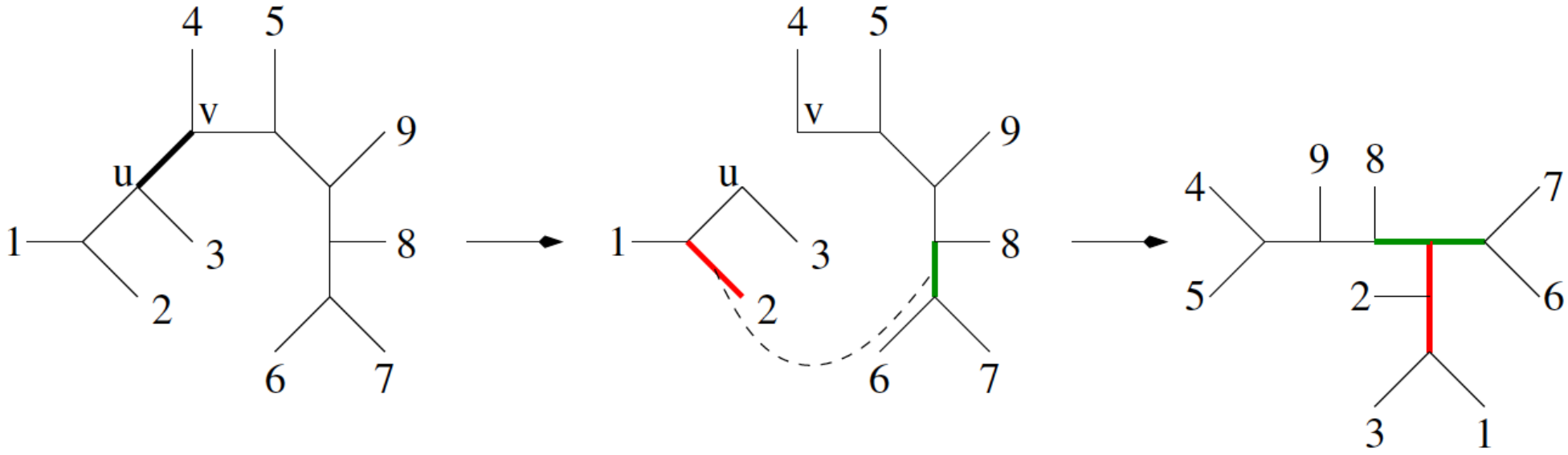
Maastricht University

Phylogenetic trees



An (*unrooted) phylogenetic tree on X* is a connected acyclic graph whose internal vertices have degree three and whose leaf set is *X*.

Tree bisection and reconnection (TBR)



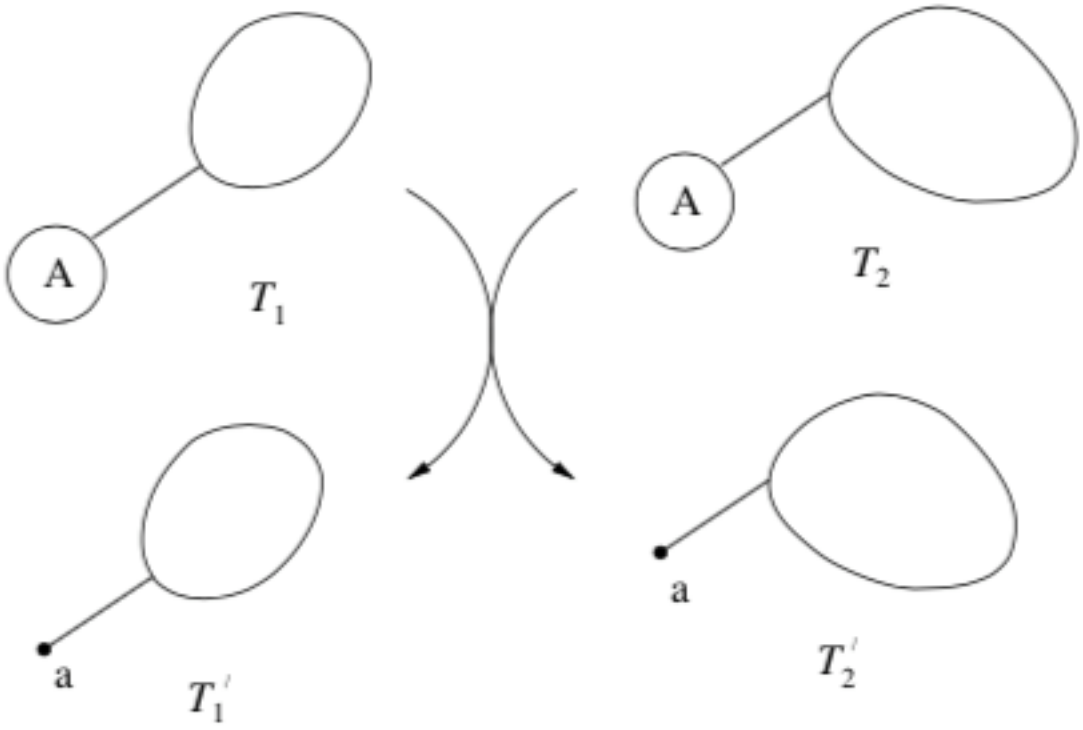
Let $d_{TBR}(T,T')$ denote the minimum number of TBR operations required to transform T into T'. Then, $d_{TBR}(T,T')$ induces a metric on the space of all unrooted phylogenetic trees with n leaves.

(Robinson, 1971; Allen and Steel, 2001).

Computing $d_{TBR}(T,T')$ is NP-hard and fixed-parameter tractable, when parameterized by $k=d_{TBR}$.

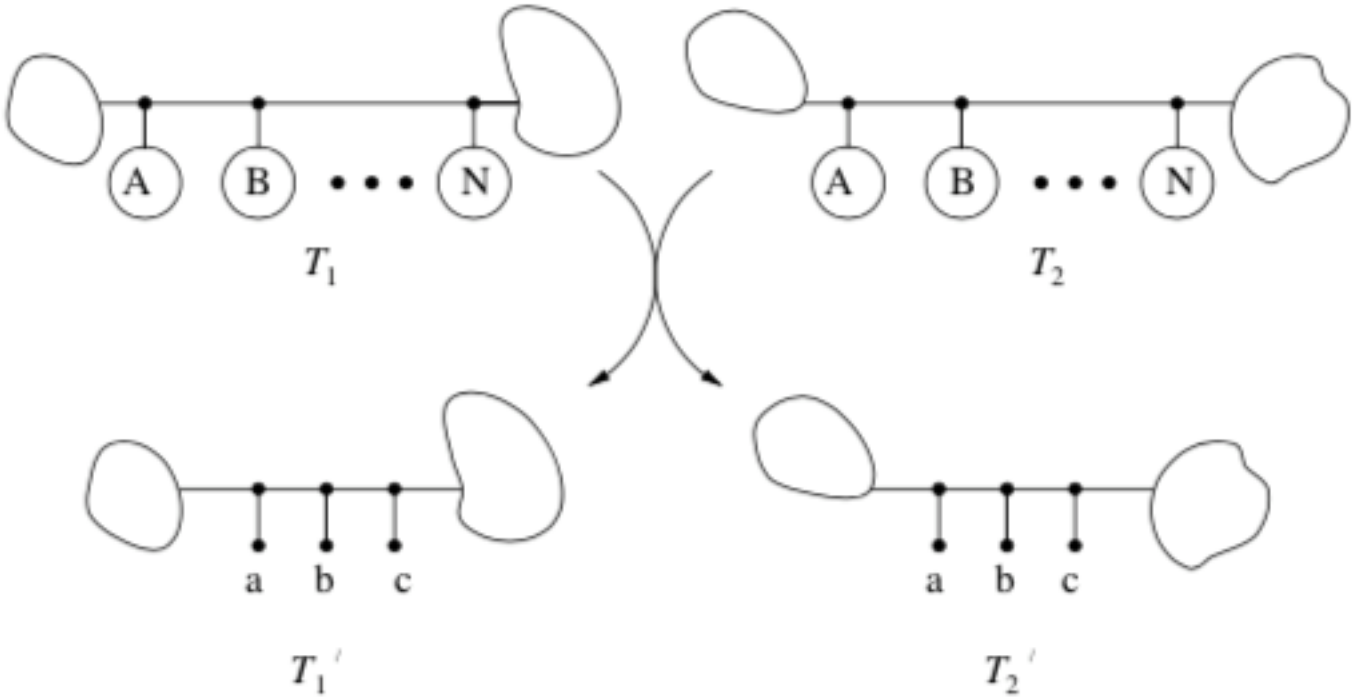
(Hein et al., 1996; Allen and Steel, 2001).

Subtree reduction



Allen and Steel, 2001

Chain reduction



Allen and Steel, 2001

Theorem. (Allen and Steel, 2001).

[Linear kernel] Let S and S' be two trees obtained from T and T' by repeated applications of the subtree and chain reduction until no further reduction is possible. Then

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

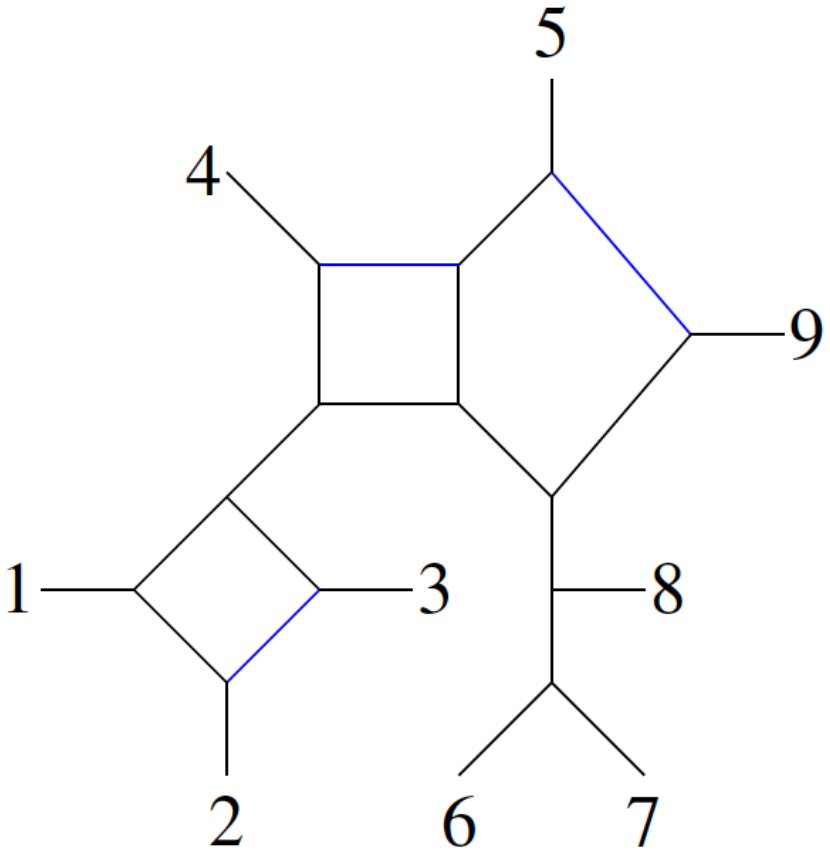
where X' is the leaf set of S and S'.

- We reanalysed Allen and Steel's kernel, and show that it is considerably smaller than they claimed: $15d_{TBR} - 9$. Moreover, this is tight. (K. & Linz, SIAM Journal on Discrete Mathematics 2019)

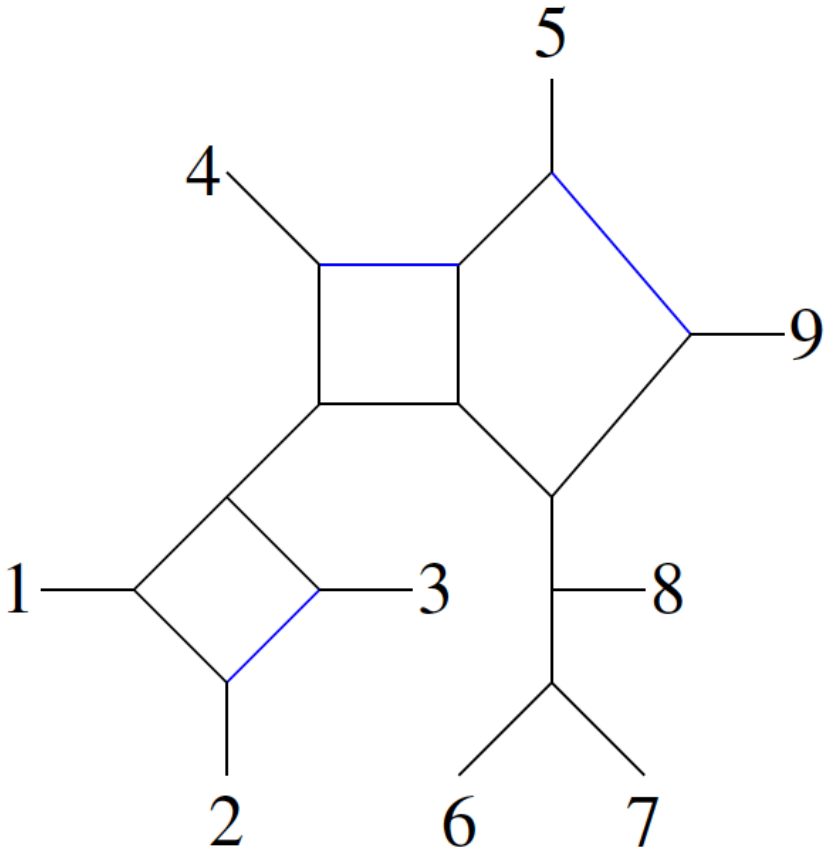
- We devised five new reduction rules which, when combined with Allen and Steel's reduction rules, yield a kernel of size: $11d_{TBR} - 9$. This is also tight. (K. & Linz, Annals of Combinatorics, 2020)

- Today: We introduce a number of 'third generation' reduction rules which reduce the kernel size to $9d_{TBR} - 8$. And, yes, *essentially* tight ☺

From trees to networks



An *unrooted phylogenetic network N on X* is a simple graph whose internal vertices have degree three and whose leaf set is *X*.



Reticulation number of *N* is

$$r(N) = |E| - (|V|-1).$$

(equal to cyclomatic number).

Example. $r(N) = 3$

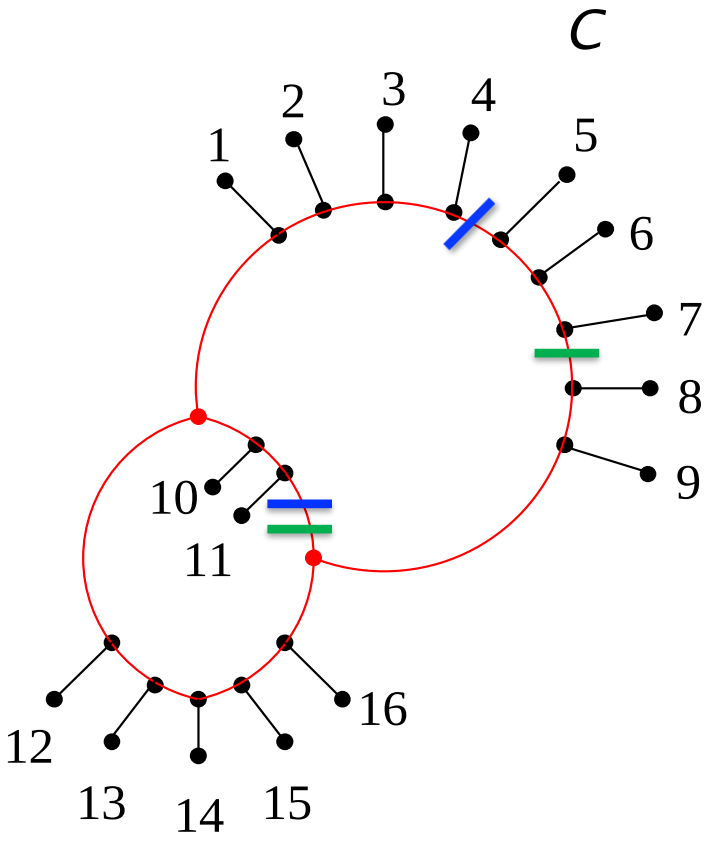
Computing d_{TBR} ≈ combining trees into networks

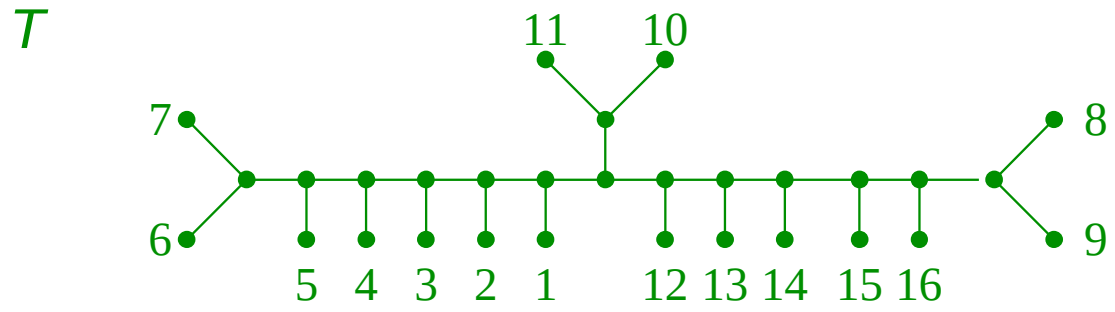
Theorem. (van Iersel et al., 2018)

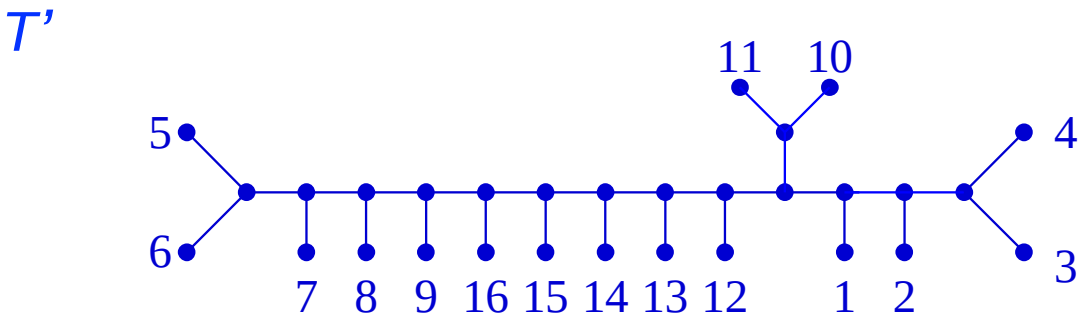
Let T and T' be two trees. Then

$d_{\mathrm{TBR}}(T,T')$ = minimum reticulation number of a network N that displays T and T'

2k breakpoints to divide across 3(k-1) sides







$r(N) = d_{TBR}(T,T') = 2$

2020: Achieving *11k…*

Idea. We described 5 (!) new reduction rules which strengthen the subtree and chain reductions and achieve the following:

- *At most 3 taxa on a side if it has no breakpoints;*
- *At most 4 taxa on a side if it has 1 breakpoint;*
- *At most 4 taxa on a side if it has 2 breakpoints.*

By dividing *2k* breakpoints across *3(k-1)* sides, we concluded that the size of the new kernel is at most…

$$4*2k + 3*(k-3) = 11k-9.$$

The correctness of these new rules requires use of the *agreement forest* characterization of d_{TBR}.

2020: Achieving *11k…*

Idea. We described 5 (!) new reduction rules which strengthen the subtree and chain reductions and achieve the following:

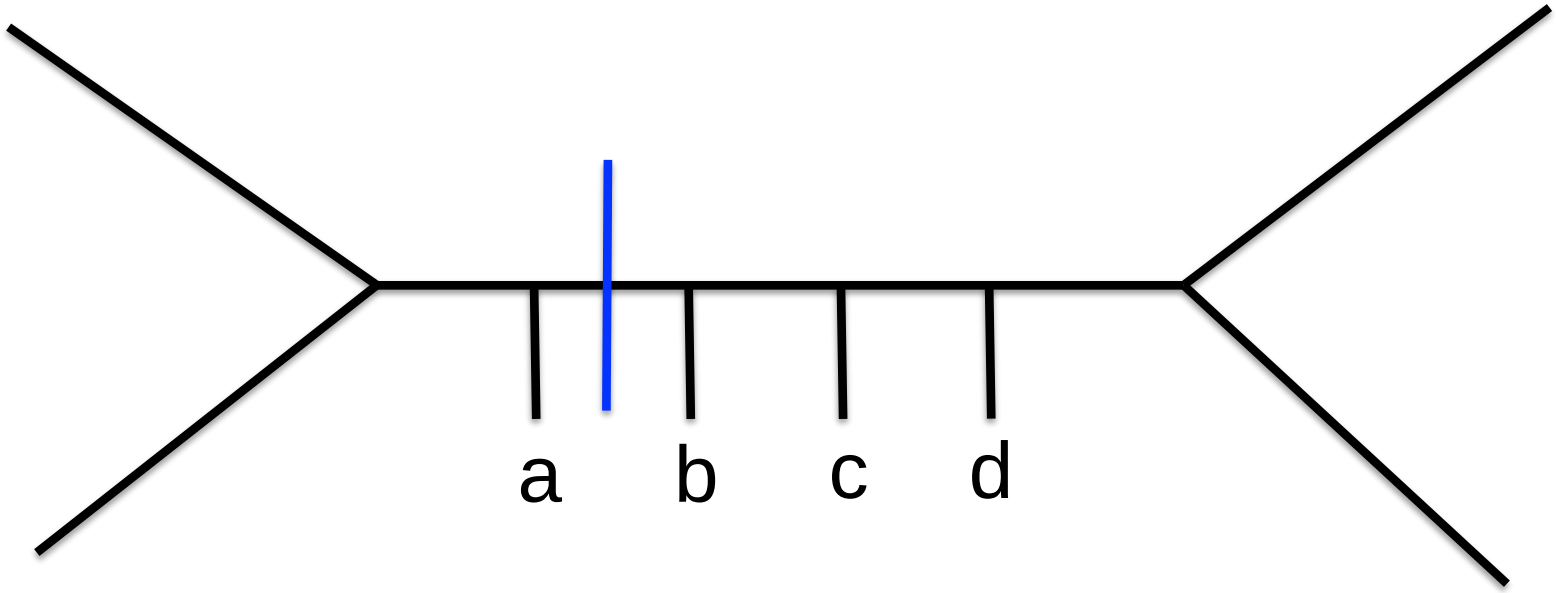
- *At most 3 taxa on a side if it has no breakpoints;*
- *At most 4 taxa on a side if it has 1 breakpoint;*
- *At most 4 taxa on a side if it has 2 breakpoints.*

By dividing *2k* breakpoints across *3(k-1)* sides, we concluded that the size of the new kernel is at most…

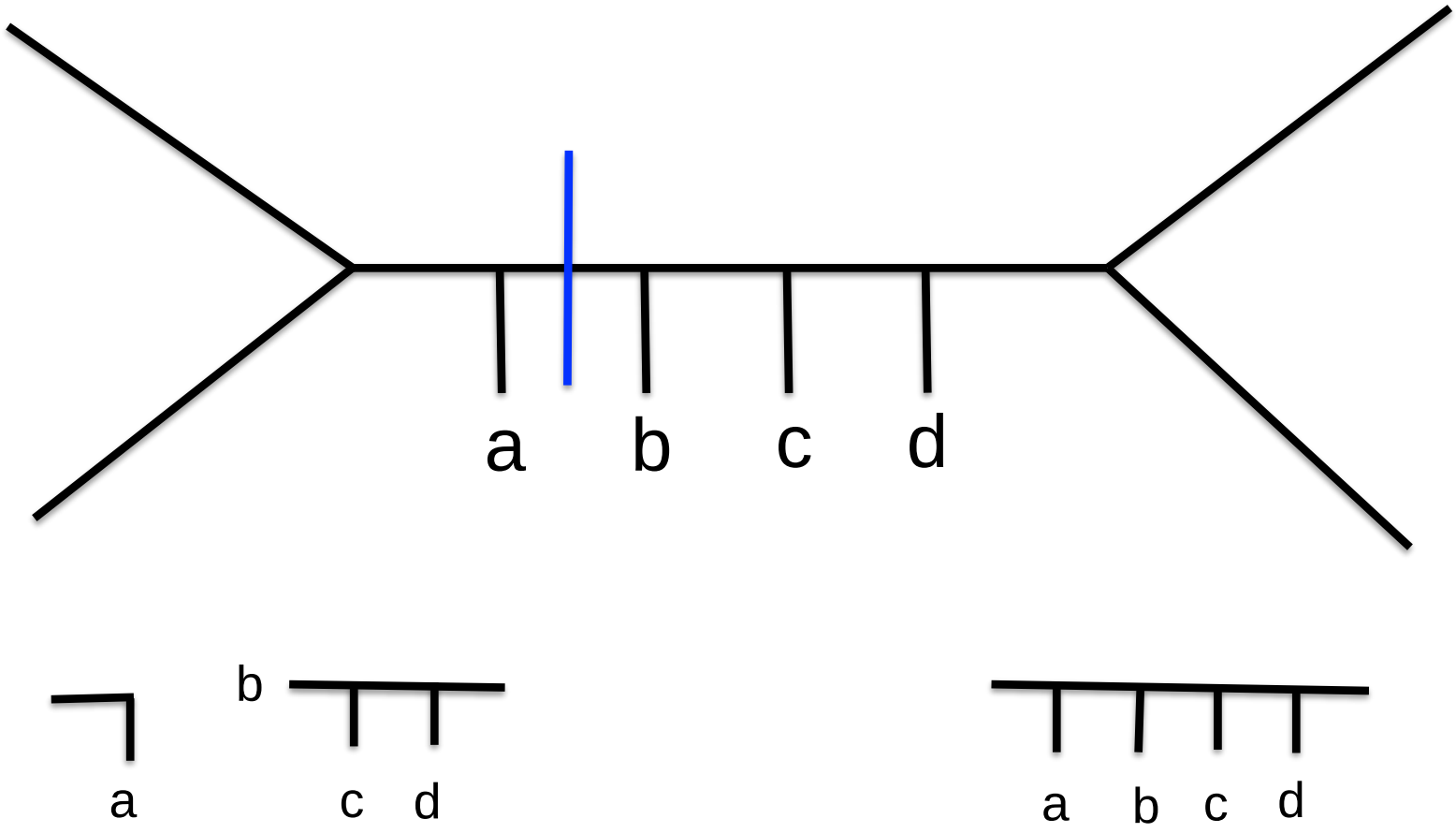
$$4*2k + 3*(k-3) = 11k-9.$$

The correctness of these new rules requires use of the *agreement forest* characterization of d_{TBR}.

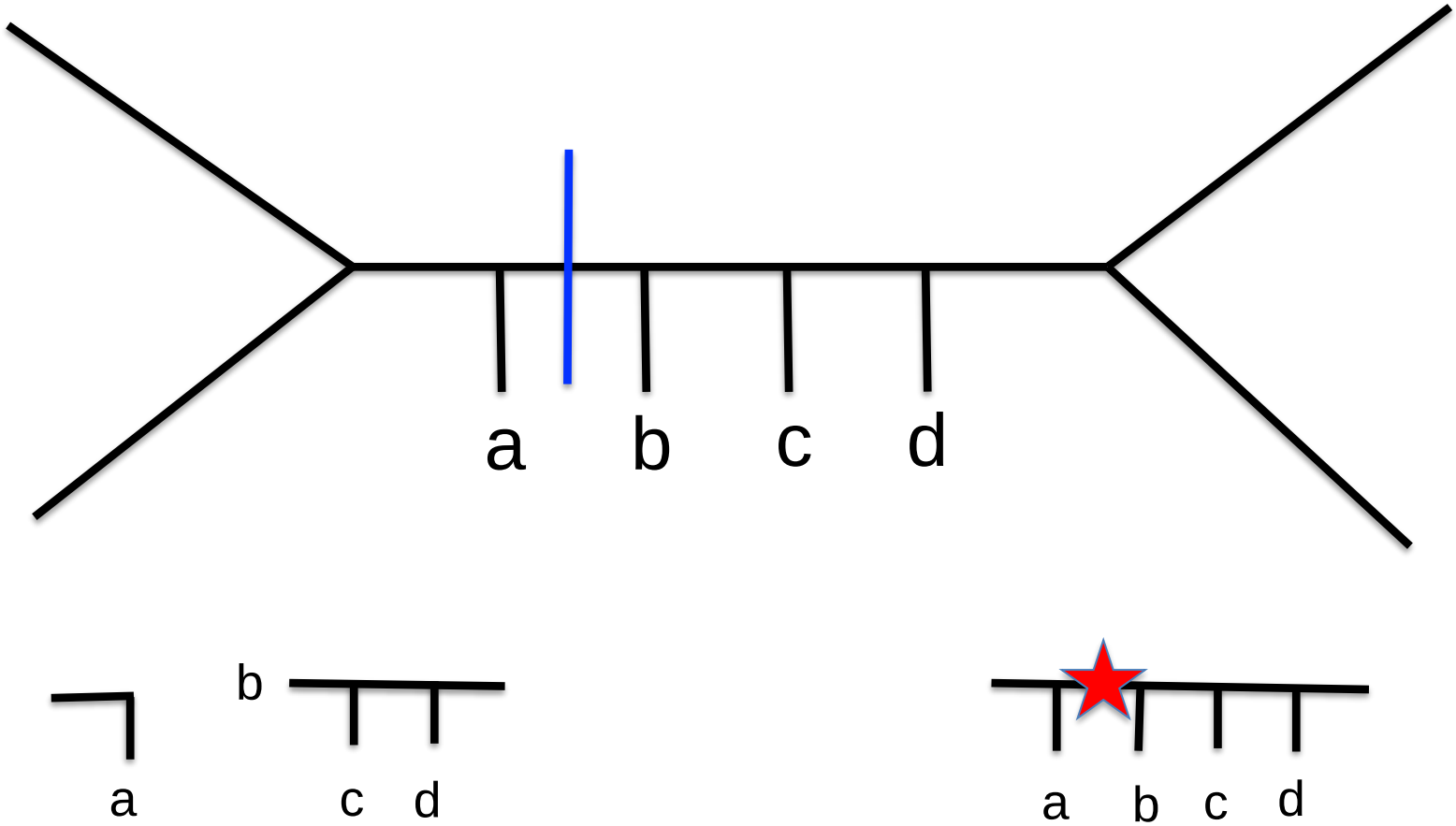
First bottleneck: a “1|3” side

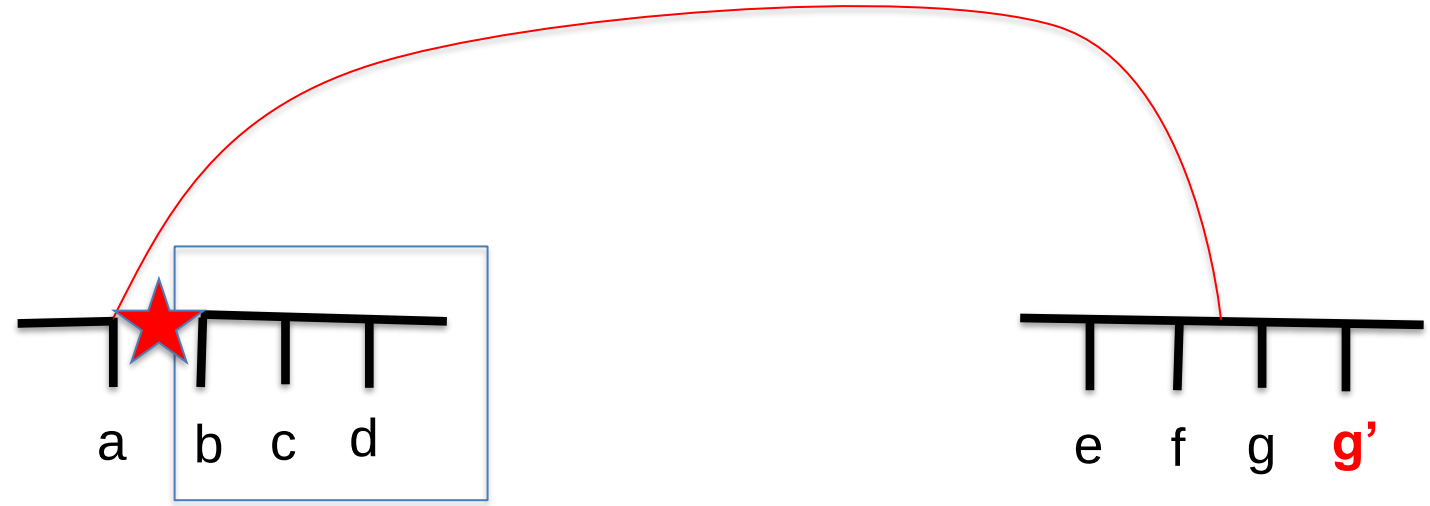


First bottleneck: a “1|3” side



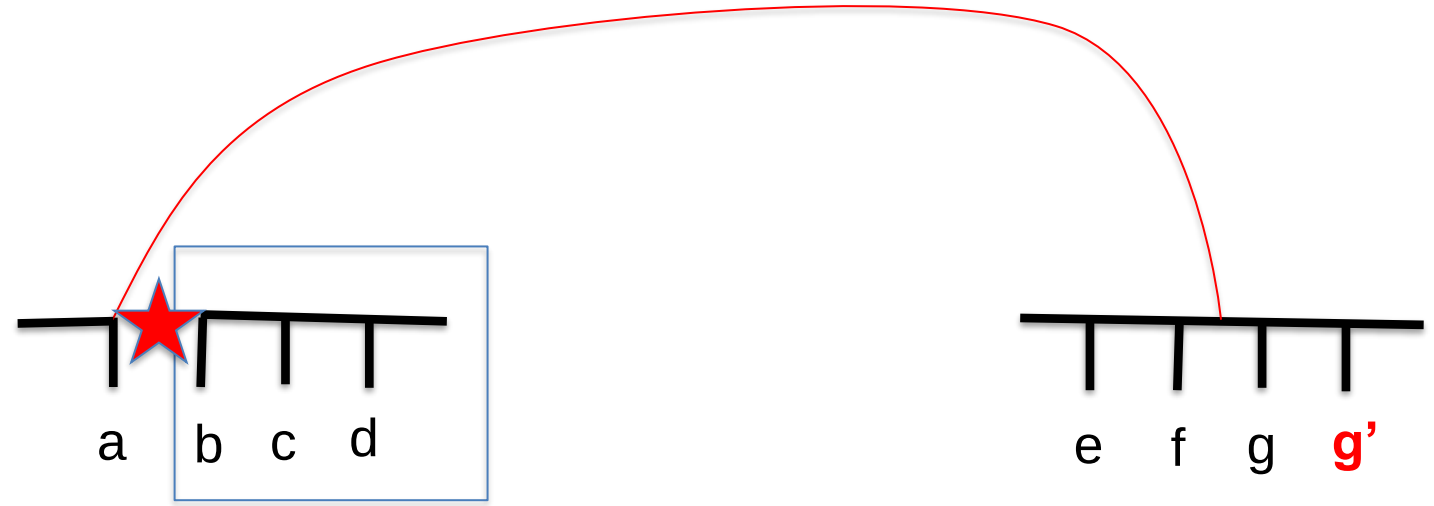
First bottleneck: a “1|3” side





Some other common 3-chain…

This is distance preserving…and creates a common subtree that we can reduce!

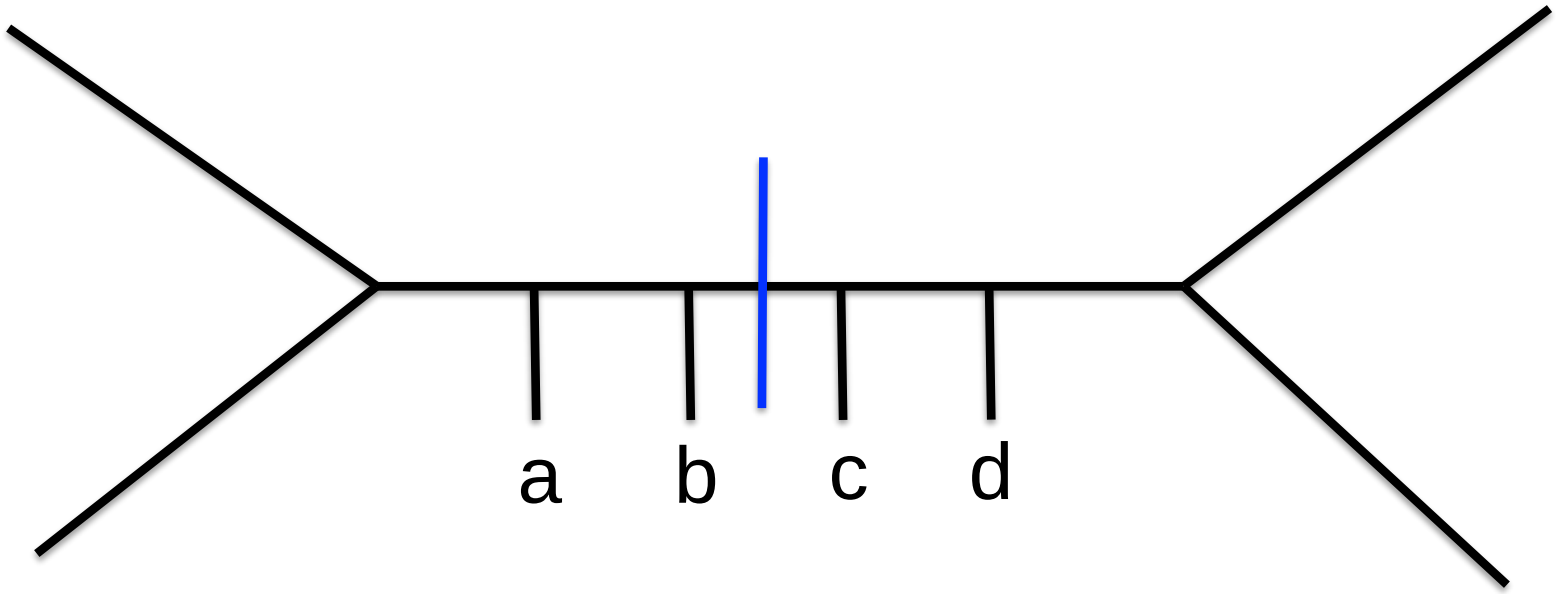


1|3 sides can eat each other…

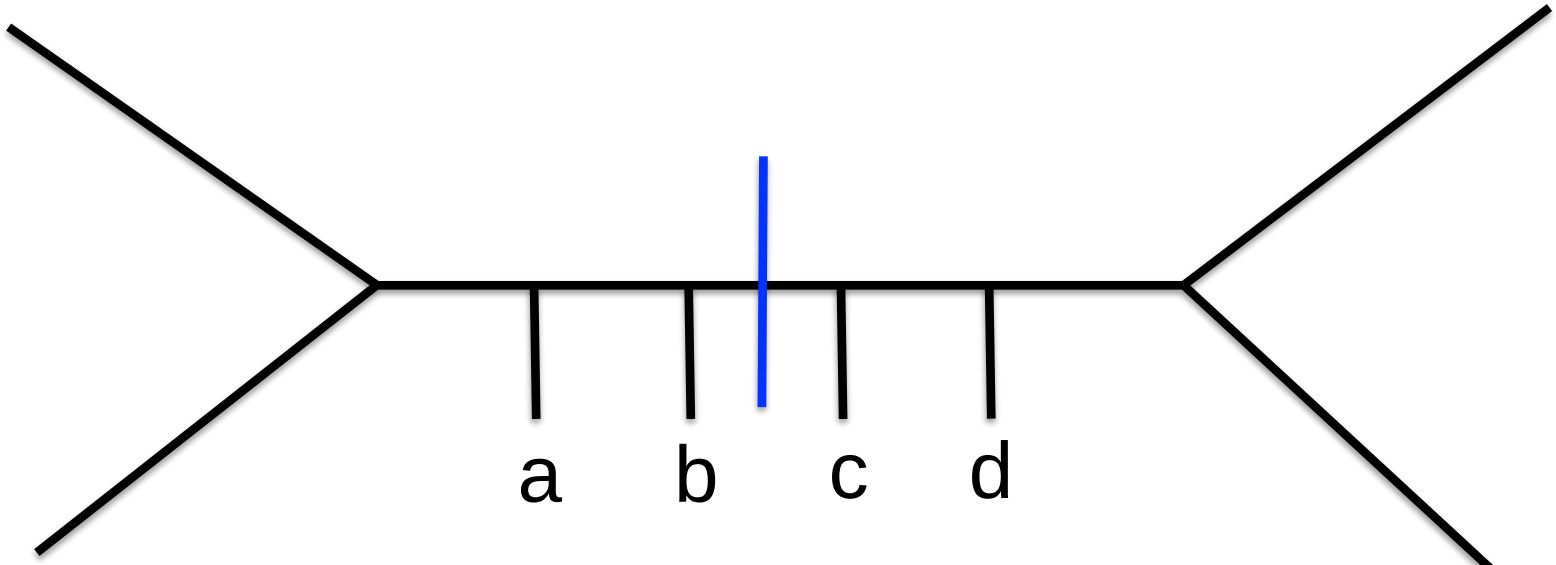
Idea. Each of these 1|3 sides contains a common chain of length 3, so you can use the chain in one 1|3 side to trigger the reduction of another 1|3 side!

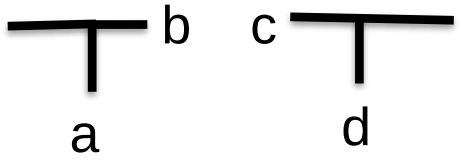
After doing this to exhaustion, there can be at most one 1|3 side.

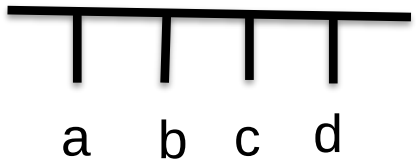
Second bottleneck: a “2|2” side



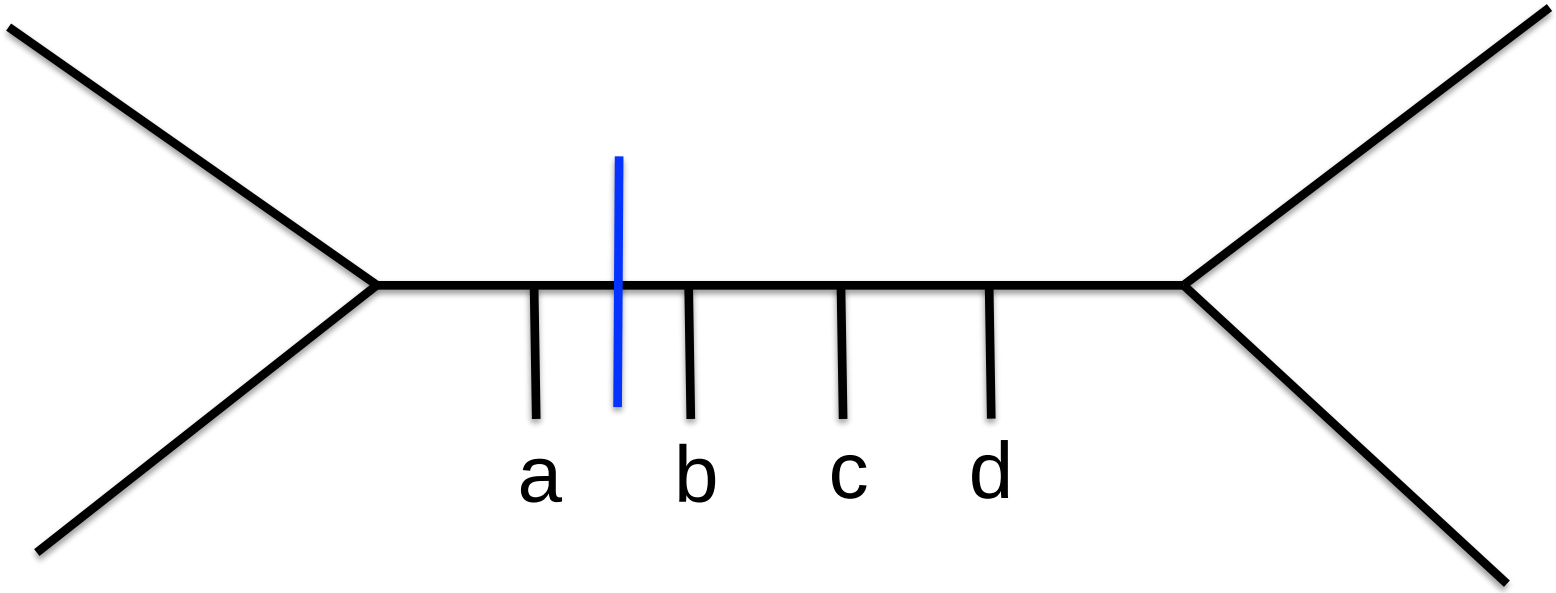
Second bottleneck: a “2|2” side

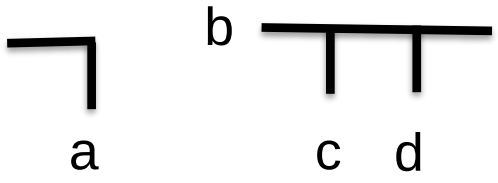


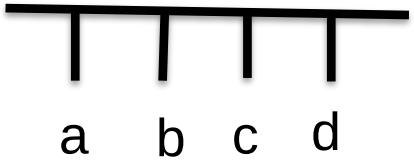




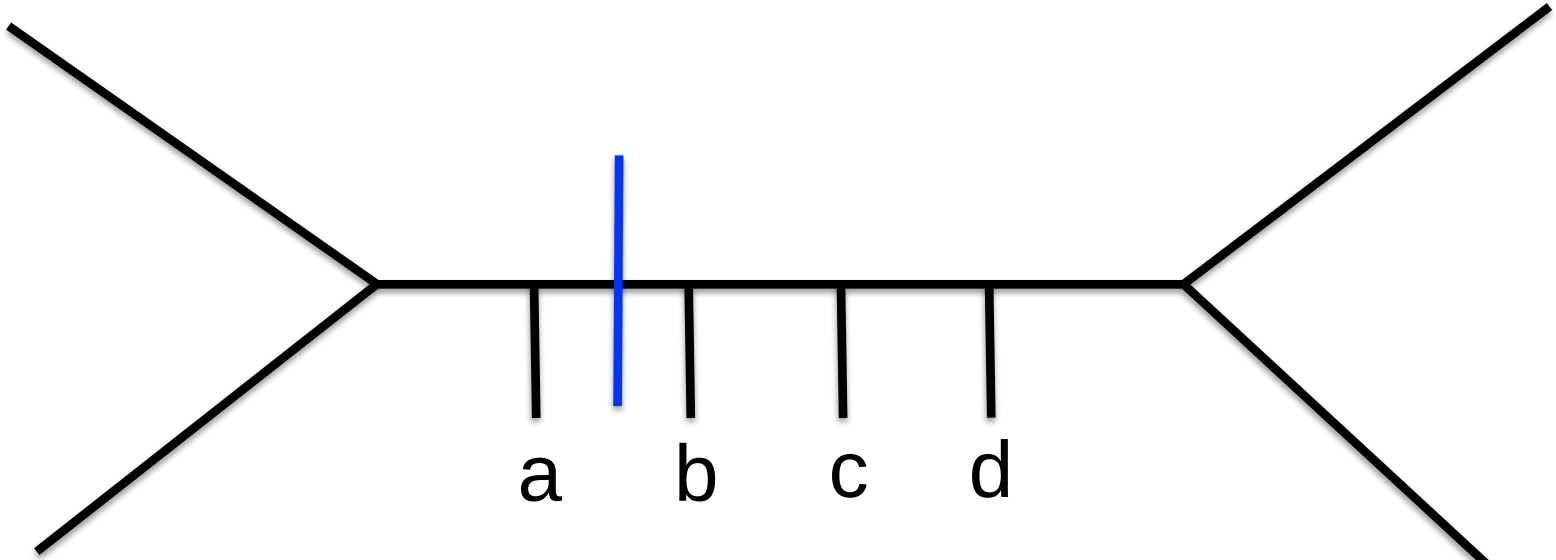
If we could turn 2|2 sides into 1|3 sides we could then use the 1|3 sides to eat each other. BUT….

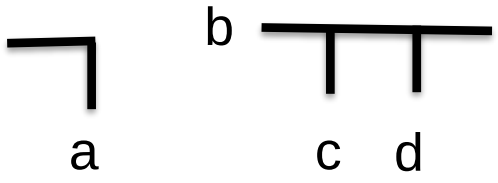


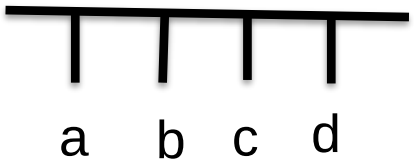




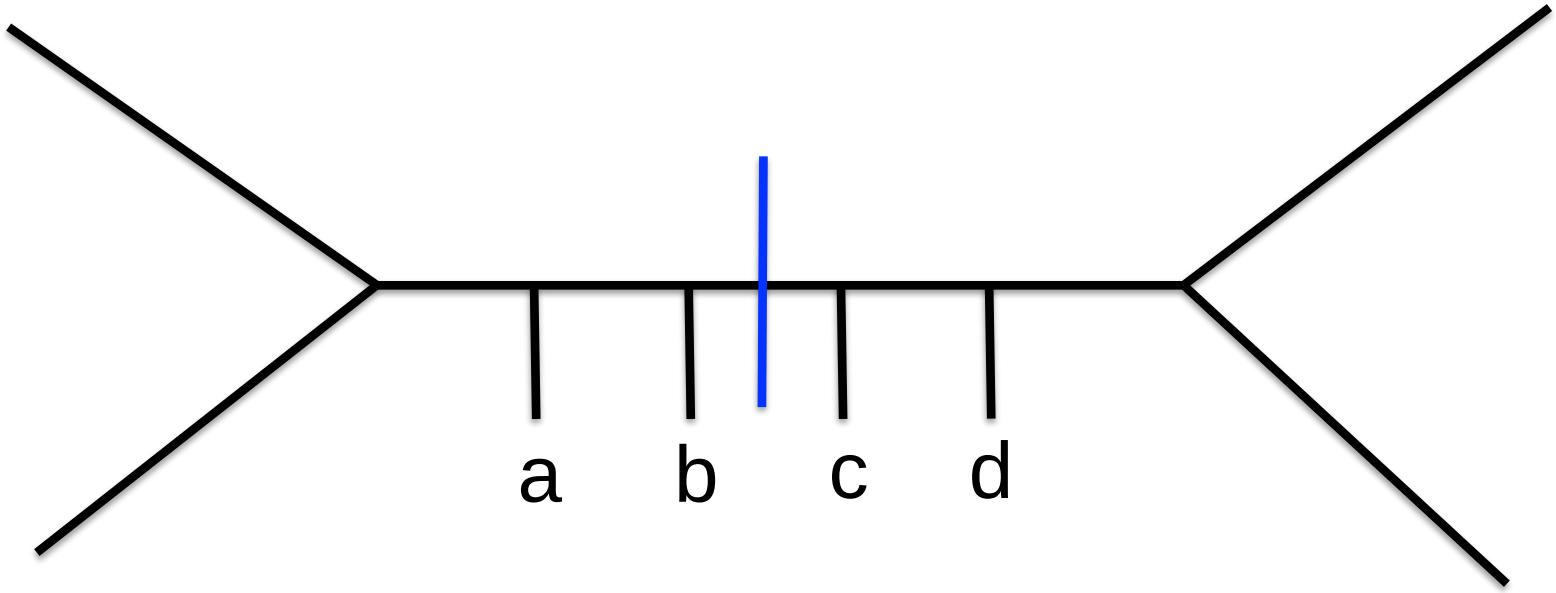
…it is not always allowed to “flip” a 2|2 side into a 1|3 side. So when is it allowed?

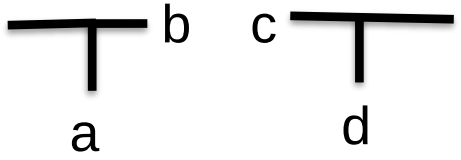


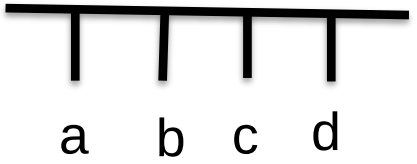




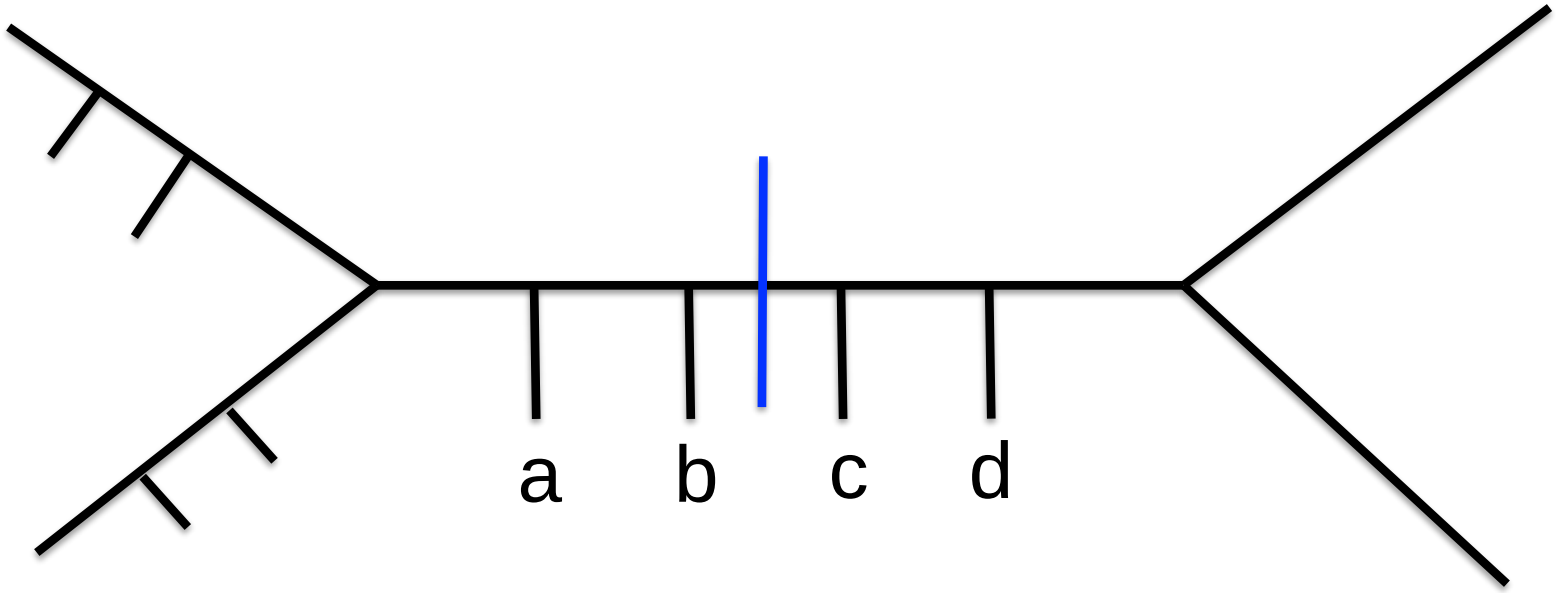
A 2|2 side can be turned into a 1|3 side when there are "many" leaves on adjacent sides.

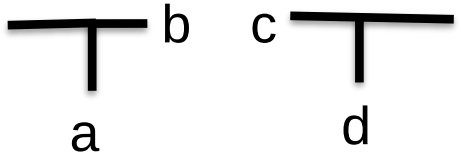


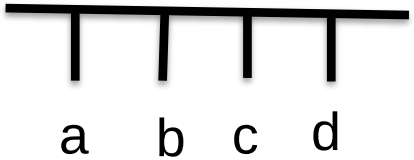




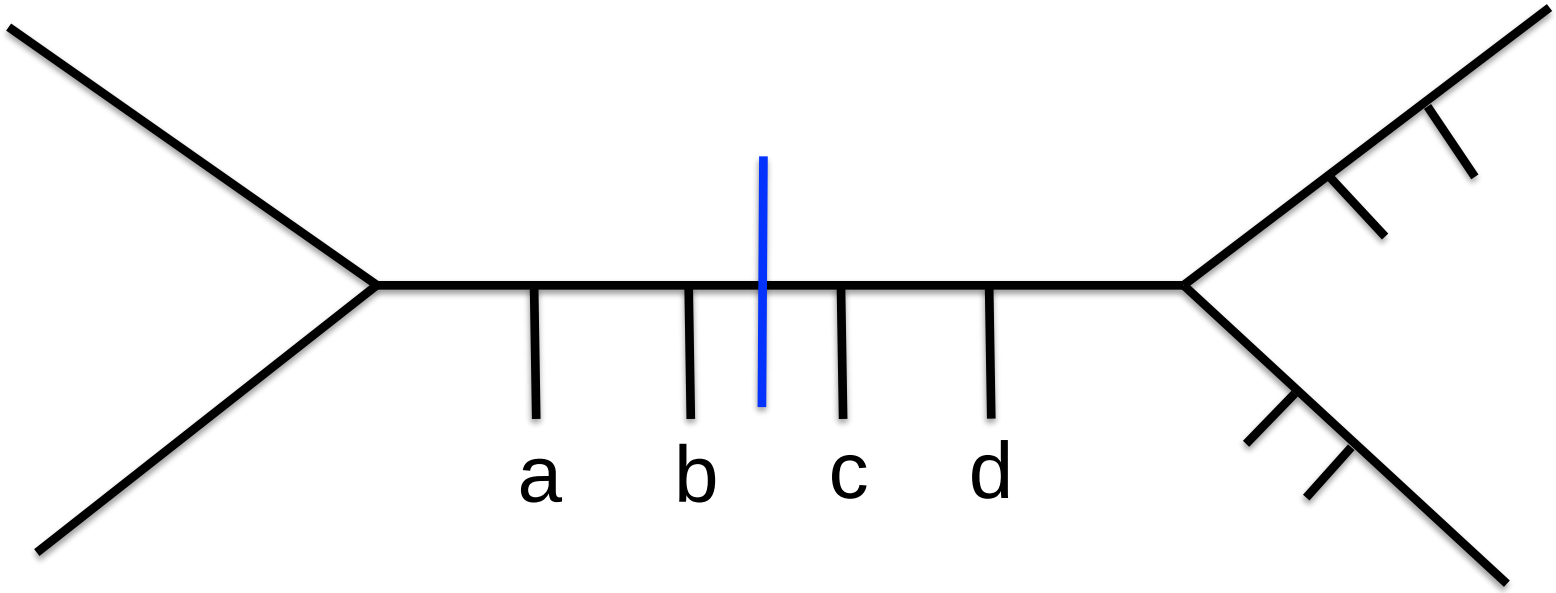
A 2|2 side can be turned into a 1|3 side when there are "many" leaves on adjacent sides.

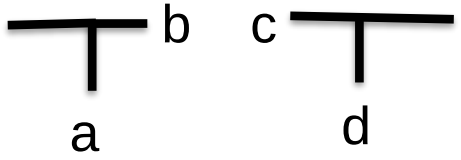


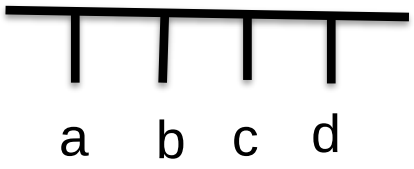




A 2|2 side can be turned into a 1|3 side when there are "many" leaves on adjacent sides.







1|3 and densely flanked 2|2 sides obliterate each other!

Idea… 2|2 sides that have many leaves on adjacent sides (“densely flanked 2|2 sides”) can be turned into 1|3 sides, which can then eat themselves.

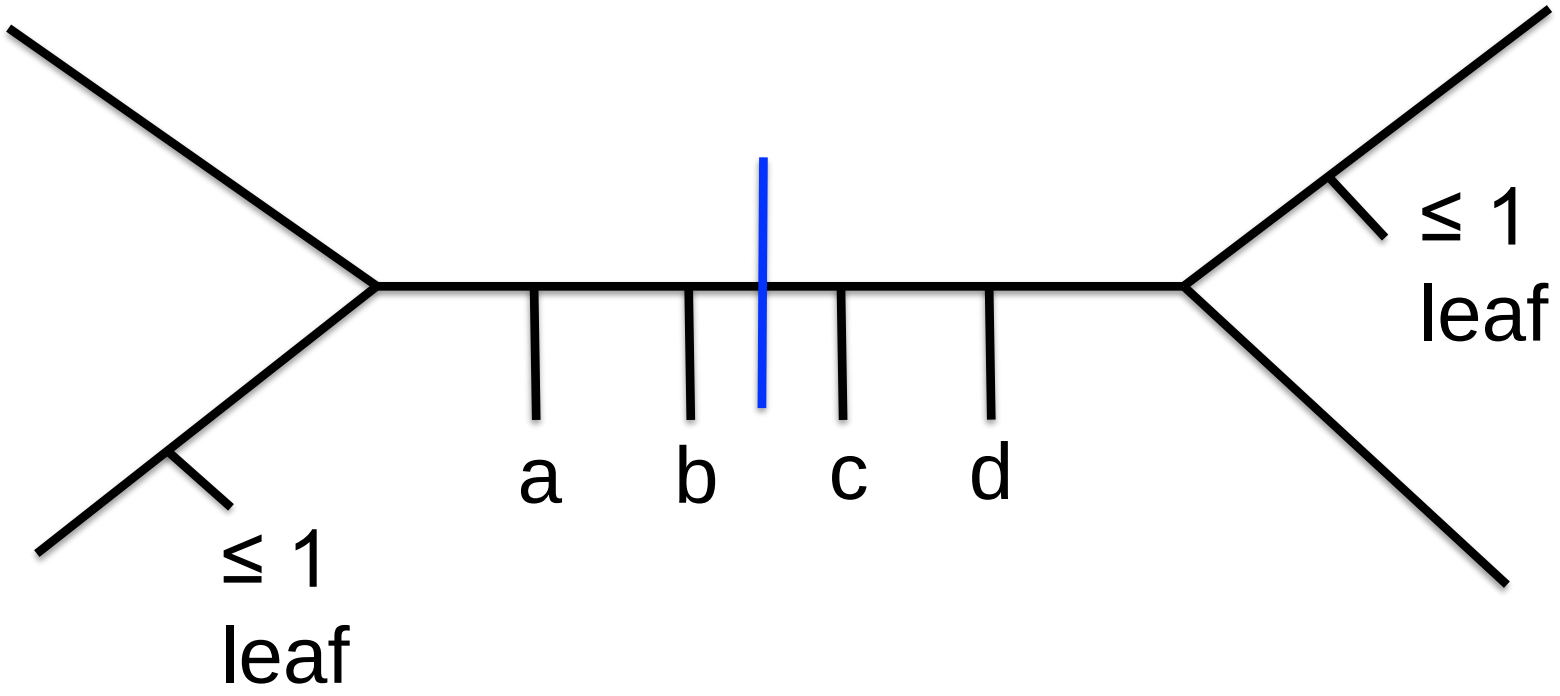
At the point that this process cannot continue anymore, all but 1 of the 1|3 sides and the densely flanked 2|2 sides have gone.

(A similar type of dense-flanking argument can be used to prove that 2|1|1 sides, a third type of bottleneck, can also be destroyed, but I won’t talk about that today.)

Insight… Apart from 1 possible exception, the only surviving sides with 4 leaves are “sparsely flanked” i.e. have relatively few leaves on adjacent sides.

So viewed together they contribute on average fewer than 4 leaves per side.

The only surviving sides
with 4 leaves are sparsely flanked:



Sketch of upper bounding argument

- We have *2k* breakpoints to divide across *3(k-1)* sides.
- We can safely assume there are no sides with 0 or 2 leaves.
- Let p, q, r be the number of sides with 4, 3 or 1 leaves.
- Crucially: all except ≤1 sides with 4 leaves are "sparsely flanked", which means they have at least two adjacent sides with 1 leaf.
- But each side with one leaf can be shared by at most 4 sides with 4 leaves, so $r \geq (2/4)p$.

Sketch of upper bounding argument

- We have *2k* breakpoints to divide across *3(k-1)* sides.
- We can safely assume there are no sides with 0 or 2 leaves.
- Let p, q, r be the number of sides with 4, 3 or 1 leaves.
- Crucially: all except ≤1 sides with 4 leaves are "sparsely flanked", which means they have at least two adjacent sides with 1 leaf.
- But each side with one leaf can be shared by at most 4 sides with 4 leaves, so $r \geq (2/4)p$.
- If we crunch the numbers this gives an upper bound of *9k-8* on the kernel ☺

Conclusions and future work

We achieved the improvement from *11k-9* to *9k-8* by introducing three new powerful reduction rules.

Can we go below *9k-8* ? Probably, but…

…auxiliary proofs and lemmas are already extremely technical ☹

Can we analytically and/or computationally (semi-)automate the search for new reduction rules, proofs of correctness and bounding arguments to keep proof complexity under control?

Can the new reduction rules be used elsewhere?

Do the new reduction rules have added value in practice? (Probably: the *11k-9* rules already work better in practice than the *15k-9* rules: Wersch, K., Linz, Stamoulis, Annals of Operations Research 2022)

Thank you for listening!

More details at:

- **Deep kernelization for the Tree Bisection and Reconnnect (TBR) distance in phylogenetics,** https://arxiv.org/abs/2206.04451 (K., Linz and Meuwese, 2022)
- **New reduction rules for the tree bisection and reconnection distance** (K. and Linz, Annals of Combinatorics 24(3), 2020)