Subtree Prune and Regraft on Ranked Trees

Lena Collienne[1,2], Chris Whidden[3], Alex Gavryushkin[1]

School of Mathematics and Statistics, University of Canterbury[1]
Department of Computer Science, University of Otago[2]
Faculty of Computer Science, Dalhousie University[3]

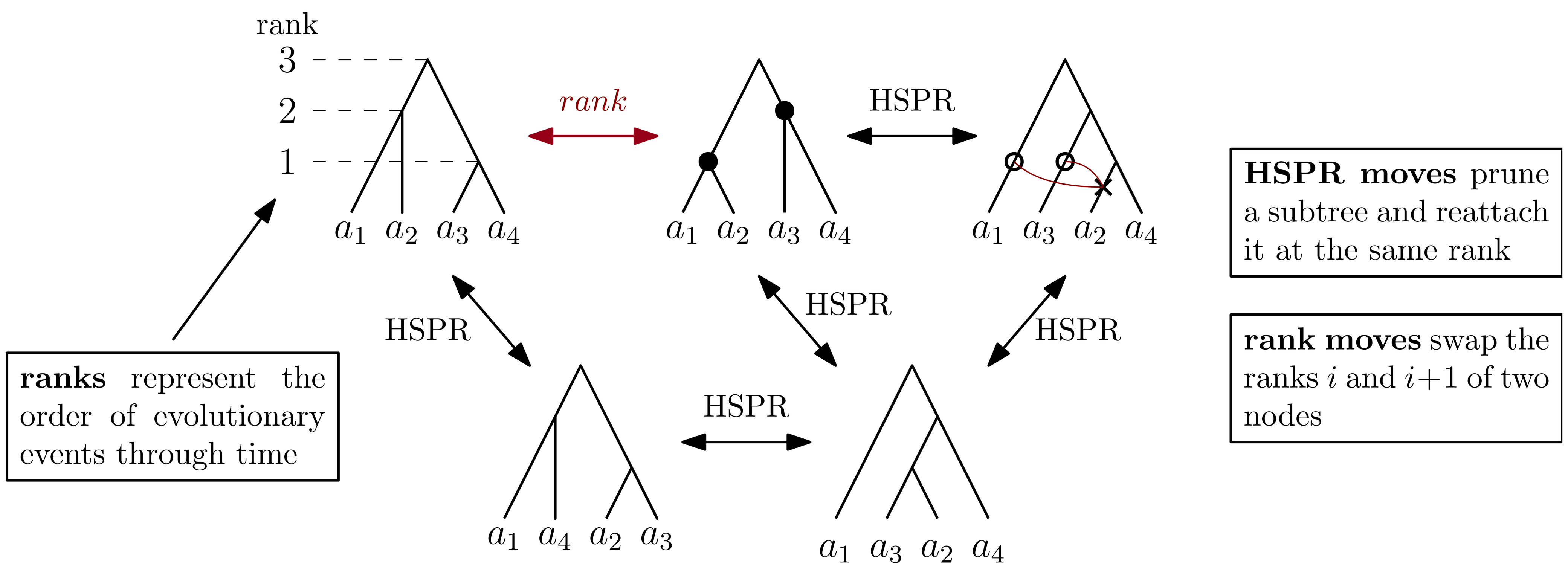
Introduction

Tree rearrangement operations like **NNI** (Nearest Neighbour Interchange) and **SPR** (Subtree Prune and Regraft) perform a local change to a given tree to produce a similar but not identical tree. These operations can be used to define a distance between two trees as the minimum number of rearrangements needed to transform one tree into another and furthermore define tree spaces, which allow statistical analyses of distribution over trees.

An SPR *move* prunes a subtree and reattaches it at a different place in the tree. These moves are motivated by biological processes like horizontal gene transfer and hybridisation. Due to its biological motivation, the SPR tree space is especially suitable for analysing a collection of gene trees and there has been much work studying this tree space on unranked trees. We consider an extension of the SPR rearrangement to ranked trees, which contain some information about times of evolutionary events, and introduce and analyse two different tree spaces based on it: **HSPR** and **RSPR** space.

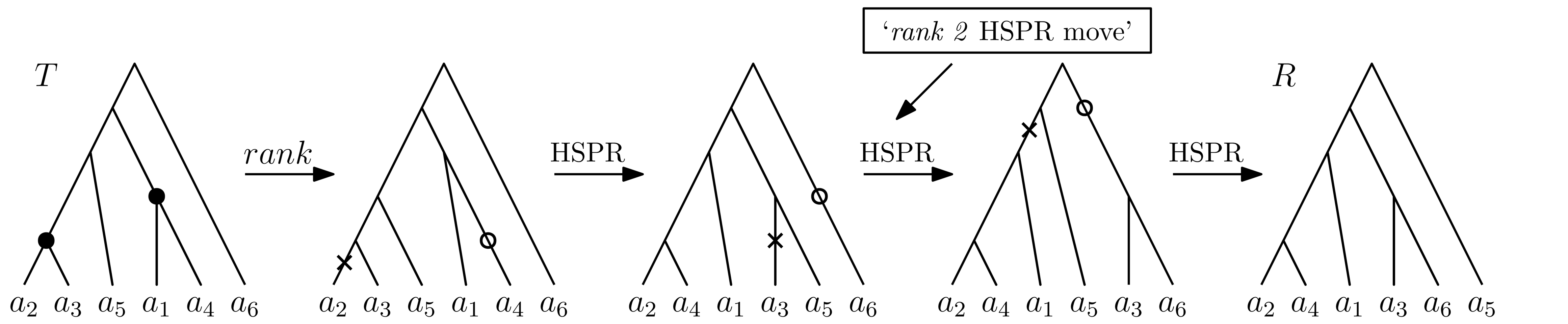
These tree spaces are of particular interest as their properties have important implications for phylogenetic inference methods (including parsimony, distance-based, likelihood, and Bayesian paradigms) on time trees.

HSPR and RSPR Space

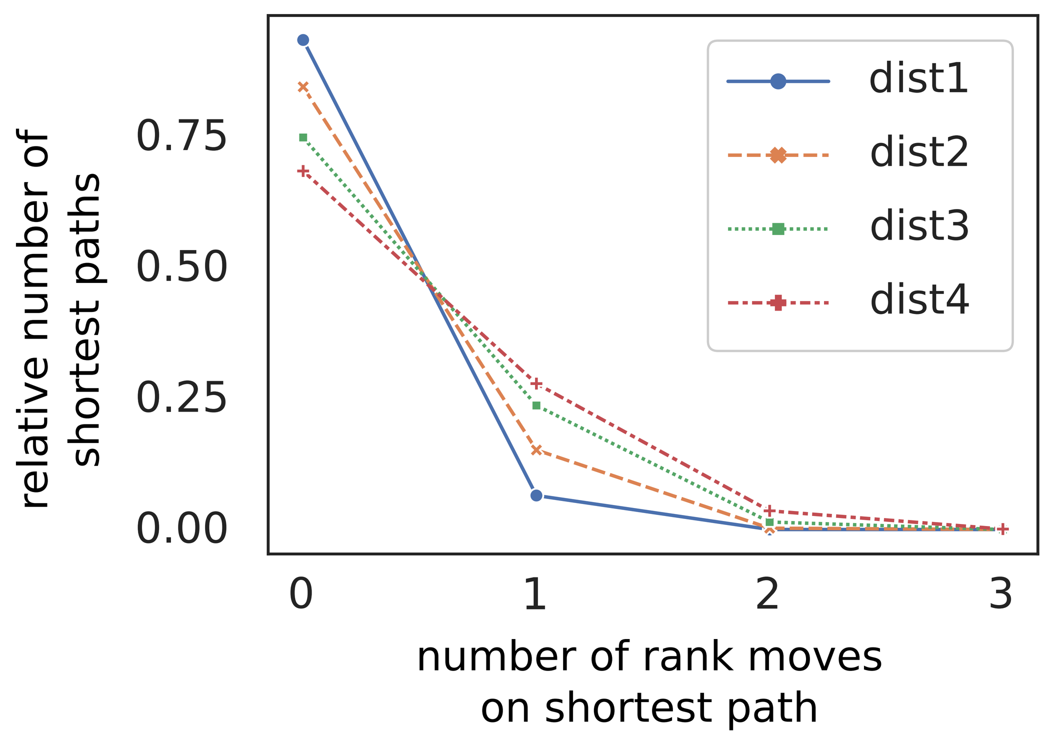


Definition: *Horizontal* SPR *(*HSPR*) space: Ranked trees connected by* HSPR *moves.*
Ranked SPR *(*RSPR*) space: Ranked trees connected by* HSPR *or rank moves.*

Shortest Paths

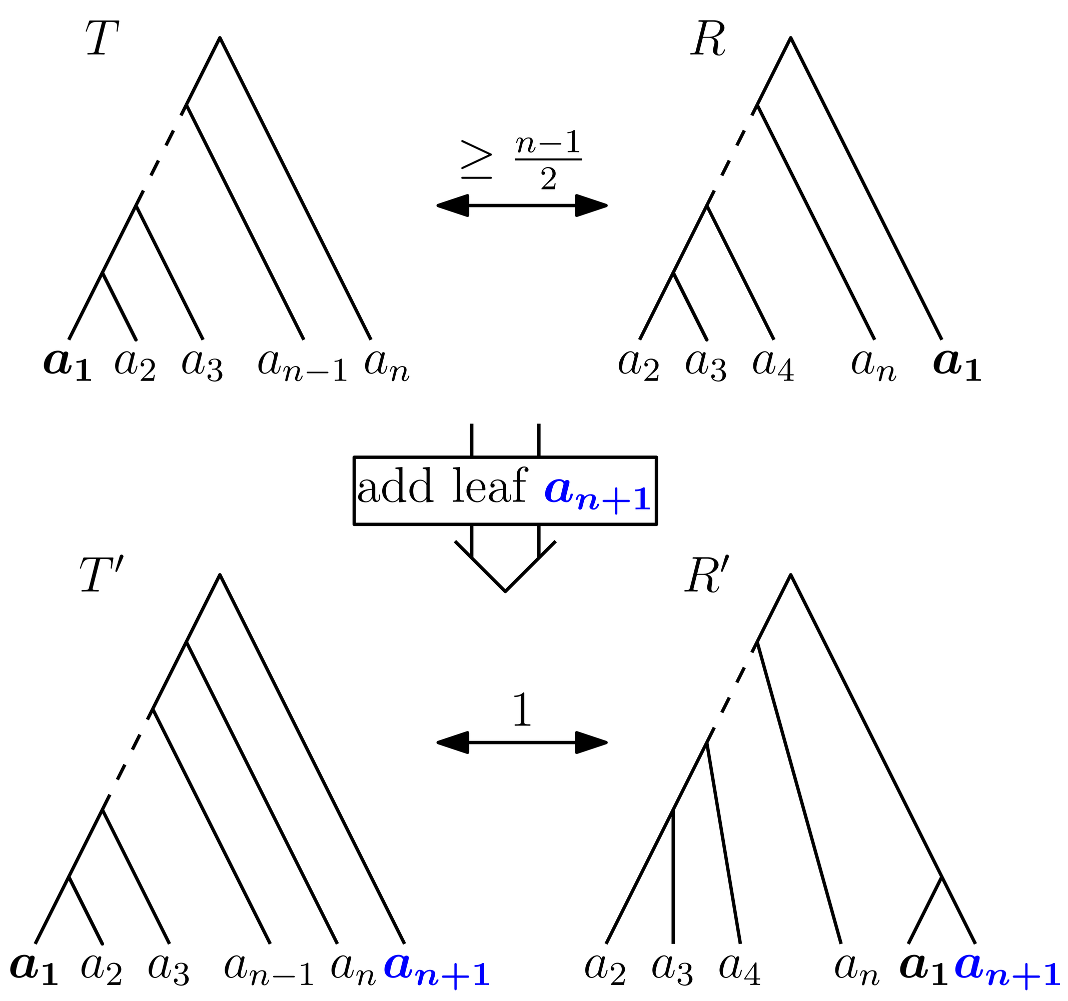


Theorem:
In RSPR*, there is always a shortest path that has all rank moves at the beginning.*
In HSPR*, there is always a shortest path on which the ranks of* HSPR *moves increase monotonically.*



Adding Leaves

Theorem: *Adding a leaf can decrease the distance by* $\frac{n-3}{2}$ *in both* HSPR *and* RSPR.



This is very surprising. There is no other tree space where adding a leaf can decrease the distance, and this might have consequences when analysing posterior samples of trees after adding more leaves.

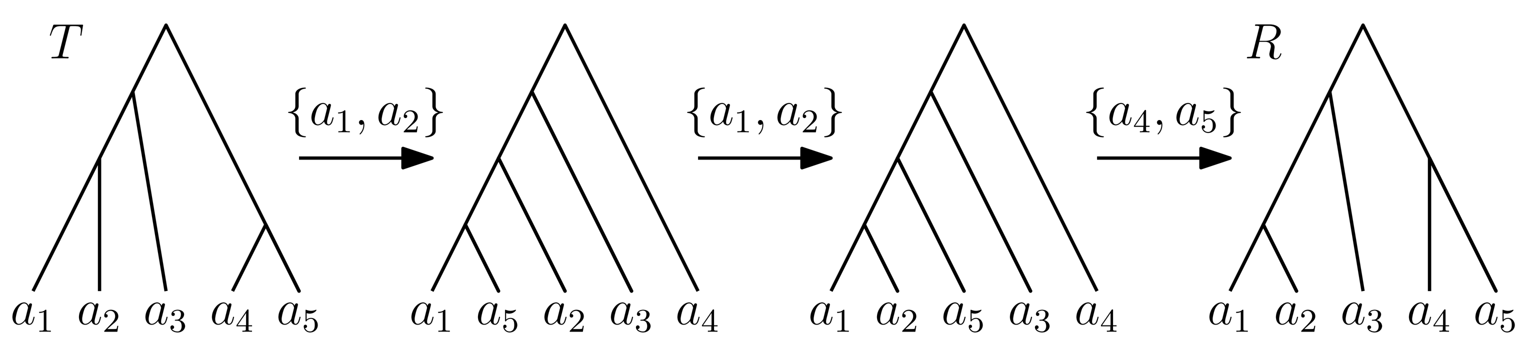
Neighbourhood sizes

Theorem: *The number* $|NH(T)|$ *of trees with distance one from a given tree* T *in* HSPR *and* RSPR *are:*

$$\text{HSPR}: \quad |NH(T)| = (n-1)(n-2)$$
$$\text{RSPR}: \quad |NH(T)| = (n-1)(n-2) + r_T$$

(r_T *= number of rank intervals in* T*)*

Diameter



(Bottom-up approach to approximate HSPR distance)

Theorem: *Lower and upper bound for the diameter, i.e. the maximum distance, in* HSPR *and* RSPR *are:*

$$\frac{n}{2} \leq \Delta(\text{HSPR}), \Delta(\text{RSPR}) \leq 2(n-2).$$

Conclusion

Known techniques for unranked SPR do not work for HSPR or RSPR:

– No cluster property
– Distances increase by linear amount when deleting leaf
– One subtree can move multiple times

⇒ **MAFs don't work**

Adding leaves can change ranked SPR distances significantly, which might have an effect on tree inference algorithm when adding more leaves.

Open questions:

➔ Complexity of computing distances. Conjecture: NP-hard
➔ Exact Diameter. Conjecture:

$$\Delta(\text{HSPR}) = \begin{cases} \frac{3}{2}(n-2) & \text{if } n \text{ is even} \\ \frac{3}{2}(n-2) - \frac{1}{2} & \text{if } n \text{ is odd} \end{cases}$$

lena@lenacoll.de
lenacoll
lenacoll.de