# Efficiently Constructing Phylogenies from Partial Characters

Bonnie Kirkpatrick* Kristian Stevens†

Determining phylogenetic relationships is of central importance to biology. Not only are the phylogenies themselves interesting, but these methods are useful for predicting functional and structural similarities between genes. Most methods for tree reconstruction under the perfect phylogeny model assume that data are given for all the taxa. This is because the general problem of finding a perfect phylogeny for incomplete or partial characters is NP-hard even for binary characters [2].

It is useful to relax the assumption of fully observed data, particularly in the case of genotype and sequence data. Furthermore, when the data has more than two states, the perfect phylogeny problem can be formulated as a binary problem with missing data.

This work discusses classes of partial characters for which there are polynomial-time algorithms for constructing the perfect phylogeny tree. Specifically, we consider the *rich data hypothesis* introduced by Halperin and Karp [1]. We introduce two broader criteria for partial characters inspired by the rich data hypothesis under which phylogenies can be found in polynomial time. One of these criteria apply to multi-state partial characters.

We also give the first enumeration algorithm for phylogenies compatible with binary partial characters satisfying the rich data hypothesis. This algorithm runs in $O(nm^2)$ time where $n$ is the number of taxa and $m$ is the number of characters. By enumerating the phylogenies, one can compute probabilities of the observed data under the coalescent model with infinite sites. This algorithm gives the first known way to compute those probabilities for missing data.

## References

[1] Eran Halperin and Richard M. Karp. Perfect phylogeny and haplotype assignment. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, pages 10–19, New York, NY, USA, 2004. ACM Press.

[2] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 1992:91–116, 1992.

*Electrical Engineering and Computer Sciences, University of California Berkeley, bbkirk@eecs.berkeley.edu

†Computer Science, University of California Davis, kastevens@ucdavis.edu