Quartets Enable Statistically Consistent Estimation of Cell Lineage Trees Under an Unbiased Error and Missingness Model

Yunheng Han
Department of Computer Science, University of Maryland, College Park, MD, USA

Erin K. Molloy[1]
Department of Computer Science, University of Maryland, College Park, MD, USA

Abstract

Cancer progression and treatment can be informed by reconstructing its evolutionary history from tumor cells [5]. Although many methods exist to estimate evolutionary trees (called phylogenies) from molecular sequences, traditional approaches assume the input data are error-free and the output tree is fully resolved. These assumptions are challenged in tumor phylogenetics because single-cell sequencing produces sparse, error-ridden data and because tumors evolve clonally [3, 12]. Here, we study the theoretical utility of methods based on quartets (four-leaf, unrooted phylogenetic trees) and triplets (three-leaf, rooted phylogenetic trees), in light of these barriers.

Quartets and triplets have long been used as the building blocks for reconstructing the evolutionary history of species [14]. The reason triplet-based methods (e.g., MP-EST [6]) and quartet-based methods (e.g., ASTRAL [7]) have garnered such success in species phylogenetics is their good statistical properties under the Multi-Species Coalescent (MSC) model [9, 10] (see [1] and [2] for identifiability results under the MSC model for quartets and triplets, respectively).

Inspired by these efforts, we study the utility of quartets and triplets for estimating cell lineage trees under a popular tumor phylogenetics model [3, 11, 15, 4] with two phases. First, mutations arise on a (highly unresolved) cell lineage tree according to the infinite sites model, and second, errors (false positives and false negatives) and missing values are introduced to the resulting mutation data in an unbiased fashion, mimicking data produced by single-cell sequencing protocols. This infinite sites plus unbiased error and missingness (IS+UEM) model generates mutations (rather than gene genealogies like the MSC model). However, a quartet (with leaves bijectively labeled by four cells) is implied by a mutation being present in two cells and absent from two cells [8, 13]; similarly, a triplet (on three cells) is implied by a mutation being present in two cells and absent from one cell.

Our main result is that under the IS+UEM, the most probable quartet identifies the unrooted model cell lineage tree on four cells, with a mild assumption: the probability of false negatives and the probability of false positives must not sum to one. Somewhat surprisingly, our identifiability result for quartets does not extend to triplets, with more restrictive assumptions being required for identifiability. These results motivate seeking an unrooted cell lineage tree such that the number of quartets shared between it and the input mutations is maximized. We prove an optimal solution to this problem is a consistent estimator of the unrooted cell lineage tree under the IS+UEM model; this guarantee includes the case where the model tree is highly unresolved, provided that tree error is defined as the number of false negative branches. We therefore conclude by outlining how quartet-based methods might be employed for tumor phylogenetics given other important challenges like copy number aberrations and doublets.

2012 ACM Subject Classification Applied computing → Molecular evolution

Keywords and phrases Tumor Phylogenetics, Cell Lineage Trees, Quartets, Supertrees, ASTRAL

Digital Object Identifier 10.4230/LIPIcs.WABI.2023.8

Category Abstract

[1] Corresponding author

23rd International Workshop on Algorithms in Bioinformatics (WABI 2023).
Editors: Djamal Belazzougui and Aïda Ouangraoua; Article No. 8; pp. 8:1–8:2
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Related Version *Full Version*: https://doi.org/10.1101/2023.04.04.535437

Funding This work was financially supported by the State of Maryland.

Acknowledgements The authors thank Michael Nute for very helpful feedback on a preliminary version of this paper, especially our notation.

References

1 Elizabeth S. Allman, James H. Degnan, and John A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62(6):833–862, 2011. doi:10.1007/s00285-010-0355-7.

2 James H. Degnan and Noah A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLOS Genetics*, 2(5):1–7, May 2006. doi:10.1371/journal.pgen.0020068.

3 Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17:86, 2016. doi:10.1186/s13059-016-0936-x.

4 Can Kizilkale, Farid Rashidi Mehrabadi, Erfan Sadeqi Azer, Eva Pérez-Guijarro, Kerrie L. Marie, Maxwell P. Lee, Chi-Ping Day, Glenn Merlino, Funda Ergün, Aydın Buluç, S. Cenk Sahinalp, and Salem Malikić. Fast intratumor heterogeneity inference from single-cell sequencing data. *Nature Computational Science*, 2:577–583, 2022. doi:10.1038/s43588-022-00298-x.

5 Bora Lim, Yiyun Lin, and Nicholas Navin. Advancing cancer research and medicine with single-cell genomics. *Cancer Cell*, 37(4):456–470, 2020. doi:10.1016/j.ccell.2020.03.008.

6 Liang Liu, Lili Yu, and Scott V. Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10:302, 2010. doi:10.1186/1471-2148-10-302.

7 Siavash Mirarab, Rezwana Reaz, Md. Shamsuzzoha Bayzid, Theo Zimmermann, Michelle S. Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014. doi:10.1093/bioinformatics/btu462.

8 Erin K. Molloy, John Gatesy, and Mark S. Springer. Theoretical and practical considerations when using retroelement insertions to estimate species trees in the anomaly zone. *Systematic Biology*, 71(3):721–740, 2021. doi:10.1093/sysbio/syab086.

9 Pekka Pamilo and Masatoshi Nei. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583, 1988. doi:10.1093/oxfordjournals.molbev.a040517.

10 Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003. doi:10.1093/genetics/164.4.1645.

11 Edith M. Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17(1):69, 2016. doi:10.1186/s13059-016-0929-9.

12 Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017. doi:10.1038/nrg.2016.170.

13 Mark S. Springer, Erin K. Molloy, Daniel B. Sloan, Mark P. Simmons, and John Gatesy. ILS-aware analysis of low-homoplasy retroelement insertions: Inference of species trees and introgression using quartets. *Journal of Heredity*, 111(2):147–168, 2019. doi:10.1093/jhered/esz076.

14 Mark Wilkinson, James A. Cotton, Chris Creevey, Oliver Eulenstein, Simon R. Harris, Francois-Joseph Lapointe, Claudine Levasseur, James O. Mcinerney, Davide Pisani, and Joseph L. Thorley. The Shape of Supertrees to Come: Tree Shape Related Properties of Fourteen Supertree Methods. *Systematic Biology*, 54(3):419–431, 2005. doi:10.1080/10635150590949832.

15 Yufeng Wu. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics*, 36(3):742–750, August 2019. doi:10.1093/bioinformatics/btz676.