Title: *Combinatorial models for reconstructing evolutionary histories*

Abstract:

Reconstructing evolutionary histories is a fundamental problem in biology, studied across life's organizational hierarchy. Often, this evolutionary reconstruction task, referred to as *phylogenetic inference*, is framed as a combinatorial optimization problem under an appropriate evolutionary model. In this work, we examine two such phylogenetic inference problems, lineage tracing and the copy number tree problem, through the lens of combinatorial optimization.

CRISPR-Cas9 based genome editing combined with single-cell sequencing enables the tracing of the history of cell divisions, or *lineage tracing*, in tissues and whole organisms. While standard phylogenetic approaches may be applied to reconstruct cellular lineage trees from this data, the unique features of the CRISPR-Cas9 editing process motivate the development of specialized models that describe the evolution of CRISPR-Cas9 induced mutations. Here, we introduce *the star homoplasy model,* a novel evolutionary model that constrains a phylogenetic character to mutate at most once along a lineage, capturing the non-modifiability property of CRISPR-Cas9 mutations. We derive a combinatorial characterization of star homoplasy phylogenies by identifying a relationship between the star homoplasy model and the binary perfect phylogeny model. We use this characterization to develop an algorithm, *Startle* (Star tree lineage estimator), that computes a maximum parsimony star homoplasy phylogeny. We demonstrate that Startle outperforms other methods at lineage reconstruction on both real and simulated data.

Low-coverage single-cell DNA sequencing technologies enable the measurement of copy number profiles from thousands of individual cells within tumors. From this data, one can infer the evolutionary history of the tumor, referred to as a *copy number phylogeny*, by modeling transformations of the genome via copy number aberrations. A widely used model to infer such copy number phylogenies is the *copy number transformation (CNT)* model. While the CNT model is useful, no efficient algorithm has been developed to find the most parsimonious phylogeny under the CNT model due to its difficult, combinatorial properties. Here, we introduce the *zero-agnostic copy number transformation (ZCNT)* model, a simplification of the CNT model that allows the amplification or deletion of genomic loci with zero copies. We use our simplified model to derive polynomial time algorithms for two natural relaxations of the small parsimony problem on copy number profiles. While the alteration of zero copy number regions allowed under the ZCNT model is not biologically realistic, we show on both simulated and real datasets that the ZCNT model is a close approximation to the CNT model. Extending our polynomial time algorithm for the ZCNT small parsimony problem, we develop an algorithm, *Lazac*, for solving the large parsimony problem on copy number profiles. We demonstrate that Lazac outperforms existing methods for inferring copy number phylogenies on both simulated and real data.

Selected Textbook Chapters:

- *“Graph Theory”* by Reinhard Diestal
    - Chapter 1: The Basics
    - Chapter 2: Matching, Covering, and Packing
    - Chapter 3: Connectivity
    - Chapter 5: Coloring
    - Chapter 6: Flows
- *“Bioinformatics Algorithms: An Active Learning Approach”* by Pavel Pezner and Phillip Compeau

Selected Papers:

[1] R. Mihaescu, D. Levy, and L. Pachter, “Why neighbor-joining works,” Algorithmica, vol. 54, pp. 1–24, 2009.

[2] W. H. E. Day, D. S. Johnson, and D. Sankoff, “The computational complexity of inferring rooted phylogenies by parsimony,” Mathematical Biosciences, vol. 81, no. 1, pp. 33–42, Sep. 1986, doi: 10.1016/0025-5564(86)90161-6.

[3] P. Bonizzoni, C. Braghin, R. Dondi, and G. Trucco, “The binary perfect phylogeny with persistent characters,” Theoretical Computer Science, vol. 454, pp. 51–63, Oct. 2012, doi: 10.1016/j.tcs.2012.05.035.

[4] M. El-Kebir, “SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error,” Bioinformatics, vol. 34, no. 17, pp. i671–i679, Sep. 2018, doi: 10.1093/bioinformatics/bty589.

[5] I. Elias, “Settling the Intractability of Multiple Alignment,” Journal of Computational Biology, vol. 13, no. 7, pp. 1323–1339, Sep. 2006, doi: 10.1089/cmb.2006.13.1323.

[6] G. Satas, S. Zaccaria, G. Mon, and B. J. Raphael, “SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses,” Cell Systems, vol. 10, no. 4, pp. 323-332.e8, Apr. 2020, doi: 10.1016/j.cels.2020.04.001.

[7] L. A. Goldberg, P. W. Goldberg, C. A. Phillips, E. Sweedyk, and T. Warnow, “Minimizing phylogenetic number to find good evolutionary trees,” Discrete Applied Mathematics, vol. 71, no. 1, pp. 111–136, Dec. 1996, doi: 10.1016/S0166-218X(96)00060-1.

[8] I. Pe’er, R. Shamir, and R. Sharan, “Incomplete Directed Perfect Phylogeny,” in Combinatorial Pattern Matching, R. Giancarlo and D. Sankoff, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2000, pp. 143–153. doi: 10.1007/3-540-45123-4_14.

[9] P. Buneman, “A Note on the Metric Properties of Trees,” Journal of Combinatorial Theory, Series B, vol. 17, no. 1, pp. 48–50, 1974, doi: 10.1016/0095-8956(74)90047-1.

[10] R. Zeira, M. Zehavi, and R. Shamir, “A Linear-Time Algorithm for the Copy Number Transformation Problem,” Journal of Computational Biology, vol. 24, no. 12, pp. 1179–1194, Dec. 2017, doi: 10.1089/cmb.2017.0060.