CTBF: Cancer Tree Biopsy Framework

UNIVERSITY OF WARSAW

Jarosław Paszek, Agnieszka Mykowiecka, Krzysztof Gogolewski

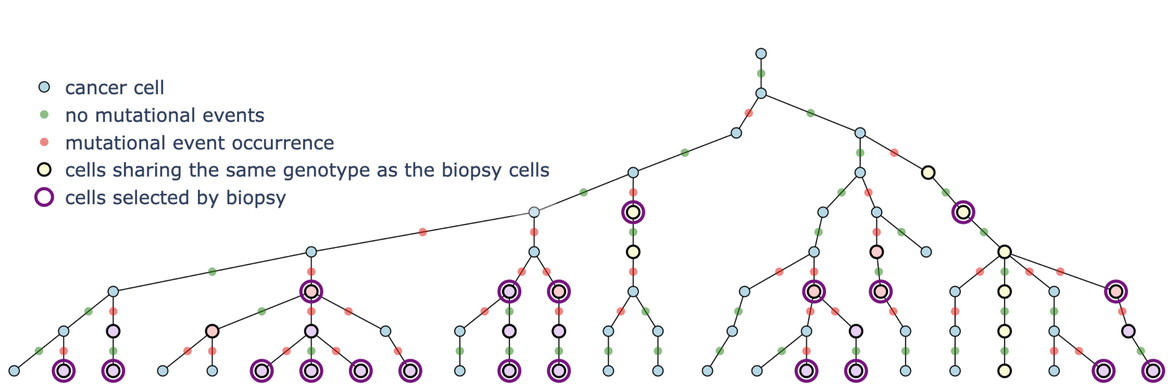
Motivation

Reconstructing cancer phylogenies clarifies tumor-progression mechanisms [1,2]. Single-cell sequencing now resolves copy-number alterations with unmatched detail, opening new avenues for evolutionary analysis [3,4].

Our contribution. We present CTBF, a three-part framework centred on biopsy-derived single-cell copy number profiles (CNPs): (1) Simulator: generates cancer cell phylogenies with tunable parameters and supports time-point-specific biopsy sampling; (2) Inference module: reconnects biopsy-sampled cells into an ancestor-descendant tree, enforcing time consistency and validating biological soundness on CNPs (3) Comparison module: adapts distance measures for trees with recurrent events to quantify reconstruction accuracy.

Evolution model

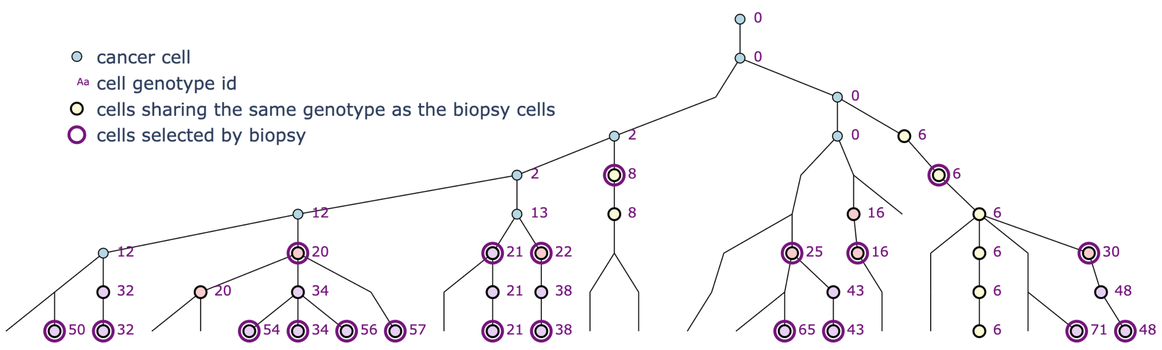
To assess phylogenetic reconstruction accuracy from observed CNPs, we developed a simulation framework modeling cancer evolution as a discrete-time branching process shaped by genomic instability and selection. Each node in the resulting tree represents a CNP vector, and edges accumulate mutational events, such as duplications, deletions, or large-scale amplifications, reflecting the gradual complexity of tumor evolution. Region-specific mutation rates (e.g., telomeric instability), chromosome-level biases, and fitness-weighted selection allow flexible modeling of diverse scenarios. Cell division follows tunable probabilistic offspring distributions (e.g., Poisson or fitness-based). Unlike models based on the infinite-sites assumption (ISA), our framework supports recurrent mutations and permits subclones with repeated genotypes. To mirror clinical conditions, biopsies are simulated by sampling cells from specific time points. This enables evaluation of reconstruction strategies across varying levels of tumor heterogeneity and temporal resolution.



Example output from CTBF's simulator illustrating tumor evolution with mutational tracking and biopsy-aware sampling. Note how the three depicted biopsies can capture both direct and recurrent genotypes, enabling realistic temporal reconstructions.

Tree reconstruction

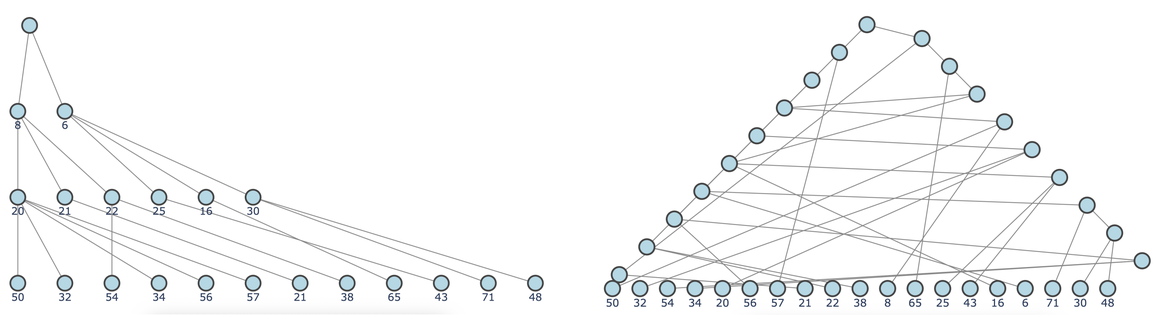
We propose a novel variant of the tree inference problem that incorporates biopsy timing information to model cell evolution. The aim is to reconstruct ancestor-descendant relationships, while addressing key challenges, such as repeated occurrence of the same subclone across time points and biologically implausible links, for example, where a descendant contains novel genetic material absent in the proposed ancestor.



Evolutionary tree generated by our simulator, restricted to biopsy-related cells.

Our method iteratively reconstructs ancestor-descendant relationships between consecutive biopsies, using an evolutionary distance threshold r.

Evolutionary distances are computed using the cnp2cnp algorithm [5], which estimates the minimum number of events required to transform one CNP into another.



Left: Tree inferred by our reconstruction algorithm. Right: Tree produced by the Neighbor-Joining (NJ) algorithm.

Tree comparison

Evaluating tree reconstruction methods requires a distance measure suitable for cancer trees, i.e. rooted trees with potentially repeated node labels. Prior approaches have considered event-labeled, mutation, and clonal trees [6–9].

GRF is based on Jaccard distance on multisets. For multisets A, B, GRF is defined by:

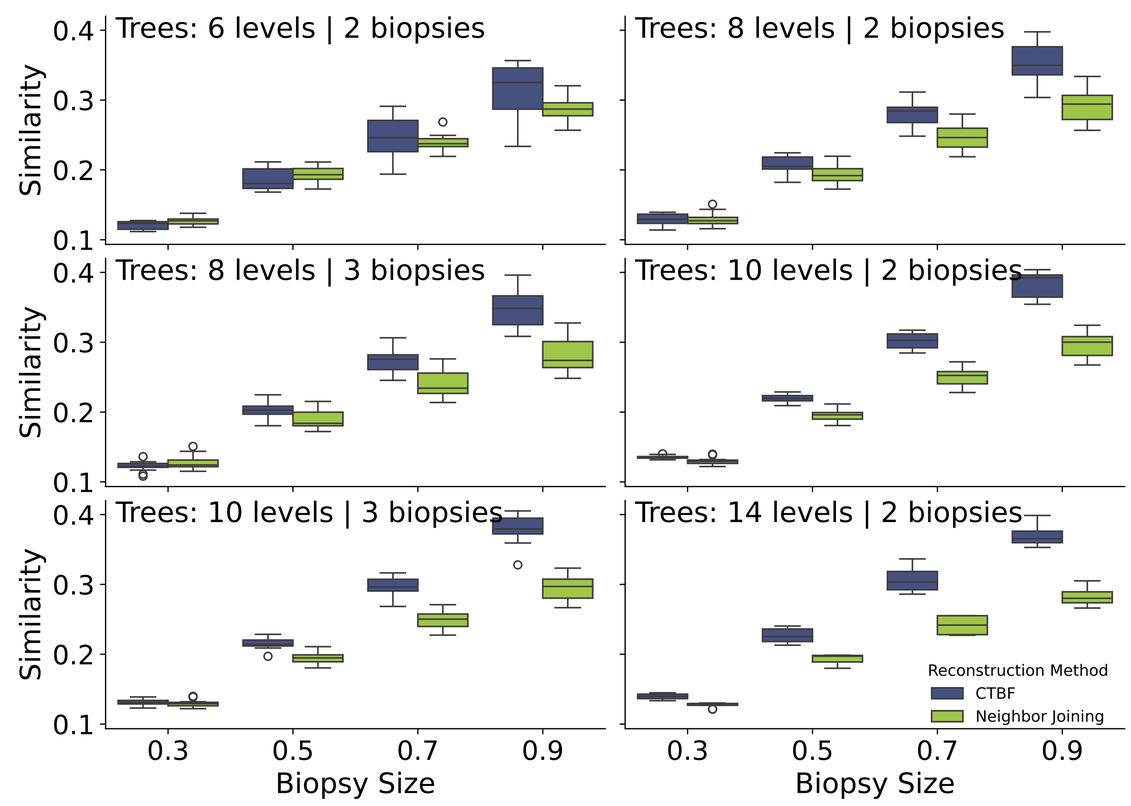
$$GRF(A,B) = \frac{\sum_{a\in A}\sum_{b\in B\setminus A}\frac{|a\triangle b|}{|a\cup b|}}{|A\cup B||A|} + \frac{\sum_{a\in A\setminus B}\sum_{b\in B}\frac{|a\triangle b|}{|a\cup b|}}{|A\cup B||B|}$$

where in each sum, each element is counted with its multiplicity in the corresponding multiset, where $\triangle$ denotes the symmetric difference between two sets.

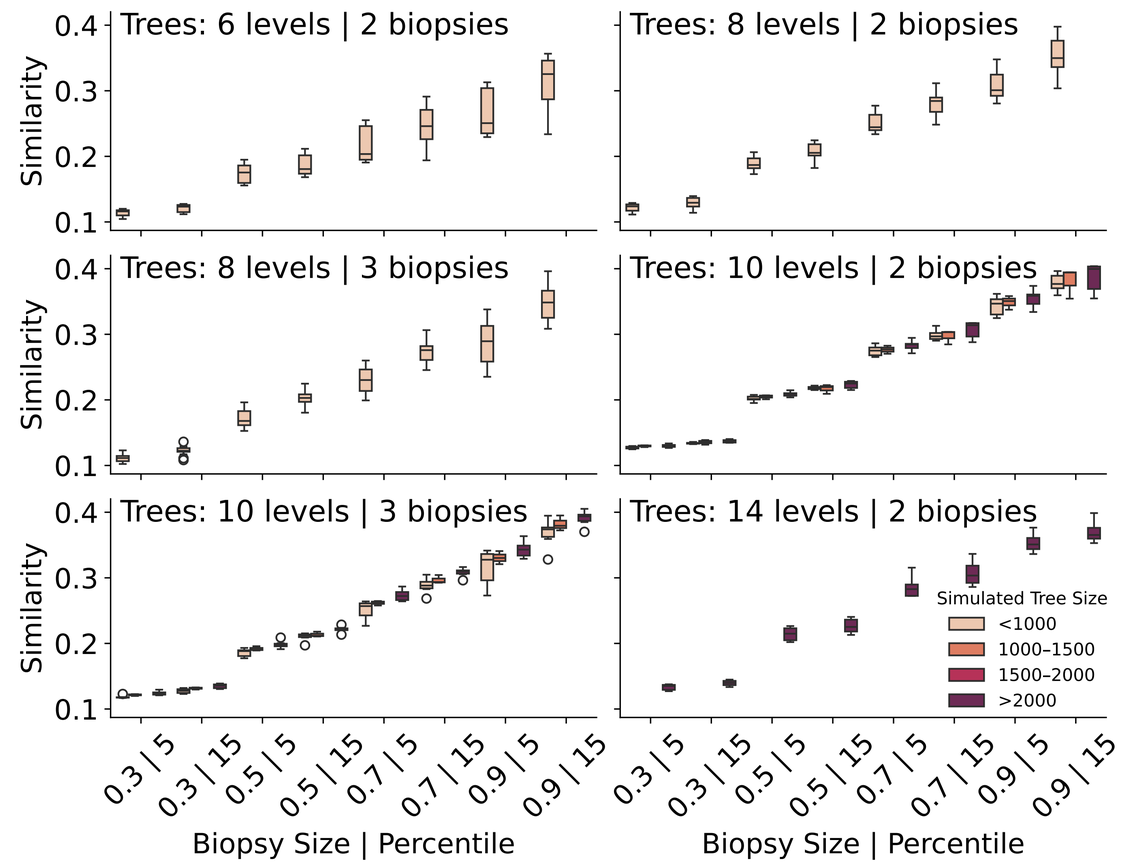
We implemented the GRF distance adapted to cancer trees and used 1-GRF as a similarity measure to evaluate reconstruction accuracy across simulations.

Results

Comparison of tree reconstruction accuracy across multiple sampling configurations and reconstruction methods. Higher similarity values indicate better agreement with the simulated tree.



Similarity scores for trees reconstructed using both CTBF and NJ.



Scores for CTBF-only reconstructions. Colors indicate bins of total number of nodes. Results are shown for various biopsy sizes and percentile thresholds.

Conclusions

Our framework advances cancer phylogenetics by integrating biopsy-aware simulation, biologically grounded reconstruction, and tailored tree comparison methods. It accounts for key complexities of tumor evolution, including recurrent mutations and region-specific genomic instability. By explicitly modeling temporal biopsy sampling and evaluating reconstructions with flexible similarity measures, the framework enables systematic benchmarking of inference strategies under realistic conditions.

Interested?

Repository: https://github.com/j-paszek/ctbf
Email: j.paszek@uw.edu.pl
Financial support: National Science Center grant no. 2020/39/D/ST6/03321.

References

1 Y. Gao et al. **Genome Research** 27.8 (2017), pp. 1312–1322.
2 E. Sollier et al. **Nature Communications** 14.1 (2023), p. 4921.
3 S. Weiner and M.S. Bansal. **Bioinformatics** 39.7 (2023), btad434.
4 S. Hui and R. Nielsen. **BMC Bioinformatics** 23.1 (2022), p. 348.
5 G. Cordonnier and M. Lafond. **BMC Genomics** 21.Suppl 2 (2020), p. 198.
6 E. Khayatian, G. Valiente, and L. Zhang. **J. Comput. Biol.** 31.4 (2024), pp. 328–344.
7 K. Jahn, N. Beerenwinkel, and L. Zhang. **Algorithms Mol. Biol.** 16.1 (2021), p. 9.
8 S.Briand, C.Dessimoz, N.El-Mabrouk, Y.Nevers. **Syst. Biol.** 71.6 (2022), pp. 1391–1403.
9 M. Llabrés, F. Rosselló, and G. Valiente. **J. Comput. Biol.** 28.12 (2021), pp. 1181–1195.