CTBF: Cancer Tree Biopsy Framework

Jarosław Paszek[1], Agnieszka Mykowiecka[1], Krzysztof Gogolewski[1]

Keywords: cancer tree reconstruction, single-cell sequencing, copy number profiles, biopsy

Introduction

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.

Methods

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 (see Fig. 1). 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.

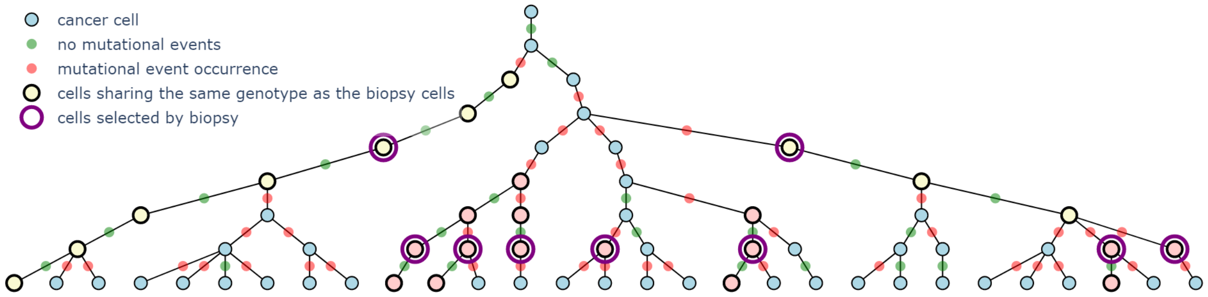
cancer cell
no mutational events
mutational event occurrence
cells sharing the same genotype as the biopsy cells
cells selected by biopsy

Figure 1: **Example output from CTBF's simulator illustrating tumor evolution with mutational tracking and biopsy-aware sampling.** Note how the two 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. Our method iteratively reconstructs ancestor-descendant relationships between consecutive biopsies, using an evolutionary distance threshold r. For all cells c in the earlier biopsy lacking a close ancestor in the next biopsy B, we insert into B a duplicate cell c' of c, and add an edge (c, c') to the reconstructed tree. Finally, we remove duplicates from the remaining cells and tree completion proceeds via neighbor joining (NJ). Evolutionary distances are computed using the cnp2cnp algorithm [5], which estimates the minimum number of events required to transform one CNP into another.

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], with the most general model allowing multisets of labels per node and label repetitions between nodes [6]. A key challenge is assessing reconstructed edges (u, v) that may represent valid ancestor-descendant relationships, even if intermediate nodes exist along the true path. Our preliminary experiments observed this limitation in the kRF distance, which compares sets of labels within k-neighborhoods around each edge [6]. The GRF distance has more potential as its value depends on the intersection of clusters of input trees, however, it was described with ISA assumed [9]. We implemented the GRF distance adapted to cancer trees and used 1-GRF as a similarity measure to evaluate reconstruction accuracy across simulations.

[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland.

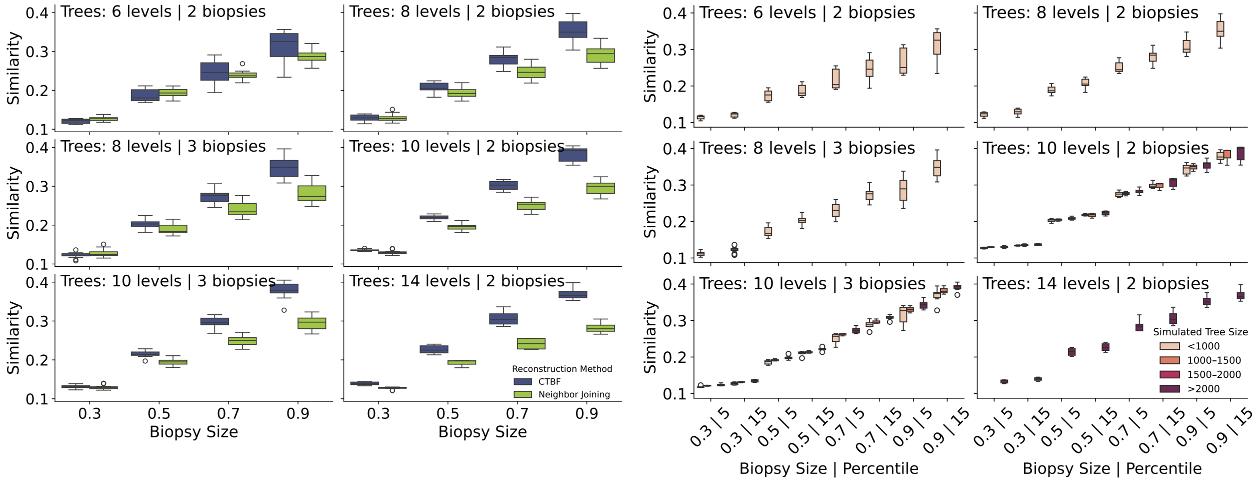

Figure 2: **Comparison of tree reconstruction accuracy across multiple sampling configurations and reconstruction methods.** Higher similarity values indicate better agreement with the simulated tree. *Left*: Similarity scores for trees reconstructed using both CTBF and NJ. *Right*: Scores for CTBF-only reconstructions. Colors indicate bins of total number of nodes. Results are shown for various biopsy sizes and percentile thresholds.

Results

To evaluate the robustness of our reconstruction method, we conducted simulation experiments across a range of tree sizes (6–14 evolutionary levels) and biopsy configurations. Each setup tested various sampling depths to reflect both balanced and extreme clinical scenarios. From each selected level, a fraction of nodes was sampled based on predefined biopsy sizes, simulating different biopsy coverages. To assess the method's sensitivity to the connectivity criterion, we varied the maximum allowed CNP distance for connecting nodes using the 5th and 15th percentiles of pairwise CNP distances. Each experimental setup was repeated 30 times to account for stochastic variation in tree generation and sampling. For each reconstructed tree, we calculated the GRF distance to the original simulated tree as a measure of reconstruction accuracy. As a baseline, we also computed GRF distances for NJ trees built from all sampled nodes, ignoring any information about their temporal origin. The results are presented in Fig. 2. The experiment highlights several key trends. As expected, GRF distances for NJ trees are comparable to those of our method for small trees and limited biopsies, likely due to NJ's placement of all nodes at the leaves, which fits shallow evolutionary structures. However, with larger trees and more biopsies, our method consistently yields higher similarity values, better capturing internal evolutionary relationships. This advantage also holds with increasing biopsy coverage: larger probes improve our method's accuracy, while NJ performance declines. Finally, the choice of the connectivity criterion significantly affects reconstruction quality. While this requires further investigation, it is evident that both overly strict and overly permissive r values result in reduced reconstruction accuracy.

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.

References

[1] Y. Gao et al. *Genome Research* 27.8 (2017), pp. 1312–1322.
[2] E. Sollier, J. Kuipers, K. Takahashi, N. Beerenwinkel, and K. Jahn. *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. *Journal of Computational Biology* 31.4 (2024), pp. 328–344.
[7] K. Jahn, N. Beerenwinkel, and L. Zhang. *Algorithms for Molecular Biology* 16.1 (2021), p. 9.
[8] S. Briand, C. Dessimoz, N. El-Mabrouk, and Y. Nevers. *Systematic Biology* 71.6 (2022), pp. 1391–1403.
[9] M. Llabrés, F. Rosselló, and G. Valiente. *Journal of Computational Biology* 28.12 (2021), pp. 1181–1195.

Acknowledgment: Financial support was provided by the National Science Center grant no. 2020/39/D/ST6/03321.