

Abstract

Tumors are highly heterogeneous populations of cells that evolve dynamically, acquiring mutations as they divide. Reconstructing this evolutionary history is essential for understanding cancer progression, metastasis, and therapy resistance. With the advent of single-cell DNA sequencing (scDNA-seq), we can now profile the genomic landscape of individual cells, which provides an unprecedented window to the formation and evolution of tumor cell populations. Specifically, single-cell whole-genome sequencing allows us to detect structural variations called *copy number aberrations* (CNAs), which are known to play a critical role in cancer development and progression. Inferring the evolutionary trees, i.e. *phylogenies*, from single-cell DNA sequences presents immense computational challenges due to both the sparsity of the data inherent to the current state of sequencing technologies, and the complexity of the underlying mutational processes.

In this thesis, we develop novel probabilistic models and algorithms to infer the evolutionary history of tumors from single-cell DNA sequencing data. The research addresses core methodological bottlenecks in tumor phylogenetics, moving from foundational distance-based methods to comprehensive joint Bayesian inference frameworks.

First, we introduce a method to estimate biologically meaningful evolutionary distances between single cells directly from noisy read counts, employing an original Hidden Markov Model to accommodate the unique noise profile of scDNA-seq and the interdependence of copy number states across the genome. Next, we extend distance-based tree inference from classical phylogenetics by presenting a scalable algorithm specifically designed for rooted trees, leveraging the biological premise that tumor evolution originates from a known healthy diploid ancestor. To enable rigorous uncertainty quantification over tree topologies, we then tackle the problem of sampling directed trees (arborescences). We present a stable, polynomial-time sampling algorithm capable of generating arborescences even on weakly connected graphs, which commonly arise when performing inference from single-cell sequences. Finally, we integrate these advancements into a comprehensive variational inference framework. This framework

efficiently achieves joint inference over clonal tree structures, branch lengths, copy number profiles, and cell-to-clone assignments.

Collectively, this thesis contributes a suite of statistically grounded, highly scalable tools that bridge the gap between noisy single-cell sequencing reads and robust insights into cancer evolution, offering a foundation for future clinical applications and oncological research.

Keywords

copy number, tumor phylogenetics, variational inference