ALGORITHMS FOR RECONSTRUCTING THE LIFE HISTORY OF TUMORS FROM GENOMICS DATA

Allocation: Director Discretionary/50 Knh
PI: Mohammed El-Kebir[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

Cancer is a genetic disease characterized by intratumor heterogeneity as well as the presence of multiple cellular populations with different sets of mutations. Cellular heterogeneity gives cancer the ability to resist treatment, and quantifying its extent is key to improving our understanding of tumorigenesis. In this project, we developed and employed novel phylogenetic (evolutionary tree-based) techniques to reconstruct the evolutionary histories of individual tumors from DNA sequencing data. Typically, this data is obtained from shotgun sequencing of tumor biopsies using either single-cell or bulk-sequencing technology. Highlights from our research outcomes include: (1) the creation of SPhyR, an algorithm that employs the k-Dollo evolutionary model to reconstruct phylogenetic trees from single-cell DNA sequencing data; and (2) a detailed analysis using simulated data on the nonuniqueness of solutions to the phylogeny estimation problem from bulk DNA sequencing data. One paper based on this work has been accepted for publication and another has been submitted.

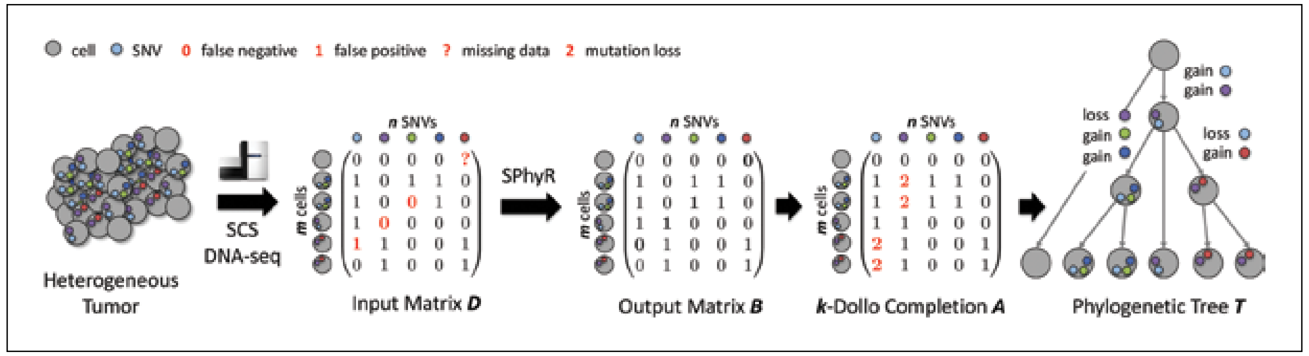
RESEARCH CHALLENGE

This research considered two current challenges in tumor phylogenetics. First, the phylogeny estimation problem from single-cell sequencing data is a variant of the classic phylogeny estimation problem with incorrect and missing data due to the sequencing technology. Current methods aim to simultaneously construct a phylogenetic tree and correct these measurement errors using either too stringent or too permissive evolutionary models. There is a need for methods that employ appropriate evolutionary models that strike a balance between being realistic and yet are sufficiently constrained.

Second, the problem of reconstructing a phylogenetic tree given bulk sequencing data from a tumor is more complicated than the classic phylogeny estimation problem, where one is given the leaves of the phylogenetic tree as input. In contrast, by using bulk sequencing data, which forms the majority of current cancer sequencing studies, we do not observe the leaves but rather are given mutation frequencies that result from mixtures of the unknown leaves of the underlying phylogenetic tree. The majority of current tumor phylogeny estimation methods that aim to infer a phylogenetic tree from mutation frequencies employ the perfect phylogeny evolutionary model. In this model, a mutation at a specific genomic site occurs only once throughout the evolutionary history of the tumor and is never subsequently lost. The underlying perfect phylogeny mixture problem is NP-complete and, importantly, from the same input data multiple distinct solutions can be inferred. This nonuniqueness has important consequences for downstream analyses by cancer biologists and clinicians, whose starting point is a single phylogenetic tree. While nonuniqueness of solutions to this computational problem has been recognized in the field, a rigorous analysis of its extent and consequences has been missing.

METHODS & CODES

We used extensive computer simulations to quantify the extent of nonuniqueness of solutions to the phylogeny estimation problem from bulk sequencing data. To solve the phylogeny estimation problem from single-cell sequencing data, we used techniques from combinatorial optimization. More specifically, we formulated the problem as an integer linear program (ILP) and designed and implemented a custom column generation and cutting plane approach for the ILP formulation. We used CPlex as the underlying ILP solver. To facilitate reproducibility of the results, the data and open source codes (Python, Bash, and C++) have been deposited in public repositories. In addition, these repositories contain Jupyter notebooks that generate the plots of the respective papers.



$$\begin{pmatrix} 0&0&0&0&?\\ 1&0&1&1&0\\ 1&0&0&1&0\\ 1&0&0&0&0\\ 1&1&0&0&1\\ 0&1&0&0&1 \end{pmatrix} \quad \begin{pmatrix} 0&0&0&0&0\\ 1&0&1&1&0\\ 1&0&1&1&0\\ 1&1&0&0&0\\ 0&1&0&0&1\\ 0&1&0&0&1 \end{pmatrix} \quad \begin{pmatrix} 0&0&0&0&0\\ 1&2&1&1&0\\ 1&2&1&1&0\\ 1&1&0&0&0\\ 2&1&0&0&1\\ 2&1&0&0&1 \end{pmatrix}$$

Figure 1: Tumors are composed of cellular populations with distinct sets of mutations. Single-cell sequencing of a tumor yields an input matrix *D* with incorrect and missing entries. SPhyR aims to simultaneously correct errors in matrix *D* and infer the evolutionary history of the cells.

RESULTS & IMPACT

Phylogeny estimation algorithms that employ appropriate evolutionary models are key to understanding the evolutionary mechanisms behind intratumor heterogeneity. One of the outcomes of this project is Single-cell Phylogeny Reconstruction (SPhyR), a method for tumor phylogeny estimation from single-cell sequencing data that employs the k-Dollo parsimony model, a relaxation of the perfect phylogeny model where a mutation may only be gained once but lost k times. In light of frequent loss of point mutations in cancer due to copy-number aberrations, the k-Dollo model is more appropriate than the evolutionary models utilized by previous methods. This project resulted in a novel combinatorial characterization of solutions to the underlying computational problem as constrained integer matrix completions, which formed the basis for the efficient integer linear programming approach utilized by SPhyR. Using simulated data, we found that SPhyR outperformed existing methods that are either based on the infinite-sites or the finite-sites evolutionary model, in terms of solution quality and runtime.

We studied the problem of counting and sampling solutions in instances of the phylogeny estimation problem from bulk sequencing data. To avoid any bias in downstream analyses it is important to know the number of solutions and to be able to sample uniformly from the solution space. As part of this project, we proved an upper bound on the number of solutions that can be computed in polynomial time. In addition, we introduced a uniform sampling algorithm based on rejection sampling that works for small problem instances. Using extensive simulations, we showed that the number of solutions increased with an increasing number of mutations, but decreased with increasing number of bulk samples from the same tumor. We observed similar trends in terms of the quality of the solutions in the solution space. Moreover, we showed that additional constraints from single-cell and long-read sequencing technology significantly reduced the number of solutions. Finally, we demonstrated that current methods are unable to sample uniformly from the solution space, often overlooking solutions. This leads to significant biases that propagate to downstream analyses.

WHY BLUE WATERS

Blue Waters was essential to the two research outcomes of this project, as they both involved extensive benchmarking and validation using simulated data. The computational resources of Blue Waters allowed us to perform these experiments at scale, enabling us to study the performance of our algorithms and the underlying problem statements in many different experimental settings. This is something that would not have been possible on other platforms.

PUBLICATIONS & DATA SETS

El-Kebir, M., SPhyR: Tumor Phylogeny Estimation from Single-Cell Sequencing Data under Loss and Error. *Bioinformatics/ECCB 2018*, accepted (2018).

Pradhan, D., and M. El-Kebir, On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture problem. Submitted (2018).

OncoLib: https://github.com/elkebir-group/OncoLib

SPhyR: https://github.com/elkebir-group/SPhyR

PPM-NonUniq: https://github.com/elkebir-group/PPM-NonUniq

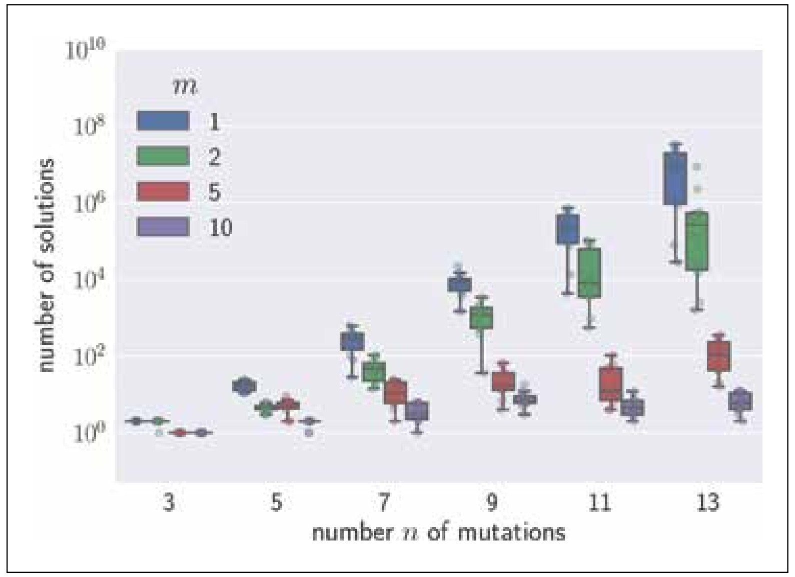


Figure 2: *In silico* simulations of tumor evolution and bulk sequencing show that the number of solutions to the phylogeny estimation problem increased with an increasing number of mutations but decreased with an increasing number of bulk samples from the same tumor.