Determining Essential Evolutionary Relations from Noisy Single-Cell Data

Robyn Burger[†] and Allison Shi [‡]

[†]Cornell University – rlb388@cornell.edu, [‡]New York University – ams9451@nyu.edu

Cancer is an evolutionary disease, progressing through the accumulation of mutations as cells multiply. A phylogenetic tree, much like a family tree, represents the genetic history of these cells. We use single-cell sampling (SCS) to collect data to reconstruct these trees. However, SCS data is error-prone, often resulting in multiple optimal trees all of which are equally likely to be correct. Our work is in determining the ancestral relationships that stay consistent in all of these models, offering greater precision in tumor modeling and analysis.

Modeling Tumor Evolution

We can represent the SCS data in a binary mutation matrix D where rows and columns represent individual cell samples and mutations, respectively. $D_{ij} = 1$ indicates that cell i contains mutation j. We say D contains a conflict if C, or any permutation of the rows and columns of C, is a submatrix of D.

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Fig. 1: The matrix C.

We assume a cell can only gain one mutation at a time, which no other cells have (until that cell, itself, divides), and that cells cannot lose mutations. As such, a conflict indicates an error in the data: either a false positive (FP) or false negative (FN). An entry must then be corrected from $1 \to 0$ or vice versa, respectively. Note that FP are ∼20x less likely than FN. We call σ the minimum number of bit flips required to make D conflict-free (CF). We call a matrix X *optimal* if it is CF and differs from D by σ bits.

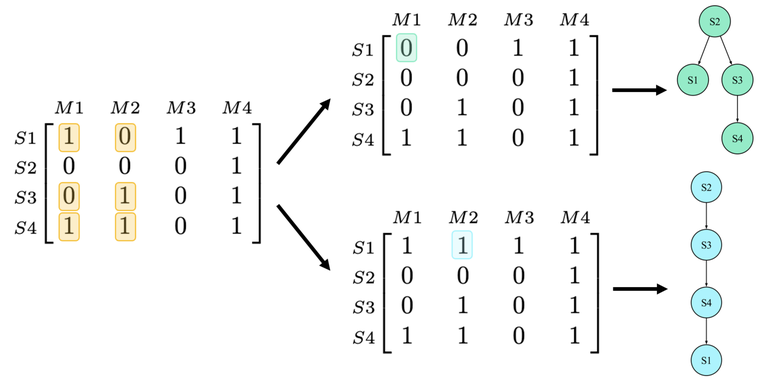

Fig. 2: From left to right: Matrix D with a conflict in rows S1, S3, and S4 and columns M1 and M2 where $\sigma = 1$; two optimal solutions; their phylogenetic trees. Notice, two distinct trees arise from the same matrix.

We classify evolutions in the following ways: in linear evolution (LE), a dominant cell out-competes all previous cells, translating to a generally linear phylogenetic tree, and in branching evolution (BE), cells diverge from a common ancestor and evolve in parallel, resulting in concurrent lineages and a multi-branching phylogenetic tree [1].

Essential Partial Order

For two cells u and v, we say that u is *essentially less than* v, notated as $u \leq^e v$, if and only if in all optimal solutions, cell v contains all mutations present in cell u (and possibly more). Equivalently, $\forall j \in [m]$, $X_{uj} \leq X_{vj}$ in all optimal solutions X. When $u \leq^e v$, we assume that u is an ancestor of v.

For example, in Fig. 2, we see that $S3 \not\leq^e S1$; in the top solution, $X_{S3} \not\leq X_{S1}$ (observe that $X_{S3,M2} > X_{S1,M2}$). But we do have $S2 \leq^e S1$ because in all possible optimal solutions, $X_{S2} \leq X_{S1}$. So we presume that $S2$ must be an ancestor of $S1$.

Integer Linear Programming

Integer Linear Programming (ILP) is a technique to meet an objective under a series of constraints where all variables are integers. We use ILP in the following algorithms:

FindOpt calculates σ. Here, X is the working conflict-free matrix. $B_{p,q,a,b} = 1$ when there exists a row i where $X_{ip} = a$ and $X_{iq} = b$, 0 otherwise. $1 \to 0$ flips are β times less likely than $0 \to 1$ flips and row i is weighted by its multiplicity, M_i, such that it is more costly (and thus less likely) to flip a bit in a row that appears more often.

$$-X_{ip} + X_{iq} \leq B_{p,q,0,1} \quad (1)$$
$$X_{ip} - X_{iq} \leq B_{p,q,1,0} \quad (2)$$
$$X_{ip} + X_{iq} - 1 \leq B_{p,q,1,1} \quad (3)$$
$$B_{p,q,0,1} + B_{p,q,1,0} + B_{p,q,1,1} \leq 2 \quad (4)$$
$$\text{Objective:} \quad \min \sum_{i,j} M_i(1 - D_{ij})X_{ij} + \beta M_i D_{ij}(1 - X_{ij}) \quad (5)$$

In *TestEss*, we add constraints 6-8, where $z_j = 1$ if and only if $X_{uj} > X_{vj}$, 0 otherwise.

$$X_{uj} - X_{vj} \leq z_j \leq \frac{X_{uj} - X_{vj} + 1}{2} \quad (6)$$
$$\sum_j z_j \geq 1 \quad (7)$$
$$\sum M_i(1 - D_{ij})X_{ij} + \beta M_i D_{ij}(1 - X_{ij}) = \sigma \quad (8)$$

TestEss tests the contrapositive of the definition of $u \leq^e v$. We assume $u > v$ and attempt to find some optimal solution X where this condition holds. If *TestEss* is infeasible, then there is no possible optimal solution where $u > v$, so in all optimal solutions, $u \leq v$, which by definition means that $u \leq^e v$.

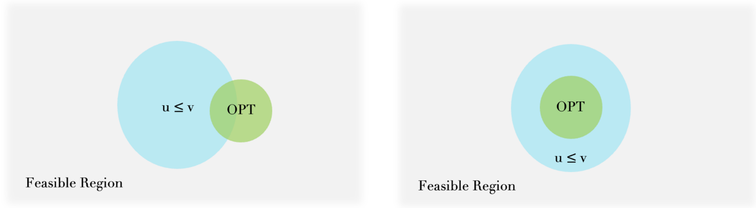

Fig. 3: (left) In some optimal solution, $u \not\leq v$, so $u \not\leq^e v$.
Fig. 4: (right) In all optimal solutions $u \leq v$, so $u \leq^e v$.

We visualize the output of *TestEss* for rows u and v of matrix D. Here, the gray region represents all feasible solutions, the green region is the subset of those solutions that are optimal, i.e. those that differ from D by at most σ bits, and the blue region is the space of all solutions where $u \leq v$. If the green region is a subset of the blue region (Fig. 3), then $u \leq v$ in all optimal solutions, so the constraints of *TestEss* are infeasible and we conclude that $u \leq^e v$. Otherwise, there is an optimal solution where $u > v$, so $u \not\leq^e v$.

Experiments

Limited SCS data is available to the public, so our work primarily relied on simulations by L. Weber and M. El-Kebir [2] of patients with acute myeloid leukemia (AML). For each simulated patient, the authors produced 10 replications; we ran *TestEss* on each replication and recorded the median results. We focused on simulated patients AML-10 and AML-67, whose phylogenetic trees were linear and branching, respectively. Further, we ran *TestEss* on ALL-2, a real patient with acute lymphoblastic leukemia (ALL) [2].

Experimental Results

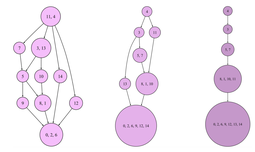

Fig. 5: $\leq^e$ graphs for AML-10, $\beta = 1, 2, 20$.

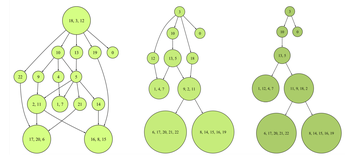

Fig. 6: $\leq^e$ graphs for AML-67, $\beta = 1, 2, 20$.

Sample	n	m	β	ILP calls	width	nodes	σ	time (s)
AML-10	8729	4	1	295	4	10	574.5	48
			2	362	3	7	584.5	68
			3	363	1	5	584.5	68
			20	363	1	5	584.5	68
AML-67	6024	7	1	542.5	6.5	14.5	449.5	462
			2	571.5	4	10	460.5	356
			3	592.5	4	8	461	433
			20	592.5	4	8	461	359
Patient 2	115	16	20	6653	4	19	381	139562

Table 1: Median results from 10 replications of AML-10 and AML-67 and results from 1 replication of ALL-2 [2].

Based on average FN and FP rates, the most realistic β value is 20. However, our simulated data are smaller than real SCS data, limiting the range of meaningful values of β: all graphs with $\beta > 2$ are identical.

As β increases, σ, the number of ILP calls, and the size of the strongly connected components increase, resulting in fewer nodes and a smaller poset width. With a more realistic β, the graphs more closely resemble the topology of their phylogenetic trees; in Fig. 5 and 6, AML-10 and AML-67 take on their respective linear and branching structures when $\beta = 20$, likely because the relative lineages in the graph are preserved in the phylogenetic tree.

In the future, we intend to test on a greater number of real datasets and analyze the constriction points, i.e. the vertices that all paths must go through.

To access the data and more detailed results, see github.com/msu-alglab/essentcell.

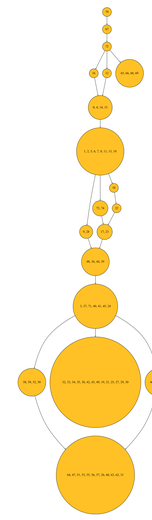

Fig. 7: $\leq^e$ graph for ALL-2, $\beta = 20$.

Acknowledgements

We would like to thank Dr. Adiesha Liyanage and Dr. Brendan Mumey from Montana State University's School of Computing for their support and guidance. This work was funded by NSF Award No. 2243010.

References

[1] Alexander Davis, Ruli Gao, and Nicholas Navin. "Tumor evolution: Linear, branching, neutral or punctuated?" In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1867.2 (2017), pp. 151–161.

[2] Leah L. Weber and Mohammed El-Kebir. "Phyolin: Identifying a Linear Perfect Phylogeny in Single-Cell DNA Sequencing Data of Tumors". In: *20th International Workshop on Algorithms in Bioinformatics, WABI 2020, September 7-9, 2020, Pisa, Italy (Virtual Conference)*. Ed. by Carl Kingsford and Nadia Pisanti. Vol. 172. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020, 5:1–5:14.