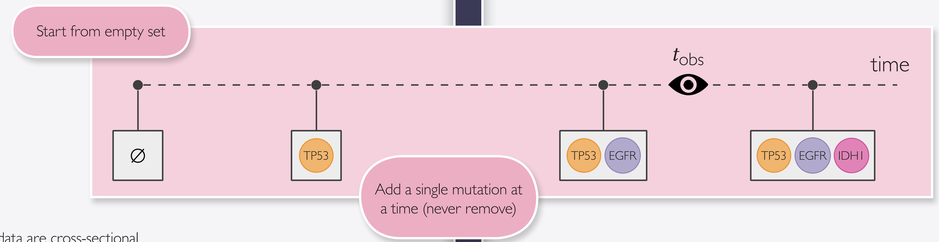
Scaling up Continuous-Time Markov Chains Helps Resolve Underspecification

Alkis Gotovos — MIT
Rebekka Burkholz — Harvard
John Quackenbush — Harvard
Stefanie Jegelka — MIT

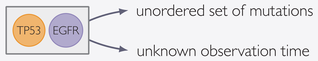
Motivation

Goal: Model the time evolution of discrete sets of items using a continuous-time Markov chain

Example application: Accumulation of genetic mutations during cancer progression



Challenge: Available data are cross-sectional



State of the art (Schill et al, 2019)

- Constrain analysis to $n \approx 20$ "important" mutations
- Run $O(2^n)$ max. likelihood

Underspecification

Can we recover the true underlying model ?

Scalability

Can we learn a model on hundreds of mutations ?

Contributions

- We show that "unimportant" mutations are valuable to resolving underspecification
- We propose an efficient approximate max. likelihood method that scales to hundreds of mutations
- We evaluate our method on synthetic and real cancer data

Problem setup

Ground set $V = \{1, \dots, n\}$ (e.g., n genetic mutations)

Define continuous-time Markov chain $\{X_t\}_{t \geq 0}$ on state space 2^V

Generator matrix

$$Q = \begin{bmatrix} q_{\emptyset \to \emptyset} & \cdots & q_{\emptyset \to R} & \cdots & q_{\emptyset \to V} \\ \vdots & & \vdots & & \vdots \\ q_{S \to \emptyset} & \cdots & q_{S \to R} & \cdots & q_{S \to V} \\ \vdots & & \vdots & & \vdots \\ q_{V \to \emptyset} & \cdots & q_{V \to R} & \cdots & q_{V \to V} \end{bmatrix} \in \mathbb{R}^{2^n \times 2^n}$$

Transition rate from S to R

Parametrization (Schill et al, 2019)

$$q_{S \to S \cup \{j\}}(\boldsymbol{\theta}) = \exp\left(\theta_{jj} + \textstyle\sum_{i \in S} \theta_{ij}\right)$$

individual rate of j — effect of i on j

Marginal likelihood over unknown observation time

$$p(S; \boldsymbol{\theta}) = \int_0^\infty p(S \mid t; \boldsymbol{\theta}) p(t) dt$$

Observation time prior — Markov chain

Tackling underspecification

Toy example $1 \to 2$ Mutation 1 (almost) always occurs before mutation 2

Data distribution $\{1\}$ $\emptyset$ $\{1,2\}$ $\emptyset$ $\{1\}$ $\{1,2\}$ $\{1\}$ $\cdots$

Proposition 1 (simplified)

There is a one-dimensional family of models with identical data distribution as above.

Insight: Assume another ground set V_+ containing i.i.d. mutations with no interaction to V

$$\Theta_{\text{full}} = \left(\begin{array}{c|c} \Theta & 0 \\ \hline 0 & \theta_+ I_m \end{array}\right)$$

Mutations in V_+ act like a clock!

Theorem 1 (simplified)

Let t^* be the true observation time. Then, the mean and variance of the posterior observation time distribution can be bounded as follows:

$$|M_{\text{post}} - t^*| \lesssim \sqrt{\frac{\log m}{m}} \quad \text{and} \quad V_{\text{post}} \lesssim \frac{1}{m}$$

Efficient likelihood maximization

Objective $F(\mathcal{D}; \boldsymbol{\theta}) = \frac{1}{N} \sum_{d=1}^{N} \log p(S^{(d)}; \boldsymbol{\theta}) - \lambda \sum_{i \neq j} |\theta_{ij}|$

L_1 regularization enforces sparsity and helps uncover block-diagonal parameter structure

Maximize $F(\mathcal{D}; \boldsymbol{\theta})$

First-order optimizer

Approximate $\nabla_{\boldsymbol{\theta}} \log p(S; \boldsymbol{\theta}) \approx \frac{1}{M} \sum_i \nabla_{\boldsymbol{\theta}} \log p(\sigma^{(i)}; \boldsymbol{\theta})$

Compute $\nabla_{\boldsymbol{\theta}} \log p(\sigma^{(i)}; \boldsymbol{\theta})$ using modified CTMC

Sample $\sigma^{(1)}, \dots, \sigma^{(M)}$ using custom M-H sampler

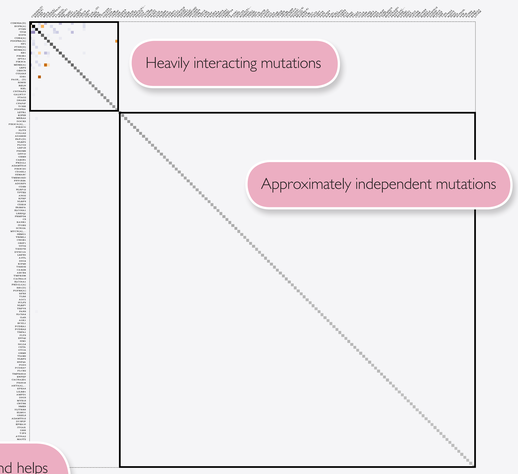
Experiments

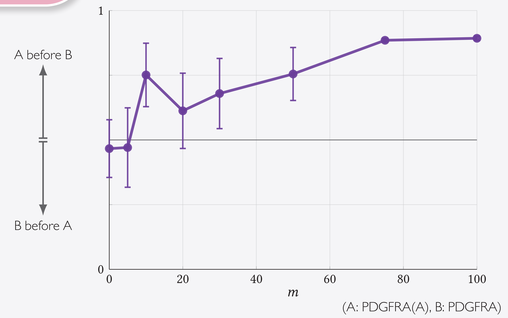
TCGA glioblastoma data: $n = 410$ mutations, amplifications, and deletions

Runtime

Method	$n = 20$	$n = 100$
(Schill et al., 2019)	121 m	–
Ours	8 s	33 m 43 s

Learned parameter matrix





(A: PDGFRA(A), B: PDGFRA)

- Schill et al. Modelling cancer progression using mutual hazard networks. Bioinformatics, 2019.
- Gerstung et al. The evolutionary history of 2,658 cancers. Nature, 2020.
- Beerenwinkel et al. Cancer evolution: Mathematical models and computational inference. Systematic Biology, 2014.