Inferring tumor progression in large datasets

M. Neyshabouri, S. Jun, J. Lagergren. *PLOS Computational Biology.*

Presented by Gillian Chu

Tumor Progression

What is tumor progression?

- Find a set of linearly ordered mutually exclusive driver pathways

Tumor Progression

What is tumor progression?

- Find a set of linearly ordered mutually exclusive driver pathways

Why is it complex?

- Driver mutations are hard to identify
- Driver gene mutations are mediated by pathways
- Set of genes provide the same effect
- Selective advantage of one gene exhausts others in a mutually exclusive manner

Overview

- Presents a probabilistic model of mutually exclusive linearly ordered pathways
- Sampling based inference algorithm to train the model
- Test it against ILP solutions

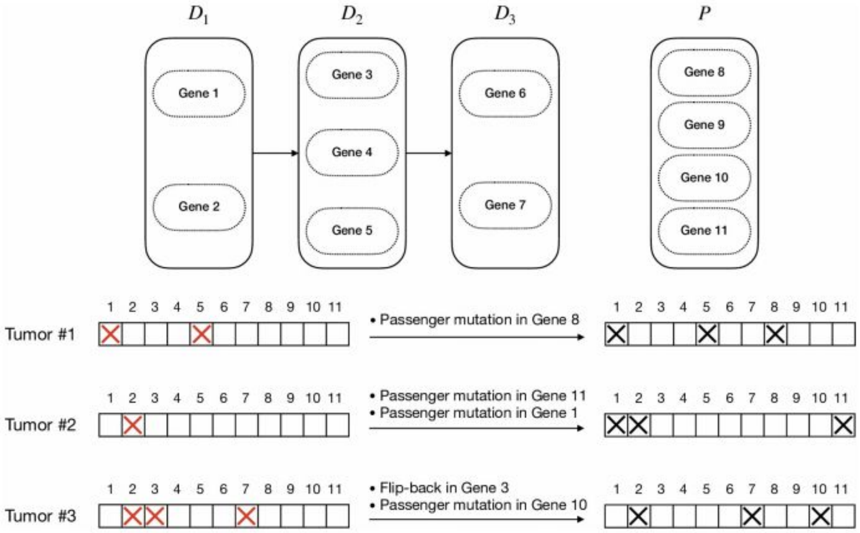
Overview

1. Linear Progression Model
2. Probabilistic Generative Process
3. Model Selection
4. Experiments on Simulated Data
5. Real Data Validation

Overview

1. Linear Progression Model
2. Probabilistic Generative Process
3. Model Selection
4. Experiments on Simulated Data
5. Real Data Validation

Linear Progression Model



D_1
D_2
D_3
P
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene 10
Gene 11
• Passenger mutation in Gene 8
• Passenger mutation in Gene 11
• Passenger mutation in Gene 1
• Flip-back in Gene 3
• Passenger mutation in Gene 10

	1	2	3	4	5	6	7	8	9	10	11
Tumor #1	☒				☒						
Tumor #2		☒									
Tumor #3		☒	☒				☒				

	1	2	3	4	5	6	7	8	9	10	11
Tumor #1	☒				☒			☒			
Tumor #2	☒	☒									☒
Tumor #3		☒					☒			☒	

- Passenger mutations
- Flip-back events

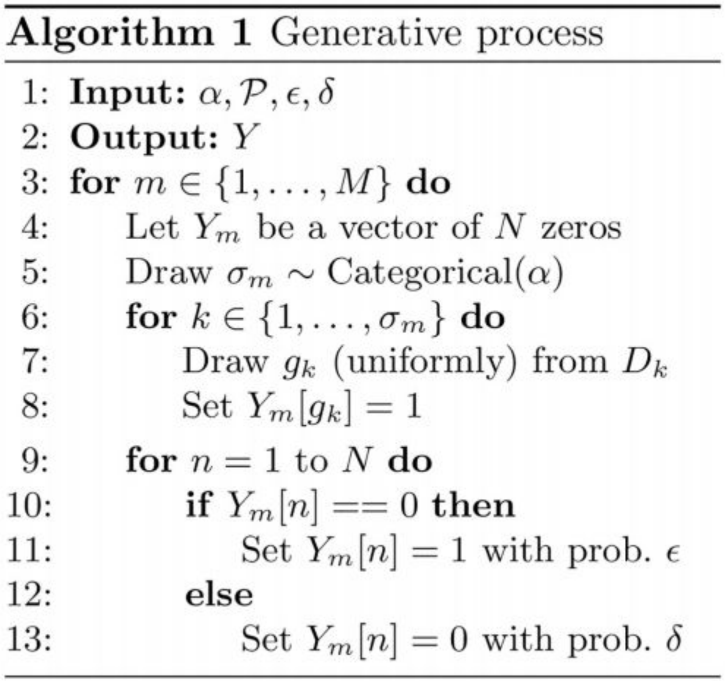
Notation

- N genes
- $\mathcal{P} = (D_1, D_2, \ldots, D_L, P)$ as an ordered partition of N. We have a linear pathway progression model of length L. D1 - DL are driver pathways, P is passengers.
- $Y \in \{0, 1\}^{M \times N}$ Mutation matrix Y for M tumors
- Noise parameters: δ, ϵ

Overview

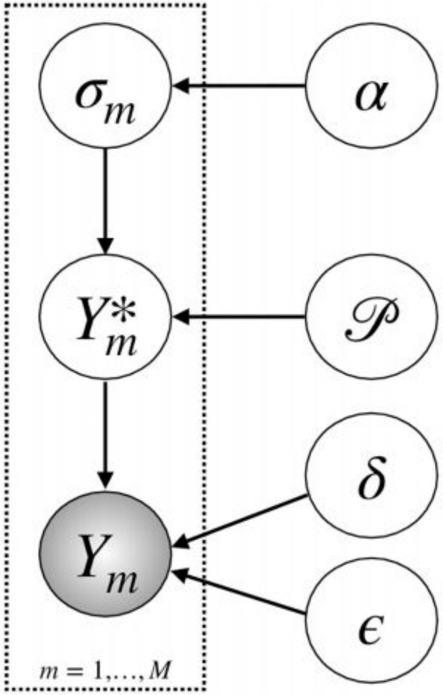
1. Linear Progression Model
2. Probabilistic Generative Process
3. Model Selection
4. Experiments on Simulated Data
5. Real Data Validation

Probabilistic Generative Model

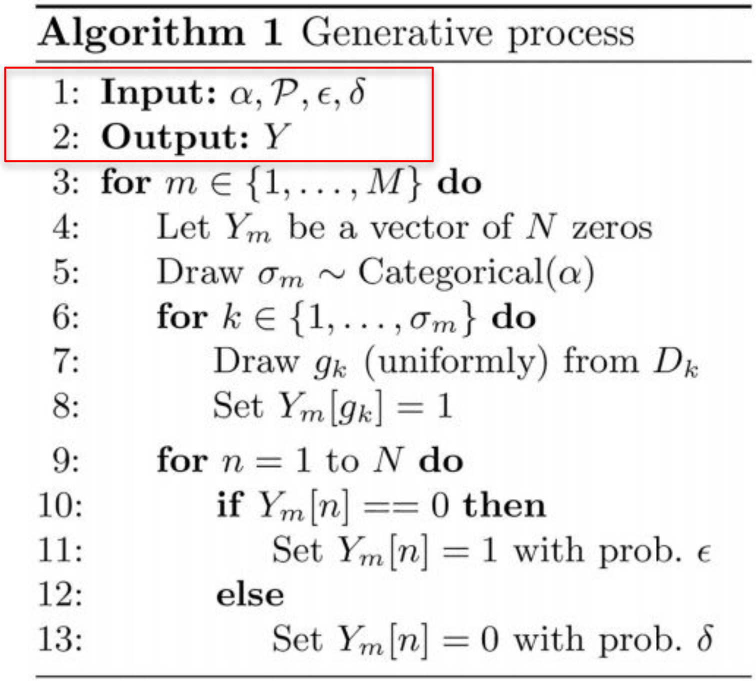


Algorithm 1 Generative process

```
**Input:** $\alpha, \mathcal{P}, \epsilon, \delta$
**Output:** $Y$
**for** $m \in \{1, \ldots, M\}$ **do**
    Let $Y_m$ be a vector of $N$ zeros
    Draw $\sigma_m \sim \text{Categorical}(\alpha)$
    **for** $k \in \{1, \ldots, \sigma_m\}$ **do**
        Draw $g_k$ (uniformly) from $D_k$
        Set $Y_m[g_k] = 1$
    **for** $n = 1$ to $N$ **do**
        **if** $Y_m[n] == 0$ **then**
            Set $Y_m[n] = 1$ with prob. $\epsilon$
        **else**
            Set $Y_m[n] = 0$ with prob. $\delta$
```

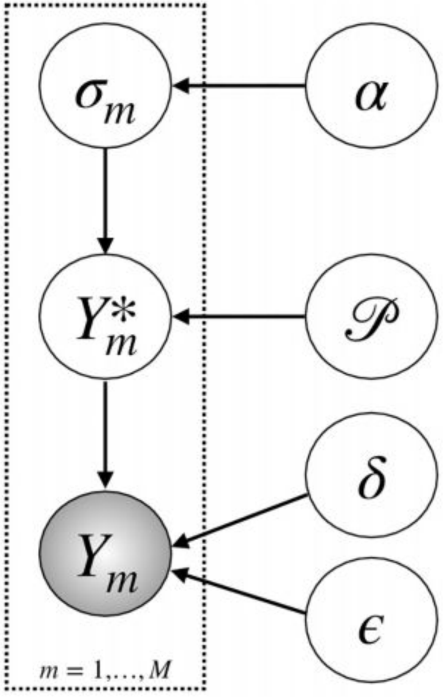


Probabilistic Generative Model



Algorithm 1 Generative process

```
Input: $\alpha, \mathcal{P}, \epsilon, \delta$
Output: $Y$
for $m \in \{1, \dots, M\}$ do
    Let $Y_m$ be a vector of $N$ zeros
    Draw $\sigma_m \sim \text{Categorical}(\alpha)$
    for $k \in \{1, \dots, \sigma_m\}$ do
        Draw $g_k$ (uniformly) from $D_k$
        Set $Y_m[g_k] = 1$
    for $n = 1$ to $N$ do
        if $Y_m[n] == 0$ then
            Set $Y_m[n] = 1$ with prob. $\epsilon$
        else
            Set $Y_m[n] = 0$ with prob. $\delta$
```



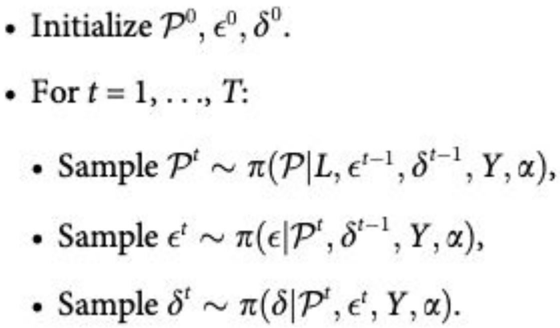
Overview

1. Linear Progression Model
2. Probabilistic Generative Process
3. Model Selection
4. Experiments on Simulated Data
5. Real Data Validation

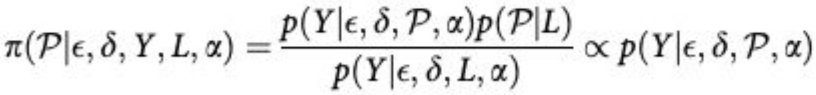
Model Likelihood

- Determine the number of driver pathways L by computing marginal likelihood given L
- $Y = \{Y_m: m \in \{1, \ldots, M\}\}$ Mutation matrix for a collection of tumors
- $\mathcal{P} = (D_1, D_2, \ldots, D_L, P)$ Pathway progression model
- Find the $p(Y|\mathcal{P}, \alpha, \epsilon, \delta)$

Model Selection w/ MCMC



- Initialize $\mathcal{P}^0, \epsilon^0, \delta^0$.
- For $t = 1, \ldots, T$:
  - Sample $\mathcal{P}^t \sim \pi(\mathcal{P}|L, \epsilon^{t-1}, \delta^{t-1}, Y, \alpha)$,
  - Sample $\epsilon^t \sim \pi(\epsilon|\mathcal{P}^t, \delta^{t-1}, Y, \alpha)$,
  - Sample $\delta^t \sim \pi(\delta|\mathcal{P}^t, \epsilon^t, Y, \alpha)$.

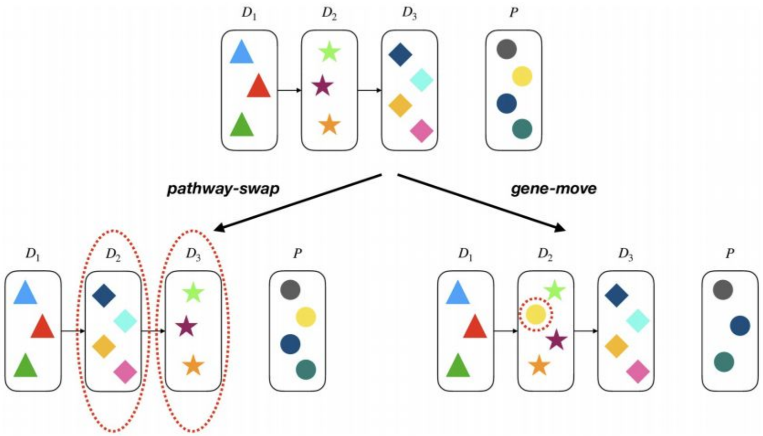


$$\pi(\mathcal{P}|\epsilon, \delta, Y, L, \alpha) = \frac{p(Y|\epsilon, \delta, \mathcal{P}, \alpha)p(\mathcal{P}|L)}{p(Y|\epsilon, \delta, L, \alpha)} \propto p(Y|\epsilon, \delta, \mathcal{P}, \alpha)$$

MCMC algorithm to generate samples from the posterior

- Given Y observations for fixed model length L
- Sample the progression model and error parameters for T iter

Progression Model Proposal



$$a(\mathcal{P}^*|\mathcal{P}^t) \;= \min\left\{1, \frac{\pi(\mathcal{P}^*|\epsilon,\delta,Y,L,\alpha)}{\pi(\mathcal{P}^t|\epsilon,\delta,Y,L,\alpha)} \times \frac{q(\mathcal{P}^t|\mathcal{P}^*)}{q(\mathcal{P}^*|\mathcal{P}^t)}\right\} = \min\left\{1, \frac{p(Y|\epsilon,\delta,\mathcal{P}^*,\alpha)}{p(Y|\epsilon,\delta,\mathcal{P}^t,\alpha)}\right\}.$$

Proposal function randomly chosen via Bernoulli:

- Pathway-swap
- gene-move

Overview

1. Linear Progression Model
2. Probabilistic Generative Process
3. Model Selection
4. Experiments on Simulated Data
5. Real Data Validation

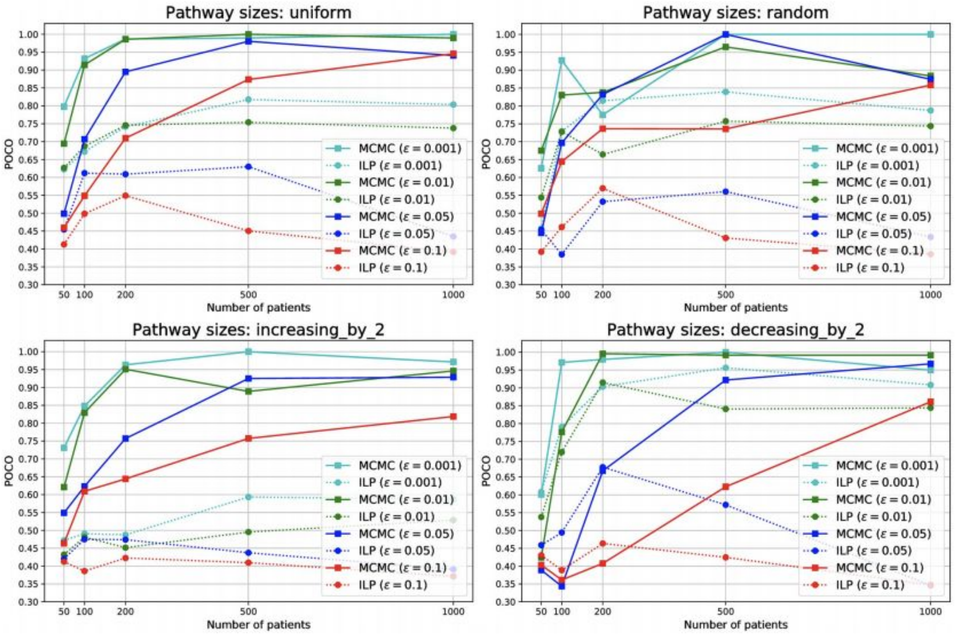
Metrics

- Model Evidence
- POCO (Percentage of Correct Ordering of Genes)
    - Given an inferred model, for each pair of genes check relative position
    - POCO is the percentage of gene pairs with correct relative position
- F1 Score

$$\text{F1} = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$$

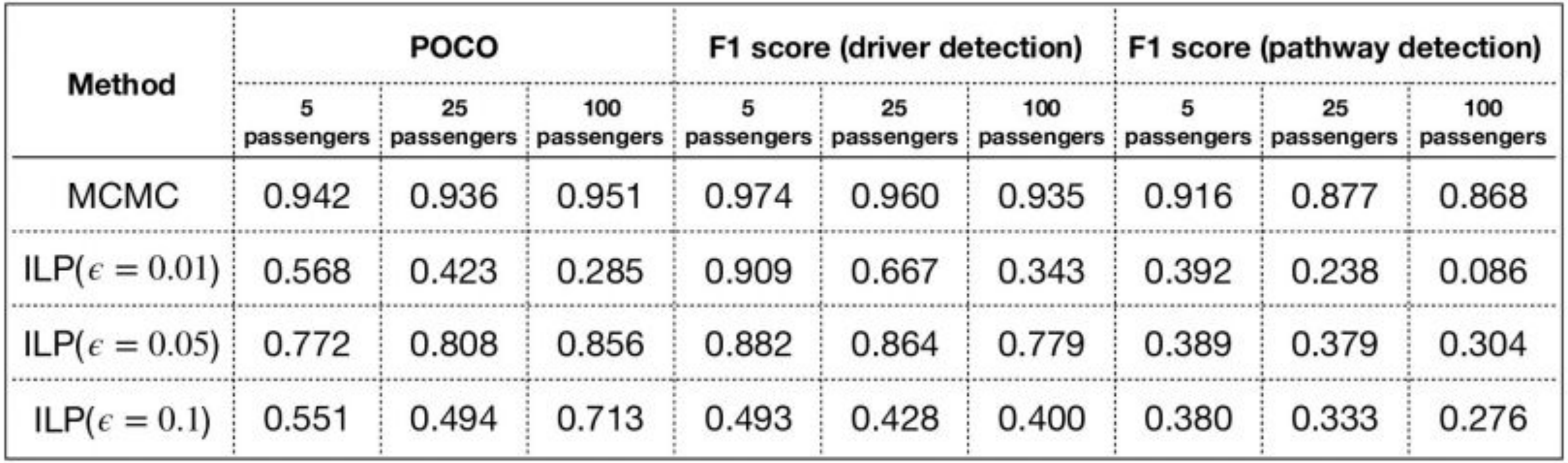
Exp 1: Known Driver Genes

- Distributed 25 genes in 5 driver pathways:



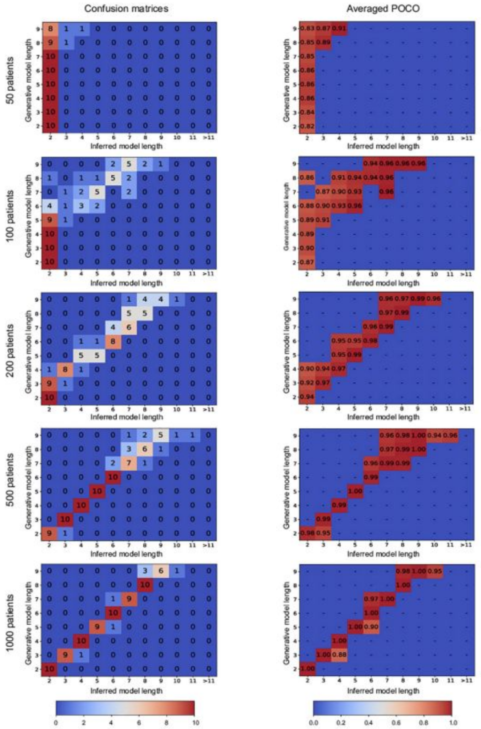
Exp 2: Unknown Driver Genes

- 10 datasets with 500 patients
- ILP requires background mutation rate as input



Method	POCO			F1 score (driver detection)			F1 score (pathway detection)		
	5 passengers	25 passengers	100 passengers	5 passengers	25 passengers	100 passengers	5 passengers	25 passengers	100 passengers
MCMC	0.942	0.936	0.951	0.974	0.960	0.935	0.916	0.877	0.868
ILP($\epsilon = 0.01$)	0.568	0.423	0.285	0.909	0.667	0.343	0.392	0.238	0.086
ILP($\epsilon = 0.05$)	0.772	0.808	0.856	0.882	0.864	0.779	0.389	0.379	0.304
ILP($\epsilon = 0.1$)	0.551	0.494	0.713	0.493	0.428	0.400	0.380	0.333	0.276

Exp 3: Model Length Selection

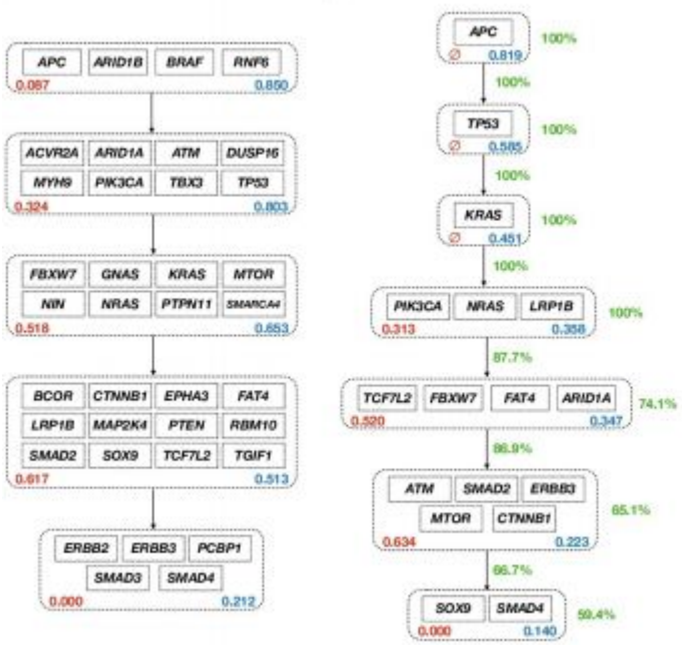


- 10 datasets from length [2, 9] with 50-1000 patients
- POCO is better than length at small patient cohort sizes

Overview

1. Linear Progression Model
2. Probabilistic Generative Process
3. Model Selection
4. Experiments on Simulated Data
5. Real Data Validation

Biological Data Analysis



$$S_{\mathrm{ME}}(g_i, g_j) = S_{\mathrm{ME}}(g_j, g_i) = \frac{\text{mutation rate of } g_i \text{ in patients with mutated } g_j}{\text{mutation rate of } g_i \text{ in all patients}}$$

ME scores of the pathways are the red numbers

Orthogonal validation of the driver genes based on domain knowledge

ILP places low mutation rate false positives, vs. MCMC focus on highly mutated genes

Conclusion

- Probabilistic generative process for tumor progression
- MCMC process for model selection
- Comparison:
    - ILP places low mutation rate genes, more false positives
    - MCMC places highly mutated genes, more passenger genes