Phyolin: Identifying a Linear Perfect Phylogeny in Single-cell DNA Sequencing Data of Tumors

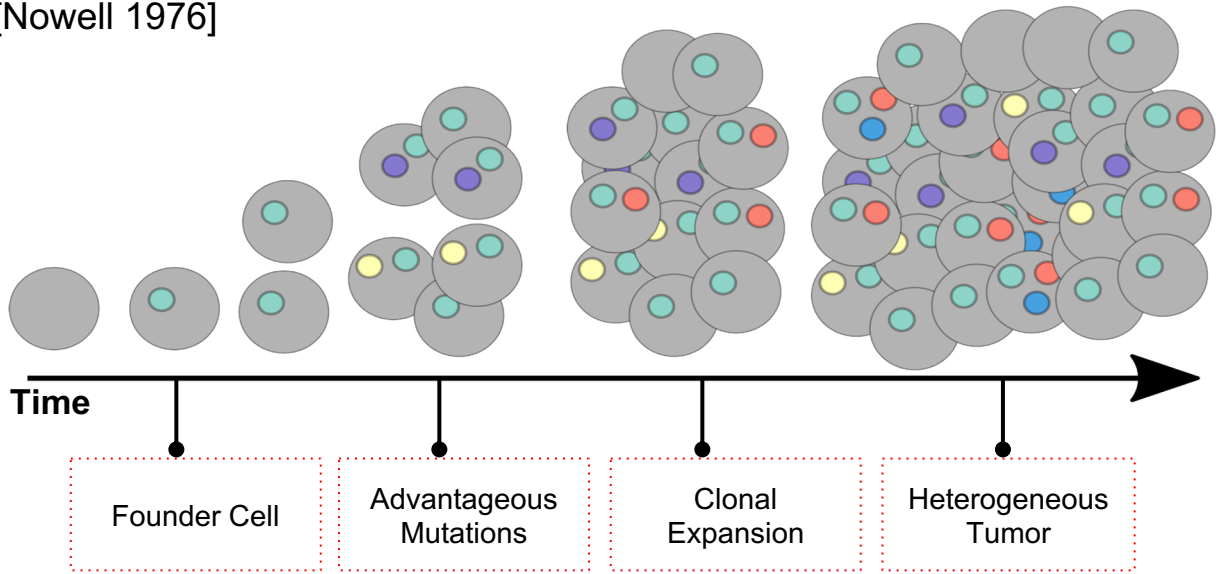
Leah Weber and Mohammed El-Kebir

University of Illinois at Urbana-Champaign, Department of Computer Science

WABI 2020
September 7-9, 2020

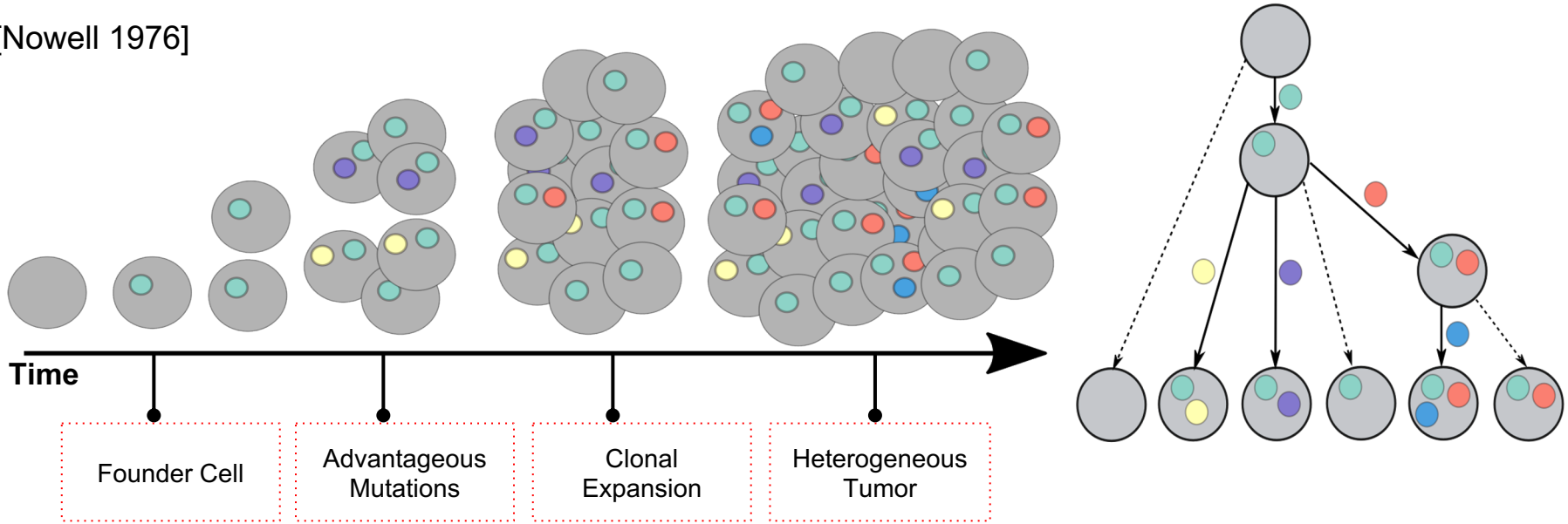
Cancer is an evolutionary process

[Nowell 1976]



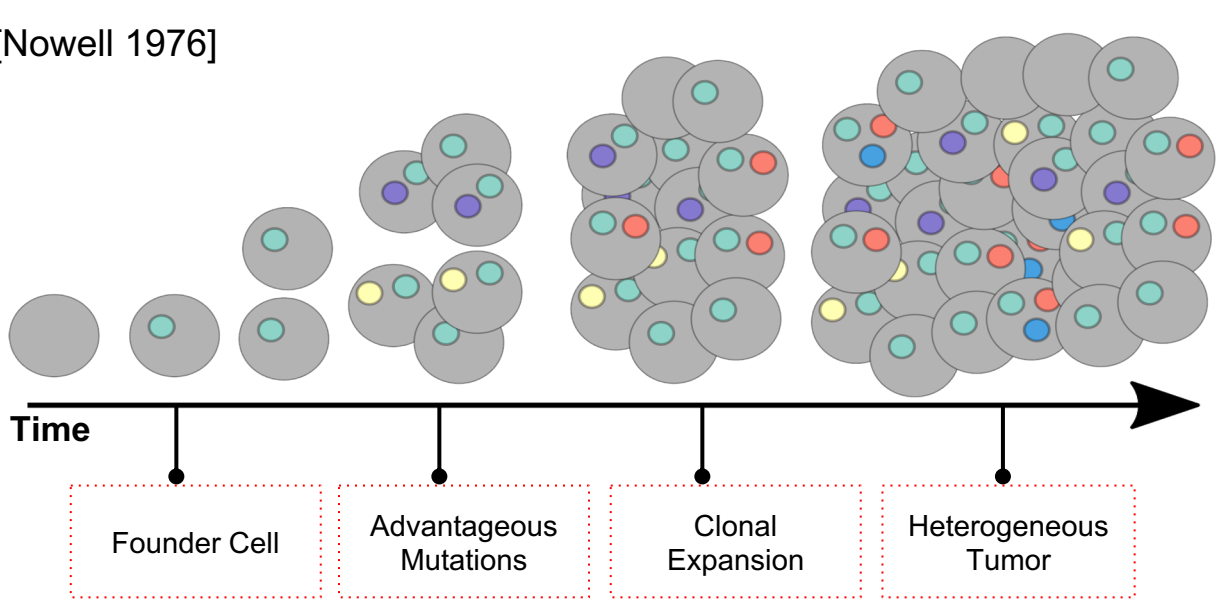
Cancer is an evolutionary process

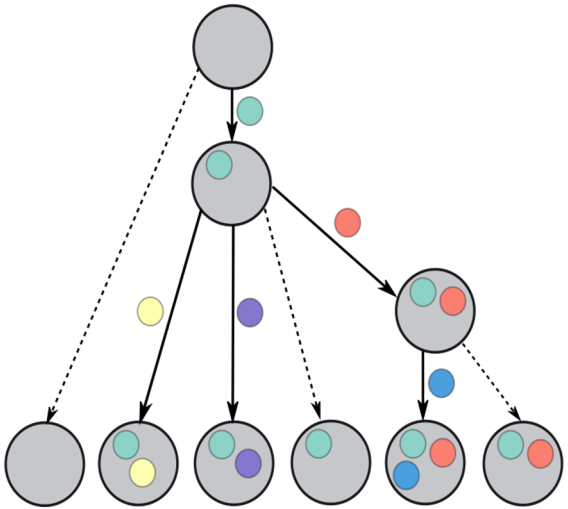
[Nowell 1976]



Cancer is an evolutionary process

[Nowell 1976]





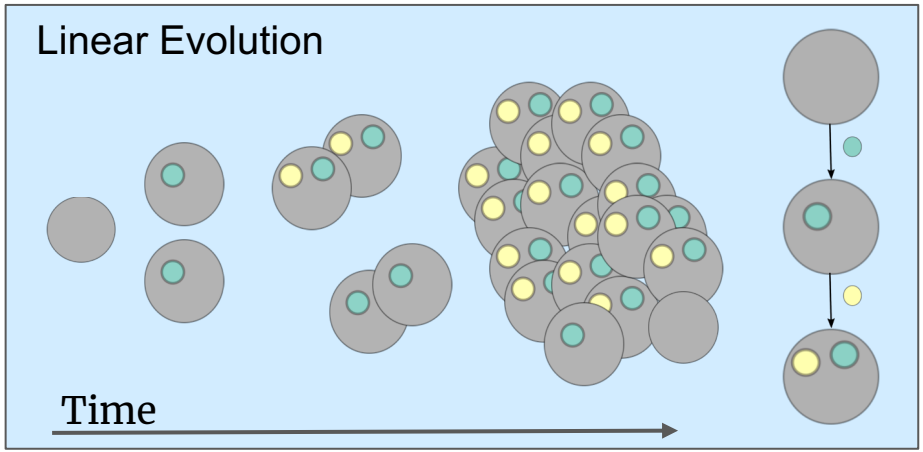
Identify treatment targets

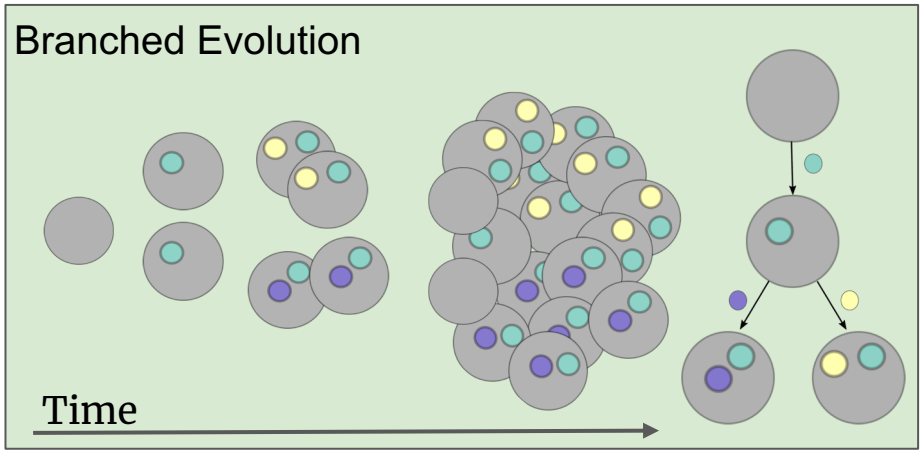
Understand metastatic development

Compare evolutionary patterns across patients

Modes of Cancer Evolution [Davis et al. 2017] [Watkins & Schwarz 2018]

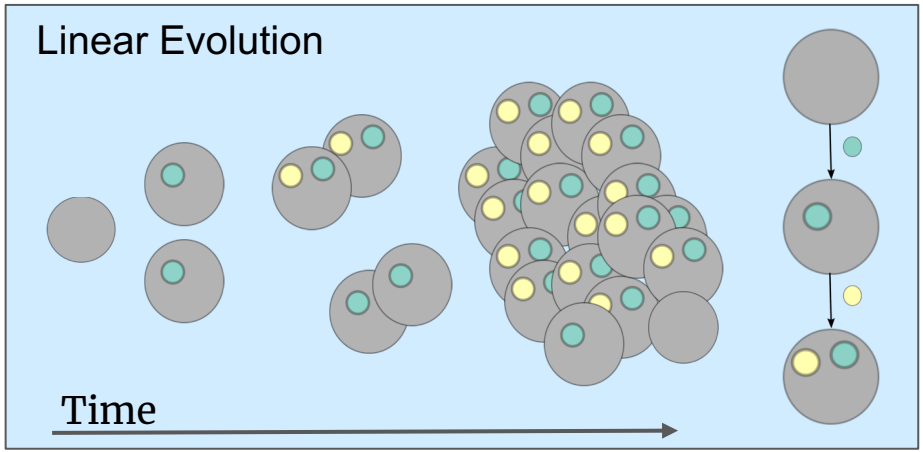
How do tumor cells respond to selective pressures and evolve over time?

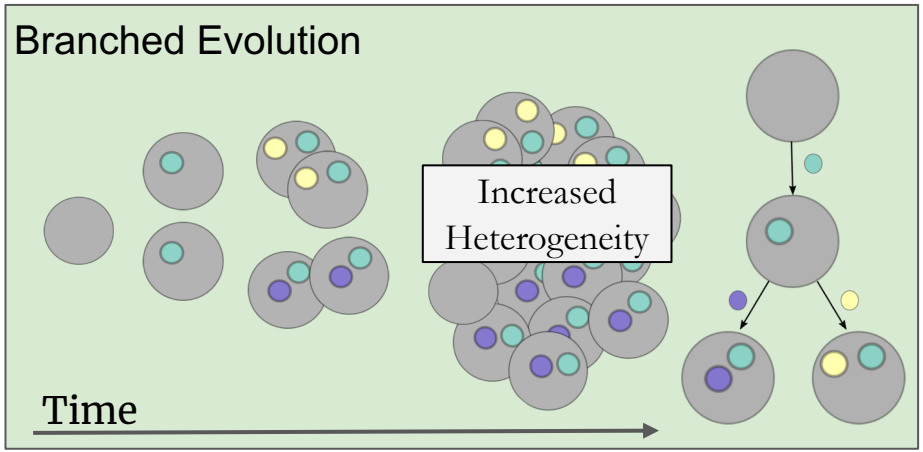




Modes of Cancer Evolution [Davis et al. 2017] [Watkins & Schwarz 2018]

How do tumor cells respond to selective pressures and evolve over time?



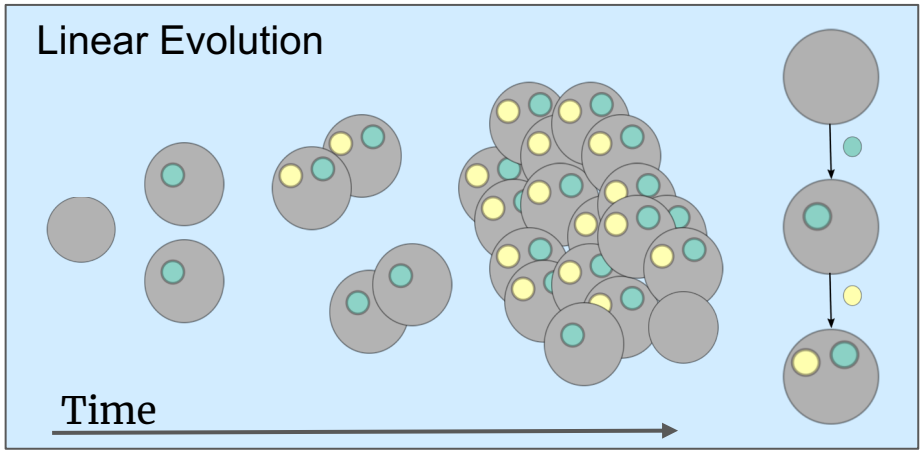


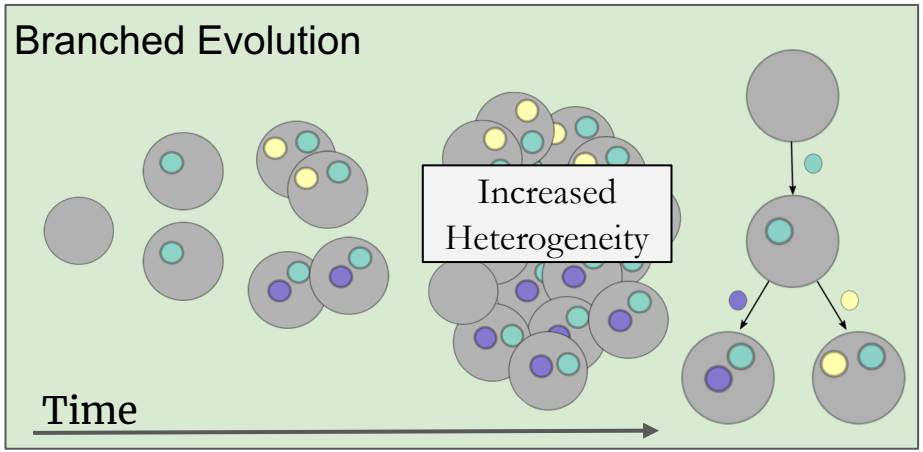
Increased Heterogeneity → Treatment Resistance [Merlo et al. 2006]

Modes of Cancer Evolution [Davis et al. 2017] [Watkins & Schwarz 2018]

How do tumor cells respond to selective pressures and evolve over time?

Given single-cell data, can we discern between linear and branched evolution?

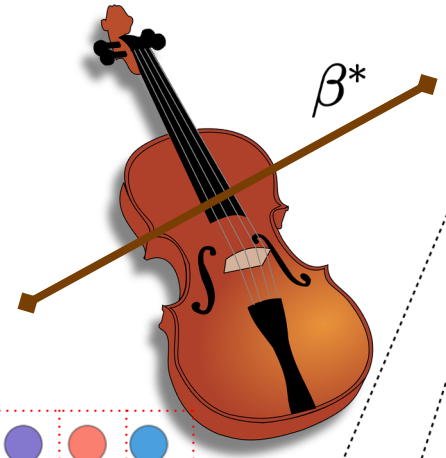




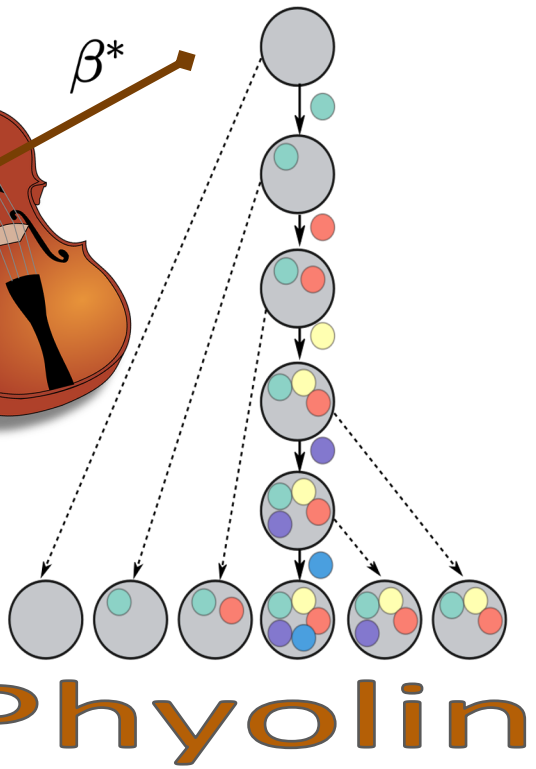
Increased Heterogeneity → Treatment Resistance [Merlo et al. 2006]

Outline

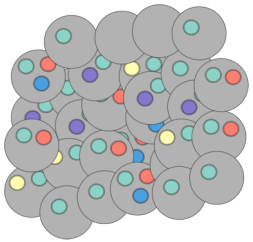
- Problem statement
- Complexity
- Methods
- Simulation study
- Application to real data
- Conclusions and future work

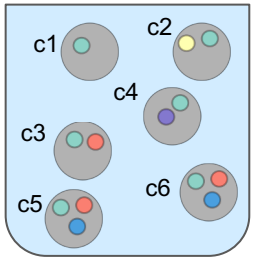


c1	1	0	0	0	0
c2	1	1	0	1	0
c3	0	0	0	1	0
c4	1	0	0	1	0
c5	1	1	0	1	1



Given single-cell data, can we reconstruct the tumor's evolution?





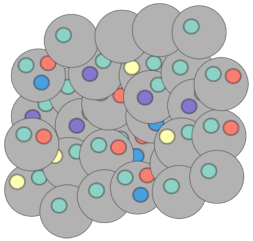
cells

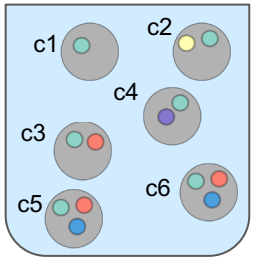
mutations

c1	1	0	0	1	0
c2	1	1	0	1	0
c3	0	0	0	1	0
c4	1	0	0	1	0
c5	1	?	0	1	1

0	False Negative
1	False Positive
?	Missing Data
1	Doublet

Given single-cell data, can we reconstruct the tumor's evolution?

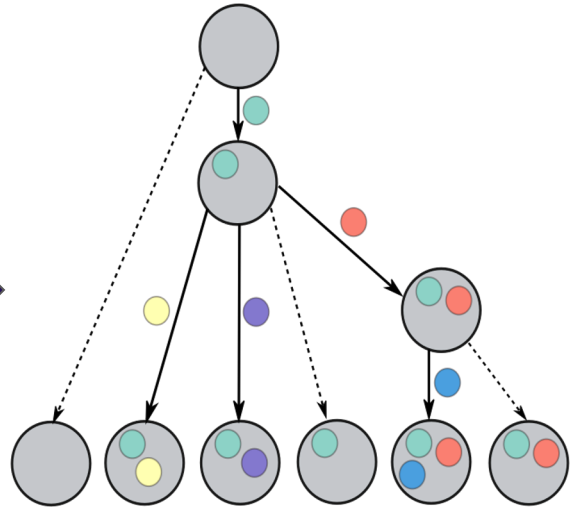




cells

mutations

c1	1	0	0	1	0
c2	1	1	0	1	0
c3	0	0	0	1	0
c4	1	0	0	1	0
c5	1	?	0	1	1

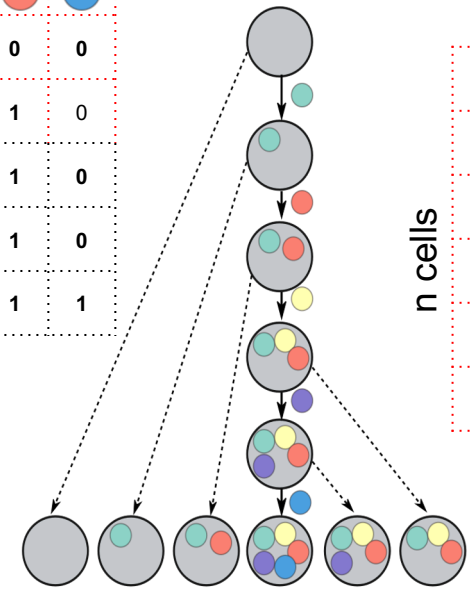


Data + Model

Model	Inference Method
Infinite sites	SCITE [Jahn et al. 2016] OncoNEM [Ross & Markowetz 2016]
Dollo parsimony	SPhyR [El-Kebir 2018] SCARLET [Satas et al. 2020]
Finite sites	SiFit/SiCloneFit [Zafar et al. 2017/2019]

Given single-cell data, can we discern between linear and branched evolution? [Azer et al. 2020]

c1	1	0	0	0	0
c2	1	0	0	1	0
c3	1	1	0	1	0
c4	1	1	1	1	0
c5	1	1	1	1	1



$B \in \{0,1\}^{n \times m}$

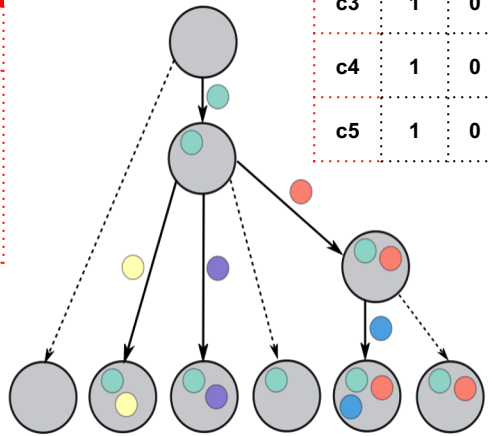
m mutations

n cells

c1	1	0	1	0	0
c2	1	1	0	0	0
c3	0	0	0	1	0
c4	1	0	0	1	0
c5	1	0	0	1	1

0 False Negative

c1	1	0	0	0	0
c2	1	1	0	0	0
c3	1	0	0	1	0
c4	1	0	1	1	0
c5	1	0	0	1	1

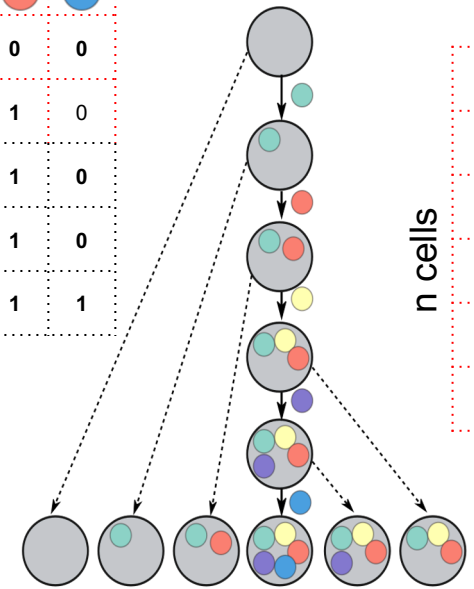


Given single-cell data, can we discern between linear and branched evolution? [Azer et al. 2020]

H_0 : B represents linear evolution

H_1 : B represents branched evolution

c1	1	0	0	0	0
c2	1	0	0	1	0
c3	1	1	0	1	0
c4	1	1	1	1	0
c5	1	1	1	1	1



$B \in \{0,1\}^{n \times m}$

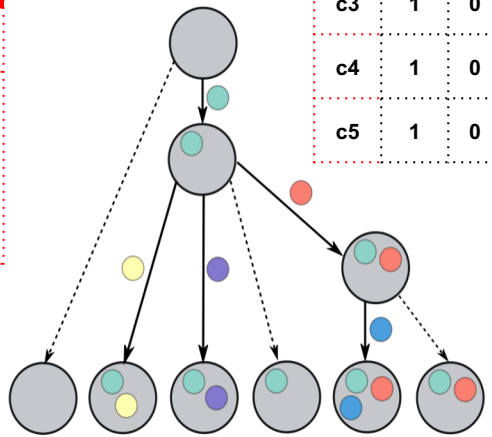
m mutations

n cells

c1	1	0	1	0	0
c2	1	1	0	0	0
c3	0	0	0	1	0
c4	1	0	0	1	0
c5	1	0	0	1	1

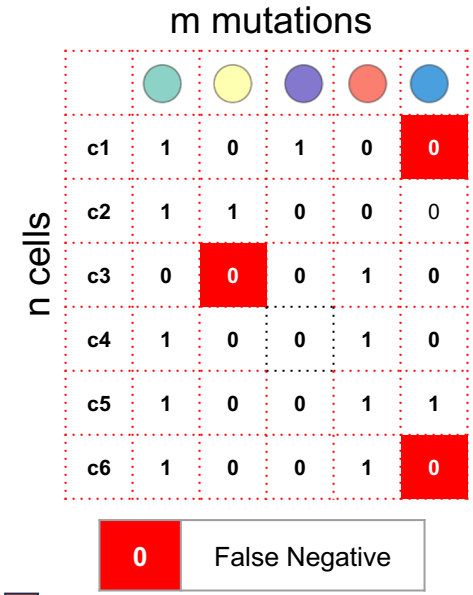
0 False Negative

c1	1	0	0	0	0
c2	1	1	0	0	0
c3	1	0	0	1	0
c4	1	0	1	1	0
c5	1	0	0	1	1



Key idea: Assessing the plausibility of linear evolution from single-cell DNA data

$$B \in \{0,1\}^{n \times m}$$

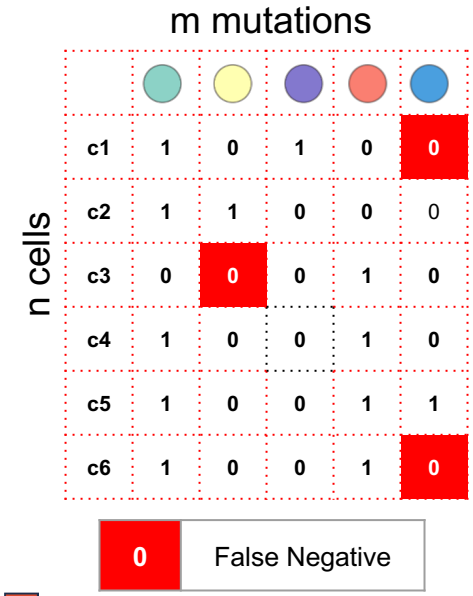


Linear Perfect Phylogeny Flipping Problem (LPPFP)

Given a matrix $B \in \{0,1\}^{n \times m}$,

Key idea: Assessing the plausibility of linear evolution from single-cell DNA data

$$B \in \{0,1\}^{n \times m}$$

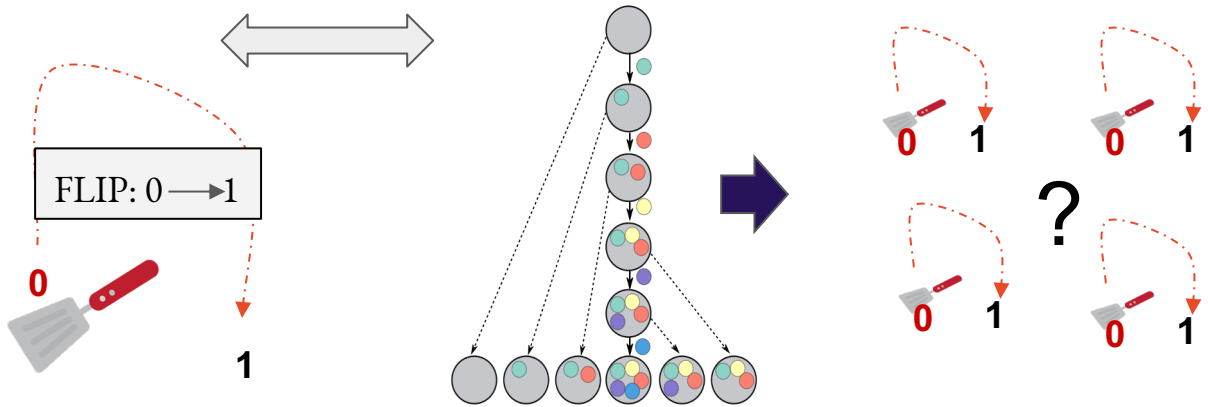


Linear Perfect Phylogeny Flipping Problem (LPFP)

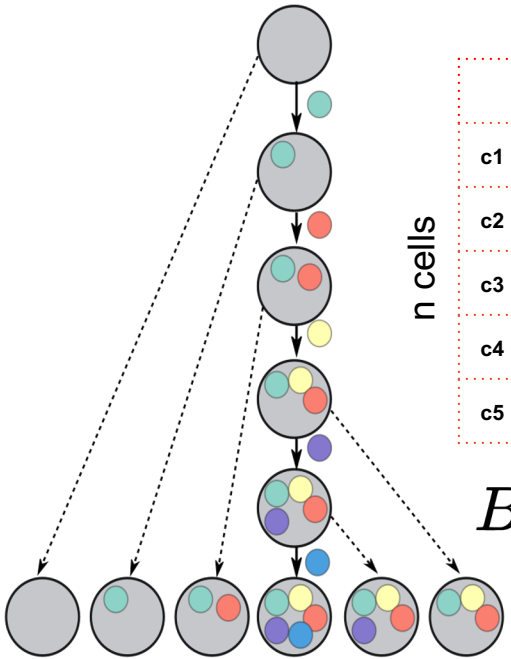
Given a matrix $B \in \{0,1\}^{n \times m}$,

find the minimum number of bit flips from 0 to1

such that B represents a linear perfect phylogeny.



Characterization of linear perfect phylogeny



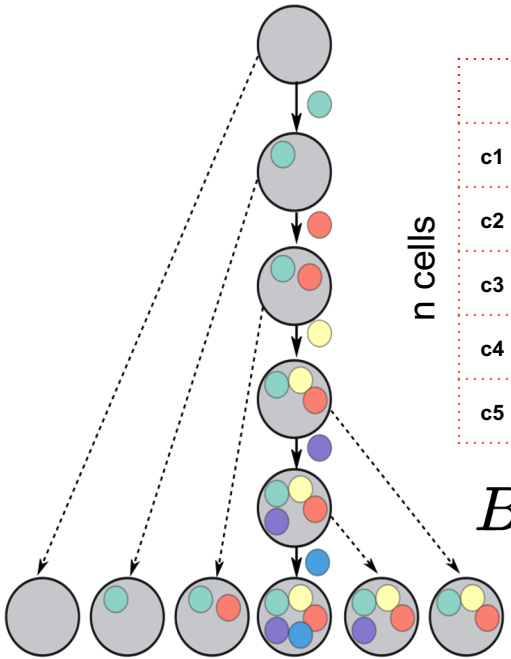
m mutations

n cells

c1	1	0	0	0	0
c2	1	1	1	0	0
c3	1	1	1	0	0
c4	1	1	1	1	0
c5	1	1	1	1	1

$$B \in \{0,1\}^{n \times m}$$

Characterization of linear perfect phylogeny



m mutations

n cells

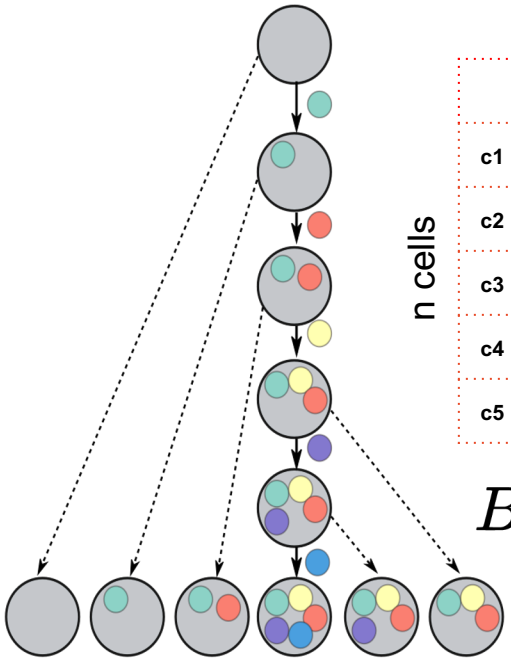
c1	1	0	0	0	0
c2	1	1	1	0	0
c3	1	1	1	0	0
c4	1	1	1	1	0
c5	1	1	1	1	1

$$B \in \{0,1\}^{n \times m}$$

The **one-state** of a mutation is the subset of cells with character state equal to 1.

$$O_{\bullet} = \{c_2, c_3, c_4, c_5\}$$

Characterization of linear perfect phylogeny



m mutations

n cells

c1	1	0	0	0	0
c2	1	1	1	0	0
c3	1	1	1	0	0
c4	1	1	1	1	0
c5	1	1	1	1	1

$$B \in \{0,1\}^{n \times m}$$

The **one-state** of a mutation is the subset of cells with character state equal to 1.

$$O_{\bullet} = \{c_2, c_3, c_4, c_5\}$$

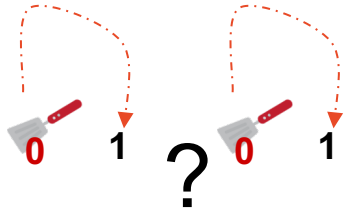
We say B **represents** a **linear perfect phylogeny** if there exists a total order on the mutation one-states with respect to the subset relation.

$$O_{\bullet} \subseteq O_{\bullet} \subseteq O_{\bullet} \subseteq O_{\bullet} \subseteq O_{\bullet}$$

Performing the hypothesis test

H_0 : B represents linear evolution

H_1 : B represents branched evolution



m mutations

n cells

c1	1	1	1	1	1
c2	1	1	0	1	1
c3	1	1	0	1	1
c4	1	0	0	0	0
c5	1	0	0	1	1

$$B' \in \{0,1\}^{n \times m}$$

Implied False Negative Rate after Flipping

$\hat{\beta}$

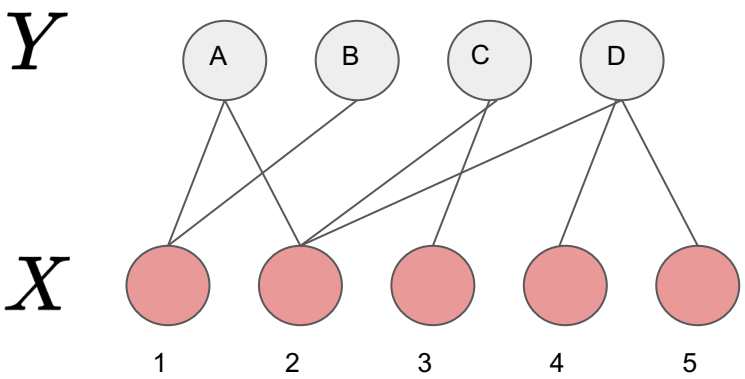
System Estimated False Negative Rate

β^*

$$\text{Reject } H_0 \text{ if } \hat{\beta} > \beta^*$$

LPPFP is NP-hard by reduction from the chain graph insertion problem (CG-IP) [Yannakakis 1981] [Chen et al. 2006]

$$G = (X \cup Y, E)$$

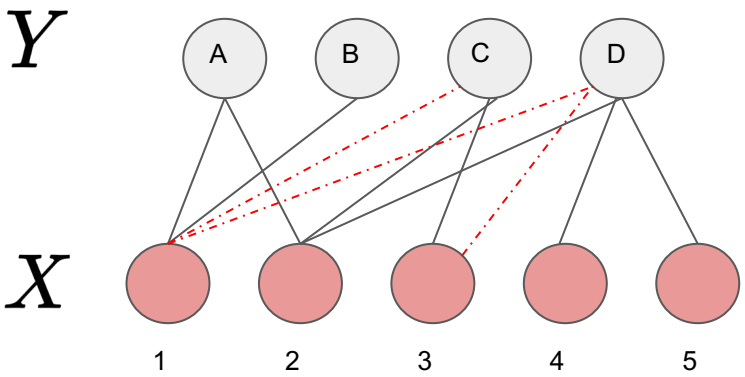


A bipartite graph G is **chain graph** if there exists a permutation $\varphi : \{1, \ldots, |Y|\} \rightarrow Y$ such that $\eta(\varphi(1)) \subseteq \eta(\varphi(2)) \subseteq \ldots \subseteq \eta(\varphi(|Y|)$ where $\eta(v) = \{w \in X : (v,w) \in E\}$ is the set of adjacent nodes of v.

$$\eta(A) = \{1, 2\}$$
$$\eta(C) = \{2, 3\}$$

LPFP is NP-hard by reduction from the chain graph insertion problem (CG-IP) [Yannakakis 1981] [Chen et al. 2006]

$$G = (X \cup Y, E)$$



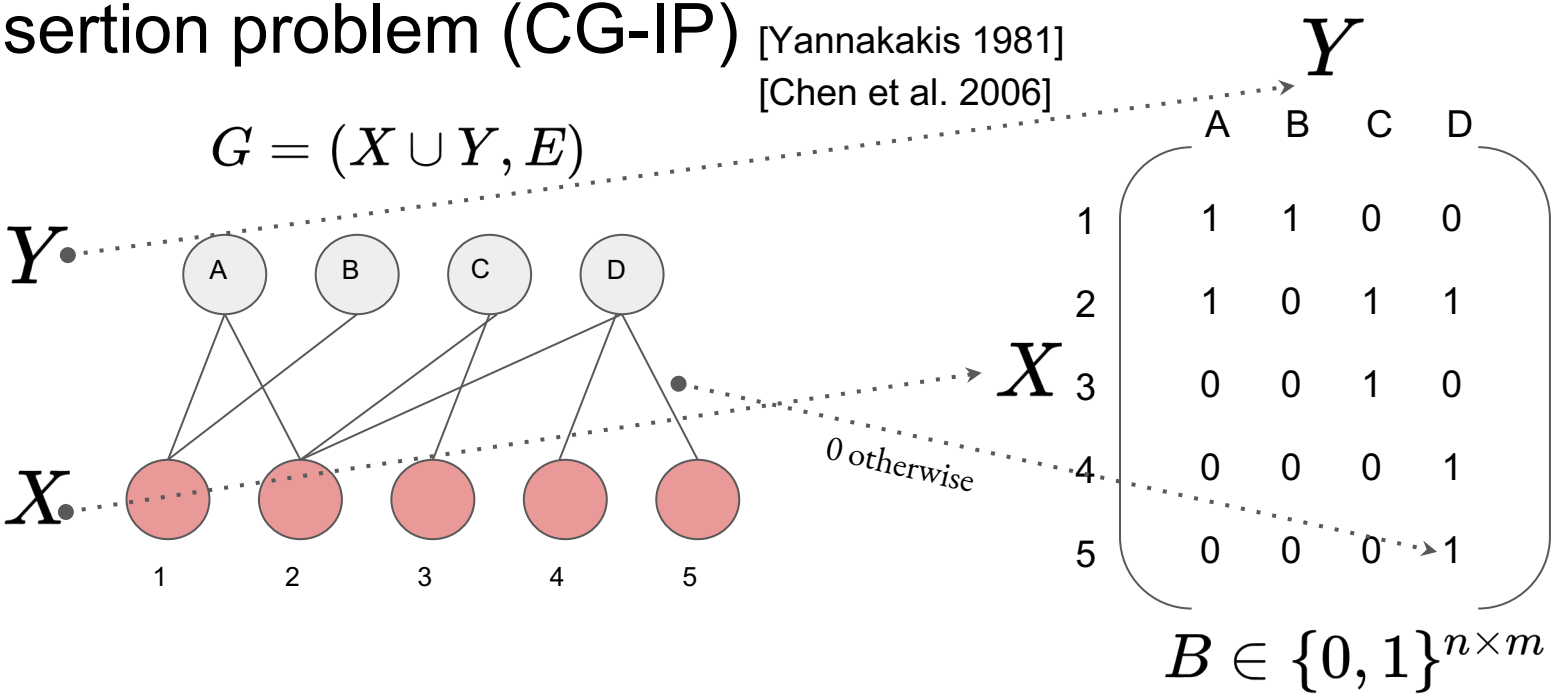
$$k = 3$$

$$\eta(A) = \{1, 2\}$$
$$\eta(B) = \{1\}$$
$$\eta(C) = \{1, 2, 3\}$$
$$\eta(D) = \{1, 2, 3, 4, 5\}$$

$$\eta(B) \subseteq \eta(A) \subseteq \eta(C) \subseteq \eta(D)$$

CG-IP: Given a bipartite graph G and an integer k, does there exists a chain graph $G' = (X \cup Y, E')$ such that $E \subset E'$ and $|E| + k = |E'|$?

LPPFP is NP-hard by reduction from the chain graph insertion problem (CG-IP) [Yannakakis 1981] [Chen et al. 2006]

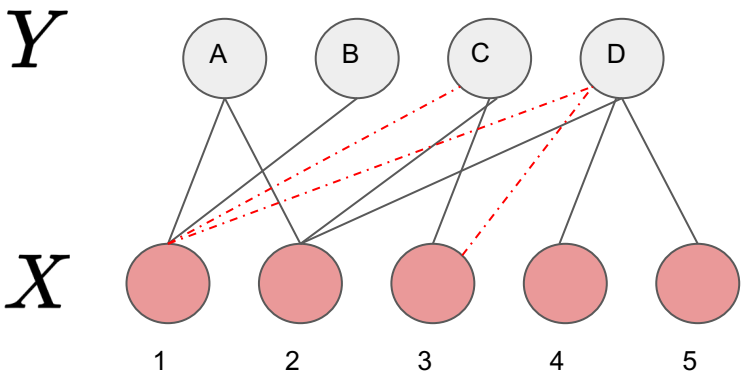


CG-IP: Given a bipartite graph G and an integer k, does there exists a chain graph $G' = (X \cup Y, E')$ such that $E \subset E'$ and $|E| + k = |E'|$?

k-LPPFP: Given a binary matrix B, does B represent a linear perfect phylogeny when k bits are flipped from 0 to 1?

LPFP is NP-hard by reduction from the chain graph insertion problem (CG-IP) [Yannakakis 1981] [Chen et al. 2006]

$$G = (X \cup Y, E)$$



$$\eta(B) \subseteq \eta(A) \subseteq \eta(C) \subseteq \eta(D)$$

Y

X	A	B	C	D
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	0	0	1
5	0	0	0	1

$$O_B \subseteq O_A \subseteq O_C \subseteq O_D$$

Lemma: A bipartite graph G with k edges inserted is a chain graph if and only if B represents a linear perfect phylogeny when k bits are flipped.

Calculating the test statistic via constraint programming

Phyolin

Objective

Minimize the total number of flips such that B' represents a linear perfect phylogeny

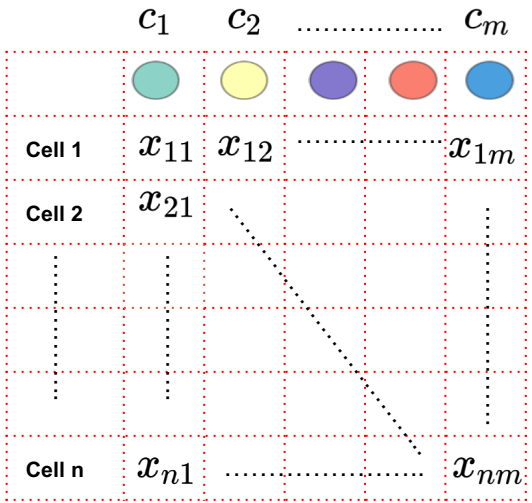
Decision Variables

The values in matrix B' after flipping.

$$x_{ij} \in \{0,1\}^{n \times m}$$

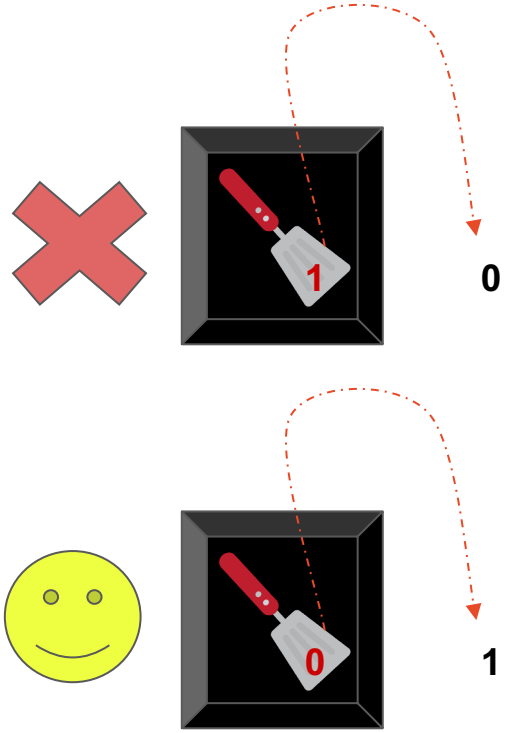
The position of the one-state of each mutation in the total order.

$$c_k \in [m]$$



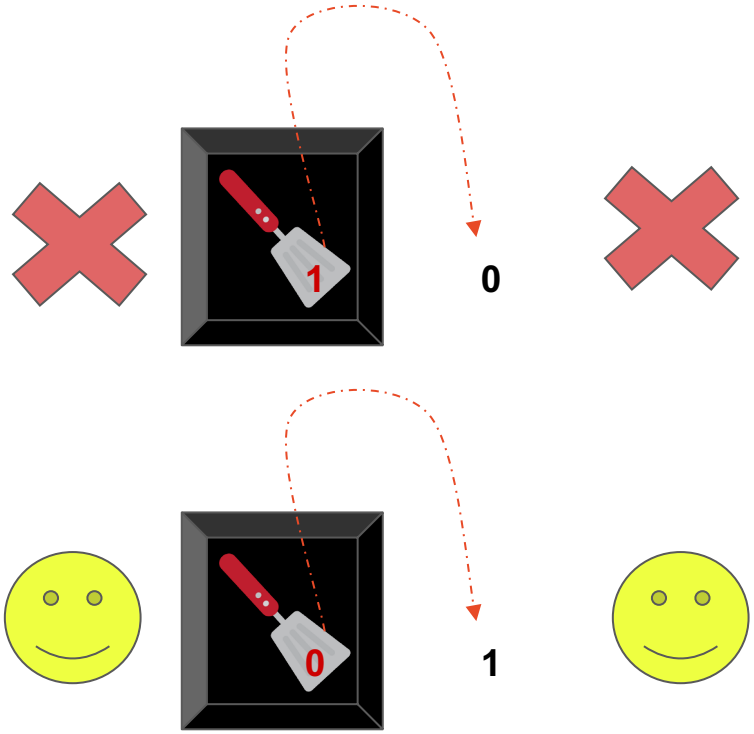
Total number of variables: nm + m

Model constraints



nm constraints

Model constraints



nm constraints

$O_i \not\subseteq O_j$

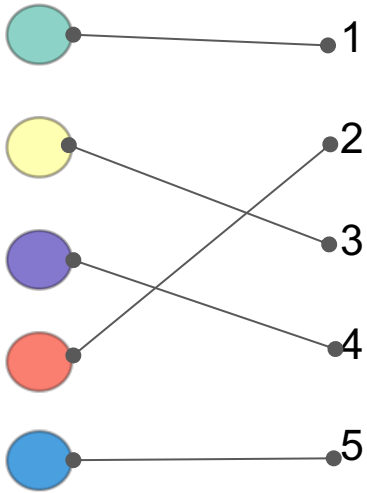
	O_i	O_j
cell $_i$	1	0
cell $_j$	0	1

$O_i \subseteq O_j$

	O_i	O_j
cell $_i$	0	1
cell $_j$	1	1

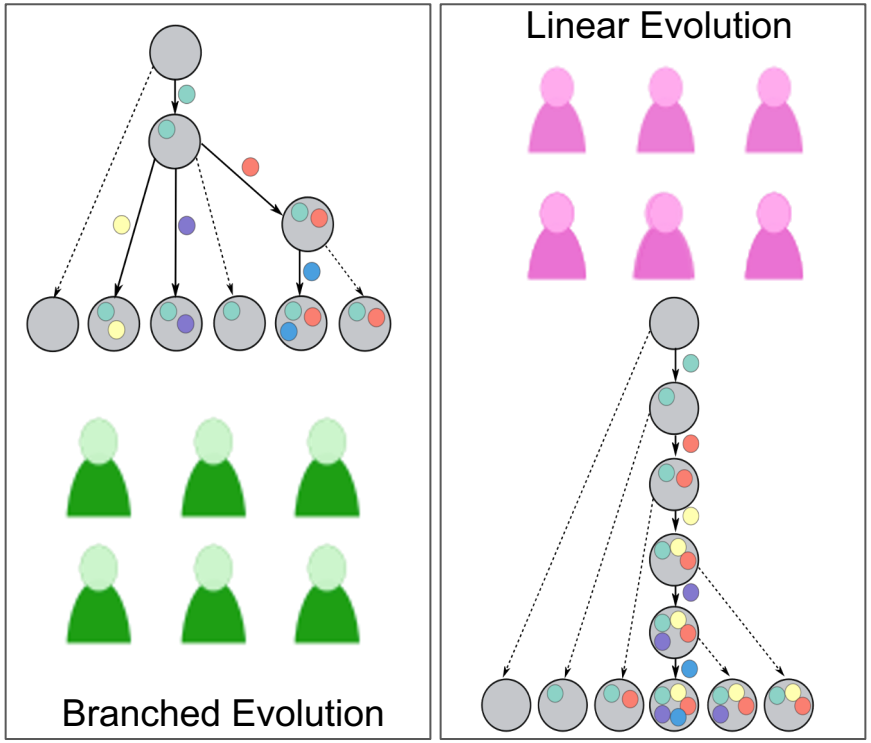
nm^2 constraints

$O_1 \subseteq O_2 \subseteq O_3 \subseteq O_4 \subseteq O_5$



1 global constraint (ALLDIFFERENT)

Simulating an acute myeloid leukemia (AML) cohort



patient	pattern	m	n [17]
AML-2	Linear	5	7931
AML-8	Linear	3	4675
AML-10	Linear	4	8729
AML-33	Linear	3	8120
AML-47	Linear	3	6491
AML-58	Linear	3	8170
AML-53	Branched	3	8013
AML-62	Branched	6	4027
AML-63	Branched	4	8347
AML-67	Branched	7	6024
AML-69	Branched	3	7462
AML-74	Branched	5	9279

m = number of mutations
n = number of cells

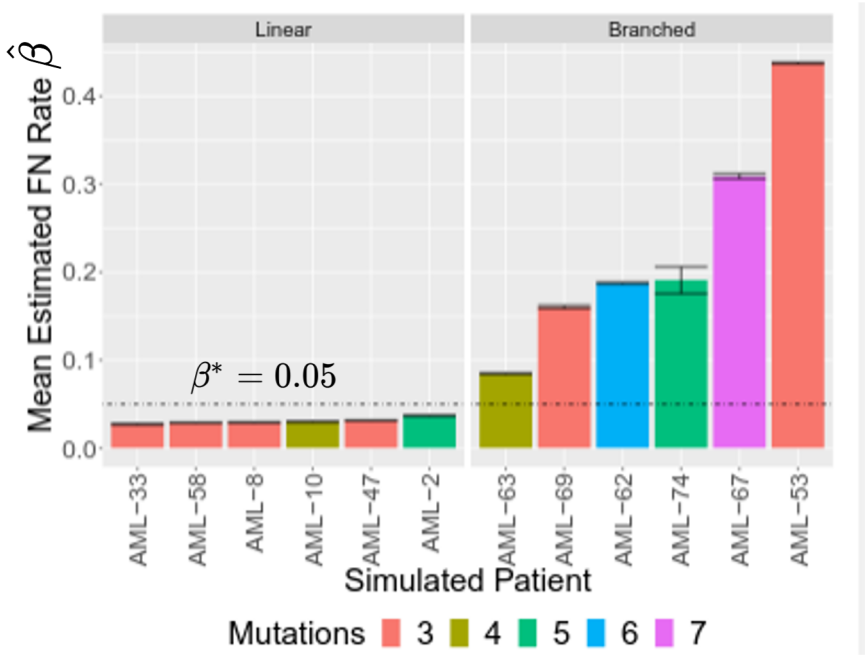
10 replications per patient

$\beta^* = 0.05$

IBM ILOG CP Optimizer with 500 s time limit

[Morita et al. 2020]

Phyolin Simulated AML Cohort Results



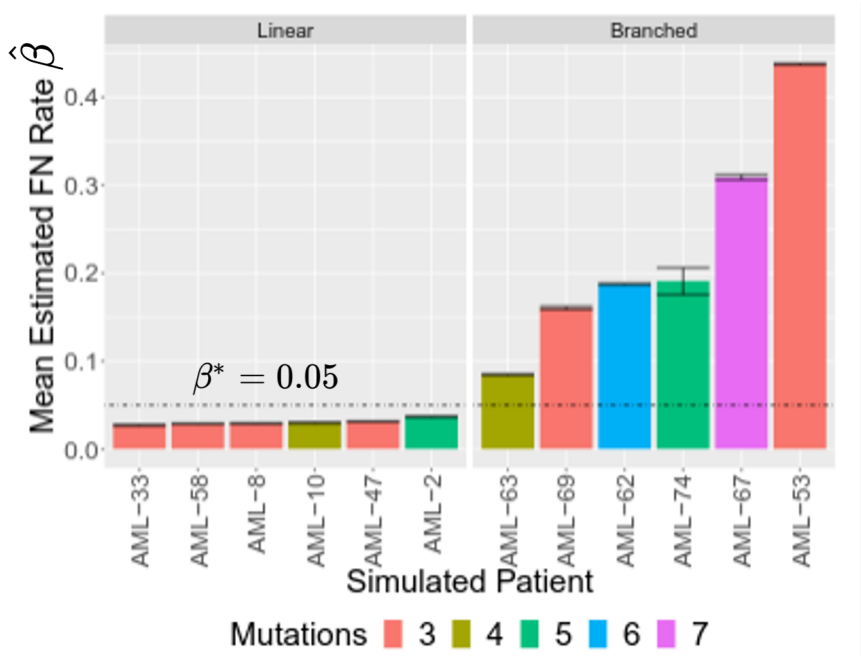
patient	pattern	m	n [17]	median flips	Phyolin median flips	median $\hat{\beta}$
AML-2	Linear	5	7931	1826	1039	0.037
AML-8	Linear	3	4675	759	294	0.029
AML-10	Linear	4	8729	1427	584	0.037
AML-33	Linear	3	8120	1091	350	0.027
AML-47	Linear	3	6491	1135	488	0.032
AML-58	Linear	3	8170	1280	472	0.029
AML-53	Branched	3	8013	544	2220	0.44
AML-62	Branched	6	4027	726.5	2299	0.19
AML-63	Branched	4	8347	1238.5	1432	0.084
AML-67	Branched	7	6024	1061.5	6440	0.31
AML-69	Branched	3	7462	651.5	2122	0.16
AML-74	Branched	5	9279	1020	294	0.17

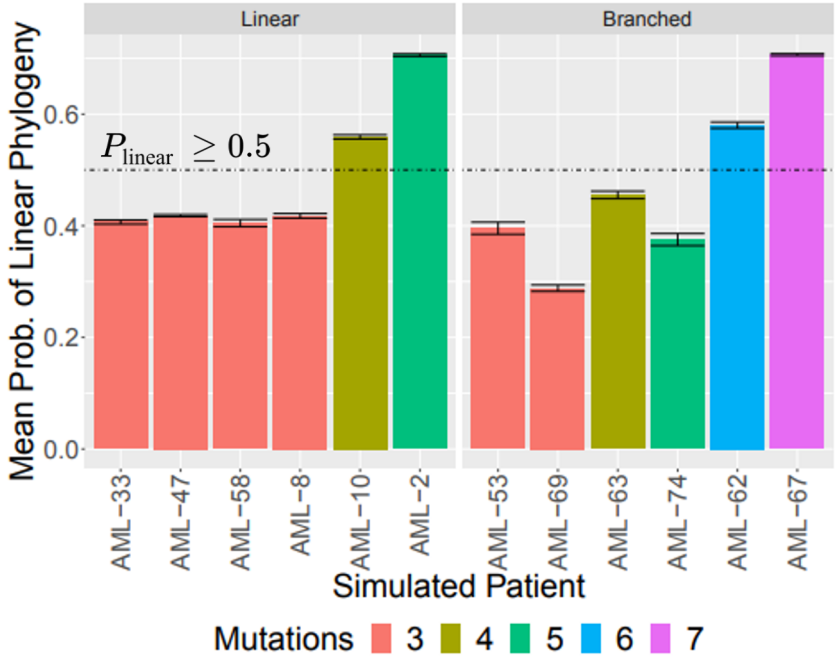
10 replications per patient

$\beta^* = 0.05$

IBM ILOG CP Optimizer with 500 s time limit

Comparison to Deep Learning Approach [Azer et al. 2020]

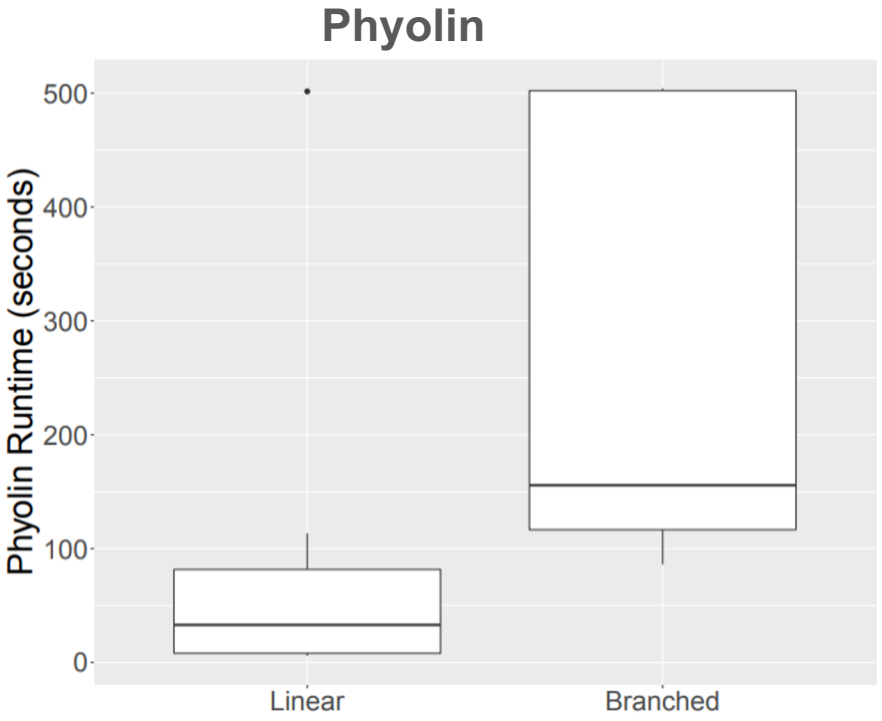




Deep Learning Approach

Retrained network for input size 9300 x 7, 10 hidden layers of 100 units and 0.9 dropout rate on 5000 training examples and 500 epochs

Runtime comparison



Deep Learning [Azer et al. 2020]

Training time with 5000 randomly generated training examples.

Epochs	Time	Training Accuracy
200	2168s (36.1 min)	64.1%
500	3997s (66.6 min)	64.8%

Application to two acute lymphoblastic leukemia patients [Gawad et al. 2010]

patient	cells sequenced [10]	mutations	Phyolin flips	$\hat{\beta}$	β^* [10]
Patient 2	115	16	403	0.36	0.18
Patient 6	146	10	191	0.15	0.18

	Phyolin	Deep Learning [Azer et al. 2020]
Patient 2	Branched	Branched

Application to two acute lymphoblastic leukemia patients [Gawad et al. 2010]

patient	cells sequenced [10]	mutations	Phyolin flips	$\hat{\beta}$	β^* [10]
Patient 2	115	16	403	0.36	0.18
Patient 6	146	10	191	0.15	0.18

	Phyolin	Deep Learning [Azer et al. 2020]
Patient 2	Branched	Branched
Patient 6	Linear	Linear

Patient 6 [Kuipers et al. 2017]

Conclusions and future work

Phyolin

- Easily and accurately assess the plausibility of a linear perfect phylogeny
- Code available at https://github.com/elkebir-group/phyolin

Future Work

- Incorporate false positives and doublets
- Consider copy number aberrations
- Explore evolutionary models beyond the infinite sites model
- Control for Type I error as a result of sampling bias

Acknowledgements

National Science Foundation (CCF-1850502)

El-Kebir group

- **Mohammed El-Kebir**
- Nuraini Aguse
- Yuanyuan Qi
- Jiaqi Wu
- Sarah Christensen
- Palash Sashittal
- Juho Kim
- Jackie Oh
- Chuanyi Zhang

Thank you to my CS598MEB Computational Cancer Genomics classmates for their helpful feedback and reviews of this work.