CS 598MEB
Computational Cancer Genomics
Lecture 3

Mohammed El-Kebir

January 22, 2019

Course Project

- 1-2 students per project
- First write a proposal, which will receive feedback from instructor and fellow students
- Then, conduct research and write a paper
- Pick venue (conference/journal) and use LaTeX style for your paper
- Students will anonymously peer review submitted papers using EasyChair (if time permits)

Lecture Outline

- Recap
- Two-state Perfect Phylogeny Mixtures

Reading

- M. El-Kebir, L. Oesper, H. Acheson-Field and B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics (Special Issue: Proceedings of ISMB), 31(12):i62-i70, 2015

- D. Pradhan and M. El-Kebir. On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem. In: Blanchette M., Ouangraoua A. (eds) Comparative Genomics. RECOMB-CG 2018. Lecture Notes in Computer Science, vol 11183. Springer, Cham.

Infinite Sites Model

The genome is large

Mutations are rare
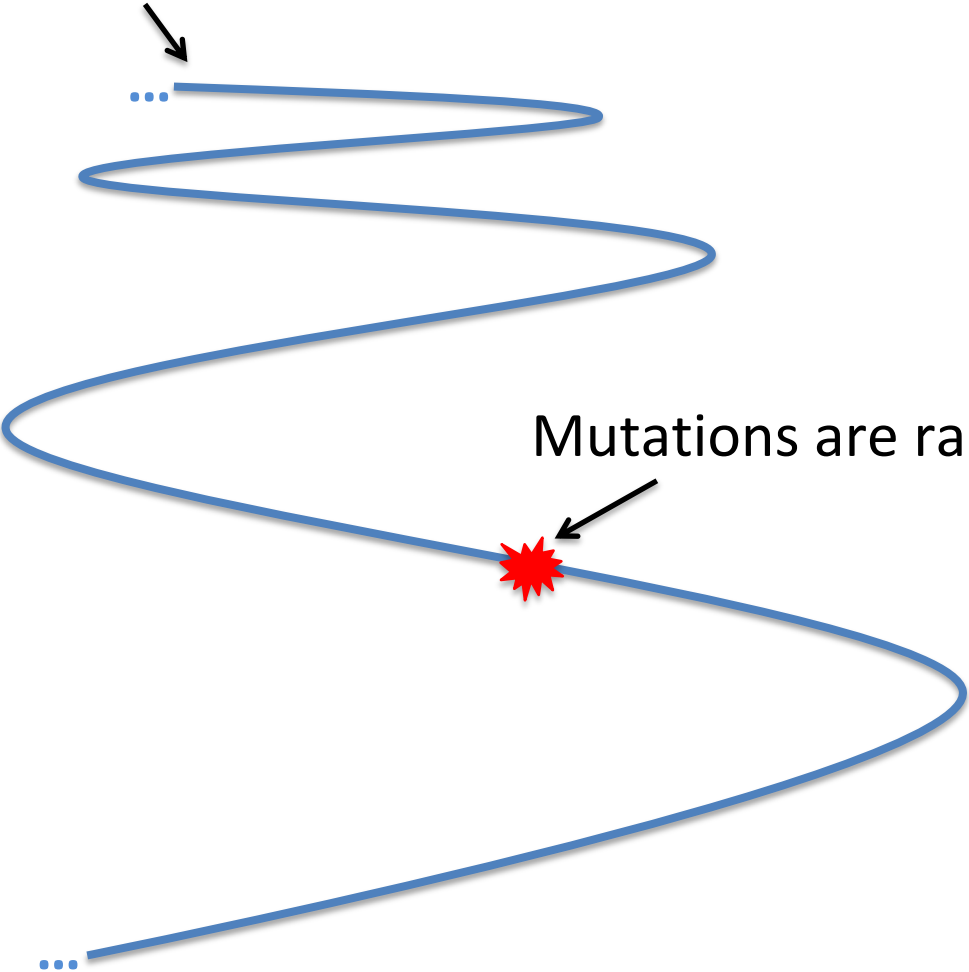
[Kimura, 1969]

Infinite sites model: multiple mutations never occur at the same position

Mutated Loci

	🔴	🔵	🟢	🟣	🟠	🟡
A	0	0	0	0	1	1
B	0	0	0	1	1	1
C	0	0	1	0	1	0
D	1	0	0	0	0	0
E	1	1	0	0	0	0

Species (cancer cells)

1: mutated
0: not

All sites are bi-allelic: mutated or not.

Progression of Somatic Mutations

Single nucleotide mutation

… CGT**A**ATTAG …
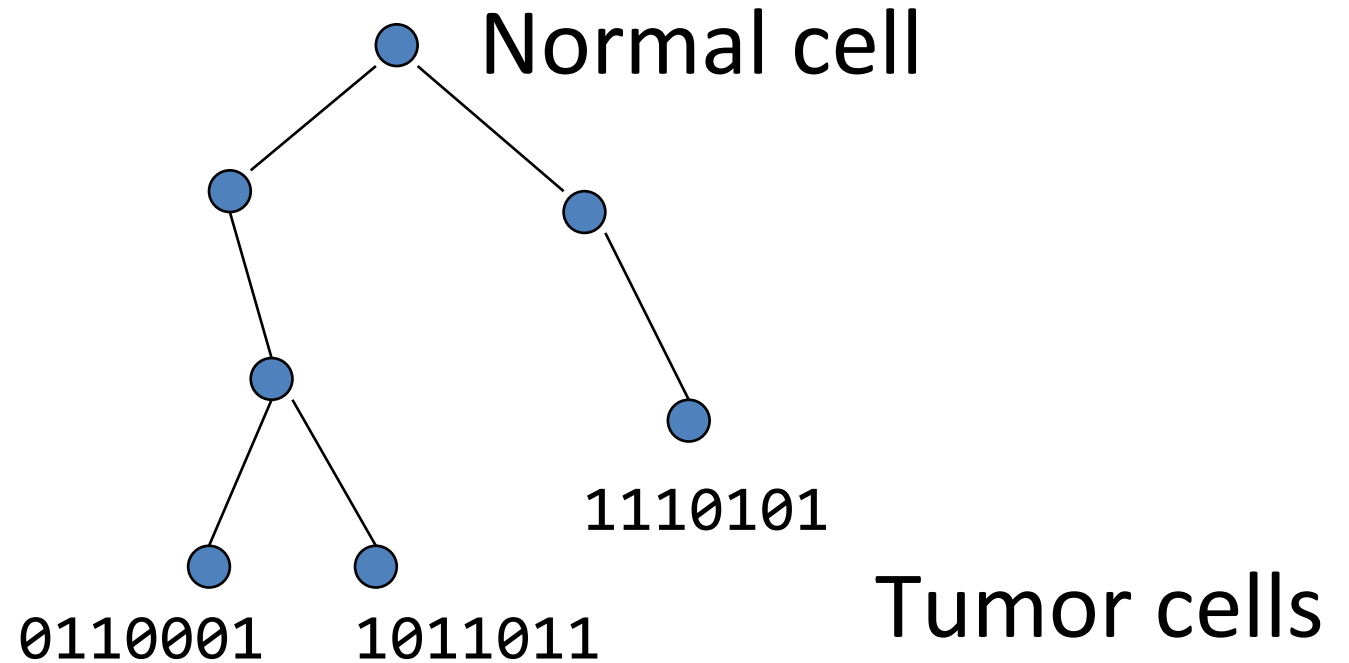
… CGT**C**ATTAG …

Normal cell

1110101

0 = normal
1 = mutated

0110001   1011011

Tumor cells

Root is the normal, founder cell and leaves are cells in tumor.

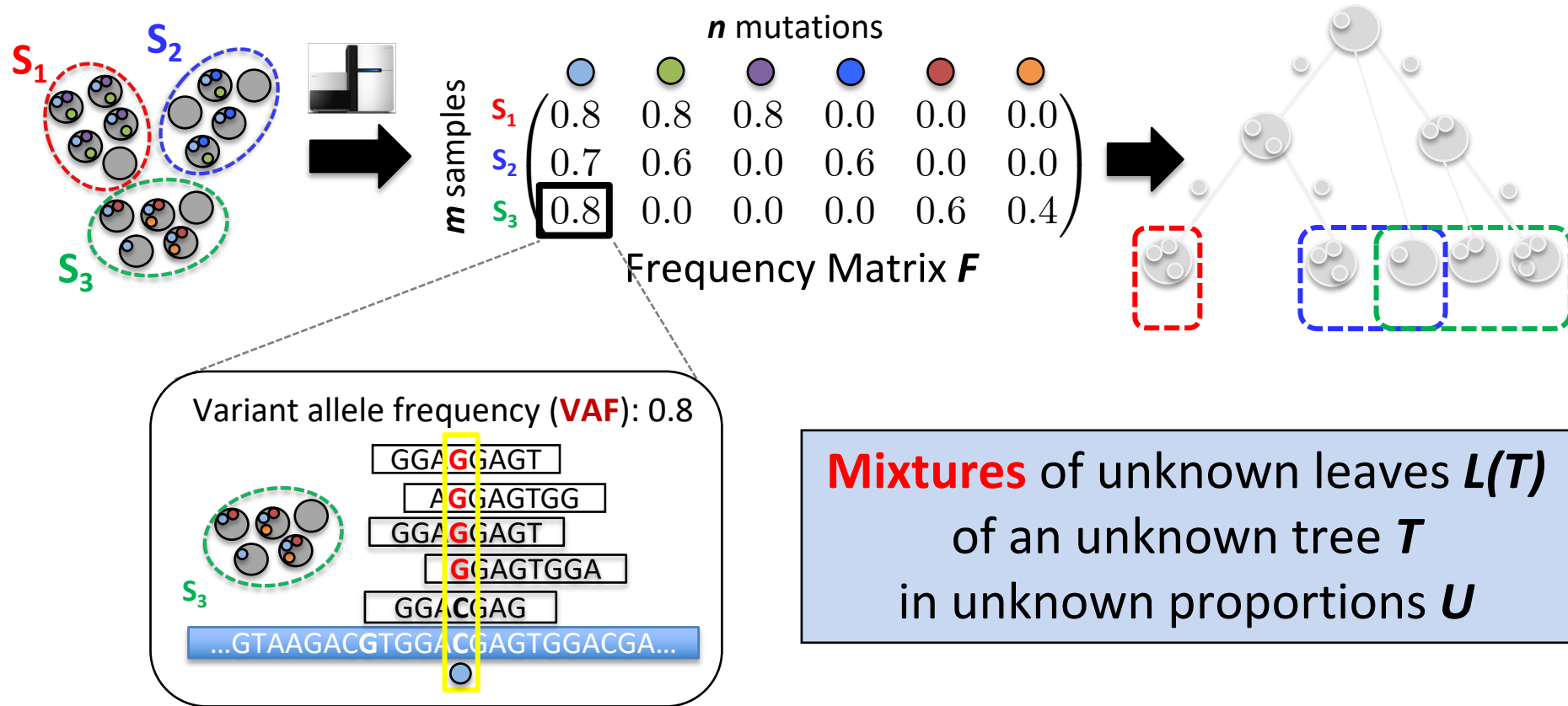**Infinite sites assumption**: each locus mutates only once.

Lecture Outline

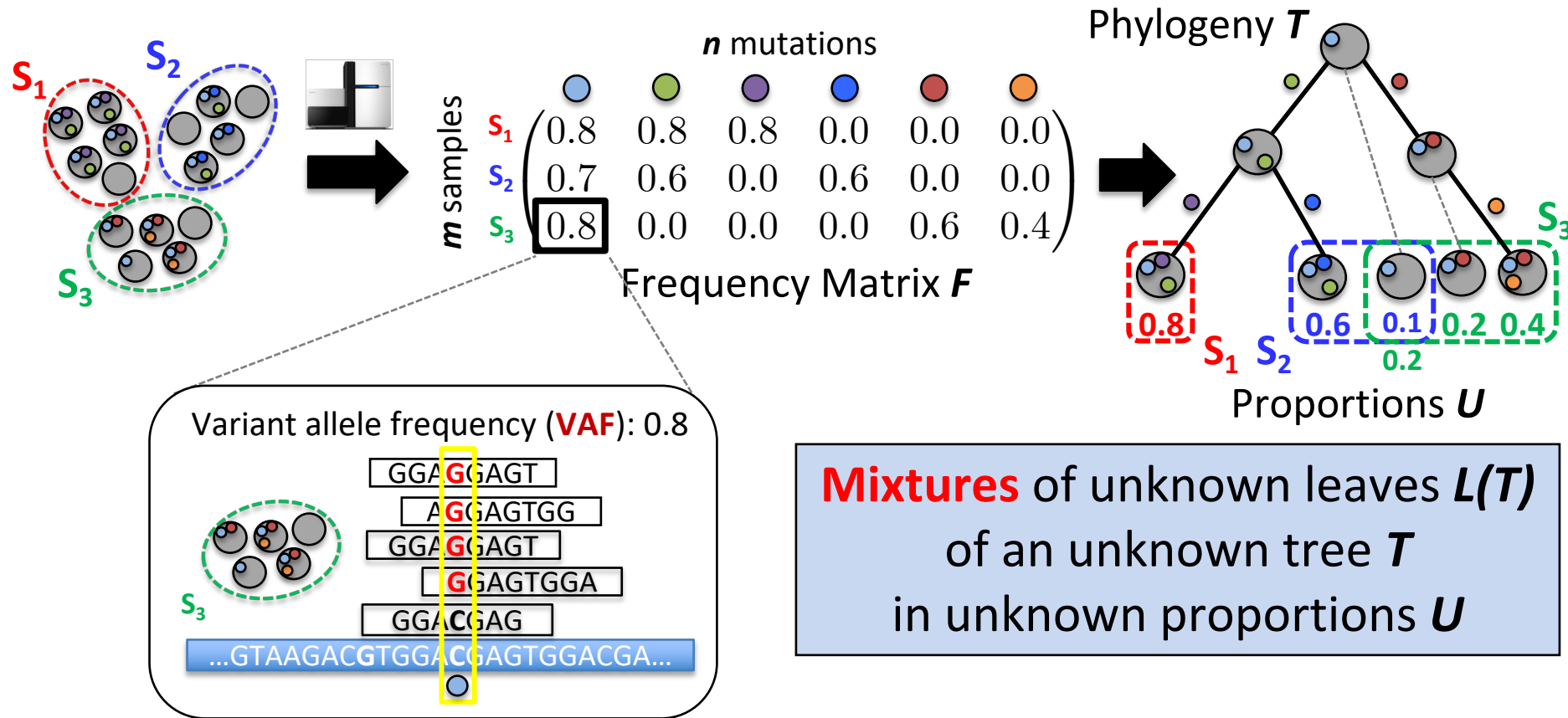- Recap
- Two-state Perfect Phylogeny Mixtures

Reading

- M. El-Kebir, L. Oesper, H. Acheson-Field and B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics (Special Issue: Proceedings of ISMB), 31(12):i62-i70, 2015

- D. Pradhan and M. El-Kebir. On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem. In: Blanchette M., Ouangraoua A. (eds) Comparative Genomics. RECOMB-CG 2018. Lecture Notes in Computer Science, vol 11183. Springer, Cham.

Sequencing and Tumor Phylogeny Inference

n mutations

m samples

$$F = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

S_1, S_2, S_3

Frequency Matrix F

Variant allele frequency (**VAF**): 0.8

GGA**G**GAGT
A**G**GAGTGG
GGA**G**GAGT
GGAGTGGA
GGA**C**GAG
…GTAAGAC**G**TGGA**C**GAGTGGACGA…

S_3

Mixtures of unknown leaves $L(T)$
of an unknown tree T
in unknown proportions U

Sequencing and Tumor Phylogeny Inference

Tumor Phylogeny Inference: Given frequencies F, find phylogeny T and proportions U
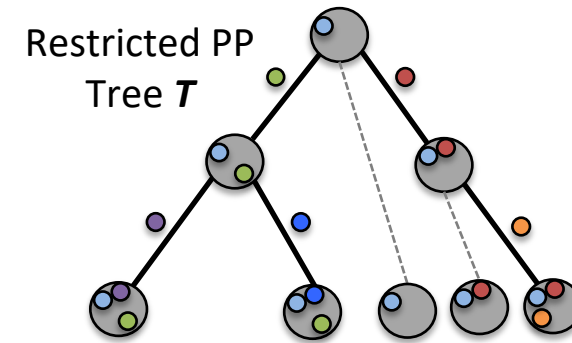
Key Challenge in Computational Biology
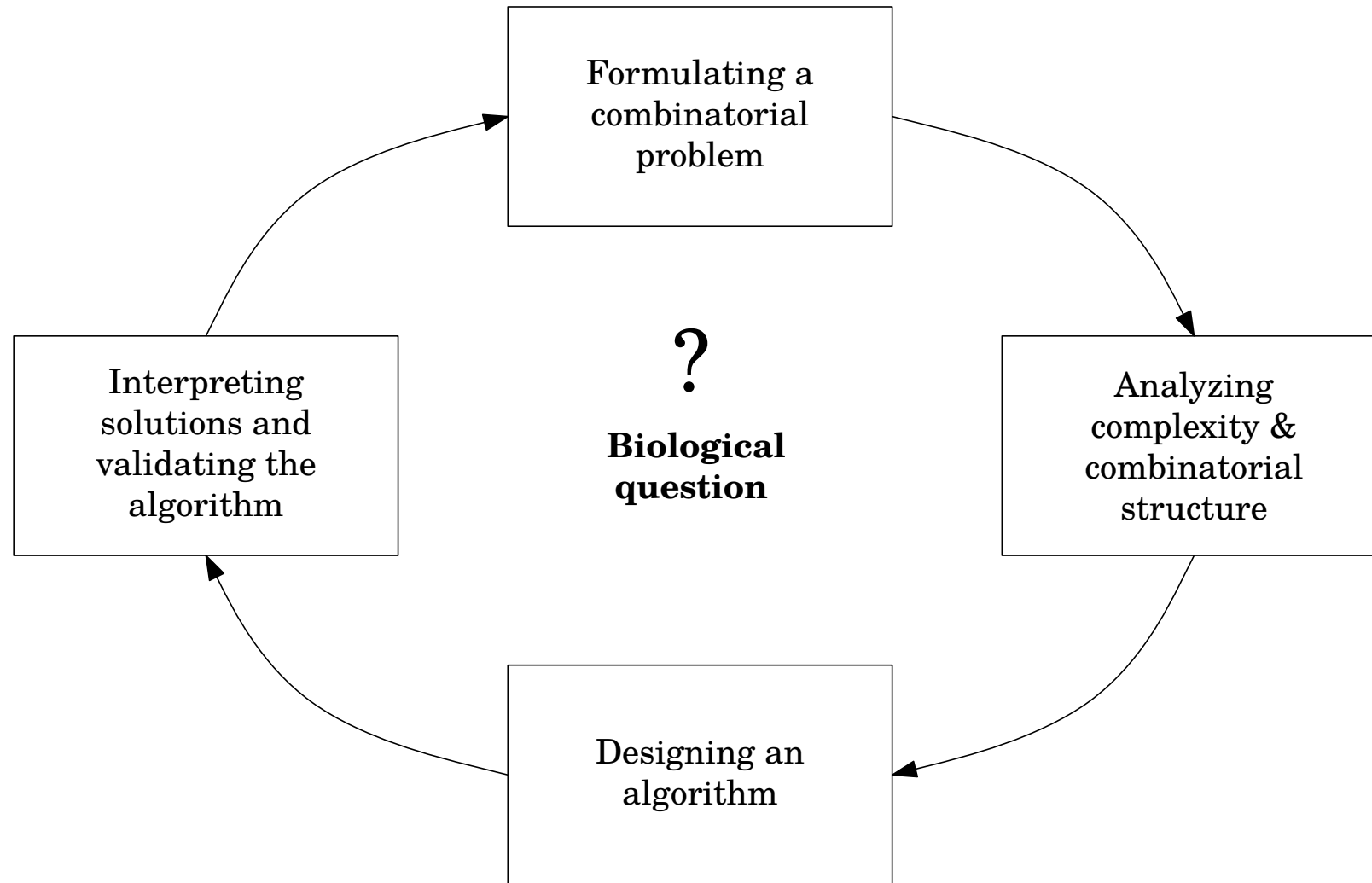
Translating a biological problem into a computational biology

Perfect Phylogeny Mixture

Restricted PP Tree T

Assumptions:
- Infinite sites assumption:
  a character changes state once
- Error-free data

1-1 ↕ **Equivalent**

n mutations

n mutations

m samples

$$\begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix} \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

Frequency Matrix F

$=$

clones

m samples

$$\begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix} \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix}$$

Mixture Matrix U

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

clones

Restricted PP Matrix B

Rows of U are proportions:
$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem
[Estabrook, 1971]
[Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
Given F, find U and B such that $F = U \, B$

Previous Work

Restricted PP
Tree T

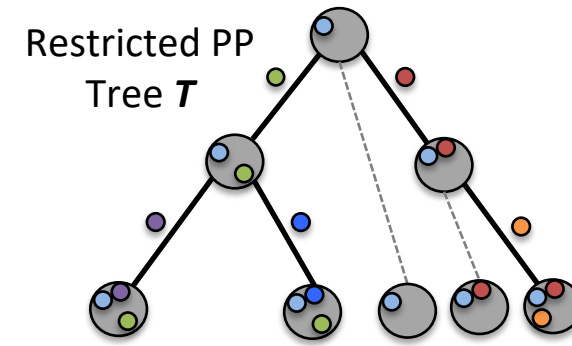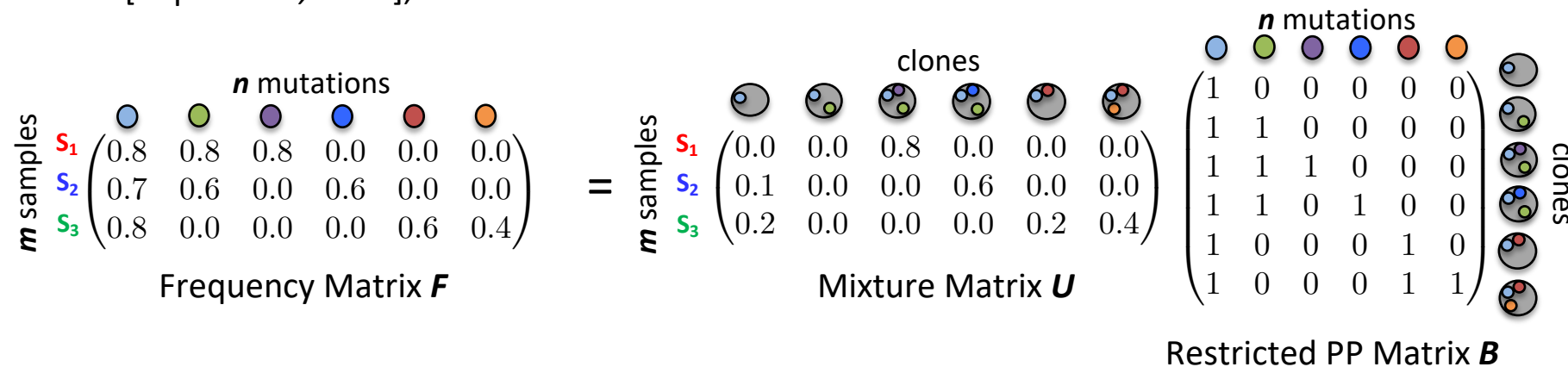**Variant of PPM:**
TrAp [Strino *et al.,* 2013], PhyloSub [Jiao *et al.*, 2014]
CITUP [Malikic *et al.*, 2015], BitPhylogeny [Yuan *et al.*, 2015]
LICHeE [Popic e*t al.,* 2015], …

1-1 $\updownarrow$ **Equivalent**

n mutations

$$
\begin{array}{c}
\text{m samples} \\
\end{array}
\begin{array}{c}
S_1 \\
S_2 \\
S_3 \\
\end{array}
\left(
\begin{array}{cccccc}
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \\
\end{array}
\right)
\quad = \quad
\begin{array}{c}
\text{m samples} \\
\end{array}
\begin{array}{c}
S_1 \\
S_2 \\
S_3 \\
\end{array}
\left(
\begin{array}{cccccc}
0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \\
\end{array}
\right)
\left(
\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 \\
\end{array}
\right)
$$

clones

n mutations

Frequency Matrix F

Mixture Matrix U

Restricted PP Matrix B

clones

Rows of U are proportions:
$u_{pj} \geq 0$ and $\displaystyle\sum_j u_{pj} \leq 1$

Perfect Phylogeny Theorem
[Estabrook, 1971]
[Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
Given F, find U and B such that $F = U\,B$

Combinatorial Characterization

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
Given F, find U and B such that $F = U B$

- **Combinatorial characterization** involves investigating what (optimal) solutions look like
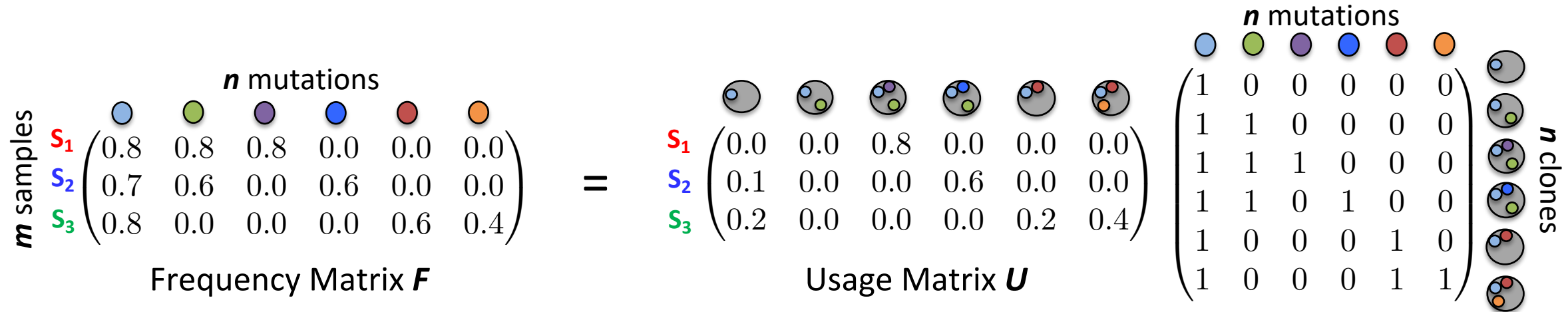- This starts by asking questions!

Given F and T (or B), is there a usage matrix U?
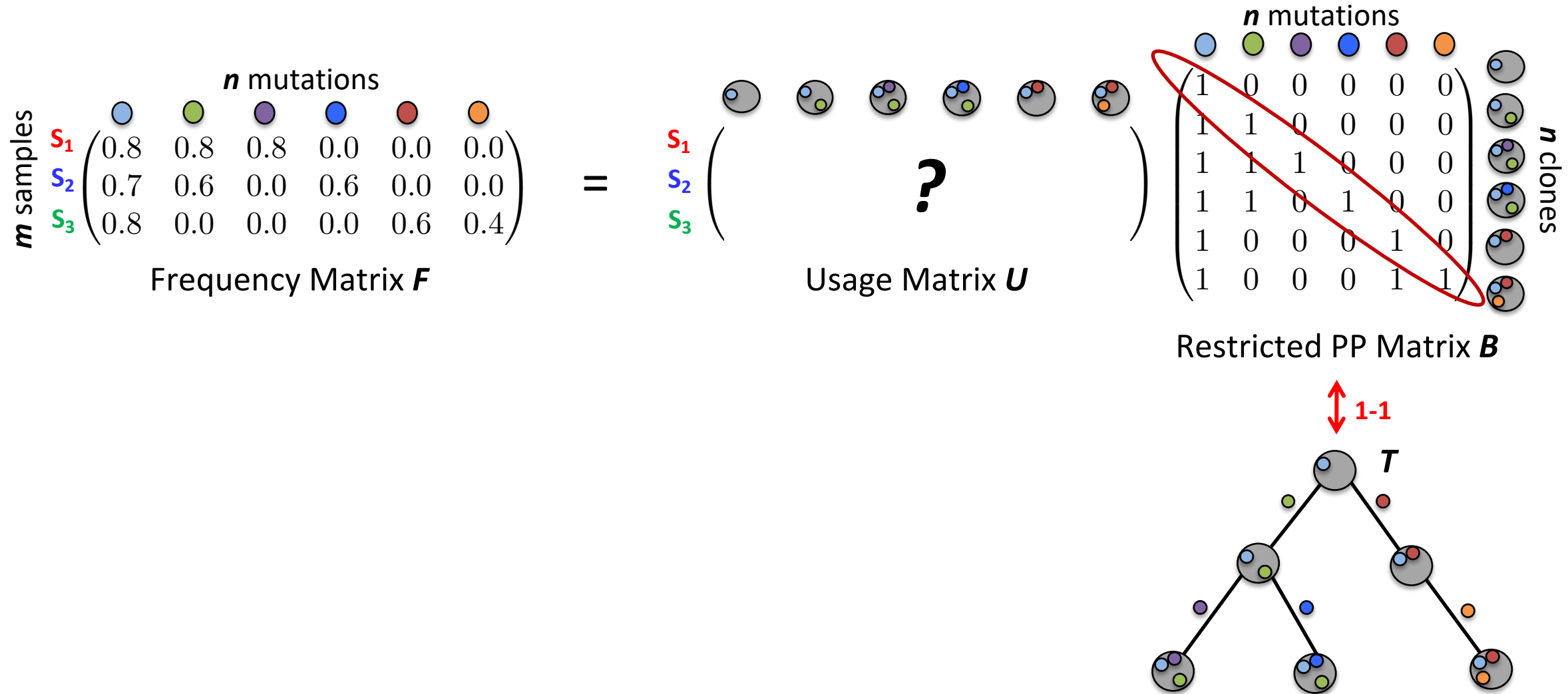
PPM: Given F, find U and B such that $F = U\,B$

Frequency Matrix F

$$S_1 \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

Usage Matrix U

Restricted PP Matrix B

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

T

1-1

13

Given *F* and *T* (or *B*), is there a usage matrix *U*?

PPM: Given *F*, find *U* and *B* such that *F* = *U B*

Frequency Matrix *F*

$$S_1 \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

n mutations

m samples

=

Usage Matrix *U*

$$S_1 \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix}$$

n mutations

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

n clones

Restricted PP Matrix *B*

Lemma: *B* is invertible

➡ Given *F* and *B*, *U* is **unique**: *U* = 2 *F B⁻¹*

Lemma:

$$u_{pj} = \boxed{f_{pj}} - \boxed{\sum_{k \text{ child of } j} f_{pk}}$$
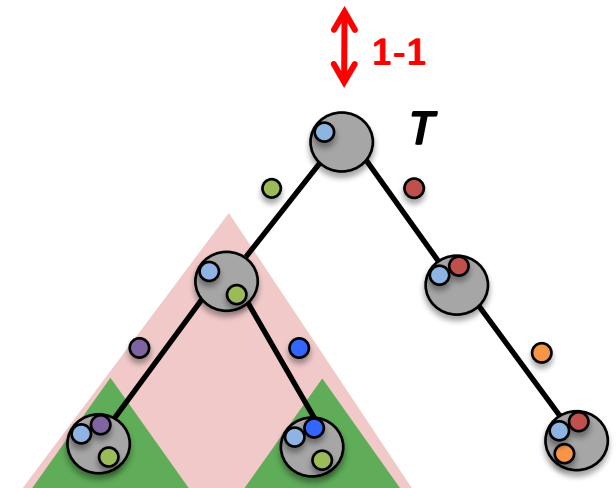
1-1

T

Given *F* and *T* (or *B*), is there a usage matrix *U*?

PPM: Given *F*, find *U* and *B* such that *F* = *U B*

$$
\begin{pmatrix} \boxed{0.15 \quad 0.1 \quad 0.06} \\ 0.2 \quad 0.05 \quad 0.04 \end{pmatrix}
\quad = \quad
\begin{pmatrix} \boxed{-0.01} \quad 0.1 \quad 0.06 \\ 0.11 \quad 0.05 \quad 0.04 \end{pmatrix}
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}
$$

Frequency Matrix *F*   Usage Matrix *U*

Restricted PP Matrix *B*

1-1

T

Lemma:

$$
u_{pj} = \boxed{f_{pj}} - \boxed{\sum_{k \text{ child of } j} f_{pk}}
$$

0.15

0.10   0.06
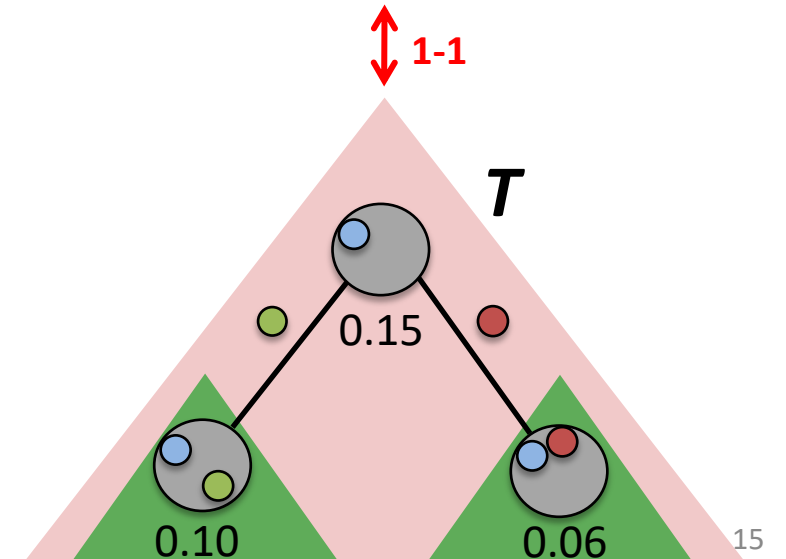
Combinatorial Characterization of Solutions

Lemma:

$$u_{pj} = \boxed{f_{pj}} - \boxed{\sum_{k \text{ child of } j} f_{pk}}$$

Lemma (Sum Condition):

Given F and T, for all samples p and mutations j, $\boxed{f_{pj}} \geq \boxed{\sum_{k \text{ child of } j} f_{pk}}$

necessary sufficient

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

T

Combinatorial Characterization of Solutions

Lemma (Sum Condition):
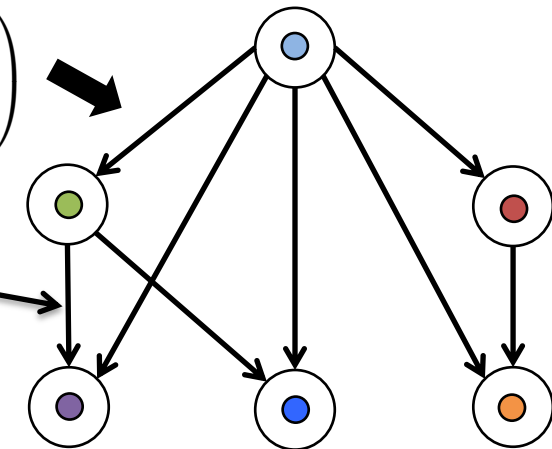Given F and T, for all samples p and mutations j,

$$f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$$

necessary *sufficient*

Lemma (Ancestry Condition):
Given F and T, for all samples p and mutations k child of j,

$$f_{pj} \geq f_{pk}$$

necessary

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

Combinatorial Characterization of Solutions

Lemma (Sum Condition):
Given F and T, for all samples p and mutations j, $$f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$$

necessary
sufficient

Lemma (Ancestry Condition):
Given F and T, for all samples p and mutations k child of j, $$f_{pj} \geq f_{pk}$$

necessary

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

potential parental relationship

Ancestry Graph $G = (V, A)$

Ancestry graph $G = (V, A)$; given F
- Vertex for every mutation
- Edge $(j, k) \in A$ iff $f_{pj} \geq f_{pk}$ for all samples p
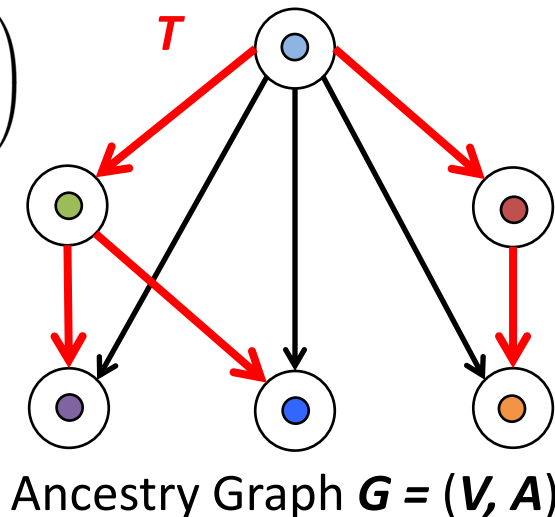
Combinatorial Characterization of Solutions

Lemma (Sum Condition):
Given **F** and **T**, for all samples p and mutations j,
$$f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$$

necessary
sufficient

Lemma (Ancestry Condition):
Given **F** and **T**, for all samples p and mutations k child of j,
$$f_{pj} \geq f_{pk}$$

necessary

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

T

Ancestry Graph **G = (V, A)**

Ancestry graph G = (V, A); given **F**
- Vertex for every mutation
- Edge $(j, k) \in A$ iff $f_{pj} \geq f_{pk}$ for all samples p

Theorem 2:
PPM is NP-complete

Theorem 1:
T is a solution to the PPM if and only if **T** is a spanning tree of **G** satisfying the Sum Condition

Solving the PPM problem: ILP formulation

$$\max \sum_{(v_j, v_k) \in A'} x_{jk}$$

Find the largest set of edges in G

$$\text{s.t.} \sum_{v_j \in \delta^+(v_r)} x_{rj} = 1$$

Exactly one root node

G = (V, A)

$$x_{kl} \leq \sum_{v_j \in \delta^-(v_k)} x_{jk} \qquad \forall (v_k, v_l) \in A \quad \textbf{Connectivity}$$

$$\sum_{v_j \in \delta^-(v_k)} x_{jk} \leq 1 \qquad \forall v_k \in V \quad \textbf{Tree}$$

$$\sum_{v_j \in \delta^-(v_k)} f_{pk} x_{jk} \geq \sum_{v_l \in \delta^+(v_k)} f_{pl} x_{kl} \quad \forall p \in [m],\, v_k \in V \quad \textbf{Sum condition}$$

$$x_{jk} \in \{0, 1\} \qquad \forall (v_j, v_k) \in A'$$
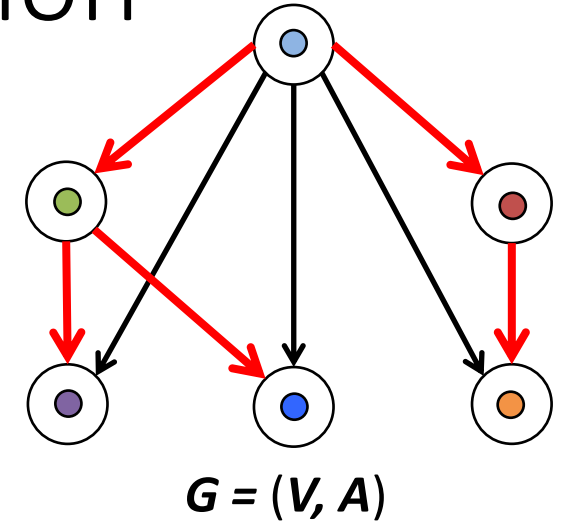
Non-uniqueness of Solutions to PPM

$$F = \begin{pmatrix} 1 & 0 & 0 & 0.06 & 0 \\ 1 & 0.75 & 0.33 & 0 & 0.25 \end{pmatrix}$$

Question 0: Reconstruct all solutions?

Non-uniqueness of Solutions to PPM

T_1
- 1: $f_{1,1} = 1.00$, $f_{2,1} = 1.00$
- 2: $f_{1,2} = 0.00$, $f_{2,2} = 0.75$
- 4: $f_{1,4} = 0.06$, $f_{2,4} = 0.00$
- 3: $f_{1,3} = 0.00$, $f_{2,3} = 0.33$
- 5: $f_{1,5} = 0.00$, $f_{2,5} = 0.25$

T_2
- 1: $f_{1,1} = 1.00$, $f_{2,1} = 1.00$
- 2: $f_{1,2} = 0.00$, $f_{2,2} = 0.75$
- 4: $f_{1,4} = 0.06$, $f_{2,4} = 0.00$
- 3: $f_{1,3} = 0.00$, $f_{2,3} = 0.33$
- 5: $f_{1,5} = 0.00$, $f_{2,5} = 0.25$

T_3
- 1: $f_{1,1} = 1.00$, $f_{2,1} = 1.00$
- 2: $f_{1,2} = 0.00$, $f_{2,2} = 0.75$
- 4: $f_{1,4} = 0.06$, $f_{2,4} = 0.00$
- 3: $f_{1,3} = 0.00$, $f_{2,3} = 0.33$
- 5: $f_{1,5} = 0.00$, $f_{2,5} = 0.25$
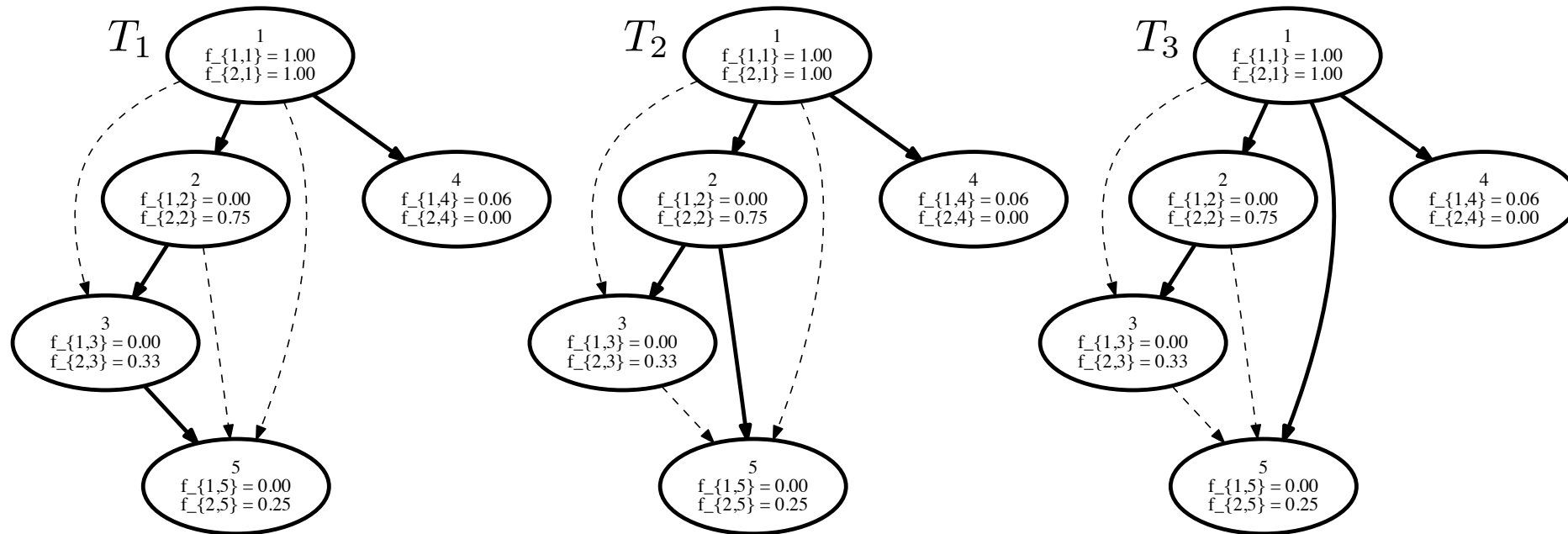
$$F = \begin{pmatrix} 1 & 0 & 0 & 0.06 & 0 \\ 1 & 0.75 & 0.33 & 0 & 0.25 \end{pmatrix}$$

Question 1: Can we determine the number of solutions?

Question 2: Can sample solutions uniformly at random?

Question 3: How to enumerate solutions?

Recall: Different Types of Problems!

Problem Π with instance X and solution set $\Pi(X)$:

- Decision problem:
  - Is $\Pi(X) = \emptyset$?
- Optimization problem:
  - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.
- Counting problem:
  - Compute $|\Pi(X)|$.
- Sampling problem:
  - Sample uniformly from $\Pi(X)$.
- Enumeration problem:
  - Enumerate all solutions in $\Pi(X)$

Algorithms:

Set of instructions for solving problem.

- Exact
- Heuristic

On the Complexity of #PPM (new results)

Question 1: Can we determine the number of solutions?

Question 2: Can sample solutions uniformly at random?

#PPM: Given F, count the number of pairs (U, B) composed of mixture matrix U and perfect phylogeny matrix B such that $F = U B$

On the Complexity of #PPM (new results)

Question 1: Can we determine the number of solutions?

Question 2: Can sample solutions uniformly at random?

#PPM: Given F, count the number of pairs (U, B) composed of mixture matrix U and perfect phylogeny matrix B such that $F = U\,B$

#P is the complexity class of counting problems whose decision problems are in NP

Every problem in #P can be reduced in polynomial time to any problem in #P-complete, preserving cardinalities

On the Complexity of #PPM (new results)

Question 1: Can we determine the number of solutions?

Question 2: Can sample solutions uniformly at random?

#PPM: Given **F**, count the number of pairs **(U, B)** composed of mixture matrix **U** and perfect phylogeny matrix **B** such that **F = U B**

#P is the complexity class of counting problems whose decision problems are in NP

Every problem in #P can be reduced in polynomial time to any problem in #P-complete, preserving cardinalities

Theorem: #PPM is #P-complete

Theorem: There is no FPRAS for #PPM

Theorem: There is no FPAUS for PPM
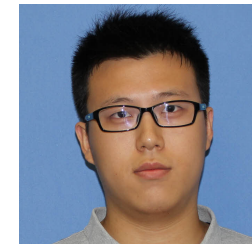
Yuanyuan Qi

Outline

Background and theory:

• Perfect Phylogeny Mixture (PPM) problem

• Combinatorial characterization of solutions

• #PPM: exact counting and uniform sampling

Simulation results:

• What contributes to non-uniqueness?

• How to reduce non-uniqueness?

• How does non-uniqueness affect current methods?
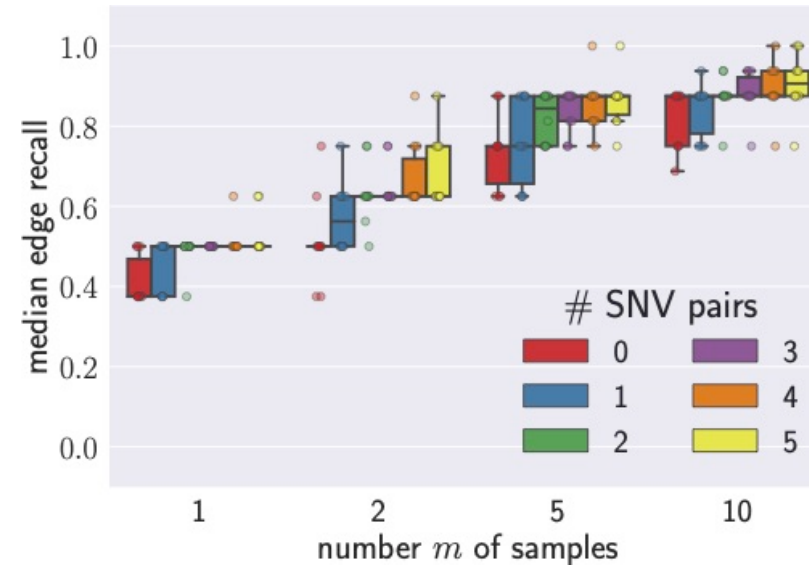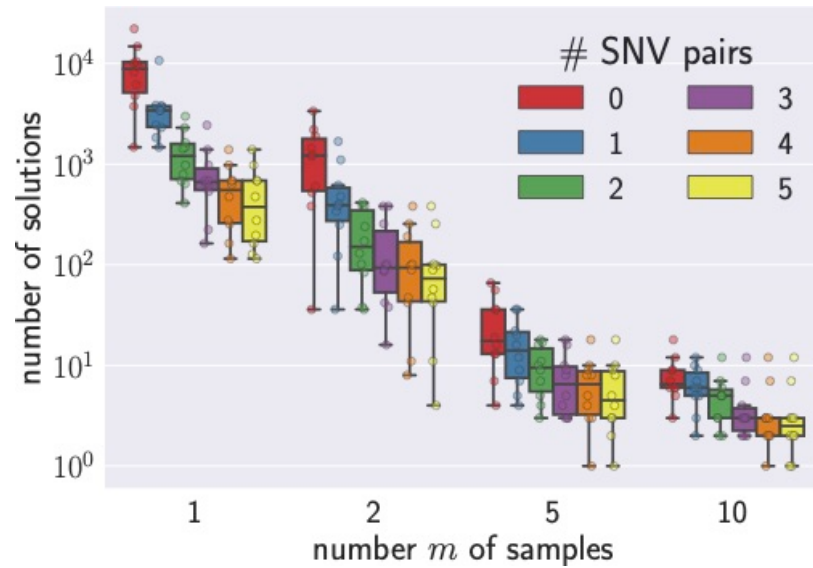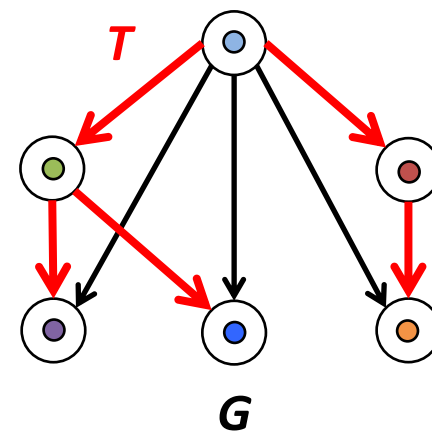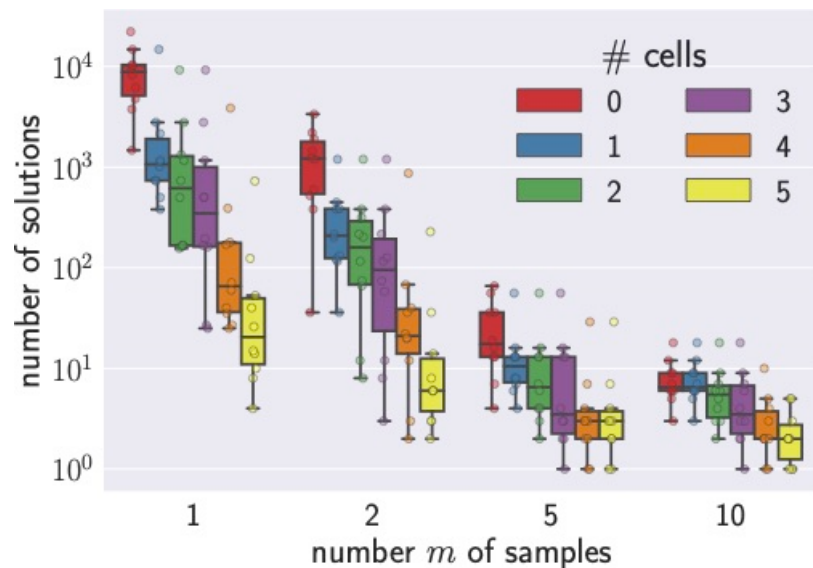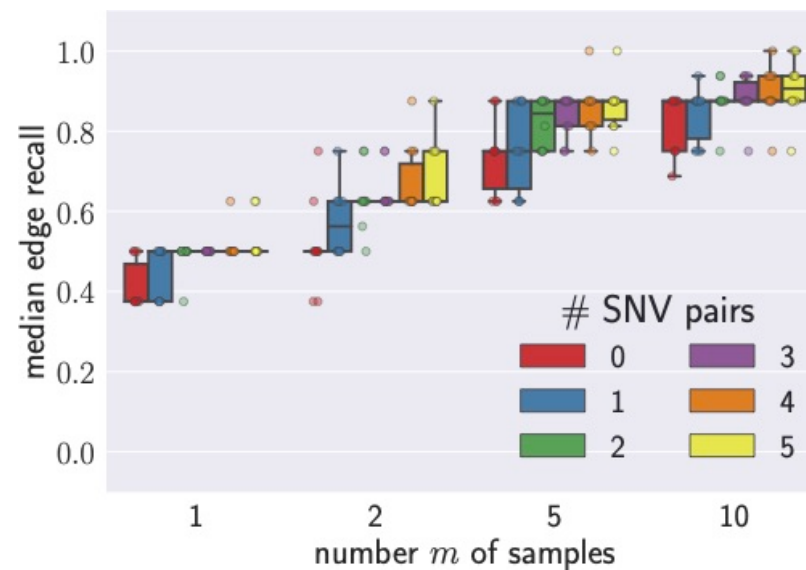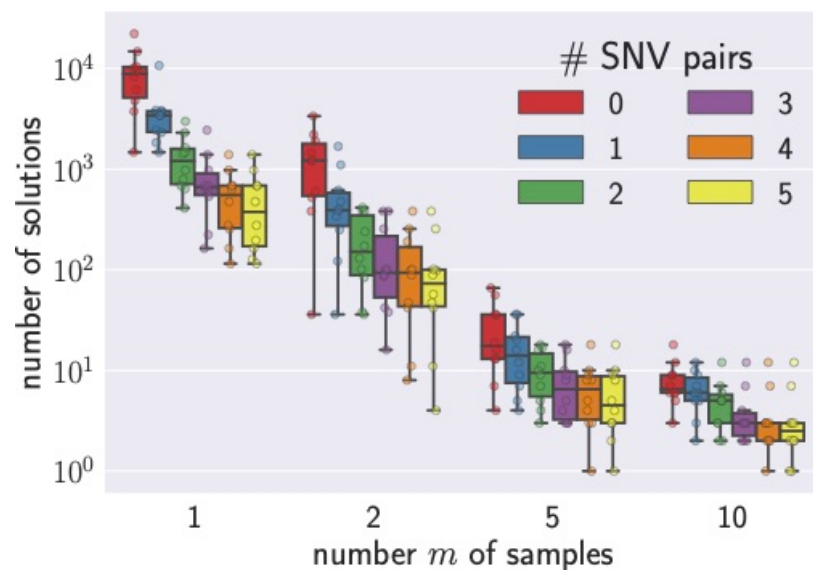
Dikshant Pradhan

What Contributes to Non-uniqueness?
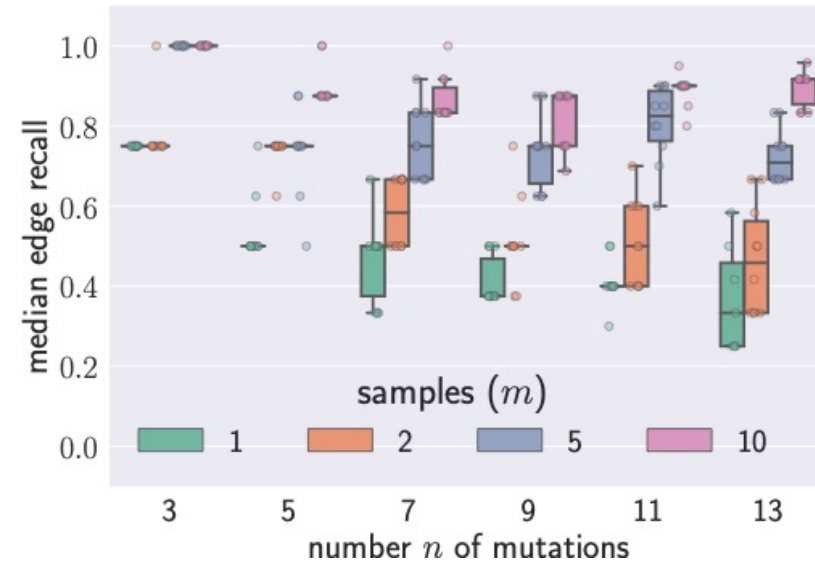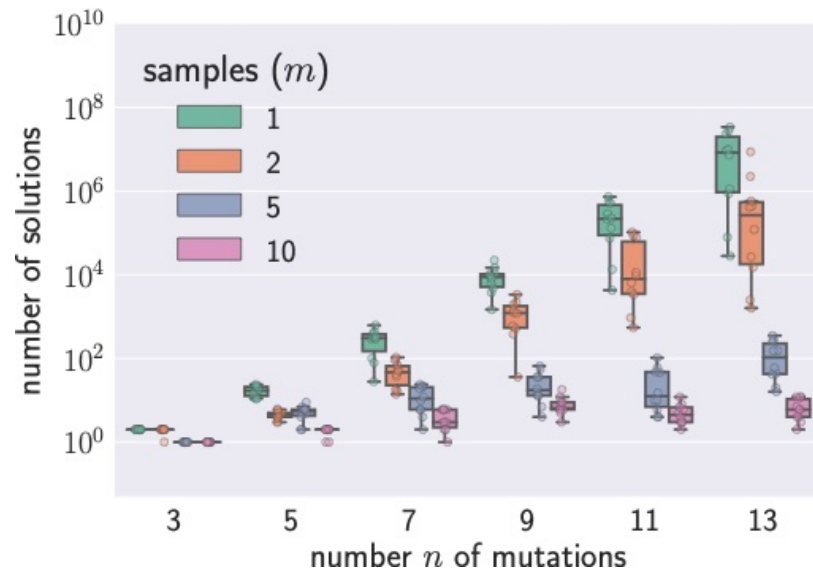
What Contributes to Non-uniqueness?
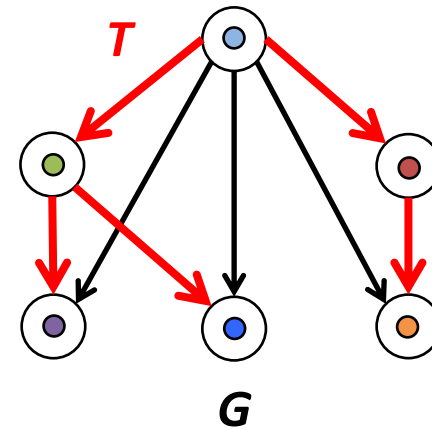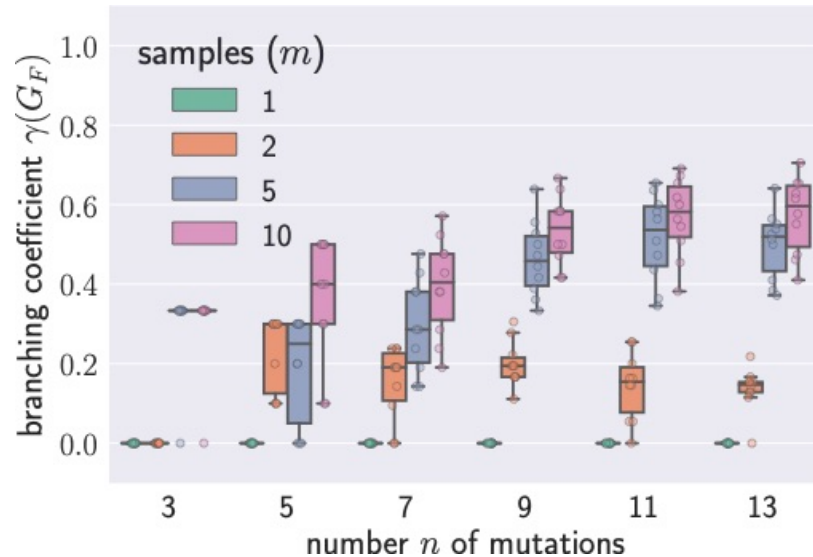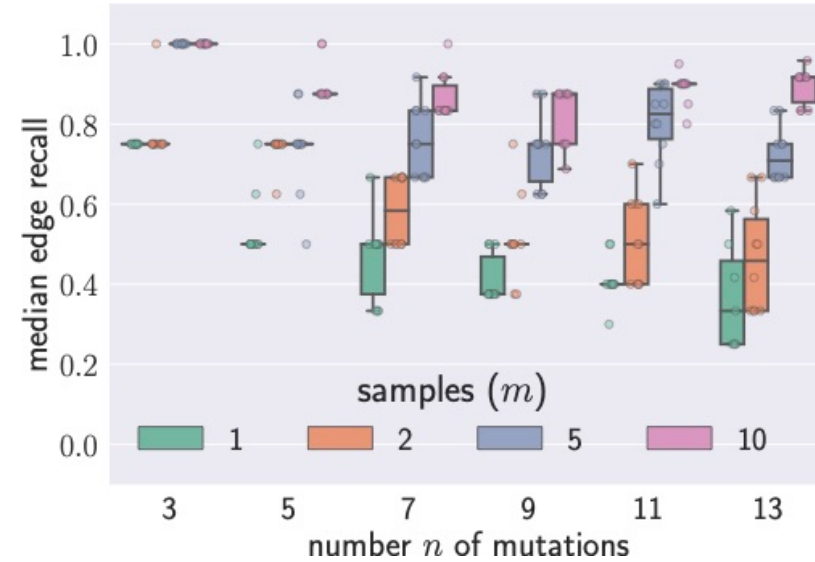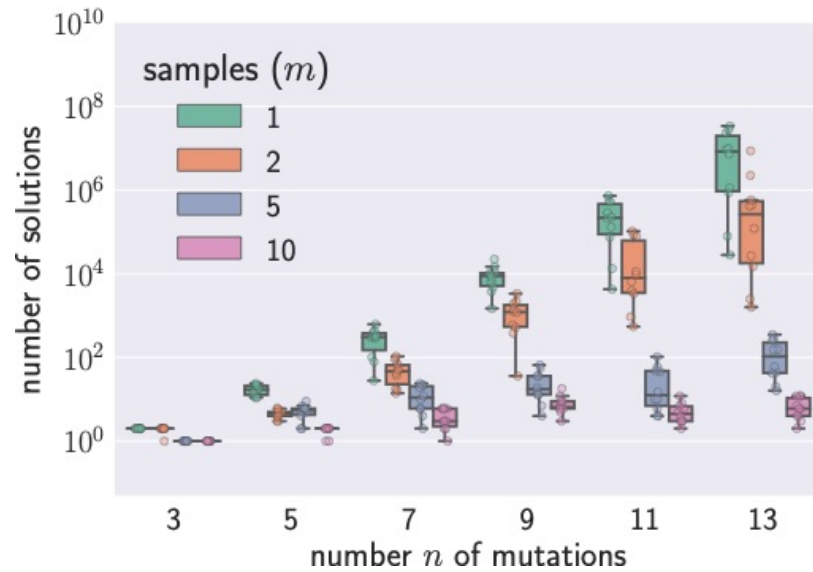
An Upper Bound for Number of Solutions
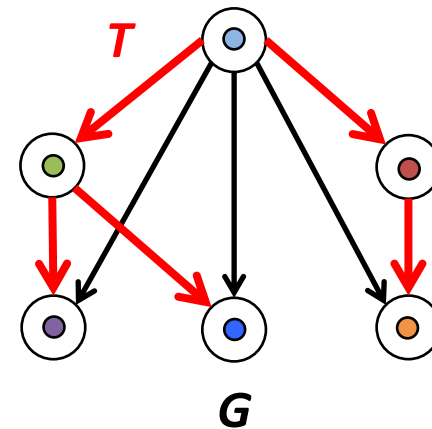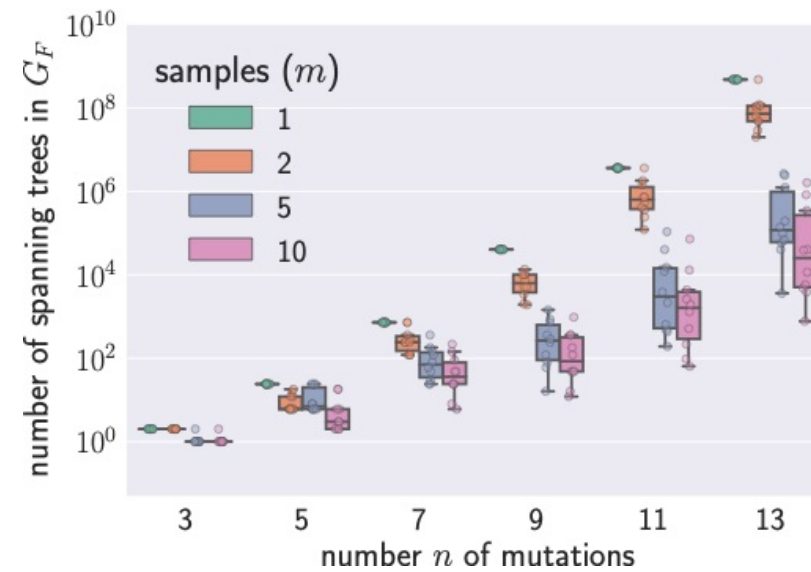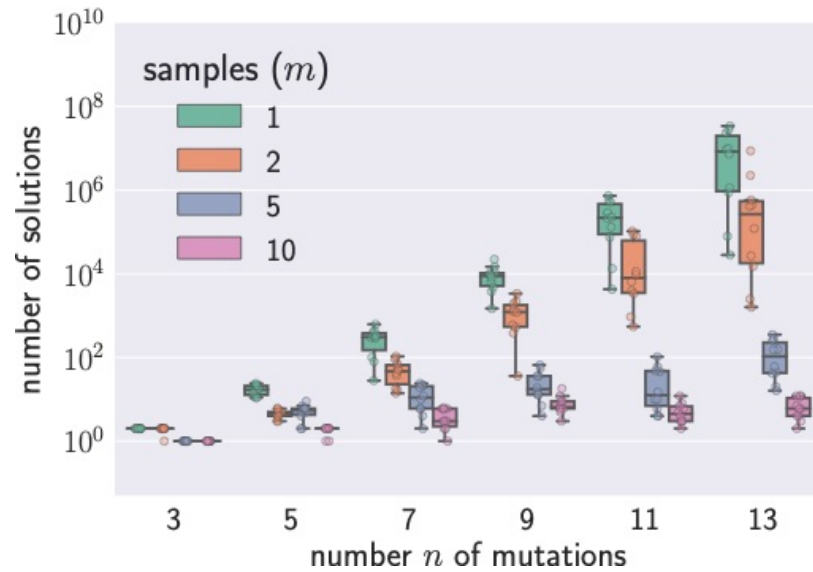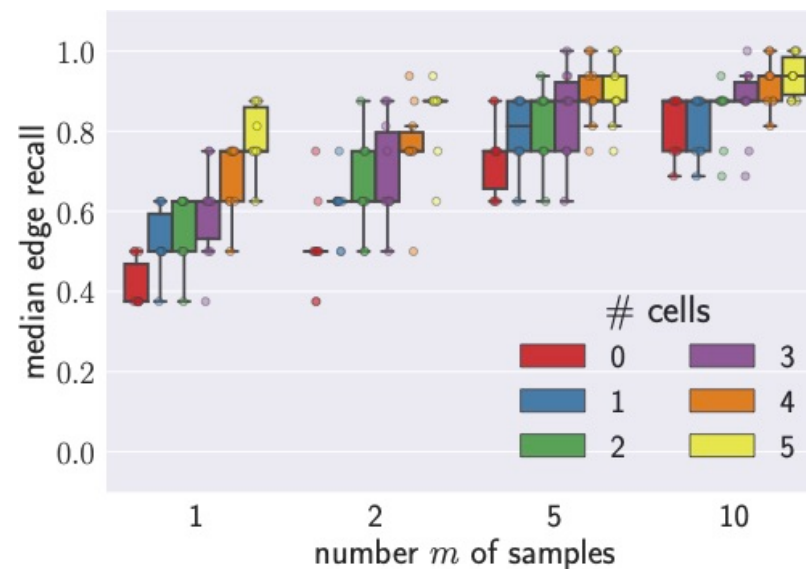
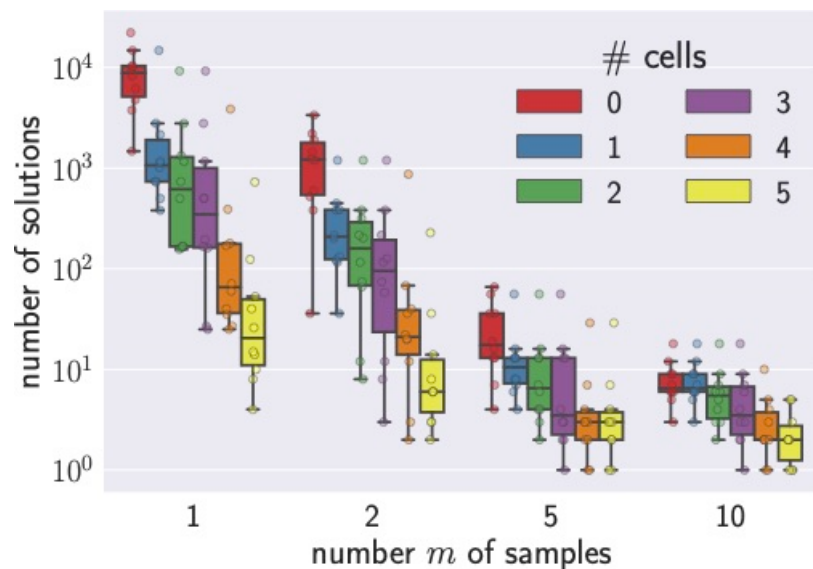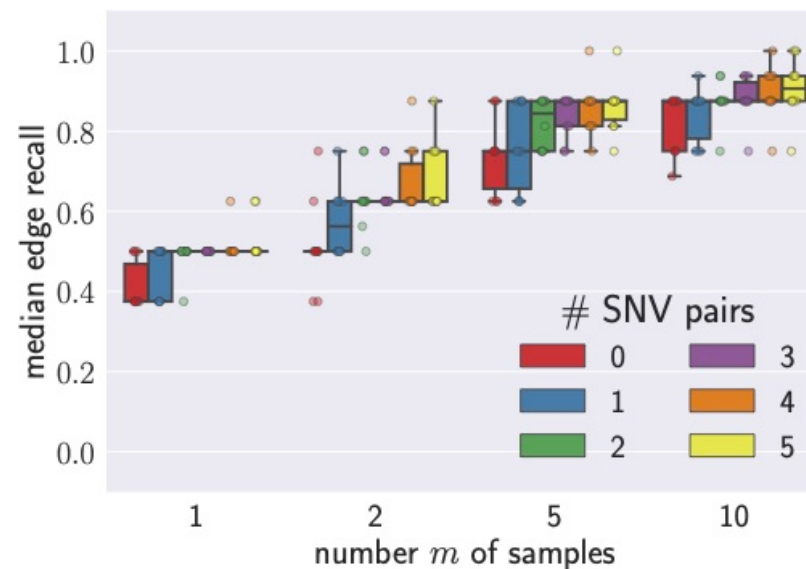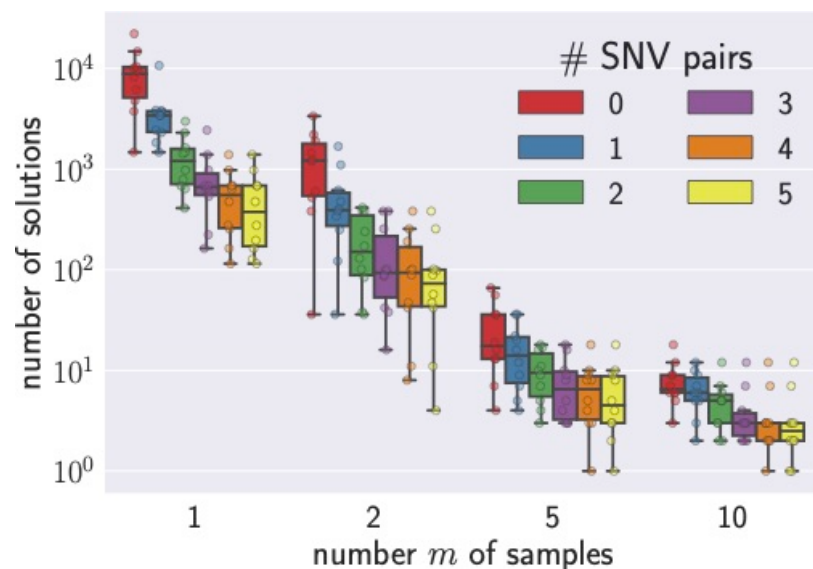# An Upper Bound for Number of Solutions
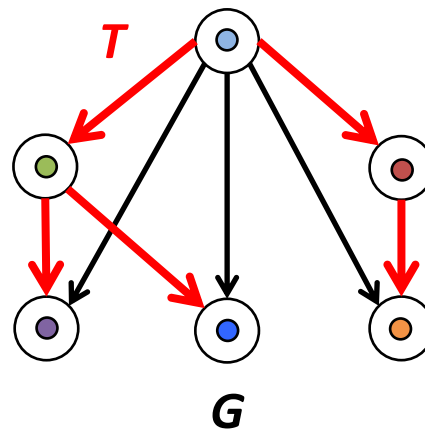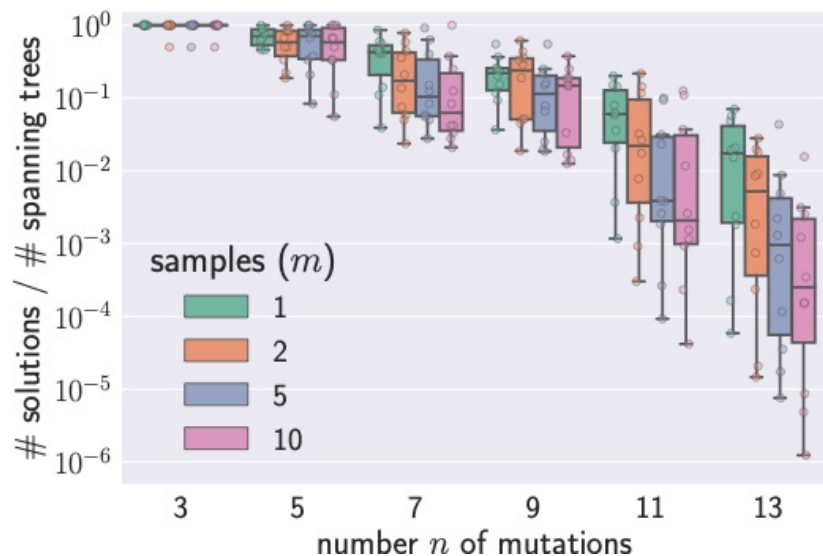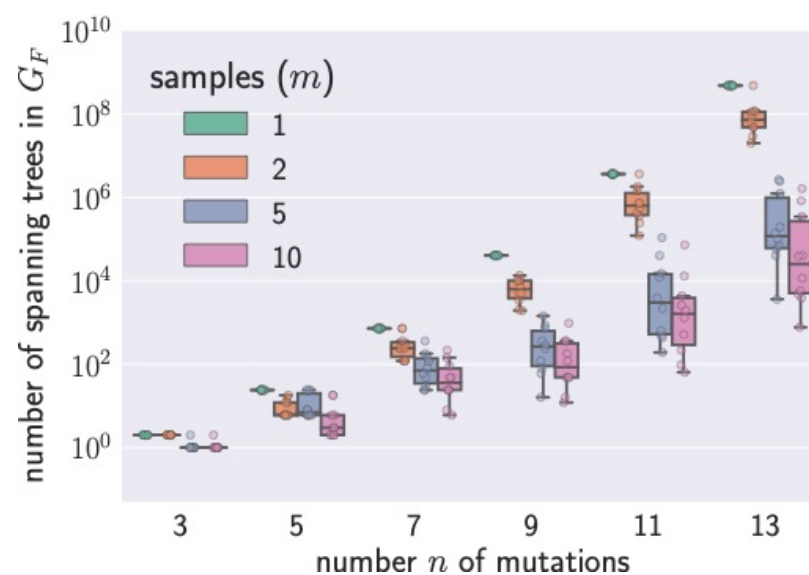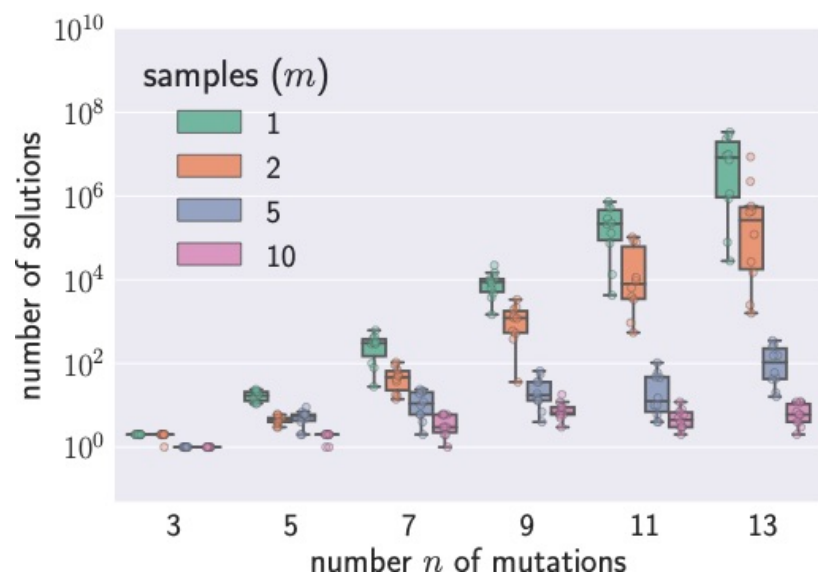
How to Reduce Non-Uniqueness?
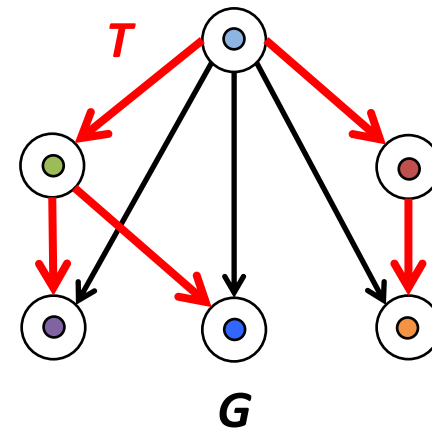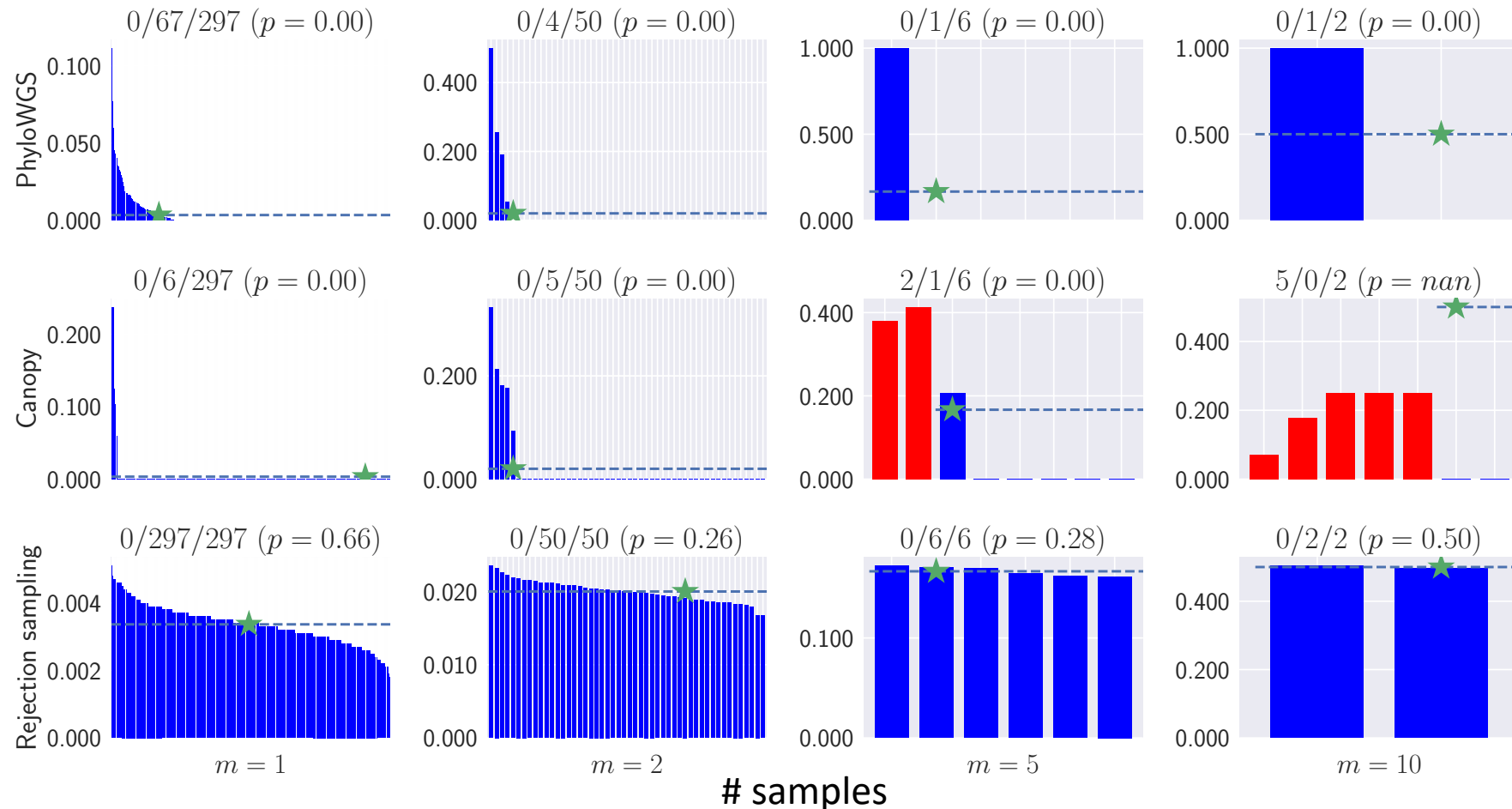
How to Reduce Non-Uniqueness?

How to Reduce Non-Uniqueness?
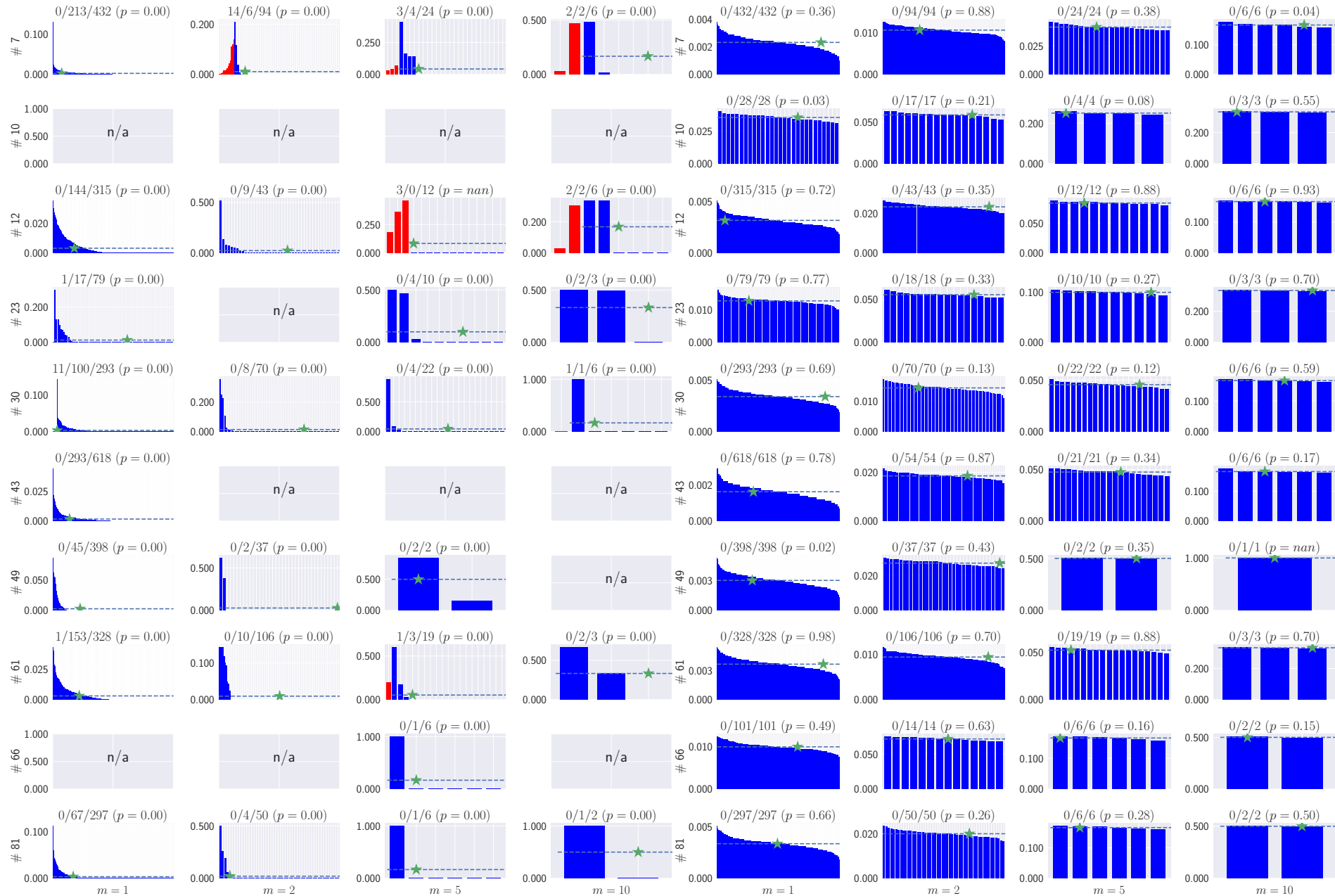
How Does Non-uniqueness affect Methods?

Two current MCMC methods using default parameters:
- PhyloWGS, Deshwar et al., Genom. Biol., 2015 [10,000 samples]
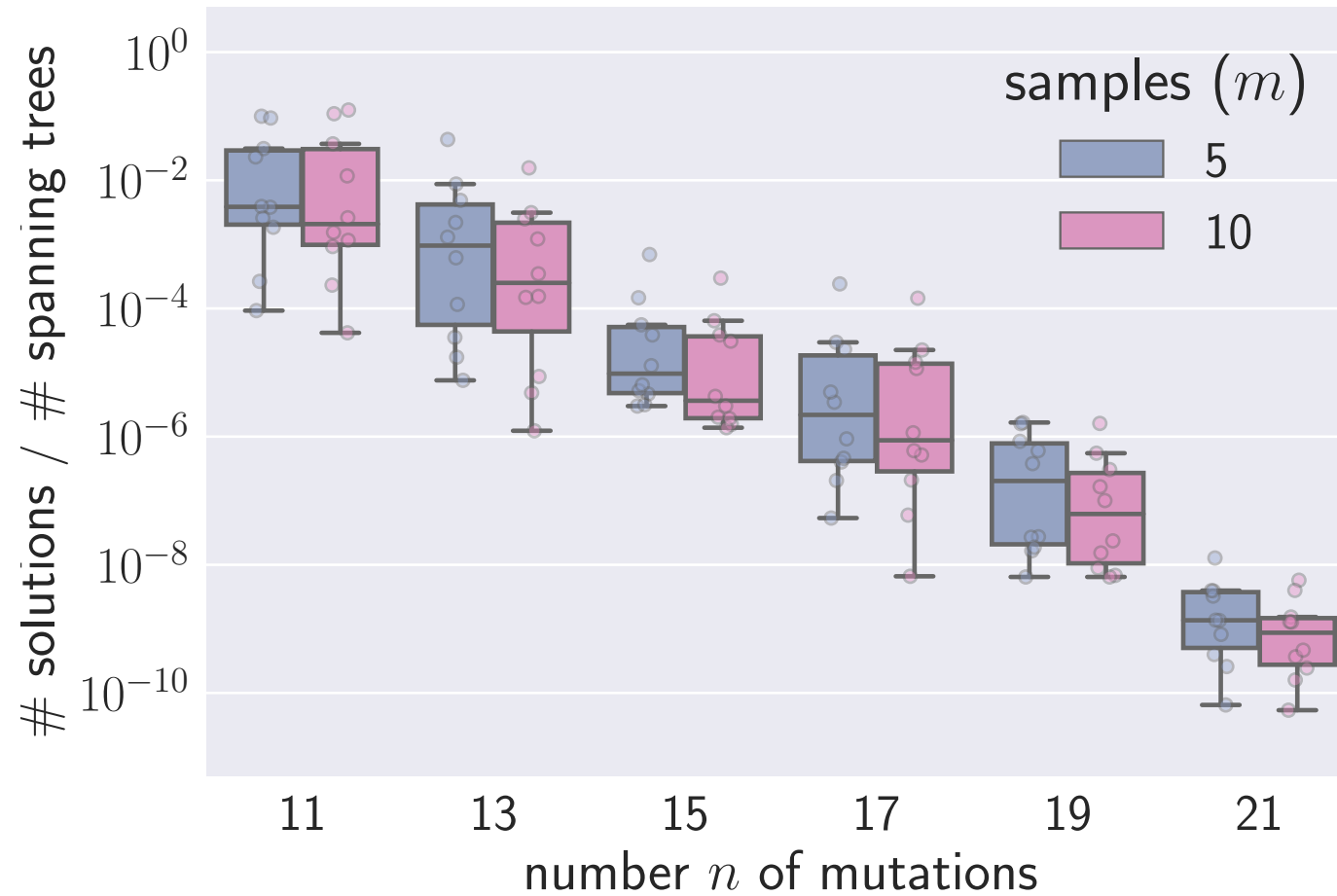- Canopy, Jiang et al., PNAS, 2016   [~300 samples]

PhyloWGS — Rejection Sampling
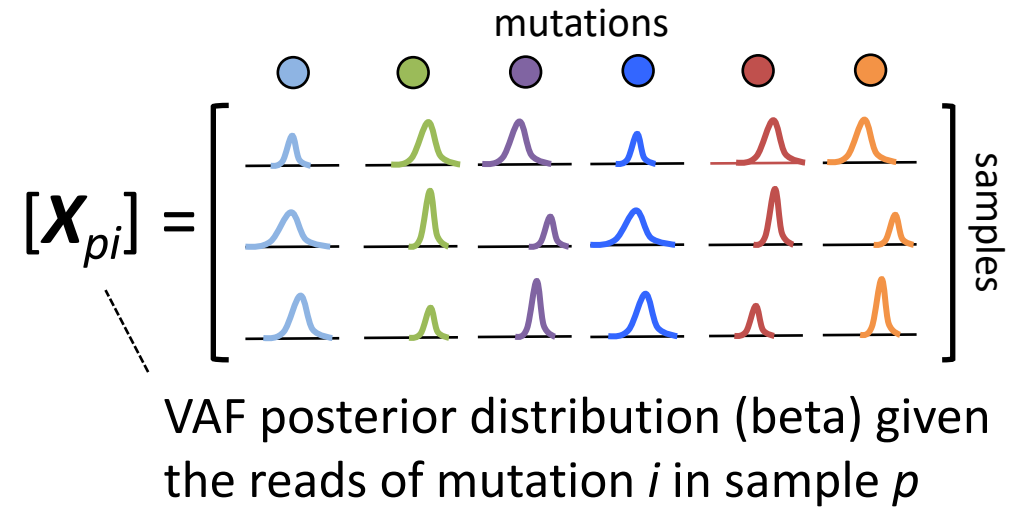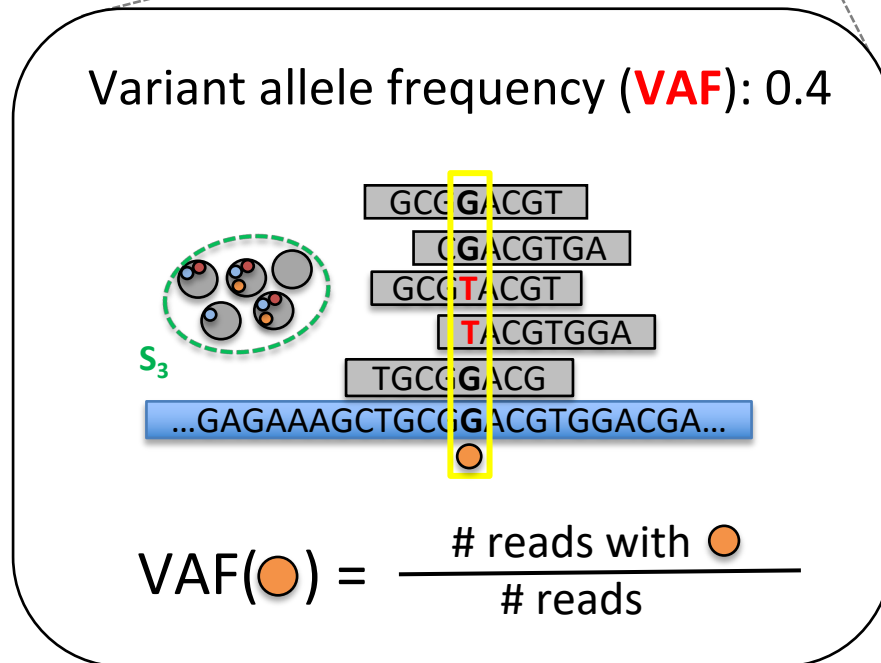
Rejection Sampling Does Not Scale

Probabilistic Model for Noisy Measurements

mutations

$$\boldsymbol{F} = [f_{pi}] = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & \boxed{0.4} \end{pmatrix} \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix}$$ samples

VAF of mutation i in sample p

Uncertainty due to:
(i) sequencing errors
(ii) mapping errors
(iii) sampling

Variant allele frequency (**VAF**): 0.4

GCC**G**ACGT
C**G**ACGTGA
GCC**T**ACGT
TACGTGGA
TGCC**G**ACG
…GAGAAAGCTGCC**G**ACGTGGACGA…

S_3

$$\text{VAF}(\bullet) = \frac{\text{\# reads with } \bullet}{\text{\# reads}}$$

mutations

$$[\boldsymbol{X}_{pi}] = \begin{bmatrix} & & & & & \\ & & & & & \\ & & & & & \end{bmatrix}$$ samples

VAF posterior distribution (beta) given the reads of mutation i in sample p

Probabilistic Model for Noisy Measurements

mutations

$$\mathbf{F}^{-} = [f^{-}_{pi}] = \begin{pmatrix} 0.75 & 0.78 & 0.77 & 0.0 & 0.0 & 0.0 \\ 0.55 & 0.43 & 0.0 & 0.54 & 0.0 & 0.0 \\ 0.56 & 0.0 & 0.0 & 0.0 & 0.57 & 0.34 \end{pmatrix} \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix}$$

samples

Consider (1 - α) confidence intervals:

mutations

$$[\mathbf{X}_{pi}] = $$

samples

Interval PPM (I-PPM)

Given $\mathbf{F}^{-}$ and $\mathbf{F}^{+}$, find $\mathbf{F}$, $\mathbf{U}$ and $\mathbf{B}$ such that $\mathbf{F} = \mathbf{U}\,\mathbf{B}$

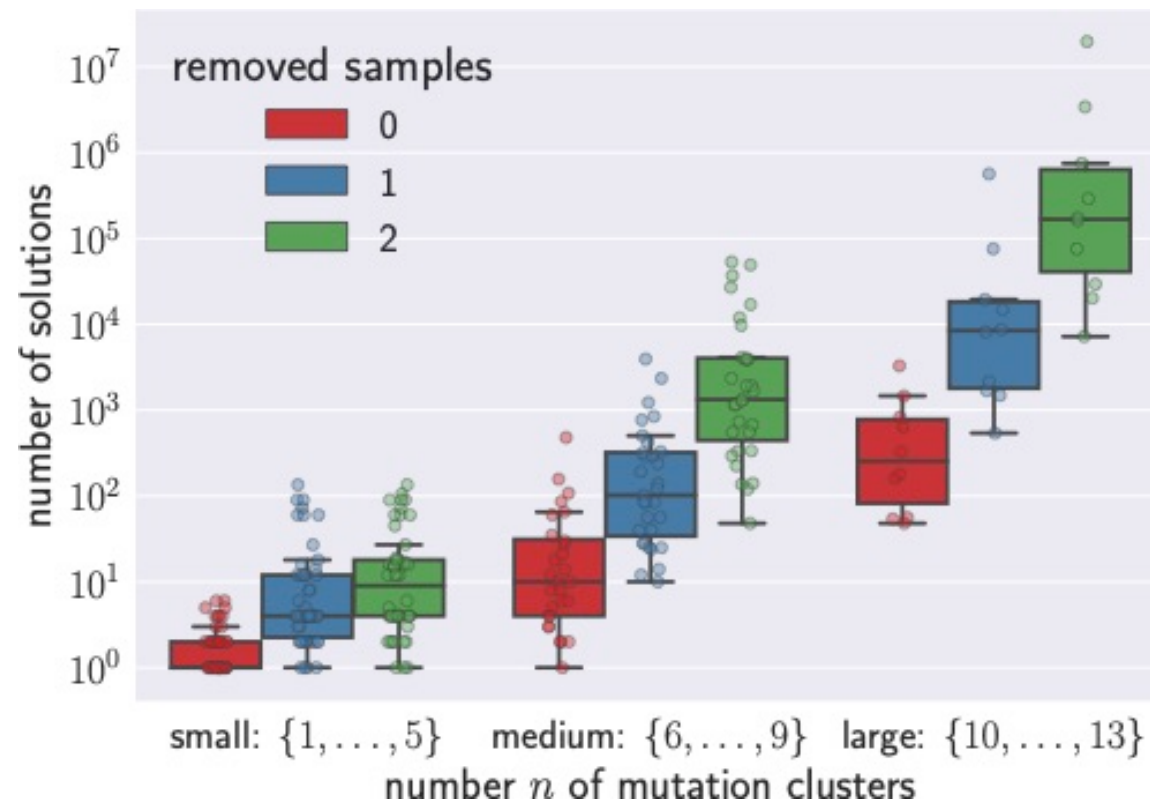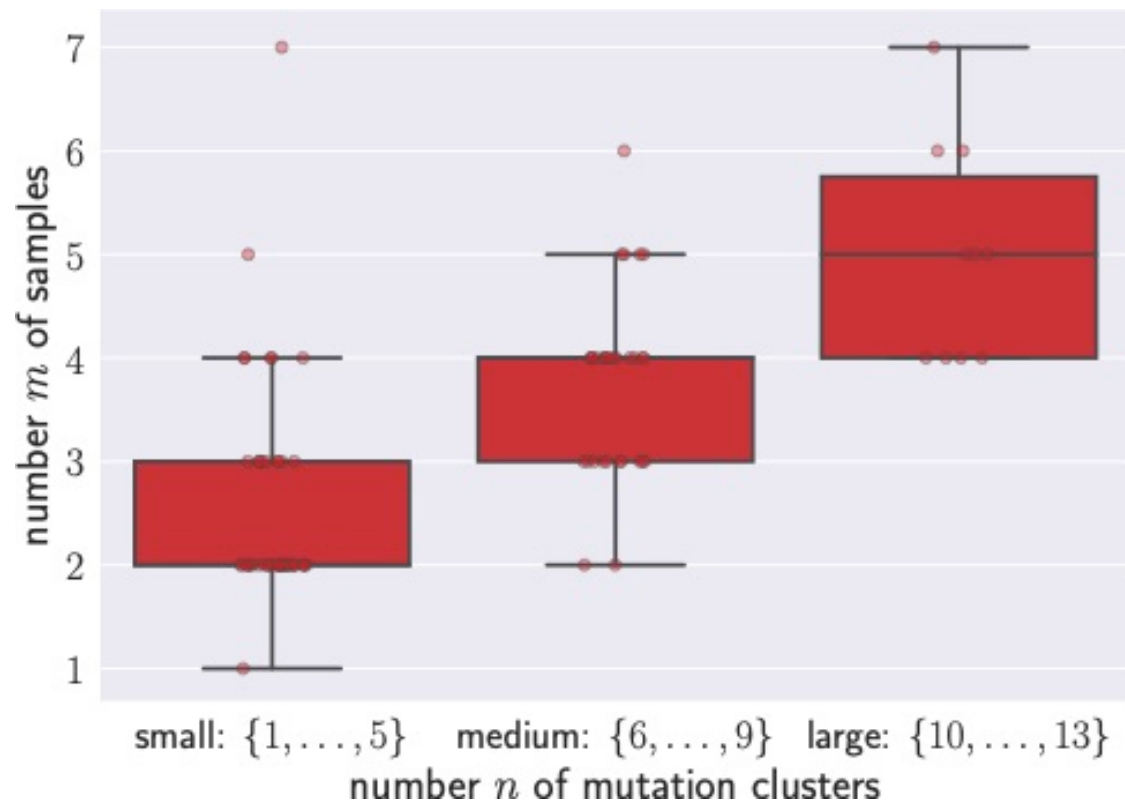and $f^{-}_{pi} \leq f_{pi} \leq f^{+}_{pi}$ for all samples p and mutations i

VAF posterior distribution (beta) given the reads of mutation i in sample p

mutations

$$\mathbf{F}^{+} = [f^{+}_{pi}] = \begin{pmatrix} 0.9 & 0.85 & 0.87 & 0.05 & 0.0 & 0.0 \\ 0.75 & 0.65 & 0.05 & 0.68 & 0.0 & 0.0 \\ 0.83 & 0.0 & 0.04 & 0.0 & 0.67 & 0.48 \end{pmatrix} \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix}$$
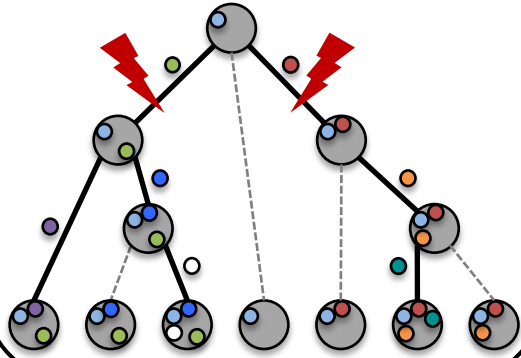
samples

Real Data

- Cohort of 100 lung cancers [Jamal-Hanjani, NEJM 2017]
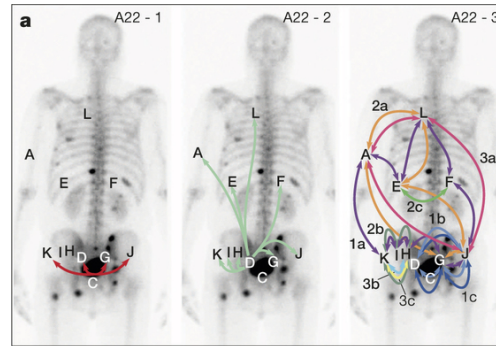- 90% confidence intervals
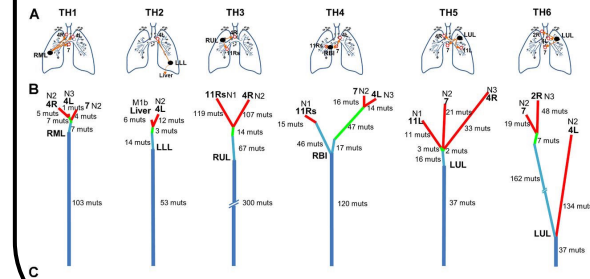
Challenges

Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients

Downstream analyses in cancer genomics **critically rely** on accurate tumor phylogeny inference

Challenge I
Novel algorithms that sample uniformly at random
from the space of PPM solutions

Challenge II
Algorithms to accurately summarizing solution space
(consensus trees)

Conclusion

Background and theory:

- Perfect Phylogeny Mixture (PPM) problem

- Combinatorial characterization of solutions

- #PPM: exact counting and uniform sampling

Simulation results:

- What contributes to non-uniqueness?

- How to reduce non-uniqueness?

- How does non-uniqueness affect current methods?

Summary of Lectures 1, 2 and 3

- DNA, RNA and proteins are sequences
  - Central dogma of molecular biology: DNA -> RNA -> protein

- Problem != algorithm

- Key challenge in computational biology is translating a biological problem into a computational problem

- Cancer is a genetic disease caused by somatic mutations

- Inter-tumor heterogeneity and intra-tumor heterogeneity:
  - *Not only is every tumor different, but so is every tumor cell…*

- **Non-uniqueness of solutions in phylogeny reconstruction from bulk DNA samples**