Implications of Non-uniqueness of Solutions in Cancer Phylogenetics

Mohammed El-Kebir

University of Illinois at Urbana Champaign, Department of Computer Science

October, 2019
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]

Founder tumor cell with somatic mutation: (e.g. BRAF V600E)
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]

Clonal expansion
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]

New clones
Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]

Intra-Tumor Heterogeneity
Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]

Intra-Tumor Heterogeneity

Phylogenetic Tree $T$

**Question:** Why are tumor phylogenies important?
Phylogenies are Key to Understanding Cancer

Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients
Phylogenies are Key to Understanding Cancer

Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients

These downstream analyses critically rely on accurate tumor phylogeny inference
Phylogenies are Key to Understanding Cancer

Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients

These downstream analyses **critically rely** on accurate tumor phylogeny inference

**Key challenge in phylogenetics:**
Accurate phylogeny inference from data at present time
Additional Challenge in Cancer Phylogenetics

- tumor
- normal

- human reference genome (3*10^9 bp)
- aligned read (100 bp)
Additional Challenge in Cancer Phylogenetics

- Tumor
- Normal

- Human reference genome (3*10^9 bp)
- Aligned read (100 bp)
- Single nucleotide variant (SNV)
Additional challenge in cancer phylogenetics:
Phylogeny inference from mixed bulk samples at present time
Outline

1. Background and theory: [RECOMB-CG 2018, AMOB 2019]
   - Perfect Phylogeny Mixture (PPM) problem
   - #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018, AMOB 2019]
   - What contributes to non-uniqueness?
   - How to reduce non-uniqueness?
   - How does non-uniqueness affect current methods?

3. Almost uniform sampling: [To be submitted]
   - Reducing PPM to SATISFIABILITY

4. Summarizing solution space: [ISMB/ECCB 2019]
   - Multiple consensus tree problem

5. Applications
   - Mutational signature dynamics [PSB 2020]
Sequencing and Tumor Phylogeny Inference

Variant allele frequency (VAF): 0.8

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

Frequency Matrix $F$

$\begin{pmatrix}
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4
\end{pmatrix}$
Sequencing and Tumor Phylogeny Inference

Tumor Phylogeny Inference: Given frequencies $F$, find phylogeny $T$ and proportions $U$
Perfect Phylogeny Mixture

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

\[
\begin{align*}
\text{Frequency Matrix } F &= \begin{pmatrix}
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \\
\end{pmatrix} \\
\text{Mixture Matrix } U &= \begin{pmatrix}
0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \\
\end{pmatrix} \\
\text{Restricted PP Matrix } B &= \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 \\
\end{pmatrix}
\end{align*}
\]

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

**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 et al., 2015], ...

\[
\begin{bmatrix}
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \\
\end{bmatrix}
\quad \begin{bmatrix}
0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \\
\end{bmatrix}
\]

- \( n \) mutations
- \( m \) samples
- \( S_1 \)
- \( S_2 \)
- \( S_3 \)
- \( u_{pj} \geq 0 \) and \( \sum_j u_{pj} \leq 1 \)

**Perfect Phylogeny Mixture:** [El-Kebir*, Oesper* et al., 2015]
Given \( F \), find \( U \) and \( B \) such that \( F = UB \)
Given $F$ and $T$ (or $B$), is there a usage matrix $U$?

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

**Lemma:** $B$ is invertible

Given $F$ and $B$, $U$ is unique: $U = F B^{-1}$

**Lemma:**

$$u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk}$$
Given $F$ and $T$ (or $B$), is there a usage matrix $U$?

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

\[
\begin{pmatrix}
0.15 & 0.1 & 0.06 \\
0.2 & 0.05 & 0.04
\end{pmatrix}
= \begin{pmatrix}
-0.01 & 0.1 & 0.06 \\
0.11 & 0.05 & 0.04
\end{pmatrix} \begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1
\end{pmatrix}
\]

**Frequency Matrix $F$**

**Usage Matrix $U$**

**Restricted PP Matrix $B$**

\[
\begin{align*}
u_{pj} &= f_{pj} - \sum_{k \text{ child of } j} f_{pk} \\
&= 0.15 - (0.2 \times 0.10 + 0.05 \times 0.06)
\end{align*}
\]
Lemma:

\[ u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk} \]

Lemma (Sum Condition):

Given \( F \) and \( T \), for all samples \( p \) and mutations \( j \),

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

PPM: Given \( F \), find \( U \) and \( B \) such that \( F = UB \)
Combinatorial Characterization of Solutions

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

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

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

\[ F = \begin{pmatrix}
0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \\
\end{pmatrix} \]
**Combinatorial Characterization of Solutions**

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

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

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

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 even if \( m = 2 \)

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

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

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

**Combinatorial Characterization of Solutions**

**PPM:** Given \( F \), find \( U \) and \( B \) such that \( F = UB \)
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 1:** Can we determine the number of solutions?

**Question 2:** Can we sample solutions uniformly at random?
Intermezzo: Problems/Algorithms in CS

Problem $\Pi$ with instance/input $X$ and feasible 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 $\Pi(X)$.
Intermezzo: Problems/Algorithms in CS

<table>
<thead>
<tr>
<th>Problem ( \Pi ) with instance/input ( X ) and feasible solution set ( \Pi(X) ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Decision problem:</td>
</tr>
<tr>
<td>• Is ( \Pi(X) = \emptyset )?</td>
</tr>
<tr>
<td>• Optimization problem:</td>
</tr>
<tr>
<td>• Find ( y^* \in \Pi(X) ) s.t. ( f(y^*) ) is optimum.</td>
</tr>
<tr>
<td>• Counting problem:</td>
</tr>
<tr>
<td>• Compute (</td>
</tr>
<tr>
<td>• Sampling problem:</td>
</tr>
<tr>
<td>• Sample uniformly from ( \Pi(X) ).</td>
</tr>
<tr>
<td>• Enumeration problem:</td>
</tr>
<tr>
<td>• Enumerate all solutions in ( \Pi(X) ).</td>
</tr>
</tbody>
</table>

**Algorithm:**
Set of instructions for solving problem.
• Exact
• Heuristic

**Running time:** How does the number of steps scale as a function of \( |X| \)?
Intermezzo: Problems/Algorithms in CS

Problem \( \Pi \) with instance/input \( X \) and feasible 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 \( \Pi(X) \).

**Algorithm:**
Set of instructions for solving problem.
- Exact
- Heuristic

**Running time:** How does the number of steps scale as a function of \( |X| \)?

Problem != algorithm

Some problems do not admit efficient algorithms
Non-uniqueness of Solutions to PPM

<table>
<thead>
<tr>
<th></th>
<th>( f_{1,1} )</th>
<th>( f_{2,1} )</th>
<th>( f_{1,2} )</th>
<th>( f_{2,2} )</th>
<th>( f_{1,3} )</th>
<th>( f_{2,3} )</th>
<th>( f_{1,4} )</th>
<th>( f_{2,4} )</th>
<th>( f_{1,5} )</th>
<th>( f_{2,5} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.75</td>
<td>0.00</td>
<td>0.33</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.25</td>
</tr>
<tr>
<td>( T_2 )</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.75</td>
<td>0.00</td>
<td>0.33</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.25</td>
</tr>
<tr>
<td>( T_3 )</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.75</td>
<td>0.00</td>
<td>0.33</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.25</td>
</tr>
</tbody>
</table>

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

**Question 1:** Can we determine the number of solutions?

**Question 2:** Can we sample solutions uniformly at random?

**Counting problem**

**Sampling problem**
On the Complexity of #PPM

**Question 1:** Can we determine the number of solutions?

**Question 2:** Can we 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

**Question 1:** Can we determine the number of solutions?

**Question 2:** Can we 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

**Question 1:** Can we determine the number of solutions?

**Question 2:** Can we 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

1. Background and theory: [RECOMB-CG 2018]
   • Perfect Phylogeny Mixture (PPM) problem
   • #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018]
   • What contributes to non-uniqueness?
   • How to reduce non-uniqueness?
   • How does non-uniqueness affect current methods?

3. Almost uniform sampling: [To be submitted]
   • Reducing PPM to SATISFIABILITY

4. Summarizing solution space: [ISMB/ECCB 2019]
   • Multiple consensus tree problem

5. Applications
   • Mutational signature dynamics [PSB 2020]
What Contributes to Non-uniqueness?
What Contributes to Non-unicateness?
How to Reduce Non-Uniqueness?
How to Reduce Non-Uniqueness?
How to Reduce Non-Uniqueness?
An Upper Bound for Number of Solutions

number of solutions

$10^{10}$

$10^8$

$10^6$

$10^4$

$10^2$

$10^0$

samples ($m$)

- 1
- 2
- 5
- 10

number $n$ of mutations

$10^{10}$

$10^8$

$10^6$

$10^4$

$10^2$

$10^0$

number of spanning trees in $G_F$

samples ($m$)

- 1
- 2
- 5
- 10

$G$

$T$
An Upper Bound for Number of Solutions

samples (m)

number of solutions

$10^0$ $10^1$ $10^2$ $10^3$ $10^4$ $10^5$ $10^6$ $10^7$ $10^8$ $10^9$ $10^{10}$

number $n$ of mutations

3 5 7 9 11 13

number of spanning trees in $G_F$

$10^0$ $10^1$ $10^2$ $10^3$ $10^4$ $10^5$ $10^6$ $10^7$ $10^8$ $10^9$ $10^{10}$

3 5 7 9 11 13

# solutions / # spanning trees

$10^{-6}$ $10^{-5}$ $10^{-4}$ $10^{-3}$ $10^{-2}$ $10^{-1}$ $10^0$

# solutions / # spanning trees

$10^{-6}$ $10^{-5}$ $10^{-4}$ $10^{-3}$ $10^{-2}$ $10^{-1}$ $10^0$

number $n$ of mutations

3 5 7 9 11 13

samples (m)

1 2 5 10

T

G
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]
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]
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]
Rejection Sampling Does Not Scale

number $n$ of mutations

$\#$ solutions / $\#$ spanning trees

samples $(m)$

5

10
Outline

1. Background and theory: [RECOMB-CG 2018]
   • Perfect Phylogeny Mixture (PPM) problem
   • #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018]
   • What contributes to non-uniqueness?
   • How to reduce non-uniqueness?
   • How does non-uniqueness affect current methods?

3. Almost uniform sampling: [To be submitted]
   • Reducing PPM to SATISFIABILITY

4. Summarizing solution space: [ISMB/ECCB 2019]
   • Multiple consensus tree problem

5. Applications
   • Mutational signature dynamics [PSB 2020]
Result

Generating 10,000 trees from the solution space almost uniformly at random.

*Figure 4*: Out method samples uniformly: (a) the same simulated instance (m=1, n=7, 297 solution trees); (b) the ranged version of PPM of CRUK0062 (90% confidence interval, m=7, n=15, 160 solutions).
Method

Unigen

An almost uniform sampler for SAT (satisfiability) problem$^4$.

Reduce PPM to SAT.

---

Boolean Satisfiability

• Goal: find a model that satisfies a propositional formula.

\[ a \land (\neg b \lor c) \rightarrow a \mapsto T, \; b \mapsto F, \; c \mapsto F \]

\[ a \land b \land (\neg b \lor \neg a) \rightarrow \text{unsatisfiable} \]

• The original NP-complete problem.
  – “As hard as any other problem in NP.”

Spanning tree of ancestry graph

Consider simple case:

- F is precise
- No repeated columns ($G_F$ is a DAG)

Variables:

- $r_p$: $r_p = 1$ iff $p$ is the root.
- $e_{p,q}$: $e_{p,q} = 1$ iff $p$ is the parent of $q$.

Restrictions:

- only one of $r_p$ for all $p \in [n]$ is true.
- for all $p \in [n]$, only one of $e_{q,p} \forall q \in [n]$ and $r_p$ is true.
Clauses for restriction "only one of $v_1, ..., v_n$ is true":

- $\bigvee_{p \in [n]} v_p$
- $\neg v_p \lor \neg v_q$, for all $p \neq q$
Method

Sum Condition

- Discretize the frequencies (multiply by \((2^N - 1)\)).
- Represent the integers with a vector of binary variables.

We need to represent addition and comparison in CNF form (clauses).
Method

Addition

half_adder\((a, b, r, c)\):

- \( r = a \oplus b \):
  - \( r \lor \neg a \lor b \)
  - \( r \lor a \lor \neg b \)
  - \( \neg r \lor a \lor b \)
  - \( \neg r \lor \neg a \lor \neg b \)
- \( c = a \lor b \):
  - \( \neg c \lor a \lor b \)
  - \( c \lor \neg a \)
  - \( c \lor \neg b \)

full_adder\((a, b, last_c, r, c)\)

Comparison

leq\((a,b)\): check the carry of \( \sim a + b + 1 \).
Method

More complicated case: given $F^-$ and $F^+$ instead of a single $F$.

Approach

- Cycle prevention: $\text{depth}(p) = \text{depth}(q) + 1$ if $e_{q,p} = 1$.
- $F$ is not determined: use the minimum frequency possible.
Result

5m \cdot |E| \cdot N \text{ variables} \\
17m \cdot |E| \cdot N \text{ clauses} \hspace{1cm} \text{3 seconds}

6m \cdot |E| \cdot N \text{ variables} \\
20m \cdot |E| \cdot N \text{ clauses} \hspace{1cm} \text{4 minutes}

Figure 4: Out method samples uniformly:
(a) the same simulated instance (m=1, n=7, 297 solution trees);
(b) the ranged version of PPM of CRUK0062 (90% confidence interval, m=7, n=15, 160 solutions)
Outline

1. **Background and theory:** [RECOMB-CG 2018]
   - Perfect Phylogeny Mixture (PPM) problem
   - #PPM: exact counting and uniform sampling

2. **Simulation results:** [RECOMB-CG 2018]
   - What contributes to non-uniqueness?
   - How to reduce non-uniqueness?
   - How does non-uniqueness affect current methods?

3. **Almost uniform sampling:** [To be submitted]
   - Reducing PPM to SATISFIABILITY

4. **Summarizing solution space:** [ISMB/ECCB 2019]
   - Multiple consensus tree problem

5. **Applications**
   - Visualizing spatial clonal architecture of tumors: [To be submitted]
   - Mutational signature dynamics [PSB 2020]
Lung Cancer Patient: CRUK0037


Authors inferred 17 trees
Lung Cancer Patient: CRUK0037


Authors inferred 17 trees

**Question:** How to summarize solution space in order to remove inference errors and identify dependencies among mutations?
Parent-child Graph: Union of all Edges

<table>
<thead>
<tr>
<th></th>
<th>$v_1 \rightarrow v_{10}$</th>
<th>$v_4 \rightarrow v_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1 \rightarrow v_7$</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$v_4 \rightarrow v_7$</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
The parent-child graph does capture patterns of mutual exclusivity
The parent-child graph does capture patterns of mutual exclusivity

**Question:** Can we infer a single consensus tree?
Single Consensus Tree: Max Weight Spanning Tree

Oesper and colleagues, ACM-BCB 2018.

<table>
<thead>
<tr>
<th></th>
<th>( v_1 \rightarrow v_{10} )</th>
<th>( v_4 \rightarrow v_{10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v_1 \rightarrow v_7 )</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>( v_4 \rightarrow v_7 )</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

62
Question: How about inferring multiple consensus trees?
Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees \( \mathcal{T} = \{T_1, \ldots, T_n\} \), find surjective clustering \( \sigma : [n] \rightarrow [k] \) and consensus trees \( \mathcal{R} = \{R_1, \ldots, R_k\} \) such that \( \sum_{i=1}^{n} d(T_i, R_{\sigma(i)}) \) is minimum
Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees $\mathcal{T} = \{T_1, \ldots, T_n\}$, find surjective clustering $\sigma: [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \ldots, R_k\}$ such that $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum.
Parent-child Distance Function

$T_1$

$T_2$
Parent-child Distance Function

\[ E(T_1) \cap E(T_2) \]

\[ E(T_1) \setminus E(T_2) \]

\[ E(T_2) \setminus E(T_1) \]

\[ T_1 \]

\[ T_2 \]
Parent-child Distance Function

Parent-child distance $d(T_1, T_2)$ is the size of the **symmetric difference** of the edge sets.

Here, $d(T_1, T_2) = |E(T_1) \setminus E(T_2)| + |E(T_2) \setminus E(T_1)| = 4.$
Single Consensus Trees (SCT):
[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, ..., T_n\}$, find consensus tree $R$
s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum
Combinatorial Characterization of Solutions

**Single Consensus Trees (SCT):**
[Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, ..., T_n\}$, find consensus tree $R$
s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum

**Theorem:** [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Solution Space $\mathcal{T}$
Consensus tree $R$
Parent-child graph $G_{\mathcal{T}}$

- 4 edges
- 3 edges
- 2 edges
- 1 edge
Combinatorial Characterization of Solutions

**Single Consensus Trees (SCT):**
[Govek et al., ACM-BCB 2018]
Given \( \mathcal{T} = \{T_1, \ldots, T_n\} \), find consensus tree \( R \) s.t. \( \sum_{i=1}^{n} d(T_i, R) \) is minimum

**Theorem:** [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences of parent-child graph \( G_{\mathcal{T}} \) are solutions to SCT

**Multiple Consensus Trees (MCT):**
[Aguse et al., ISMB 2019]
Given \( \mathcal{T} = \{T_1, \ldots, T_n\} \) and \( k > 0 \), find surjective clustering \( \sigma : [n] \to [k] \) and consensus trees \( \mathcal{R} = \{R_1, \ldots, R_k\} \) s.t. \( \sum_{i=1}^{n} d(T_i, R_{\sigma(i)}) \) is minimum

Solution Space \( \mathcal{T} \)
Combinatorial Characterization of Solutions

**Single Consensus Trees (SCT):**

[Aguse et al., ISMB 2019]

Given $\mathcal{T} = \{T_1, ..., T_n\}$, find consensus tree $R$

$s.t. \sum_{i=1}^{n} d(T_i, R)$ is minimum

**Theorem:** [Govek et al., ACM-BCB 2018]

Max weight spanning arborescences of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

**Multiple Consensus Trees (MCT):**

[Aguse et al., ISMB 2019]

Given $\mathcal{T} = \{T_1, ..., T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees $\mathcal{R} = \{R_1, ..., R_k\}$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum

**Proposition:** [Aguse et al., ISMB 2019]

Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into $k$ independent SCT instances
Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):
[Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, ..., T_n\}$, find consensus tree $R$
s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT):
[Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, ..., T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees
$\mathcal{R} = \{R_1, ..., R_k\}$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into $k$ independent SCT instances
**Combinatorial Characterization of Solutions**

**Single Consensus Trees (SCT):**

[Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, ..., T_n\}$, find consensus tree $R$

s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum

**Theorem:** [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

**Multiple Consensus Trees (MCT):**

[Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, ..., T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees

$\mathcal{R} = \{R_1, ..., R_E\}$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

**Proposition:** [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into $k$ independent SCT instances

**Question:** How to find $\sigma^*$?
Multiple Consensus Trees (MCT):
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}\sigma(i)}$

Theorem: MCT is NP-hard for general $k$ (by reduction from CLIQUE).
Alternating optimization heuristic

1. Fix clustering $\sigma$ at random
2. Compute consensus tree $R_s$ for each cluster $s$
3. Reassign each input tree $T_i$ to cluster $s$ where $d(T_i, R_s)$ is minimum

Same clustering?

- Yes
- No

Done
Heuristic finds optimal solutions efficiently

<table>
<thead>
<tr>
<th>Number of instances solved by MILP to provable optimality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small, medium, and large simulated instances</strong></td>
</tr>
<tr>
<td>#clusters $k$</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>small (16)</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>large (15)</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>large (14)</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
Heuristic finds optimal solutions efficiently

<table>
<thead>
<tr>
<th>#clusters $k$</th>
<th>MILP (1 h)</th>
<th>CA (100 r.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (large 16)</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>2 (small 15)</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2 (medium 14)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Small, medium, and large simulated instances

Number of instances where heuristic returned MILP’s optimal solution
Heuristics finds optimal solutions efficiently

<table>
<thead>
<tr>
<th>#clusters $k$</th>
<th>MILP (1 h)</th>
<th>CA (100 r.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Small, medium, and large simulated instances

Number of instances where heuristic returned MILP’s optimal solution

**Question:** How to determine $k$?
Bayesian Information Criterion determines the number of clusters for each solution space


Jamal-Hanjani et al. inferred 8 trees for patient CRUK0013
Bayesian Information Criterion determines the number of clusters for each solution space

Jamal-Hanjani et al. (2017). *NEJM.*

Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037
Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence.

These edges are mutually exclusive:

- $v_1 \rightarrow v_7$
- $v_4 \rightarrow v_{10}$
Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence

These edges tend to co-occur in the trees in the solution space
Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees $\mathcal{T} = \{T_1, ..., T_n\}$, find surjective clustering $\sigma: [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, ..., R_k\}$ such that $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum

- Characterize combinatorial structure of optimal solutions
- Show that MCT is NP-hard for general $k$
- Introduce a heuristic that returns optimal solution in most cases
- Model selection for $k$
Outline

1. **Background and theory**: [RECOMB-CG 2018]
   - Perfect Phylogeny Mixture (PPM) problem
   - #PPM: exact counting and uniform sampling

2. **Simulation results**: [RECOMB-CG 2018]
   - What contributes to non-uniqueness?
   - How to reduce non-uniqueness?
   - How does non-uniqueness affect current methods?

3. **Almost uniform sampling**: [To be submitted]
   - Reducing PPM to SATISFIABILITY

4. **Summarizing solution space**: [ISMB/ECCB 2019]
   - Multiple consensus tree problem

5. **Applications**
   - Mutational signature dynamics [PSB 2020]
Mutational Signatures

There are $4 \times 6 \times 4 = 96$ SNV categories:
AC>AA, ... GT>GG

Distinct mutational processes =>
distinct patterns of SNV categories
Mutational Signatures – NMF

The figure illustrates the non-negative matrix factorization (NMF) approach, which decomposes a feature matrix $P$ into a signature matrix $S$ and an exposure matrix $E$. The NMF equation is given by:

$$P \approx S \times E$$

where $P$ is the feature matrix of size $m \times n$, $S$ is the signature matrix of size $m \times r$, and $E$ is the exposure matrix of size $r \times n$.

Alexandrov et al. [Nature, 2013] performed nonnegative matrix factorization on a large patient cohort ($n = 10,000$) identifying $r = 30$ signatures and exposures.
Alexandrov et al. [Nature, 2013] performed nonnegative matrix factorization on a large patient cohort (n = 10,000) identifying r = 30 signatures and exposures.
Clone-specific exposure inference problems

Single exposure inference

1+2+3

Independent exposure inference

- Alexandrov et al., Cell reports 2013
- Rosenthal et al., Genome Biology 2016
- Huang et al., Bioinformatics, 2017
- ...

- McPherson et al., Nature Genetics, 2016
- Jamal-Hanjani et al., NEJM 2017
- ...

89
Tree-constrained exposure (TE) inference

**TE problem**: [Christensen, Leiserson and El-Kebir, PSB 2020]:
Given phylogenetic tree $T$ and feature matrix $P$, find a small number of exposure shifts along edges of $T$
PhySigs solves the TE problem to optimality

**TE problem:**
Given feature matrix $P$, corresponding count matrix $C$, signature matrix $S$, phylogenetic tree $T$ and integer $k \geq 1$, find relative exposure matrix $D$ such that $||P - SDC||_F$ is minimum and $D$ is composed of $k$ sets of identical columns, each corresponding to a connected subtree of $T$. 
PhySigs identifies accurate exposures without overfitting in a lung cancer cohort.
PhySigs identifies an exposure shift supported by a subclonal driver

Fig. 5. PhySigs detects a large increase in DNA mismatch repair-associated Signature 6 (orange) along one branch (clusters 2 and 3; green) of the CRUK 0064 tree. In support of this finding, the branch includes a subclonal driver mutation to DNA mismatch repair gene \textit{MLH1}. 
PhySigs enables prioritization of trees in solution space

(a) CRUK0025

<table>
<thead>
<tr>
<th>$k$</th>
<th>Tree $T_1$</th>
<th>Tree $T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,344</td>
<td>1,344</td>
</tr>
<tr>
<td>2</td>
<td>1,199</td>
<td>1,199</td>
</tr>
<tr>
<td>$k^* = 3$</td>
<td>1,106</td>
<td>1,122</td>
</tr>
<tr>
<td>4</td>
<td>1,096</td>
<td>1,105</td>
</tr>
<tr>
<td>5</td>
<td>1,095</td>
<td>1,095</td>
</tr>
<tr>
<td>6</td>
<td>1,095</td>
<td>1,095</td>
</tr>
<tr>
<td>7</td>
<td>1,094</td>
<td>1,094</td>
</tr>
</tbody>
</table>

(b) Tree: $T_1$, $k = 3$

(c) Tree: $T_2$, $k = 3$
Tree-constrained exposure (TE) inference

**TE problem**: [Christensen, Leiserson and El-Kebir, PSB 2020]: Given phylogenetic tree $T$ and feature matrix $P$, find $k$ exposure shifts along edges of $T$

- **Key idea**: exposure may change along edges of phylogenetic tree
- TE interpolates between single exposure (SE) and independent exposure (IE) problems
- Model selection for $k$
Conclusion

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

Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients
Conclusion

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

1. Theory and background of perfect phylogeny mixture (PPM) problem
2. Simulation study to assess factors contributing to and impact of non-uniqueness
3. Almost uniform sampling of solutions to PPM
4. Summarizing solution space using multiple consensus trees
5. Example of downstream application: Mutational signature dynamics
Acknowledgments

El-Kebir group
• Yuanyuan Qi
• Nuraini Aguse
• Sarah Christensen
• Dikshant Pradhan
• Jiaqi Wu
• Juho Kim
• Yerong Li
• Chuanyi Zhang

• Experiments were run on NCSA’s Blue Waters supercomputer
• This work was supported by:
  • UIUC Center for Computational Biotechnology and Genomic Medicine (grant: CSN 1624790)
  • National Science Foundation (CCF 18-50502)
Somatic Mutations Occur at Different Genomic Scales

© Gryte Satas