Non-uniqueness of Solutions in Phylogenetic Deconvolution of Bulk DNA Samples of Tumors

Mohammed El-Kebir

University of Illinois at Urbana Champaign, Department of Computer Science

CISS 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

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

- 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.

Additional challenge in cancer phylogenetics:

- Tumor
- Normal
- Human reference genome (3*10^9 bp)
- Aligned read (100 bp)
- Single nucleotide variant (SNV)
Outline

1. **Background and theory:** [RECOMB-CG 2018]
   - Perfect Phylogeny Mixture (PPM) problem
   - Combinatorial characterization of solutions
   - \#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. **Summarizing solution space:** [ISMB 2019]
   - Multiple consensus tree problem
Sequencing and Tumor Phylogeny Inference

$m$ samples

$n$ mutations

Variant allele frequency (VAF): 0.4

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 $U$.
Perfect Phylogeny Mixture

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

Frequency Matrix $F$

Clones

Restricted PP Matrix $B$

Equivalent

Rows of $U$ are proportions:

$u_{pj} \geq 0$ 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 $F = UB$
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], ...

Frequency Matrix $F$

\[
\begin{pmatrix}
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4
\end{pmatrix}
\]

Mixture Matrix $U$

\[
\begin{pmatrix}
0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4
\end{pmatrix}
\]

Restricted PP Matrix $B$

\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}
\]

Rows of $U$ are proportions:

$u_{pj} \geq 0$ and $\sum_j u_{pj} \leq 1$

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

Given $F$, find $U$ and $B$ such that $F = UB$
Combinatorial Characterization

- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced $i$
- Usage $u_{p,i}$ is mass of node that introduced $i$

**Theorem 1:**
$T$ is a solution to the PPM if and only if $T$ is a spanning tree of $G$ satisfying the sum condition

**Theorem 2:**
PPM is NP-complete even for $m=2$

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

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

Question 2: Can sample solutions uniformly at random?

#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
   - Combinatorial characterization of solutions
   - \#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. Summarizing solution space: [ISMB 2019]
   - Multiple consensus tree problem

Dikshant Pradhan
What Contributes to Non-uniqueness?
What Contributes to Non-uniqueness?

samples \((m)\)

- 1
- 2
- 5
- 10

number of solutions

\[10^{10} \quad 10^8 \quad 10^6 \quad 10^4 \quad 10^2 \quad 10^0\]

number \(n\) of mutations

3 5 7 9 11 13

median edge recall

0.0 0.2 0.4 0.6 0.8 1.0

samples \((m)\)

- 1
- 2
- 5
- 10

branching coefficient \(\gamma(G_F)\)

0.0 0.2 0.4 0.6 0.8 1.0

number \(n\) of mutations

3 5 7 9 11 13

\(T\)

\(G\)
How to Reduce Non-Uniqueness?

Graph $G$ and $T$.
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]
Outline

1. Background and theory: [RECOMB-CG 2018]
   • Perfect Phylogeny Mixture (PPM) problem
   • Combinatorial characterization of solutions
   • #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. Summarizing solution space: [ISMB 2019]
   • Multiple consensus tree problem

Yuanyuan Qi
Nuraini Aguse
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

![Diagram of a parent-child graph with various nodes and edges labeled with numbers.]
Parent-child Graph: Union of all Edges

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?
Oesper and colleagues, ACM-BCB 2018.

Single Consensus Tree: Max Weight Spanning Tree

<table>
<thead>
<tr>
<th></th>
<th>$v_4 \rightarrow v_5$</th>
<th>$v_8 \rightarrow v_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1 \rightarrow v_{10}$</td>
<td>$v_4 \rightarrow v_{10}$</td>
<td>$v_1 \rightarrow v_{10}$</td>
</tr>
<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 (b)</td>
<td>5 (e)</td>
</tr>
</tbody>
</table>
Inaccurate summary for diverse solution spaces

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

Given trees $\mathcal{T} = \{T_1, ..., T_n\}$, find surjective clustering $
\sigma : [n] \rightarrow [k]$ and consensus trees $\mathcal{R} = \{R_1, ..., R_k\}$ 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.

- Characterize combinatorial structure of optimal solutions
- Show that MCT is NP-hard for general $k$
- Introduce an MILP for solving the problem for small instance sizes
- Introduce a heuristic that returns optimal solution in most cases
Conclusion

1. Background and theory: [RECOMB-CG 2018]
   • Perfect Phylogeny Mixture (PPM) problem
   • Combinatorial characterization of solutions
   • #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. Summarizing solution space: [ISMB 2019]
   • Multiple consensus tree problem
Outlook

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:**
Novel algorithms that sample **uniformly at random** from the space of PPM solutions
Acknowledgments

• Yuanyuan Qi
• Nuraini Aguse
• Dikshant Pradhan

• 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)
An Upper Bound for Number of Solutions

![Graph showing number of solutions and spanning trees](image)

**Graph $G$**

- $T$ is a spanning tree of $G$

---

45
An Upper Bound for Number of Solutions

The diagrams illustrate the number of solutions and spanning trees in a graph $G_F$ with varying numbers of mutations and samples $m$. The x-axis represents the number $n$ of mutations, and the y-axis shows the number of solutions or spanning trees. The color bars indicate different sample sizes $m$, ranging from 1 to 10. The graphs display box plots for each combination of $n$ and $m$, showing the distribution of the number of solutions or spanning trees.
Rejection Sampling Does Not Scale

Number of mutations $n$

# solutions / # spanning trees

Number $n$ of mutations

Samples ($m$)

- $5$
- $10$

$10^{-10}$ $10^{-8}$ $10^{-6}$ $10^{-4}$ $10^{-2}$ $10^0$
Somatic Mutations Occur at Different Genomic Scales

© Gryte Satas