Multi-State Perfect Phylogeny Mixture Deconvolution and Applications to Cancer Sequencing

Mohammed El-Kebir
Tumor Evolution as a Two-State Perfect Phylogeny

Given:

\[
M = \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}
\]

SNVs

States:

0 : non-mutated
1 : mutated

Assumptions:

• No copy number aberrations
• Infinite sites assumption

Find:

Two-state perfect phylogeny tree \( T \)

Seq. method Mixing Inferring \( T \)
single-cell no two-state perfect phylogeny [Gusfield, 1991]
Tumor Evolution as a Two-State Perfect Phylogeny

**Find:**

Two-state perfect phylogeny tree $T$

Mixing proportions $U$

**Given:**

Variant Allele Frequency (VAF): Fraction of reads covering position of single-nucleotide variant (SNV) that contain variant allele

**Seq. method** | **Mixing** | **Inferring $T$**
--- | --- | ---
single-cell | no | two-state perfect phylogeny
 | | [Gusfield, 1991]
bulk | yes | TrAp [Strino et al., 2013]
 | | Rec-BTP [Hajirasouliha et al., 2014]
 | | PhyloSub [Jiao et al., 2014]
 | | Clomial [Zare et al., 2014]
 | | Binary F [Hajirasouliha et al., 2014]
 | | CITUP [Malikic et al., 2015]
 | | BitPhylogeny [Yuan et al., 2015]
 | | LICHeE [Popic et al., 2015]
 | | AncesTree [El-Kebir, Oesper et al., 2015]

$F = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \ 0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \ 0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0 \ \end{bmatrix}$

$\text{VAF} = 2/5 = 0.4$
Tumor Evolution as a Two-State Perfect Phylogeny

Given:
- VAFs \( F = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \\ 0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \\ 0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0 \end{bmatrix} \)
- Mutations

Find:
- Two-state perfect phylogeny tree \( T \)
- Mixing proportions \( U \)

States:
- 0: non-mutated
- 1: mutated
- 2: CN loss-of-heterozygosity
- 3: amplification

States: rescale VAFs to CCFs

Mixing proportions:
- \( U = \begin{bmatrix} 0.8 & 0.6 & 0.2 & 0.2 & 0.4 & 0.0 \end{bmatrix} \)

Inferring \( T \):
- Single-cell no
- Two-state perfect phylogeny yes
- Gusfield, 1991

Seq. method:
- TrAp [Strino et al., 2013]
- Rec-BTP [Hajirasouliha et al., 2014]
- PhyloSub [Jiao et al., 2014]
- Clomial [Zare et al., 2014]
- Binary F [Hajirasouliha et al., 2014]
- CITUP [Malikic et al., 2015]
- BitPhylogeny [Yuan et al., 2015]
- LICHeE [Popic et al., 2015]
- AncesTree [El-Kebir, Oesper et al., 2015]
- ...
Tumor Evolution as a **Multi**-State Phylogeny

**Given:**

- **Mutations**
  - VAFs \( F = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \\ 0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \\ 0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0 \end{bmatrix} \)  
- Samples: \( S_1, S_2, S_3 \)

**Find:**

- Two-state perfect phylogeny tree \( T \)
- Mixing proportions \( U \)

**States:**

- 0: non-mutated
- 1: mutated
- 2: CN loss-of-heterozygosity
- 3: amplification
- ... more than > 2 states

**NP-complete**

**Seq. method** | **Mixing** | **Inferring T**
--- | --- | ---
Single-cell | no | two-state perfect phylogeny [Gusfield, 1991]
Bulk | yes | TrAp [Strino et al., 2013]
Rec-BTP [Hajirasouliha et al., 2014]
PhyloSub [Jiao et al., 2014]
Clomial [Zare et al., 2014]
Binary F [Hajirasouliha et al., 2014]
CITUP [Malikic et al., 2015]
BitPhylogeny [Yuan et al., 2015]
LICH EE [Popic et al., 2015]
AncesTree [El-Kebir, Oesper et al., 2015]
...
Outline

• Problem Statement
• Combinatorial Characterization of Solutions
• Application to Cancer Sequencing
Problem Statement

Two-State Perfect Phylogeny:

Infinite sites assumption: a character changes state once

\[ F = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.6 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix} \]
Problem Statement

**Two-State Perfect Phylogeny:**

Infinite sites assumption: a character changes state once

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

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

**Usage Matrix** \(U\)

\(U = [u_{pj}]\) is a usage matrix iff \(u_{pj} \geq 0\) and \(\sum_{j} u_{pj} \leq 1\)

Complete Two-State Perfect Phylogeny \(B/T\)
Problem Statement

Two-State Perfect Phylogeny:
Infinite sites assumption: a character changes state once

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

Usage Matrix \( U \)

\( U = [u_{pj}] \) is a usage matrix iff
\[ u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1 \]

VAF Factorization Problem (VAFFP): [El-Kebir, Oesper et al., 2015]
Given \( F \), find \( U \) and \( B \)
such that \( F = U B \)
Problem Statement

Two-State Perfect Phylogeny:
Infinite sites assumption: a character changes state once

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

Usage Matrix \( U \)

\[ U = [u_{pj}] \text{ is a usage matrix iff } u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1 \]

VAF Factorization Problem (VAFFP): [El-Kebir, Oesper et al., 2015]

Given \( F \), find \( U \) and \( B \) such that \( F = U B \)

Multi-State Perfect Phylogeny:
Infinite alleles assumption: a character changes to a state once

\[ F_0 = \begin{pmatrix}
0.1 & 0.8 \\
0.7 & 0.0 \\
\end{pmatrix} \]

\[ F_1 = \begin{pmatrix}
0.2 & 0.2 \\
0.3 & 0.4 \\
\end{pmatrix} \]

\[ F_2 = \begin{pmatrix}
0.7 & 0.0 \\
0.0 & 0.6 \\
\end{pmatrix} \]
**Problem Statement**

**Two-State Perfect Phylogeny:**  
Infinite sites assumption: a character changes state once

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

**Usage Matrix**  
\[ U = [u_{pj}] \text{ is a usage matrix iff } u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1 \]

**VAF Factorization Problem (VAFFP):**  
[El-Kebir, Oesper et al., 2015]  
Given \( F \), find \( U \) and \( B \) such that \( F = UB \)

**Multi-State Perfect Phylogeny:**  
Infinite alleles assumption: a character changes to a state once

\[ F_0 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad F_1 = \begin{pmatrix} 0.7 & 0.0 \\ 0.0 & 0.6 \end{pmatrix} \]

**Usage Matrix**  
\[ U \]

**Perfect Phylogeny Mixture Deconvolution Problem (PPMDP)**  
[El-Kebir et al., 2016]: Given \( F \), find \( U \) and \( A \) such that \( F_i = U A_i \) for all states \( i \)
Two-State Perfect Phylogeny:
- A character changes state once
  - Once a mutation happens it persists
- Thus $T_{(c,1)} = \bar{T}_{(c,1)}$ — subtree rooted at $V_{(c,1)}$
Combinatorial Characterization

Two-State Perfect Phylogeny:

- A character changes state once
  - Once a mutation happens it persists
- Thus $T_{(c,1)} = \bar{T}_{(c,1)}$ — subtree rooted at $V_{(c,1)}$

\[
\sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \geq f_{p,(c,1)}
\]
Combinatorial Characterization

Two-State Perfect Phylogeny:
• A character changes state once
  • Once a mutation happens it persists
• Thus $T_{(c,1)} = \overline{T}_{(c,1)}$ — subtree rooted at $V_{(c,1)}$

Multi-State Perfect Phylogeny:
• A character changes to a state once
• Thus, $T_{(c,i)} \neq \overline{T}_{(c,i)}$
• Instead:
  $$\overline{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$$

Descendant set
$D_{(c,1)} = \{1, 2\}$

Sum Condition (SC)
$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)}$$
Combinatorial Characterization

Two-State Perfect Phylogeny:
- A character changes state once
  - Once a mutation happens it persists
- Thus $T_{(c,1)} = \overline{T}_{(c,1)}$ — subtree rooted at $V_{(c,1)}$

Multi-State Perfect Phylogeny:
- A character changes to a state once
- Thus, $T_{(c,i)} \neq \overline{T}_{(c,i)}$
- Instead:
  $$\overline{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$$

Descendant set $D_{(c,1)} = \{1,2\}$

Sum Condition (SC)
$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)}$$

Multi-State Sum Condition (MSSC) [El-Kebir et al., 2016]
- Cumulative frequency
  $$f^+_p(D_{(c,i)}) \geq \sum_{(d,j) \in \delta(c,i)} f^+_p(D_{(d,j)})$$
Spanning Trees in Ancestry Graph

**Two-State Perfect Phylogeny:**

- Simple directed graph (DAG)
- Vertices are characters
- Edges are potential ancestral relationships

### Theorem 1
[El-Kebir, Oesper et al., 2015; Popic et al., 2015]
Solutions are spanning trees that satisfy (SC)

### Theorem 2
[El-Kebir, Oesper et al., 2015]
VAFFP is NP-complete for $m = O(n)$
Spanning Trees in Ancestry Graph

Two-State Perfect Phylogeny:

- Simple directed graph (DAG)
- Vertices are characters
- Edges are potential ancestral relationships

Multi-State Perfect Phylogeny:

- Directed multi-graph
- Vertices are character-state pairs
- Edges are labeled by valid descendant set pairs

Theorem 1 [El-Kebir, Oesper et al., 2015; Popic et al., 2015]
Solutions are spanning trees that satisfy (SC)

Theorem 2 [El-Kebir, Oesper et al., 2015]
VAFFP is NP-complete for $m = O(n)$

Theorem 1 [El-Kebir et al., 2016]
Solutions are threaded spanning trees satisfying (MSSC)

Theorem 2 [El-Kebir et al., 2016]
PPMDP is NP-complete even for $m = 2$ and $k = 2$
Application to Cancer Sequencing

**Input**
- Read-depth ratio
- B-allele frequencies
- Variant allele frequencies

**Model**
- Character is a genomic position (SNV)
- State is a triple $(x, y, z)$ where
  - $x$ is # maternal copies
  - $y$ is # paternal copies
  - $z$ is # mutated copies
- Cladistic characters

\[ \begin{align*}
SCA &= (2,1,\cdot) \\
SCD &= (1,0,\cdot) \\
CN - LOH &= (2,0,\cdot)
\end{align*} \]
Application to Cancer Sequencing

**Input**
- Read-depth ratio
- B-allele frequencies
- Variant allele frequencies

**Model**
- Character is a genomic position (SNV)
- State is a triple \((x, y, z)\) where
  - \(x\) is \# maternal copies
  - \(y\) is \# paternal copies
  - \(z\) is \# mutated copies
- Cladistic characters

\[
\begin{align*}
\text{SCA} &= (2, 1, \cdot) \\
\text{SCD} &= (1, 0, \cdot) \\
\text{CN - LOH} &= (2, 0, \cdot)
\end{align*}
\]
Application to Cancer Sequencing

**Input**
- Read-depth ratio
- B-allele frequencies
- Variant allele frequencies

**Model**
- Character is a genomic position (SNV)
- State is a triple \((x, y, z)\) where
  - \(x\) is \# maternal copies
  - \(y\) is \# paternal copies
  - \(z\) is \# mutated copies
- Cladistic characters
Conclusions

• Generalization of infinite sites model for SNVs is infinite alleles model for SNVs + CNAs

• Introduced Perfect Phylogeny Mixture Deconvolution Problem (PPMDP) for multi-state characters

• Combinatorial characterization of solutions

• PPMDP is NP-complete for $k = 2$ and $m = 2$

• Application to cancer sequencing
  • Metagenomics, somatic hypermutations, mtDNA, ...
Acknowledgements

Research Group

Benjamin J. Raphael
Gryte Satas
Layla Oesper
Dora Erdos
Matthew Reyna
Ashley Conard
Cyrus Cousins
Rebecca Elyanow
Hsin-Ta Wu

Funding

Preprint will be available soon on arXiv