SPhyR: Tumor Phylogeny Estimation from Single-Cell Sequencing Data under Loss and Error

Mohammed El-Kebir

Accepted at ECCB 2018
http://doi.org/10.1093/bioinformatics/bty589
Tumorigenesis: (i) Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]
Clonal Evolution Theory of Cancer
[Nowell, 1976]

Tumorigenesis: (i) Cell Mutation & (ii) Cell Division

Heterogeneous Tumor
Tumorigenesis: (i) Cell Mutation & (ii) Cell Division

**Goal:** Given single-cell DNA sequencing data, find phylogenetic tree $T$

**Requirement:** Evolutionary model
Somatic Mutations Occur at Different Genomic Scales

- Single Nucleotide Variant (SNV)
- Small Insertion / Deletion (indel)
- Copy-Number Aberration (CNA)
- Structural Variant (SV)
- Whole-Genome Duplication (WGD)
Infinite Sites Assumption is too Restrictive for SNVs

- **Single Nucleotide Variant (SNV)**
- **Small Insertion / Deletion (indel)**
- **Copy-Number Aberration (CNA)**
- **Structural Variant (SV)**
- **Whole-Genome Duplication (WGD)**

**Infinite sites assumption:**
- No parallel evolution of SNVs
- No loss of SNVs
- SCITE [Jahn et al. 2016]
- OncoNEM [Ross and Markowetz, 2016]

SNVs can be **lost** due to CNAs
Outline

• Perfect data (error free)
  • Problem statement
  • Combinatorial characterization of solutions
  • Exact algorithm
  • Results

• Real data (with errors)
  • Problem statement
  • Heuristic algorithm
  • Results

• Conclusions
**$k$-Dollo Phylogeny ($k$-DP) Problem**

**Definition 1.** A $k$-Dollo phylogeny $T$ is a rooted, node-labeled tree subject to the following conditions.

1. Each node $v$ of $T$ is labeled by a vector $b_v \in \{0, 1\}^n$.
2. The root $r$ of $T$ is labeled by vector $b_r = [0, \ldots, 0]^T$. 
**k-Dollo Phylogeny (k-DP) Problem**

**Definition 1.** A *k-Dollo phylogeny* $T$ is a rooted, node-labeled tree subject to the following conditions.

1. Each node $v$ of $T$ is labeled by a vector $b_v \in \{0, 1\}^n$.
2. The root $r$ of $T$ is labeled by vector $b_r = [0, \ldots, 0]^T$.
3. For each character $c \in [n]$, there is exactly one *gain edge* $(v, w)$ in $T$ such that $b_{v, c} = 0$ and $b_{w, c} = 1$.
4. For each character $c \in [n]$, there are at most $k$ *loss edges* $(v, w)$ in $T$ such that $b_{v, c} = 1$ and $b_{w, c} = 0$. 
**k-Dollo Phylogeny (k-DP) Problem**

**Definition 1.** A *k-Dollo phylogeny* $T$ is a rooted, node-labeled tree subject to the following conditions.

1. Each node $v$ of $T$ is labeled by a vector $b_v \in \{0, 1\}^n$.
2. The root $r$ of $T$ is labeled by vector $b_r = [0, \ldots, 0]^T$.
3. For each character $c \in [n]$, there is exactly one gain edge $(v, w)$ in $T$ such that $b_{v,c} = 0$ and $b_{w,c} = 1$.
4. For each character $c \in [n]$, there are at most $k$ loss edges $(v, w)$ in $T$ such that $b_{v,c} = 1$ and $b_{w,c} = 0$.

**k-Dollo Phylogeny problem (k-DP).** Given a binary matrix $B \in \{0, 1\}^{m \times n}$ and parameter $k \in \mathbb{N}$, determine whether there exists a $k$-Dollo phylogeny for $B$, and if so construct one.
**Theorem 3.** Let $B \in \{0, 1\}^{m \times n}$. The following statements are equivalent.

1. There exists a $k$-Dollo phylogeny $T$ for $B$.
2. There exists a $k$-Dollo completion $A$ of $B$.
3. There exists a $k$-completion $A$ of $B$, and perfect phylogeny $T$ for $A$ whose characters are consistent with $S[k]$.

---

**Combinatorial Characterization of $k$-DP**

**Input Matrix $B$**

```
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

**$k$-Dollo Completion $A$**

```
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

**$k$-Dollo State Tree $S[k]$**

**$k$-Dollo Phylogeny $T$**
Forbidden Submatrices in Solutions $A$ to $k$-DP

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

$$\begin{pmatrix}
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 2 \\
1 & 1 & 0 & 2 \\
1 & 1 & 1 & 2
\end{pmatrix}$$

$$\begin{pmatrix}
2 & 0 & 2 & 0 \\
0 & 1 & 0 & 2 \\
1 & 1 & 1 & 2 \\
1 & 1 & 2 & 2
\end{pmatrix}$$

$$\begin{pmatrix}
1 & 1 & 2 & 1 \\
0 & 2 & 0 & 2 \\
1 & 2 & 2 & 2 \\
2 & 2 & 2 & 2
\end{pmatrix}$$

$$\begin{pmatrix}
2 & 1 \\
1 & 2 \\
2 & 2
\end{pmatrix}$$

$k = 0$

$k = 1$

Number of forbidden submatrices is $4k^4 + 8k^3 + 8k^2 + 4k + 1$

Open question: Hardness of deciding whether $B$ admits a $k$-Dollo completion $A$
Results for $k$-DP

- Naive ILP does not scale and has $O(mnk)$ variables and $O(m^3n^2k^4)$ constraints
- Column and cutting plane generation
  - Introduce variables and constraints only when needed
- Simulations with 60 instances for each each $m$, $n$ and $k$
Outline

• Perfect data (error free)
  • Problem statement
  • Combinatorial characterization of solutions
  • Exact algorithm
  • Results

• Real data (with errors)
  • Problem statement
  • Heuristic algorithm
  • Results

• Conclusions
**$k$-Dollo Phylogeny Flip and Cluster ($k$-DPFC) problem.** Given matrix $D \in \{0, 1, ?\}^{m \times n}$, error rates $\alpha, \beta \in [0, 1]$, integers $k, s, t \in \mathbb{N}$, find matrix $B \in \{0, 1\}^{m \times n}$ and tree $T$ such that: (1) $B$ has at most $s$ unique rows and at most $t$ unique columns; (2) $\Pr(D \mid B, \alpha, \beta)$ is maximum; and (3) $T$ is a $k$-Dollo phylogeny for $B$.

$$\Pr(D \mid B, \alpha, \beta) = \prod_{p=1}^{m} \prod_{c=1}^{n} \begin{cases} \alpha, & d_{p,c} = 1 \text{ and } b_{p,c} = 0 \\ 1 - \alpha, & d_{p,c} = 1 \text{ and } b_{p,c} = 1, \\ \beta, & d_{p,c} = 0 \text{ and } b_{p,c} = 1, \\ 1 - \beta, & d_{p,c} = 0 \text{ and } b_{p,c} = 0, \\ 1, & d_{p,c} = ? \end{cases}$$
SPhyR: Single-cell Phylogeny Reconstruction

• Coordinate ascent:
  1. k-Means with random seed to obtain cell clustering $\pi$ and SNV clustering $\psi$
  2. ILP to obtain maximum likelihood $k$-Dollo completion $A$ given $D$, $\pi$ and $\psi$
  3. Identify maximum likelihood $\pi$ given $A$ and $\psi$
  4. Identify maximum likelihood $\psi$ given $A$ and $\pi$
  5. Repeat until convergence

• Available on Github: https://github.com/elkebir-group/SPhyR
Simulation Results \( (m = 50, n = 50, k = 1) \)
Simulation Results ($m = 50$, $n = 50$, $k = 1$)
Colorectal patient CRC1 [Leung et al., 2017]

\[ m = 178 \text{ cells} \]

\[ n = 17 \text{ SNVs} \]

\[ \alpha = 0.0152 \]

\[ \beta = 0.0789 \]

**SCITE:**

likelihood = -413.38

**SPhyR:**

likelihood = -450.70
Conclusions

• $k$-Dollo parsimony model strikes a balance between realistic and yet sufficiently constrained
• Solutions are integer matrix completions
• SPhyR outperformed existing methods

Future work:
• Include $\alpha$ and $\beta$ into optimization
• Model selection for $s$, $t$ and $k$
• Hardness is open