Cancer Phylogenetics Pipeline

(i) Alignment

(ii) Somatic variant calling

(iii) SNV clustering

(iv) Phylogeny inference

(v) Downstream analysis

reference genome
aligned maternal read
aligned paternal read

normal

(i) Alignment

(ii) Somatic variant calling

single nucleotide variant (SNV)

(copy number state (# maternal, # paternal copies)

SNV cluster

variant allele frequency

SNV count

normal
tumor

aligned read
Cancer Phylogenetics Pipeline

(i) Alignment

(ii) Somatic variant calling

(iii) SNV clustering

(iv) Phylogeny inference

(v) Downstream analysis

reference genome
aligned maternal read
aligned paternal read

aligned read

single nucleotide variant (SNV)

SNV cluster

(copy number state (# maternal, # paternal copies)

SNV count

variant allele frequency

normal

(2,2) (1,2) (0,2)

tumor

(x, y)

aligned maternal read
aligned paternal read

SNV count

variant allele frequency

(iii) SNV clustering

(iv) Phylogeny inference

(v) Downstream analysis
Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants

Reading:


Tumorigenesis: (i) Cell Mutation

Clonal Theory of Cancer
[Nowell, 1976]
Tumorigenesis: (i) Cell Mutation, (ii) Cell Division

Clonal Theory of Cancer
[Nowell, 1976]

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

Primary Tumor

Brain Metastasis

Liver Metastasis
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

**Question:**
How to reconstruct pattern of metastasis?

Primary Tumor

Brain Metastasis

Liver Metastasis
Key Challenge in Computational Biology

- Translating a biological problem into computer science
- Formulating a combinatorial problem
- Designing an algorithm
- Interpreting solutions and validating the algorithm
- Analyzing complexity & combinatorial structure

Biological question

Translating a biological problem into computer science
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

Cell Tree

Phylogenetic Tree
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

**Goal:** Given phylogenetic tree $T$, find *parsimonious* vertex labeling $\ell$ with fewest migrations.
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

**Goal:** Given phylogenetic tree $T$, find *parsimonious* vertex labeling $\ell$ with fewest migrations

**Question:** Find parsimonious vertex labeling
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

**Goal:** Given phylogenetic tree $T$, find **parsimonious** vertex labeling $\ell$ with fewest migrations.
Outline

• Metastasis
• Maximum parsimony
• Problem statement
• Complexity
• Algorithm & results
• Problem variants

Reading:


Character-Based Phylogeny Reconstruction: Criterion

Parsimony: minimize number of changes on edges of tree
A Small and a Large Problem

**Small Maximum Parsimony Phylogeny Problem:**
Given $m \times n$ matrix $A = [a_{i,j}]$ and tree $T$ with $m$ leaves, find assignment of character states to each internal vertex of $T$ with minimum parsimony score.

**Large Maximum Parsimony Phylogeny Problem:**
Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree $T$ with $m$ leaves labeled according to $A$ and an assignment of character states to each internal vertex of $T$ with minimum parsimony score.

**Question:** Are both problems easy (i.e. in $P$)?
Small Maximum Parsimony Phylogeny Problem

Key observations: (1) Characters can be solved independently.
(2) Optimal substructure in subtrees.
Small Maximum Parsimony Phylogeny Problem:
Given rooted tree $T$ whose leaves are labeled by $\sigma : L(T) \rightarrow \Sigma$, find assignment of states to each internal vertex of $T$ with minimum parsimony score.
Recurrence for Small Maximum Parsimony Problem

Small Maximum Parsimony Phylogeny Problem:
Given rooted tree $T$ whose leaves are labeled by $\sigma : L(T) \to \Sigma$, find assignment of states to each internal vertex of $T$ with minimum parsimony score.

Let $\mu(v, s)$ be the minimum number of mutations in the subtree rooted at $v$ when assigning state $s$ to $v$.

Let $\delta(v)$ be the set of children of $v$. 
Recurrence for Small Maximum Parsimony Problem

Small Maximum Parsimony Phylogeny Problem:
Given rooted tree $T$ whose leaves are labeled by $\sigma : L(T) \rightarrow \Sigma$, find assignment of states to each internal vertex of $T$ with minimum parsimony score.

Let $\mu(v, s)$ be the minimum number of mutations in the subtree rooted at $v$ when assigning state $s$ to $v$.

\[
c(s, t) = \begin{cases} 
0, & \text{if } s = t \\
1, & \text{if } s \neq t,
\end{cases}
\]

\[
\mu(v, s) = \min \begin{cases} 
\infty, & \text{if } v \in L(T) \text{ and } s \neq \sigma(v), \\
0, & \text{if } v \in L(T) \text{ and } s = \sigma(v), \\
\sum_{w \in \delta(v)} \min_{t \in \Sigma} \{c(s, t) + \mu(w, t)\}, & \text{if } v \not\in L(T).
\end{cases}
\]

Let $\delta(v)$ be the set of children of $v$. 
Filling out DP Table and Traceback

Let \( r(T) \) be the root vertex

**Filling out \( \pi \)**

\[ \text{Fill}(T, \pi(T), \Sigma, \Sigma) \]

- \( O(m|\Sigma|^2) \)
- If \( v \in \mathcal{L}(T) \)
  - For \( s \in \Sigma \)
    - If \( s = \pi(v) \)
      - \( \pi(v, s) = 0 \)
    - Else
      - \( \pi(v, s) = \infty \)
  - Else
    - For \( w \in \mathcal{S}(v) \)
      - \( \pi(T, w, \Sigma) \)  // children
      - \( \pi(v, s) = \infty \)
      - For \( s \in \mathcal{S}(v) \)
        - \( \pi(v, s) = \min_{t \in \Sigma} \{ \sum_{t \in \Sigma} c(s, t) + \pi(w, t) \} \)

**Backtrace**

\[ \text{Backtrace}(T, v, \pi) \]

- If \( v = r(T) \)
  - \( \pi(r(T)) = \arg \min_{s \in \Sigma} \{ \sum_{s \in \Sigma} c(s, t) + \pi(v, t) \} \)
  - Else let \( u \) be the parent of \( v \) and let \( s \) be the state of \( u \)
    - \( \pi(u) = \arg \min_{s \in \Sigma} \{ \sum_{s \in \Sigma} c(s, t) + \pi(v, t) \} \)
- For \( w \in \mathcal{S}(v) \)
  - \( \text{Backtrace}(T, w, \pi) \)
Outline

• Metastasis
• Maximum parsimony
• Problem statement
• Complexity
• Algorithm & results
• Problem variants

Reading:


Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

Cell Tree

primary tumor $P$

metastasis $M_2$

metastasis $M_1$

Phylogenetic Tree $T$

Instance of small maximum parsimony problem

Vertex labeling $\ell$

Goal: Given phylogenetic tree $T$, find \textit{parsimonious} vertex labeling $\ell$ with fewest migrations

Minimum Migration Analysis in Ovarian Cancer


- Instance of the maximum parsimony small phylogeny problem [Fitch, 1971; Sankoff, 1975]
Minimum Migration Analysis in Ovarian Cancer


- Instance of the maximum parsimony small phylogeny problem [Fitch, 1971; Sankoff, 1975]

\[ \mu^* = 13 \]

The migration graph \( G \) represents the spread and mixing of tumor cells across anatomical sites. The figure illustrates the different routes of migration from the right ovary (ROv) to other sites such as the small bowel (SBwl), omentum (Om), left ovary (LOv), and left fallopian tube (LFTB). The graph highlights the complexity of tumor growth and spread in ovarian cancer.

- \( m = 7 \) anatomical sites
- SBwl: Small Bowel
- LFTB: Left Fallopian Tube
- LOv: Left Ovary
- Om: Omentum
- ApC: Appendix
- RFTA: Right Fallopian Tube
Minimum Migration History is **Not** Unique

- Enumerate all minimum-migration vertex labelings in the backtrace step

\[ \mu^* = 13 \]

---

ApC: Appendix  
LFB: Left Fallopian Tube  
LOv: Left Ovary  
RFTA: Right Fallopian Tube  
ROv: Right Ovary  
SBwl: Small Bowel  
Om: Omentum
Comigrations: Simultaneous Migrations of Multiple Clones

- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number $\gamma$ of comigrations is the number of multi-edges in migration graph $G^\dagger$

$\mu^* = 13$
$\gamma = 10$

Clone Tree $T$

Migration Graph $G$

† Not necessarily true in the case of directed cycles
Comigrations: Simultaneous Migrations of Multiple Clones

- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number $\gamma$ of comigrations is the number of multi-edges in migration graph $G^\dagger$

$\mu^* = 13$
$\gamma = 10$

$\mu^* = 13$
$\gamma = 11$

$\mu^* = 13$
$\gamma = 11$

$\mu^* = 13$
$\gamma = 7$

$\mu^* = 13$
$\gamma = 7$

$\mu^* = 13$
$\gamma = 7$

$\mu^* = 13$
$\gamma = 7$

- ApC: Appendix
- LFTB: Left Fallopian Tube
- LOv: Left Ovary
- RFTA: Right Fallopian Tube
- ROv: Right Ovary
- SBwl: Small Bowel
- Om: Omentum
Tradeoffs between Migrations, Comigrations and Migration Pattern

<table>
<thead>
<tr>
<th></th>
<th>single-source seeding (S)</th>
<th>multi-source seeding (M)</th>
<th>reseeding (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>monoclonal (m)</strong></td>
<td>tree</td>
<td>directed acyclic graph</td>
<td>directed graph</td>
</tr>
<tr>
<td><strong>polyclonal (p)</strong></td>
<td>multi-tree</td>
<td>directed acyclic graph</td>
<td>directed multi-graph</td>
</tr>
</tbody>
</table>

Migration patterns can be distinguished using two criteria. First, by the number of clones that migrate between two anatomical sites: with monoclonal (m) seeding, every metastasis is seeded by a single clone, whereas with polyclonal (p) seeding, multiple clones migrate from one anatomical site to another. Second, by the migration topology: with single-source seeding (S), every metastasis is seeded from a single anatomical site, with multi-source seeding (M), a metastasis may be seeded by multiple anatomic sites, and with reseeding (R), clones migrate back and forth between anatomical sites, resulting in cycles of seeding.

Suboptimal solutions include monoclonal reseeding (mR) and polyclonal single-source seeding (pS). Montoclonal multi-source seeding (pM) is also considered.
**Constrained Multi-objective Optimization Problem**

**Parsimonious Migration History (PMH):** Given a phylogenetic tree \( T \) and a set \( \mathcal{P} \subseteq \{ \text{S, M, R} \} \) of allowed migration patterns, find vertex labeling \( \ell \) with minimum migration number \( \mu^*(T) \) and smallest comigration number \( \hat{\gamma}(T) \).

![Diagram](https://via.placeholder.com/150)

- **a.** Single-source seeding (S)
  - \( \mathcal{P} = \{ \text{S} \} \)
  - \( (\mu^*, \hat{\gamma}) = (6, 2) \)
  
- **b.** Multi-source seeding (M)
  - \( \mathcal{P} = \{ \text{S, M} \} \)
  - \( (\mu^*, \hat{\gamma}) = (5, 3) \)

- **c.** Reseeding (R)
  - \( \mathcal{P} = \{ \text{S, M, R} \} \)
  - \( (\mu^*, \hat{\gamma}) = (4, 4) \)

---

Outline

• Metastasis
• Maximum parsimony
• Problem statement
• Complexity
• Algorithm & results
• Problem variants

Reading:


Results [El-Kebir, WABI 2018]

**Parsimonious Migration History (PMH):** Given a phylogenetic tree $T$ and a set $\mathcal{P} \subseteq \{S, M, R\}$ of allowed migration patterns, find vertex labeling $\ell$ with minimum migration number $\mu^*(T)$ and smallest comigration number $\hat{\gamma}(T)$.

**Theorem 1:** PMH is NP-hard when $\mathcal{P} = \{S\}$

**Theorem 2:** PMH is fixed parameter tractable in the number $m$ of locations when $\mathcal{P} = \{S\}$
PMH is NP-hard when $\mathcal{P} = \{S\}$

**3-SAT:** Given $\varphi = \bigwedge_{i=1}^{k}(y_{i,1} \lor y_{i,2} \lor y_{i,3})$ with variables $\{x_1, \ldots, x_n\}$ and $k$ clauses, find $\phi : [n] \rightarrow \{0, 1\}$ satisfying $\varphi$

$$\Sigma = \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, c_1, \ldots, c_k, \bot\}$$
PMH is NP-hard when $\mathcal{P} = \{S\}$

**3-SAT:** Given $\varphi = \bigwedge_{i=1}^{k}(y_{i,1} \lor y_{i,2} \lor y_{i,3})$ with variables $\{x_1, \ldots, x_n\}$ and $k$ clauses, find $\phi : [n] \rightarrow \{0,1\}$ satisfying $\varphi$

Three ideas:

1. Ensure that $(x, \neg x) \in E(G)$ or $(\neg x, x) \in E(G)$
2. Ensure that $\ell^* (r(T)) = \bot$
3. Ensure that $\varphi$ is satisfiable if and only if $\ell^*$ encodes a satisfying truth assignment
PMH is NP-hard when $\mathcal{P} = \{S\}$

**3-SAT:** Given $\varphi = \bigwedge_{i=1}^{k}(y_{i,1} \lor y_{i,2} \lor y_{i,3})$ with variables $\{x_1, \ldots, x_n\}$ and $k$ clauses, find $\phi : [n] \to \{0,1\}$ satisfying $\varphi$

$\Sigma = \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, c_1, \ldots c_k, \bot\}$

Three ideas:

1. Ensure that $(x, \neg x) \in E(G)$ or $(\neg x, x) \in E(G)$
2. Ensure that $\ell^*(r(T)) = \bot$
3. Ensure that $\varphi$ is satisfiable if and only if $\ell^*$ encodes a satisfying truth assignment

**Lemma:** Let $B > 10k + 1$ and $A > 2Bn + 27k$. Then, $\varphi$ is satisfiable if and only if $\mu^*(T) = (B + 1)n + 25k$
PMH is NP-hard when $\mathcal{P} = \{S\}$

$\varphi = (x_1 \lor x_2 \lor \neg x_3) \land (\neg x_1, \neg x_2, \neg x_3)$

$k = 2, n = 3$

$B = 10k + 2 = 22$

$A = 2Bn + 27k + 1 = 187$

$\pm \theta$ $\perp$ $x_1$ $x_2$ $\neg x_3$

$B + 5$ $3$ $B + 8$ $B + 5$ $3$

$\neg x_1$ $c_1$ $\neg x_2$ $x_3$ $c_2$

$\Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \perp \}$

$\mu^*(T) = (B + 1)n + 25k$

$= 23 \cdot 3 + 50 \cdot 2 = 119$

**Lemma:** Let $B > 10k + 1$ and $A > 2Bn + 27k$.

Then, $\varphi$ is satisfiable if and only if $\mu^*(T) = (B + 1)n + 25k$
PMH is FPT in number $m$ of locations when $\mathcal{P} = \{S\}$

**Lemma:** If there exists labeling $\ell$ consistent with $\hat{G}$ then
\[
d_T(u, v) \geq d_{\hat{G}}(\text{lca}_{\hat{G}}(u), \hat{l}(v)) \quad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \tag{1}
\]

\[
\ell^*(v) = \begin{cases} 
\text{lca}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\
\sigma(\ell^*(\pi(v)), \text{lca}_{\hat{G}}(v)), & \text{if } v \neq r(T), 
\end{cases}
\]

where $\sigma(s, t) = s$ if $s = t$ and otherwise $\sigma(s, t)$ is the unique child of $s$ that lies on the path from $s$ to $t$ in $\hat{G}$.

**Lemma:** If (1) holds then $\ell^*$ is a minimum migration labeling consistent with $\hat{G}$. 38
PMH is FPT in number $m$ of locations when $\mathcal{P} = \{S\}$

Lemma: If there exists labeling $\ell$ consistent with $\hat{G}$ then

$$d_T(u, v) \geq d_{\hat{G}}(\text{lca}_{\hat{G}}(u), \hat{\ell}(v)) \quad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

$$\ell^*(v) = \begin{cases} \text{lca}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \text{lca}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

where $\sigma(s, t) = s$ if $s = t$ and otherwise $\sigma(s, t)$ is the unique child of $s$ that lies on the path from $s$ to $t$ in $\hat{G}$.

Lemma: If (1) holds then $\ell^*$ is a minimum migration labeling consistent with $\hat{G}$. 

O($nm^m$) time
Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants

Reading:


Simulations

Available on: https://github.com/elkebir-group/PMH-S
Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants

Reading:


Resolving Clone Tree Ambiguities

Clone Tree $T$

Polytomy

PMH Problem
Parsimonious Migration History

Allowed patterns $\mathcal{P}$
$\{S, M, R\}$

PMH-TR Problem
Parsimonious Migration History with Tree Resolution

Polyclonal single source seeding (pS)

Monoclonal single source seeding (mS)

Resolved polytomy
Resolving Clone Tree Ambiguities

PMH-TI Problem
Parsimonious Migration History with Tree Inference

$F^- = P \begin{bmatrix} 0.75 & 0 & 0 \\ 0.82 & 0.15 & 0.56 \\ 0.66 & 0 & 0.68 \end{bmatrix}$

$F^+ = P \begin{bmatrix} 0.86 & 0 & 0 \\ 0.92 & 0.30 & 0.68 \\ 0.78 & 0 & 0.77 \end{bmatrix}$

$\mu = 2$
$\gamma = 2$

monoclonal single source seeding (mS)
Clone Tree
Migration Graph

Label ancestral vertices by anatomical sites
Resolve clone tree ambiguities
Cell Division and Mutation History
Cell Migration History

Standard Phylogenetic Techniques

Sample Tree

Sequencing and Mutation Calling

Mutation Matrix

Tumor Phylogenetic Techniques

Clone Tree

MACHINA

Cellular History of Metastatic Cancer

Mutation Calling

unobserved clones

inferred extant clones

homoplasy

mutations
MACHINA accurately infers clone trees and migration histories on simulated data.

(a) MACHINA
   Neighbor joining
   Treeomics
   Treeomics-sub
   PhyloWGS
   AncesTree

   clone tree distance

(b) Migration graph $F_1$ score
   mS
   pS
   pM
   pR

(c) Migrating clones $F_1$ score
   mS
   pS
   pM
   pR
Applying MACHINA to Metastatic Breast Cancer


- **Triple negative, basal-like breast cancer presenting with Stage IIIA disease**
- **Treated with neoadjuvant AC-T achieving stable disease, followed by mastectomy and radiation**
- **After 17 months, patient presented with Stage IV disease with 7 distant metastases**
- **Died of disease in 25 months**
- **Six tumors for WGS: primary, rib, kidney, brain, liver, and lung**

**Reported**

\( (\mu, \gamma) = (12, 6) \)

\( 1 \) (Supp.)

- Resolved monoclonal single-source seeding (mS)
- Polyclonal single-source seeding (pS)
- Monoclonal multi-source seeding (mM)
- Polyclonal multi-source seeding (pM)
- Inferred

\( (\mu, \gamma) = (5, 5) \)

\( 1 \) (Supp.)

- Resolved polytomy

- Resolved polytomy

- Inferred polytomy

\( \gamma_{min} \) and \( \mu_{min} \)