Course Announcements

HW 3 due Oct 29 by 11:59pm

Office hour after class in SC 3216
Outline

• Recap character-based phylogeny
• Application of small phylogeny maximum parsimony problem to cancer
• Compatibility

Reading:
• Lecture notes
Character-Based Tree Reconstruction

• Characters may be morphological features
  • Shape of beak \{generalist, insect catching, ...\}
  • Number of legs \{2, 3, 4, ..\}
  • Hibernation \{yes, no\}

• Character may be nucleotides/amino acids
  • \{A, T, C, G\}
  • 20 amino acids

• Values of a character are called states
  • We assume discrete states
Character-Based Phylogeny Reconstruction

**Input characters**  
**Output optimal tree**

**Question:** What is optimal?

**Want:** Optimization criterion

**Question:** How to optimize this criterion?

**Want:** Algorithm
Character-Based Phylogeny Reconstruction: Criterion

**Parsimony**: minimize number of changes on edges of tree

(a) *Parsimony Score=3*  
(b) *Parsimony Score=2*
Again, 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.
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 $v$.

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

\[
\mu(v, s) = \min \left\{ \begin{array}{ll}
\infty, & \text{if } v \in L(T) \text{ and } s \neq \sigma(v), \\
0, & \text{if } v \in L(T) \text{ and } s = \sigma(v), \\
\sum_{w \in \delta(v)} \min_{t \in \Sigma} \{c(s, t) + \mu(w, t)\}, & \text{if } v \notin L(T).
\end{array} \right.
\]

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

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

\[ \text{Fill}(T, r(T), \sigma, \Sigma) \]
\[ O(|\Sigma|^2) \]

\[
\begin{align*}
\text{if } v \in \mathcal{L}(T) & \text{ then} \\
\quad \text{for } s \in \Sigma & \\
\quad \quad \text{if } s = \sigma(v) & \text{ then} \\
\quad \quad \quad \mu(v, s) = 0 & \\
\quad \quad \text{else} & \mu(v, s) = \infty \\
\text{else} & \\
\quad \text{for } w \in \mathcal{S}(v) & \\
\quad \quad \text{Fill}(T, w, \sigma, \Sigma) & \\
\quad \quad \mu(v, s) = 0 & \\
\quad \text{for } w \in \mathcal{S}(v) & \\
\mu(v, s) & = \min_{t \in \Sigma} [c(s, t) + \mu(w, t)]
\end{align*}
\]

Backtrace \( (T, v, \mu) \)

\[
\begin{align*}
\text{if } v = r(T) & \\
\sigma(r(T)) & = \arg \min_{s \in \Sigma} \{\mu(r(T), s)\} \\
\text{else} & \\
\text{let } u & \text{ be the parent of } v \text{ and let } s \text{ be the state} \\
\sigma(v) & = \arg \min_{t \in \Sigma} \{c(s, t) + \mu(v, t)\} \\
\text{for } w \in \mathcal{S}(v) & \\
\text{Backtrace } (T, w, \mu)
\end{align*}
\]
Outline

• Recap character-based phylogeny
• Application of small phylogeny maximum parsimony problem to cancer
• Compatibility

Reading:
• Lecture notes
Tumorigenesis: (i) Cell Mutation

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

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

Cell Tree

primary tumor $P$

metastasis $M_2$

metastasis $M_1$
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

Cell Tree

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

**Goal:** Given phylogenetic tree $T$, find **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]

$\mu^* = 13$

$m = 7$ anatomical sites
Minimum Migration History is **Not** Unique

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

\[ \mu^* = 13 \]
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 $G$

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

Migration Graph $G$

Clone Tree $T$

Not necessarily true in the case of directed cycles

<table>
<thead>
<tr>
<th></th>
<th>ApC</th>
<th>Appendix</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFTB</td>
<td>Left Fallopian Tube</td>
<td></td>
</tr>
<tr>
<td>LOv</td>
<td>Left Ovary</td>
<td></td>
</tr>
<tr>
<td>RFTA</td>
<td>Right Fallopian Tube</td>
<td></td>
</tr>
<tr>
<td>ROv</td>
<td>Right Ovary</td>
<td></td>
</tr>
<tr>
<td>SBwl</td>
<td>Small Bowel</td>
<td></td>
</tr>
<tr>
<td>Om</td>
<td>Omentum</td>
<td></td>
</tr>
</tbody>
</table>
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 $G^+$

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

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

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

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

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

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

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

---

ApC  Appendix
LFTB  Left Fallopian Tube
LOv   Left Ovary
RFTA  Right Fallopian Tube
ROv   Right Ovary
SBwl  Small Bowel
Om    Omentum
Constrained Multi-objective Optimization Problem

**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 $\hat{\gamma}(T)$.

![Diagram showing single-source seeding (S), multi-source seeding (M), and reseeding (R).]
Theorem 1: PMH is NP-hard when $\mathcal{P} = \{S\}$.

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

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 $\gamma(T)$. 

**Figure a:** Diagram of a phylogenetic tree $T$ and leaf labeling $\hat{\ell}$. The migration graph $G_A$ is shown with single-source seeding $\mathcal{P} = \{S\}$, where $\mu^*(\hat{\ell}), \gamma(\hat{\ell}) = (6, 2)$. The vertex labeling $\ell_A$ is also shown for a specific migration pattern.
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 $\varphi$

$\Sigma = \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, c_1, \ldots c_k, \bot\}$
**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 \( \varphi \)

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

\[ \Sigma = \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, c_1, \ldots c_k, \bot\} \]
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 $\phi : [n] \to \{0, 1\}$ 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 \( \mathcal{P} = \{S\} \)**

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

\( k = 2, n = 3 \)

\( B = 10k + 2 = 22 \)

\( A = 2Bn + 27k + 1 = 187 \)

\[ \Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \bot\} \]

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

Then, \( \varphi \) is satisfiable if and only if \( \mu^*(T) = (B + 1)n + 25k \)

\[ \mu^*(T) = (B + 1)n + 25k = 23 \cdot 3 + 50 \cdot 2 = 119 \]
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 $\hat{G}$. 

---

![Phylogenetic tree $T$](image1)

![Migration tree $\hat{G}$](image2)

![Phylogenetic tree $T$](image3)
PMH is FPT in number $m$ of locations when $\mathcal{P} = \{S\}$

\[ O(nm^m) \] time

**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 $\hat{G}$. 
Simulations

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

• Recap character-based phylogeny
• Application of small phylogeny maximum parsimony problem to cancer
• Compatibility

Reading:
• Lecture notes
Maximum Parsimony

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.
## Binary Characters

Characters only have two possible states

<table>
<thead>
<tr>
<th>Species</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0 1 1 0 0</td>
</tr>
<tr>
<td>B</td>
<td>0 0 1 1 0</td>
</tr>
<tr>
<td>C</td>
<td>1 1 1 1 0</td>
</tr>
<tr>
<td>D</td>
<td>1 1 0 1 1</td>
</tr>
</tbody>
</table>

Possible Encoding:
- 0: not-mutated
- 1: mutated

Possible Encoding:
- 0: no wings
- 1: wings
### Binary Characters

<table>
<thead>
<tr>
<th>Species</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0 1 1 0 0</td>
</tr>
<tr>
<td>B</td>
<td>0 0 1 1 0</td>
</tr>
<tr>
<td>C</td>
<td>1 1 1 1 0</td>
</tr>
<tr>
<td>D</td>
<td>1 1 0 1 1</td>
</tr>
</tbody>
</table>

**Characters only have two possible states**

Possible Encoding:
- 0: not-mutated
- 1: mutated

Possible Encoding:
- 0: no wings
- 1: wings

**Question**: Given $n$ binary characters, what is the smallest parsimony score?
“typical tumor”: ~10 driver mutations
100’s – 1000’s of passenger mutations
Somatic Mutations and Cancer

“typical tumor”: ~10 driver mutations
100’s – 1000’s of passenger mutations
Progression of Somatic Mutations

Single nucleotide mutation
... CGTATTAG ...
... CGTCATTAG ...

0 = normal
1 = mutated

Root is the normal, founder cell and leaves are cells in tumor.
Progression of Somatic Mutations

Single nucleotide mutation
... CGTATTAG ...
... CGTCATTAG ...

0 = normal
1 = mutated

Root is the normal, founder cell and leaves are cells in tumor.

Infinite sites assumption: each locus mutates only once.
The genome is large

Mutations are rare

Infinite sites model: multiple mutations never occur at the same position

Mutated Loci

<table>
<thead>
<tr>
<th>Species (cancer cells)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: mutated</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0: not</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1: mutated</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0: not</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1: mutated</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0: not</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

All sites are bi-allelic: mutated or not.

[Kimura, 1969]
Two-state Perfect Phylogeny

Matrix $M \in \{0, 1\}^{n \times m}$ has $n$ taxa and $m$ characters

- Taxon $f$ has state 1 for character $c$ if $f$ possesses character $c$

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$c_4$</th>
<th>$c_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$r_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$r_3$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$r_4$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$r_5$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Definition

A perfect phylogeny for $M$ is a rooted tree $T$ with $n$ leaves such that:

1. Each taxon labels only one leaf
2. Each character labels only one edge
3. Character possessed by a taxon are on unique path to root

Root node is all zero ancestor
Two-state Perfect Phylogeny – Alternative Definitions

1. Each taxon labels exactly one leaf
2. Each character labels exactly one edge
3. Character possessed by a taxon are on unique path to root

1. Each taxon labels exactly one leaf
2. Each node is labeled by \( \{0, 1\}^m \)
3. Nodes labeled with state \( i \) for character \( c \) form a connected subtree

1. Each taxon labels exactly one leaf
2. \( T_c(i) \) is smallest subtree connecting all leaves labeled with state \( i \) for character \( c \)
3. \( T_c(0) \) and \( T_c(1) \) are disjoint for all \( c \)
Two-state Perfect Phylogeny Problem

**Input:**

Matrix $M \in \{0, 1\}^{n \times m}$ has $n$ taxa and $m$ characters

- Taxon $f$ has state 1 for character $c$  
  $\iff$ $f$ possesses character $c$

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$c_4$</th>
<th>$c_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$r_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$r_3$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$r_4$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$r_5$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Problem**

*Given* $M \in \{0, 1\}^{n \times m}$ *does* $M$ *have a perfect phylogeny?*
Try it yourself!

Only one of these matrices can be used to build a perfect phylogeny.

(1) As a group, **decide on an approach** to try to determine which one is which.
(2) Try out your approach to see if you can construct the tree.
(3) What did you learn from your attempt?

\[
\begin{array}{ccccc}
\text{Characters} & C_1 & C_2 & C_3 & C_4 & C_5 \\
\text{Species} & A & B & C & D & E \\
A & 0 & 1 & 0 & 0 & 0 \\
B & 0 & 0 & 1 & 0 & 0 \\
C & 1 & 1 & 0 & 0 & 0 \\
D & 0 & 0 & 1 & 1 & 0 \\
E & 1 & 1 & 0 & 0 & 1 \\
\end{array}
\]

\[
\begin{array}{ccccc}
\text{Characters} & C_1 & C_2 & C_3 & C_4 & C_5 \\
\text{Species} & A & B & C & D & E \\
A & 0 & 0 & 1 & 1 & 0 \\
B & 0 & 0 & 1 & 0 & 1 \\
C & 1 & 1 & 0 & 0 & 1 \\
D & 1 & 1 & 0 & 0 & 0 \\
E & 0 & 1 & 0 & 0 & 1 \\
\end{array}
\]
The Perfect Phylogeny Problem – Preliminaries

Problem

Given $M \in \{0, 1\}^{n \times m}$ does $M$ have a perfect phylogeny?

Definition

$I(c)$ is the set of taxa that possess character $c$; and $\sigma(f)$ is the set of characters possessed by taxon $f$.

\[
\begin{array}{c|cccccc}
 & c_1 & c_2 & c_3 & c_4 & c_5 \\
 r_1 & 1 & 1 & 0 & 0 & 0 \\
r_2 & 0 & 0 & 1 & 0 & 0 \\
r_3 & 1 & 1 & 0 & 0 & 1 \\
r_4 & 0 & 0 & 1 & 1 & 0 \\
r_5 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\Rightarrow
\begin{array}{c|cccccc}
 & c_1 & c_2 & c_3 & c_4 & c_5 \\
 r_1 & 1 & 1 & 0 & 0 & 0 \\
r_2 & 0 & 0 & 1 & 0 & 0 \\
r_3 & 1 & 1 & 0 & 1 & 0 \\
r_4 & 0 & 0 & 1 & 0 & 1 \\
r_5 & 1 & 0 & 0 & 0 & 0 \\
\end{array}
\]

$I(c_1) = \{r_1, r_3\}$

$\sigma(r_1) = \{c_1, c_2\}$

Sort columns of $M$ s.t. $c < d$ iff $|I(c)| \geq |I(d)|$. Break ties arbitrarily.
Consider rows of $M$ iteratively
  ▶ $T_i$ is tree of first $i$ rows of $M$

$T_1$ is a path graph
  ▶ Terminal nodes $r$ and 1
  ▶ $|\sigma(1)| + 1$ edges labeled by $\sigma(1)$

\[ c < d \text{ iff } |l(c)| \geq |l(d)| \]

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$c_4$</th>
<th>$c_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$r_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$r_3$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$r_4$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$r_5$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Consider rows of $M$ iteratively
  - $T_i$ is tree of first $i$ rows of $M$

$T_1$ is a path graph
  - Terminal nodes $r$ and $1$
  - $|\sigma(1)| + 1$ edges labeled by $\sigma(1)$

$T_{i+1}$ is a supertree of $T_i$
  - Let $v$ be last node on walk from $r$
    matching characters $\sigma(i + 1)$
    - Character $d$ is the last match
    - Unmatched characters $\tau(i + 1)$

\[
\begin{array}{cccccc}
  & c_1 & c_2 & c_3 & c_4 & c_5 \\
r_1 & 1 & 1 & 0 & 0 & 0 \\
r_2 & 0 & 0 & 1 & 0 & 0 \\
r_3 & 1 & 1 & 0 & 1 & 0 \\
r_4 & 0 & 0 & 1 & 0 & 1 \\
r_5 & 1 & 0 & 0 & 0 & 0 \\
\end{array}
\]
Consider rows of $M$ iteratively
  - $T_i$ is a tree of first $i$ rows of $M$

$T_1$ is a path graph
  - Terminal nodes $r$ and 1
  - $|\sigma(1)| + 1$ edges labeled by $\sigma(1)$

$T_{i+1}$ is a supertree of $T_i$
  - Let $v$ be last node on walk from $r$
    - matching characters $\sigma(i+1)$
      - Character $d$ is the last match
      - Unmatched characters $\tau(i+1)$
  - Extend $T_i$ with path $\Pi$
    - $\Pi$ has terminals $v$ and $i+1$
    - $\Pi$ has $|\tau(i+1)| + 1$ edges labeled by $\tau(i+1)$

$c < d$ iff $|I(c)| \geq |I(d)|$
Consider rows of $M$ iteratively
  - $T_i$ is tree of first $i$ rows of $M$

$T_1$ is a path graph
  - Terminal nodes $r$ and 1
  - $|\sigma(1)| + 1$ edges labeled by $\sigma(1)$

$T_{i+1}$ is a supertree of $T_i$
  - Let $v$ be last node on walk from $r$
    matching characters $\sigma(i + 1)$
    - Character $d$ is the last match
    - Unmatched characters $\tau(i + 1)$
  - Extend $T_i$ with path $\Pi$
    - $\Pi$ has terminals $v$ and $i + 1$
    - $\Pi$ has $|\tau(i + 1)| + 1$ edges labeled by $\tau(i + 1)$

\[
c < d \iff |I(c)| \geq |I(d)|\]
Consider rows of $M$ iteratively
- $T_i$ is tree of first $i$ rows of $M$

$T_1$ is a path graph
- Terminal nodes $r$ and $1$
- $|\sigma(1)| + 1$ edges labeled by $\sigma(1)$

$T_{i+1}$ is a supertree of $T_i$
- Let $v$ be last node on walk from $r$
  - matching characters $\sigma(i+1)$
    - ★ Character $d$ is the last match
    - ★ Unmatched characters $\tau(i+1)$
- Extend $T_i$ with path $\Pi$
  - ★ $\Pi$ has terminals $v$ and $i+1$
  - ★ $\Pi$ has $|\tau(i+1)| + 1$ edges labeled by $\tau(i+1)$

Lemma

Let $M_i \in 0,1^{i \times m}$ be a submatrix of $M$. If $M$ is conflict-free then $T_i$ is a perfect phylogeny for $M_i$. 

\[
c < d \iff |I(c)| \geq |I(d)|\]
Outline

• Recap character-based phylogeny
• Application of small phylogeny maximum parsimony problem to cancer
• Compatibility

Reading:
• Lecture notes

HW 3 due Oct 29 by 11:59pm