CS 466
Introduction to Bioinformatics
Lecture 14

Mohammed El-Kebir

October 24, 2018

Course Announcements

HW 3 due Oct 29 by 11:59pm

Outline

- Recap additive distance
- Neighbor joining
- Character-based phylogeny (small)
- Application to cancer

Reading:

- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner

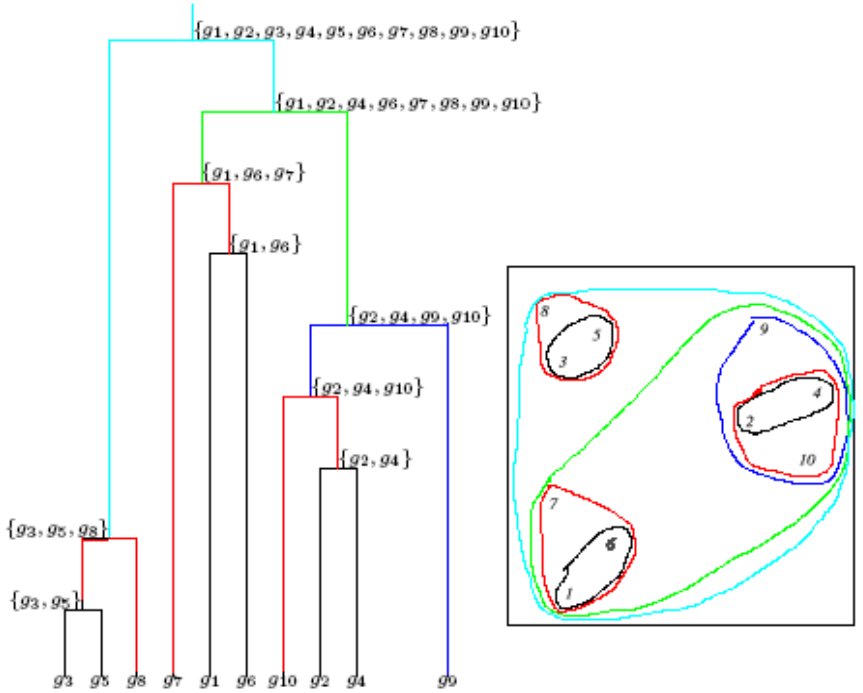
Hierarchical Clustering

1. Hierarchical Clustering (***D*** , *n*)
2. Form *n* clusters each with one element
3. Construct a graph ***T*** by assigning one vertex to each cluster
4. **while** there is more than one cluster
5. Find the two closest clusters C_1 and C_2
6. Merge C_1 and C_2 into new cluster *C* with $|C_1| + |C_2|$ elements
7. **Compute distance from *C* to all other clusters**
8. Add a new vertex ***C*** to ***T*** and connect to vertices C_1 and C_2
9. Remove rows and columns of ***D*** corresponding to C_1 and C_2
10. Add a row and column to ***D*** corresponding to the new cluster ***C***
11. return ***T***

Definition of distance between clusters (or, linkage criterion) affects clustering!

Organize elements into a tree s.t.:

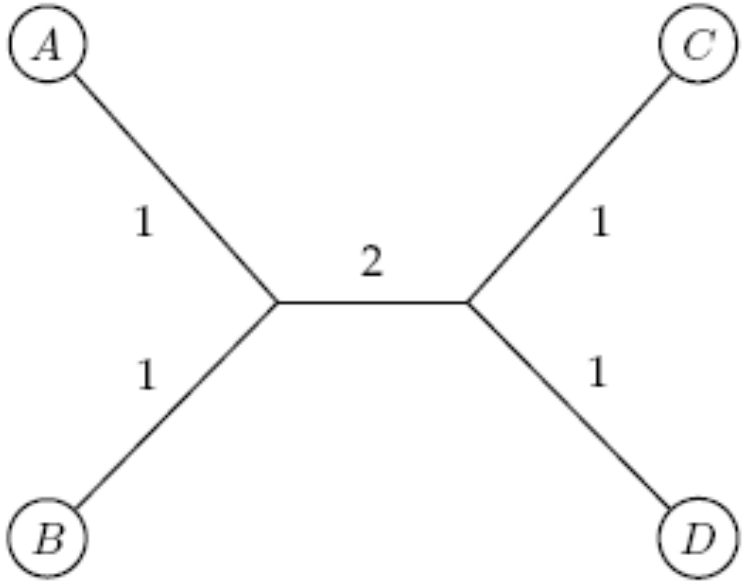
- Leaves are elements
- Paths between leaves represent pairwise element distance
- Similar elements lie within same subtrees



Additive Distance Matrices

Matrix *D* is ADDITIVE if there exists a tree *T* with $d_{ij}(T) = D_{ij}$

	A	B	C	D
A	0	2	4	4
B	2	0	4	4
C	4	4	0	2
D	4	4	2	0



NON-ADDITIVE otherwise

	A	B	C	D
A	0	2	2	2
B	2	0	3	2
C	2	3	0	2
D	2	2	2	0

?

This is a constructive definition

A Small and a Large Problem

Small Additive Distance Phylogeny Problem:

Given $n \times n$ distance matrix $D = [d_{i,j}]$ and unweighted tree T with n leaves, determine edge weights such that $d_T(i,j) = d_{i,j}$

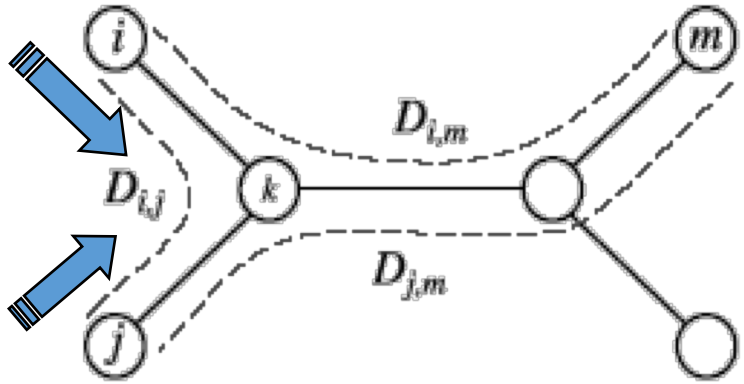
Large Additive Distance Phylogeny Problem:

Given $n \times n$ distance matrix $D = [d_{i,j}]$, find tree T with n leaves **and** edge weights such that $d_T(i,j) = d_{i,j}$

Both problems can be solved in polynomial time

Small Additive Distance Problem

1. Find neighboring leaves i and j with parent k
2. Remove the rows and columns of i and j
3. Add a new row and column corresponding to k, where the distance from k to any other leaf m is computed as

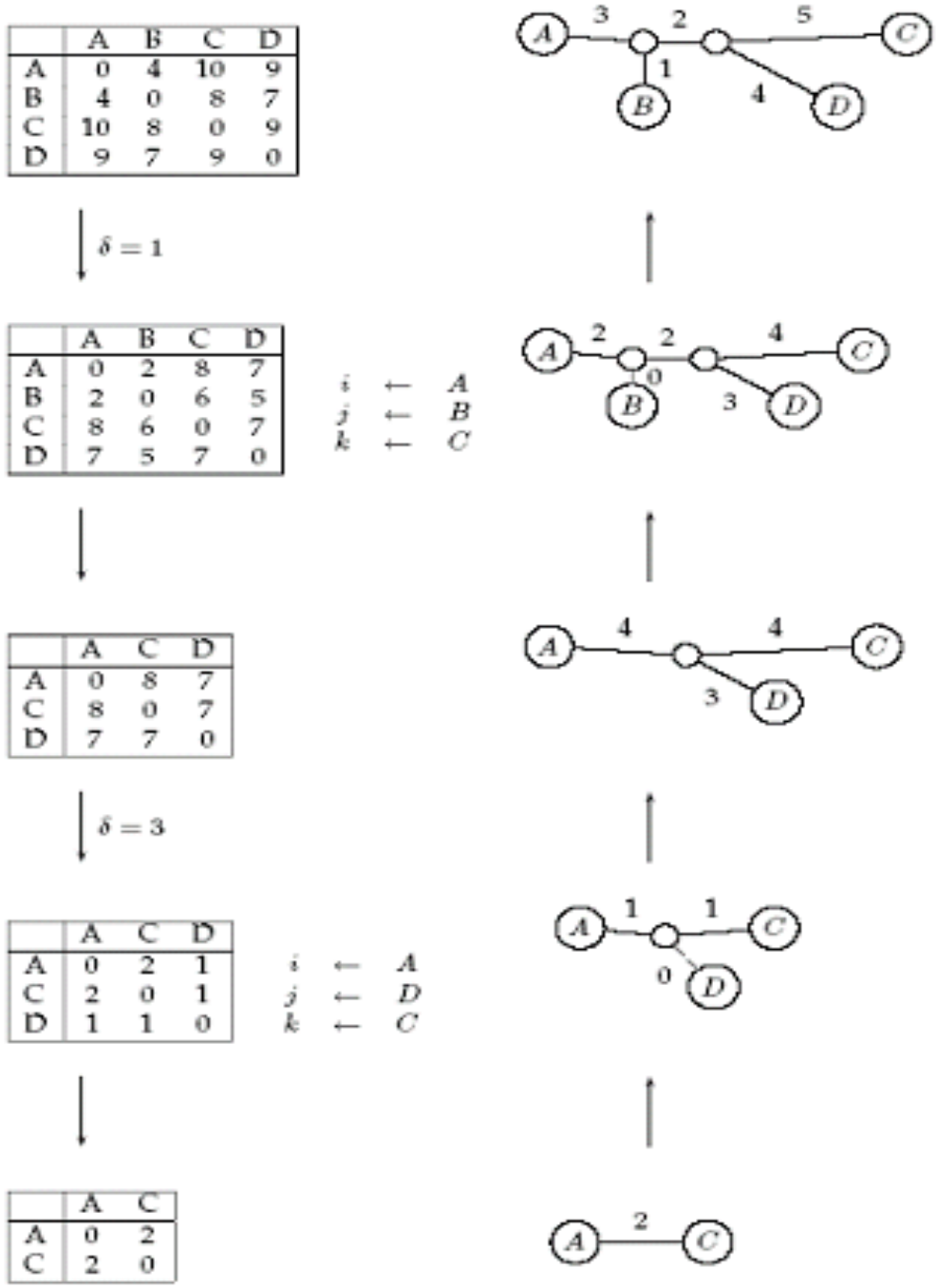


$$d_{k,m} = \frac{(d_{i,m} + d_{j,m} - d_{i,j})}{2}$$

4. Repeat steps 1-3 until tree has only two vertices

Additive Phylogeny

AdditivePhylogeny(D)
  if D is a 2 x 2 matrix
    T = tree of a single edge of length $D_{1,2}$
    return T
  if D is non-degenerate
    Compute trimming parameter δ
    Trim(D, δ)
 Find a triple i, j, k in D such that $D_{ij} + D_{jk} = D_{ik}$
  $x = D_{ij}$
  Remove j^{th} row and j^{th} column from D
  T = AdditivePhylogeny(D).
Add a new vertex v to T at distance x from i to k
  Add j back to T by creating an edge (v,j) of length 0
  for every leaf l in T
    if distance from l to v in the tree $\neq D_{l,j}$
      output "matrix is not additive"
      return
  Extend all "hanging" edges by length δ
  return T



Additive Distance Matrix

Four point condition of matrix $D = [d_{i,j}]$:

Every four leaves (quartet) can be labeled as (i, j, k, l) such that

$$d_{i,j} + d_{k,l} \leq d_{i,k} + d_{j,l} = d_{i,l} + d_{j,k}$$

Theorem: Let D be an $n \times n$ matrix. The following statements are equivalent.

1. Matrix D is additive.
2. There exists a unique tree T (modulo isomorphism) s.t. $d_{i,j} = d_T(i, j)$ for all $(i, j) \in n^2$.
3. Four point condition holds for every quartet $(i, j, k, l) \in [n]^4$.

Outline

- Recap additive distance
- Neighbor joining
- Character-based phylogeny (small)
- Application to cancer

Reading:

- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner

Distance Based Phylogeny Problem

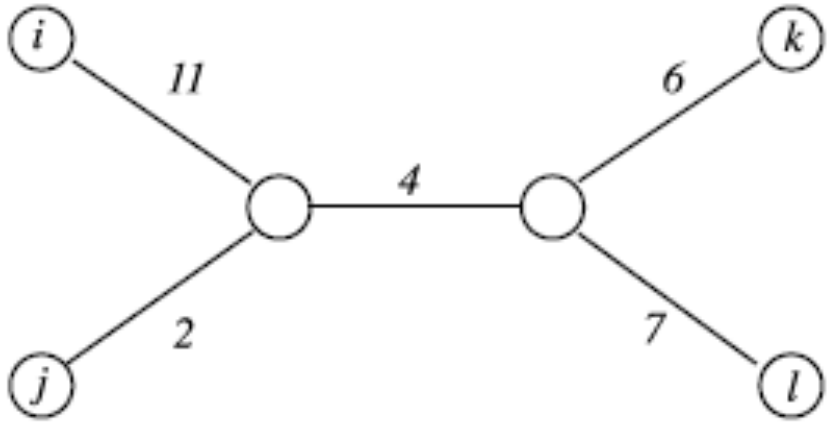
Large Additive Distance Phylogeny Problem:

Given $n \times n$ matrix $D = [d_{i,j}]$, find tree T with n leaves **and** edge weights such that $\max_{(i,j) \in [n]^2} |d_T(i,j) - d_{i,j}|$ is minimum.

Equivalently, find additive matrix D' closest to input matrix D

Neighbor Joining Algorithm (Saitou and Nei 1987)

- Constructs binary unrooted trees.
- Recall: leaves a and b are neighbors if they have a common parent
- Recall: closest leaves are not necessarily neighbors
- NJ: Find pair of leaves that are "close" to each other but "far" from other leaves

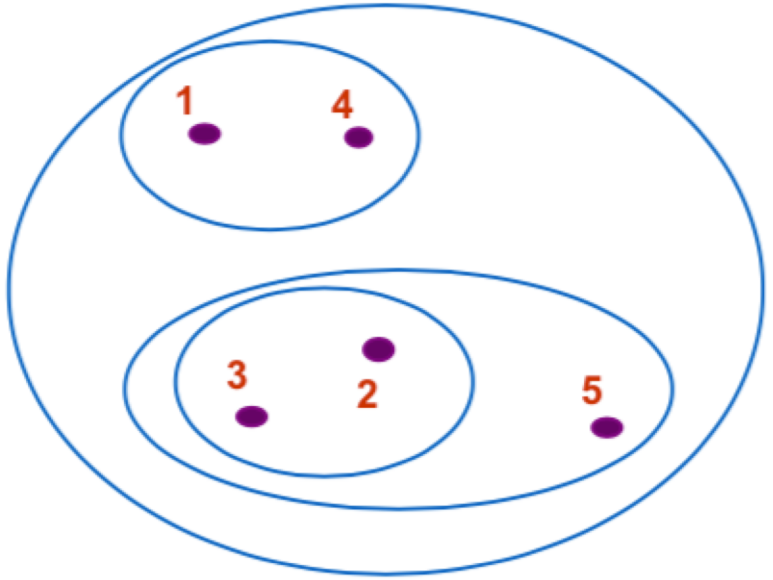


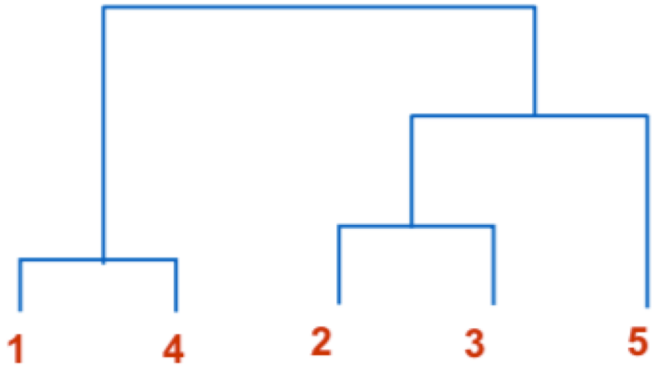
Two advantages: (1) reproduces correct tree for additive matrix, and (2) otherwise gives good approximation of correct tree

Distance Trees as Hierarchical Clustering

Leaves = Data points.

Data points clustered/grouped into hierarchy according to some distance criterion.

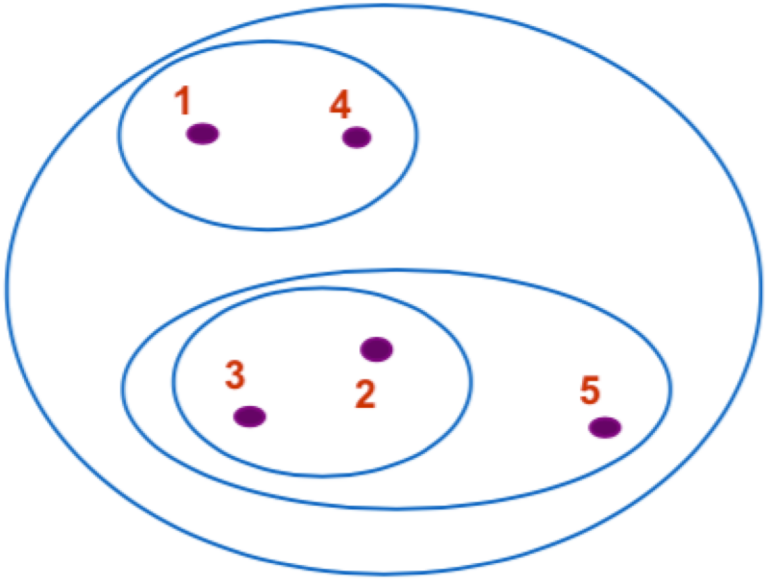


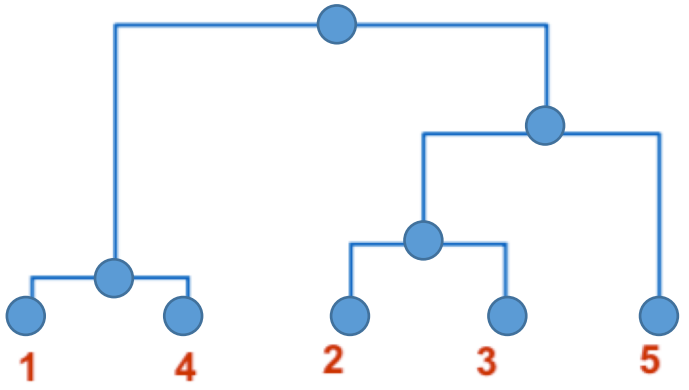


Distance Trees as Hierarchical Clustering

Leaves = Data points.

Data points clustered/grouped into hierarchy according to some distance criterion.

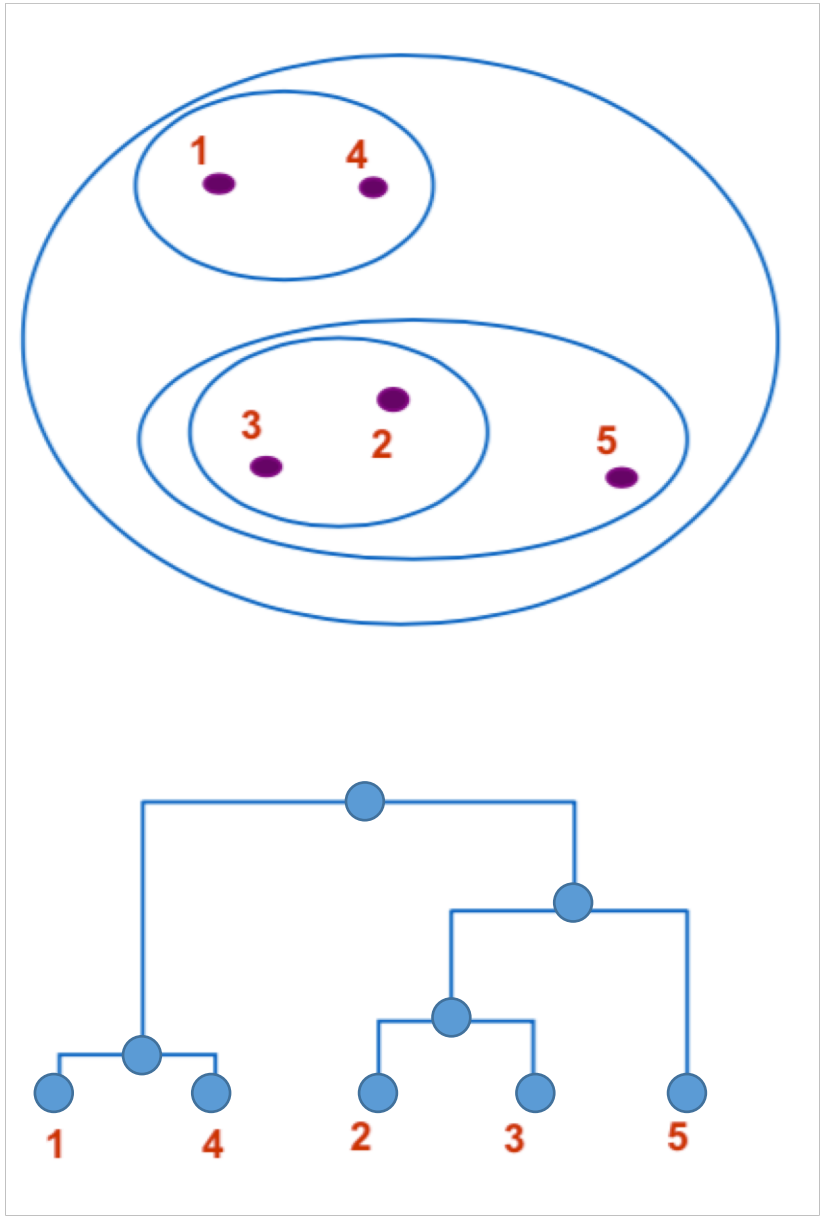




Distance Trees as Hierarchical Clustering

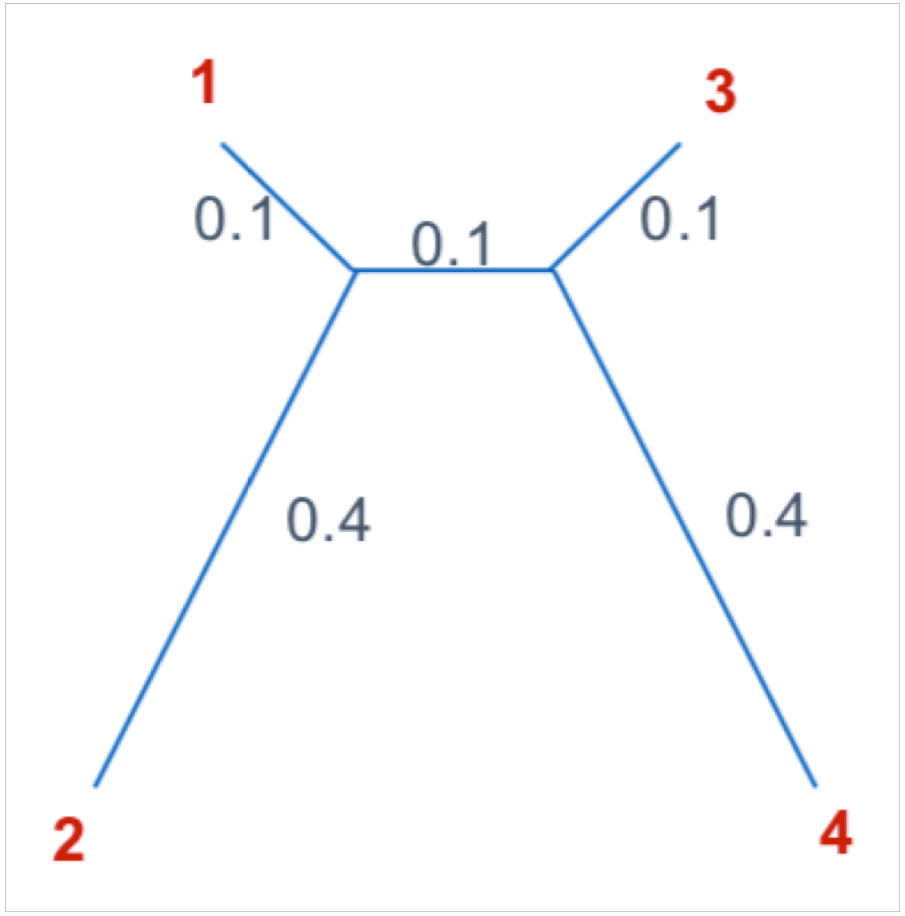
1. Hierarchical Clustering (***D*** , *n*)
2. Form *n* clusters each with one element
3. Construct a graph ***T*** by assigning one vertex to each cluster
4. **while** there is more than one cluster
5. Find the two closest clusters C_1 and C_2
6. Merge C_1 and C_2 into new cluster *C* with $|C_1| + |C_2|$ elements
7. **Compute distance from *C* to all other clusters**
8. Add a new vertex ***C*** to ***T*** and connect to vertices C_1 and C_2
9. Remove rows and columns of ***D*** corresponding to C_1 and C_2
10. Add a row and column to ***D*** corresponding to the new cluster ***C***
11. return ***T***

Selection criterion: distance between clusters affects clustering!



Neighbor Joining: Selection Criterion

$u_1 = D(1,3) + D(1,2) + D(1,c_1)$



Let $\boldsymbol{C} = \{1, \ldots, n\}$ be current clusters/leaves.

Define: $u_i = \sum_k D(i, k)$.

Intuitively, u_i measures separation of i from other leaves.

Goal: Minimize $D(i, j)$ and maximize $u_i + u_j$.

homogeneity — seperation

Solution: Find pair (i, j) that minimizes:

$$S_D(i, j) = (n - 2)\, D(i, j) - u_i - u_j$$

Claim: Given additive matrix D.

$S_D(x, y) = \min S_D(i, j)$ if and only if x and y are neighbors in tree T with $d_T = D$.

Neighboring Joining: Algorithm

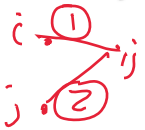
Initialization:

Form n clusters $C_1, C_2, \ldots, C_n$, one for each leaf node.

Define tree T to be the set of leaf nodes, one per sequence.

Iteration: (D is $m \times m$)

Pick i, j such that $S_D(i, j) = (m - 2)\, D(i, j) - u_i - u_j$ is minimal.



Merge i and j into new node $[ij]$ in T.

Assign length $\frac{1}{2}\,(D(i, j) + 1/(m\text{-}2)\,(u_i - u_j))$ to edge $(i, [ij])$

Assign length $\frac{1}{2}\,(D(i, j) + 1/(m\text{-}2)\,(u_j - u_i))$ to edge $(j, [ij])$

Remove rows and columns from D corresponding to i and j.

Add row and column to D for new vertex $[ij]$.

Set $D([ij], m) = \frac{1}{2}\,[D(i, m) + D(j, m) - D(i, j)]$

Termination:

When only one cluster

Neighboring Joining: Example

Initialization:

Form n clusters $C_1, C_2, \ldots, C_n$, one for each leaf node.

Define tree T to be the set of leaf nodes, one per sequence.

Iteration: (D is $m \times m$)

Pick i, j such that $S_D(i, j) = (m - 2)\, D(i, j) - u_i - u_j$ is minimal.

Merge i and j into new node $[ij]$ in T.

Assign length $\frac{1}{2}\,(D(i, j) + 1/(m\text{-}2)\,(u_i - u_j))$ to edge $(i, [ij])$

Assign length $\frac{1}{2}\,(D(i, j) + 1/(m\text{-}2)\,(u_j - u_i))$ to edge $(j, [ij])$

Remove rows and columns from D corresponding to i and j.

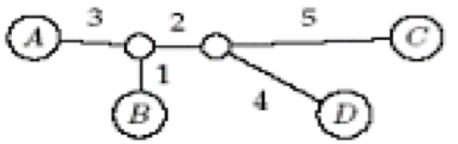
Add row and column to D for new vertex $[ij]$.

Set $D([ij], m) = \frac{1}{2}\,[D(i, m) + D(j, m) - D(i,j)]$

Termination:

When only one cluster

	A	B	C	D
A	0	4	10	9
B	4	0	8	7
C	10	8	0	9
D	9	7	9	0



Advantages of Neighbor Joining

Theorem: Let D be an $n \times n$ matrix. If matrix D is additive then neighbor joining produces the unique phylogenetic tree T (modulo isomorphism) such that $d_{i,j} = d_T(i,j)$ for all $(i,j) \in n^2$.

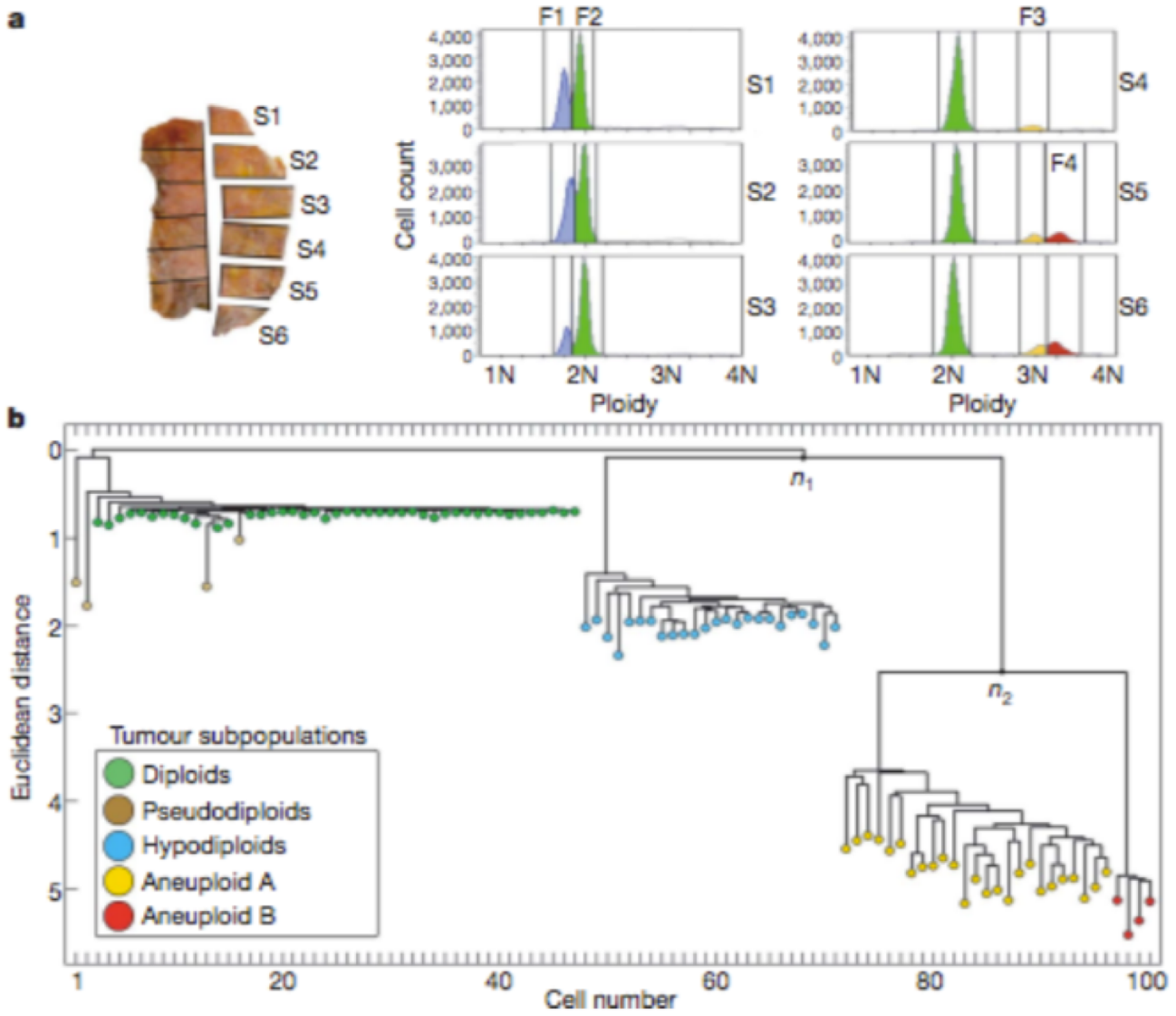
$\max_{(i,j)} \cdot |d_{ij} - d'_{ij}|$

Theorem: Let D be an $n \times n$ matrix. If there exists an additive matrix D' such that $|D - D'|_\infty \leq 0.5$ then neighbor joining applied to D reconstructs the unique tree T (modulo isomorphism) such that $d'_{i,j} = d_T(i,j)$ for all $(i,j) \in n^2$.

Is this normalized.

Atteson 1991

Neighbor Joining in Practice



[Navin et al, Nature 2011]

Neighbor Joining tree relating copy number profiles from single cells in a tumor.

Outline

- Recap additive distance
- Neighbor joining
- Character-based phylogeny (small)
- Application to cancer

Reading:

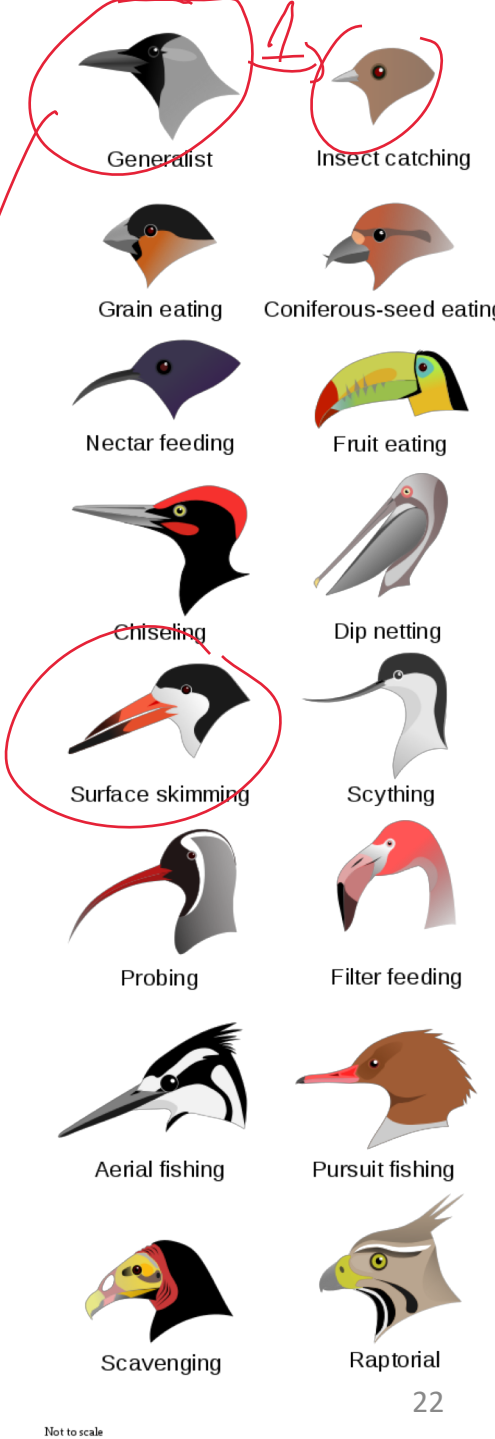
- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner

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

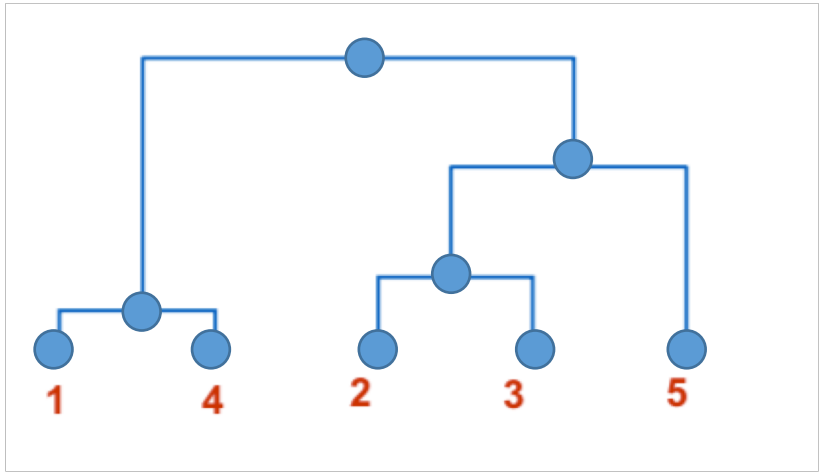


Character-Based Phylogeny Reconstruction

Input characters → Output optimal tree

Question: What is optimal?

Want: Optimization criterion



Character-Based Phylogeny Reconstruction

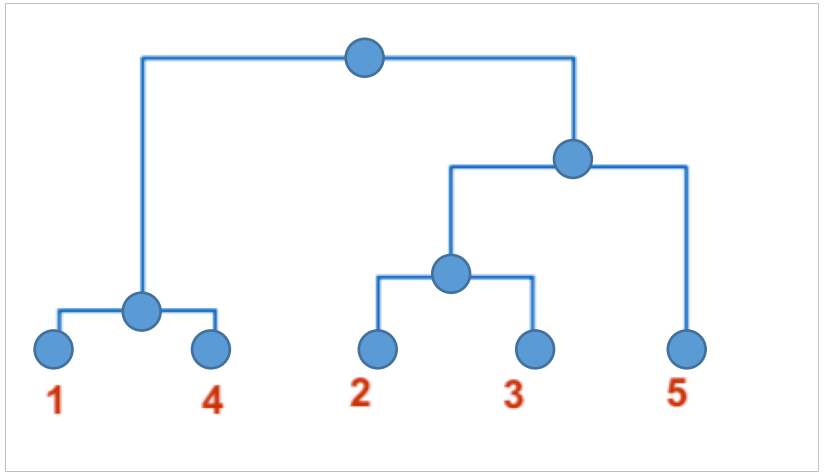
Input characters → Output optimal tree

Question: What is optimal?

Want: Optimization criterion

Question: How to optimize this criterion?

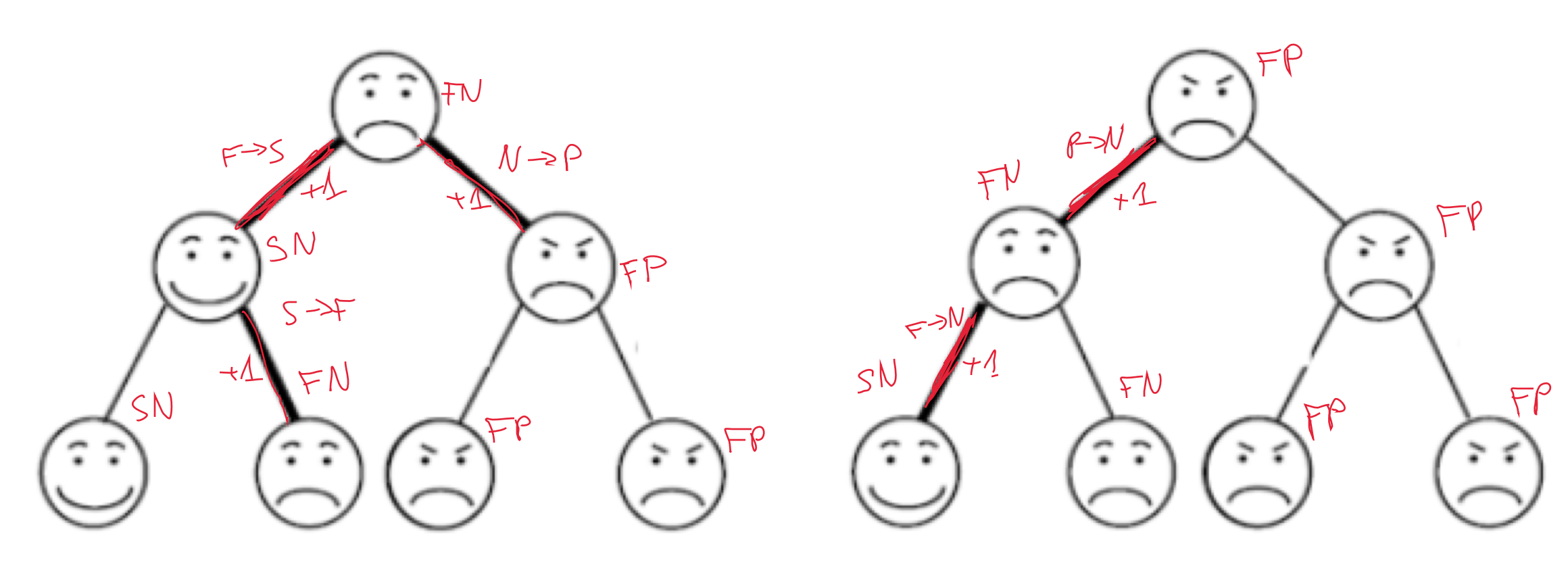
Want: Algorithm



Character-Based Phylogeny Reconstruction: Input

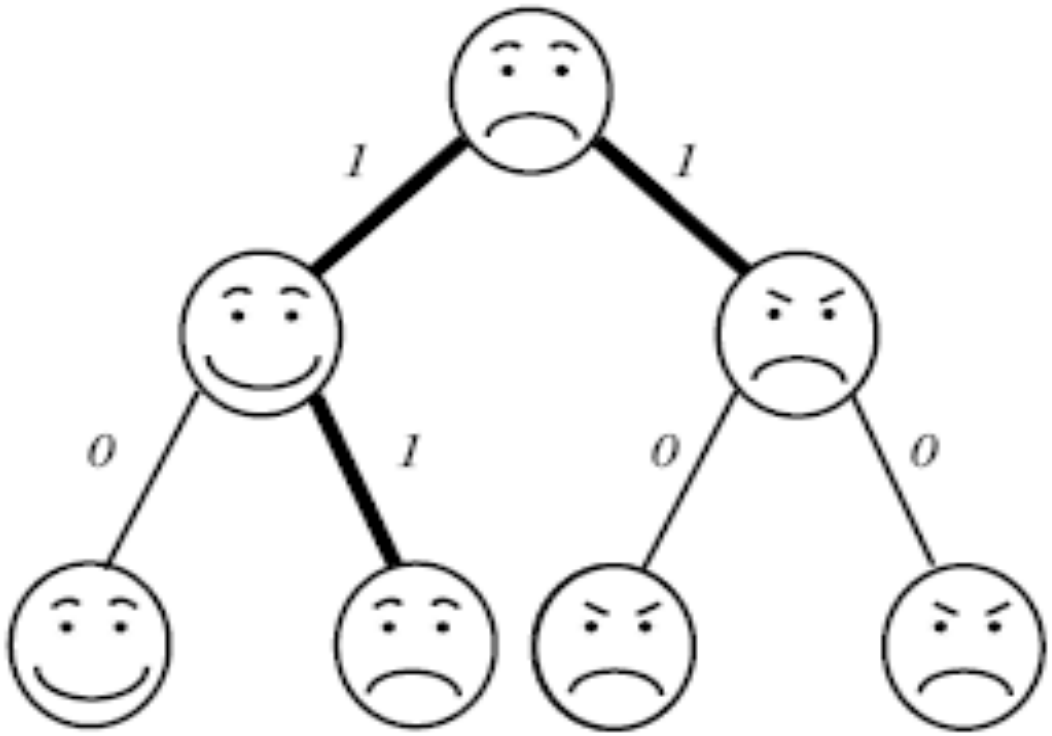
Characters / states	State 1	State 2
Mouth M	Smile S	Frown F
Eyebrows E	Normal N	Pointed P

Character-Based Phylogeny Reconstruction: Criterion

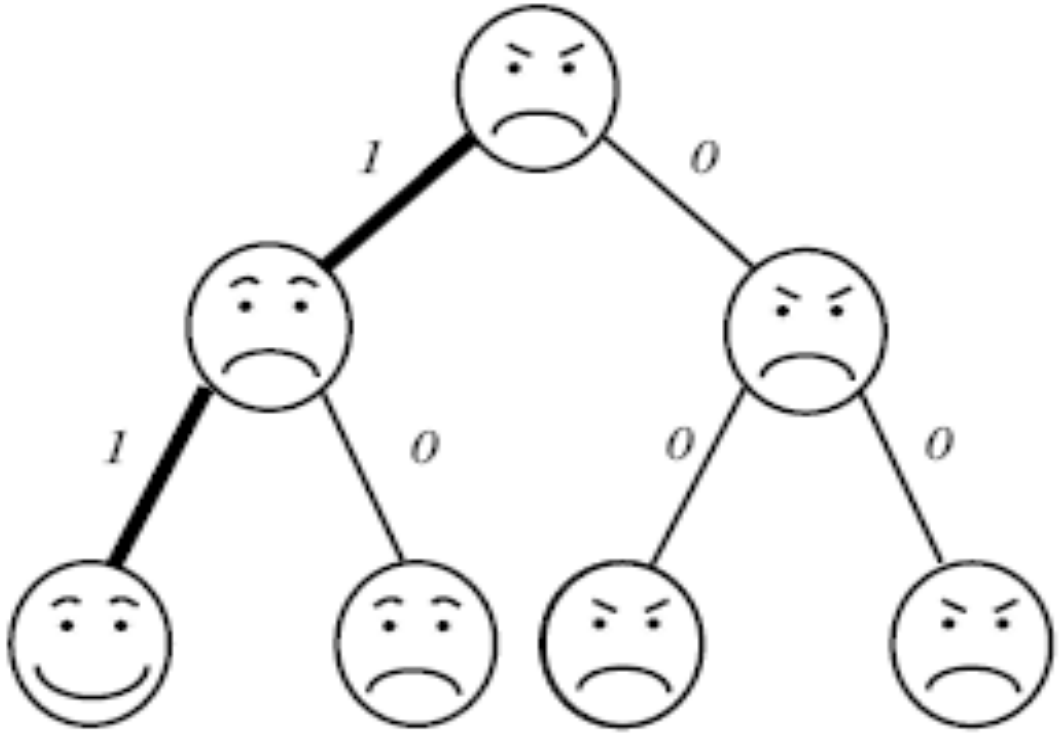


Question: Which tree is better?

Character-Based Phylogeny Reconstruction: Criterion



(a) *Parsimony Score=3*

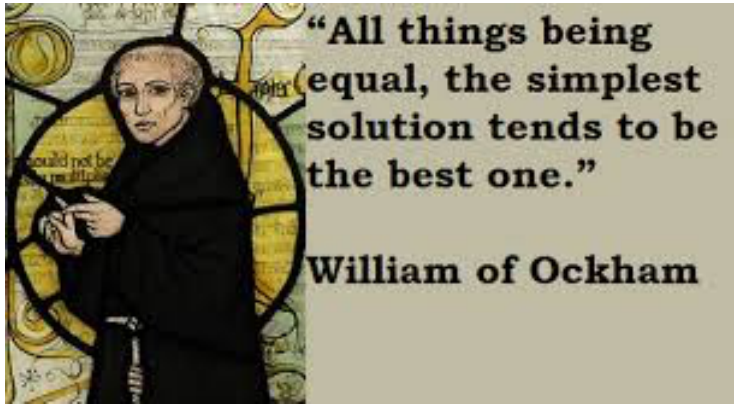


(b) *Parsimony Score=2*

Parsimony: minimize number of changes on edges of tree

Why Parsimony?

- Ockham's razor: "simplest" explanation for data
- Assumes that observed character differences resulted from the fewest possible mutations
- Seeks tree with the lowest **parsimony score**, i.e. the sum of all (costs of) mutations in the tree.



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 ~~Additive~~ Parsimony Distance 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)?

Again, a Small and a Large Problem

Small Maximum Parsimony Phylogeny Problem:

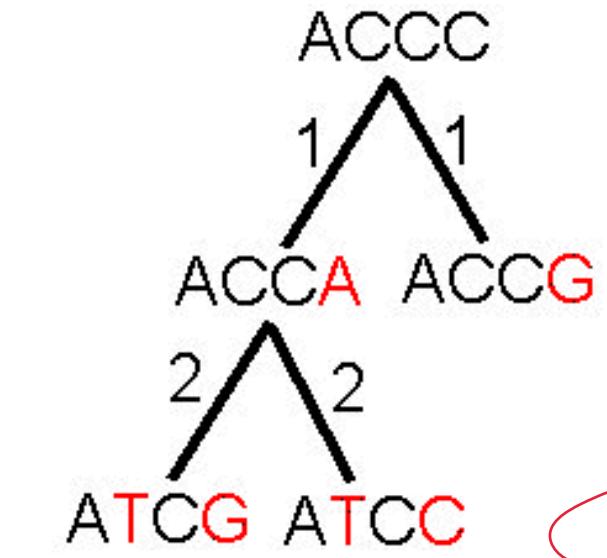
Given $m \times n$ matrix $A = [a_{i,j}]$ and tree T with m leaves, find assignment of character states to each internal vertex of T with minimum parsimony score.

Large Additive Distance Phylogeny Problem:

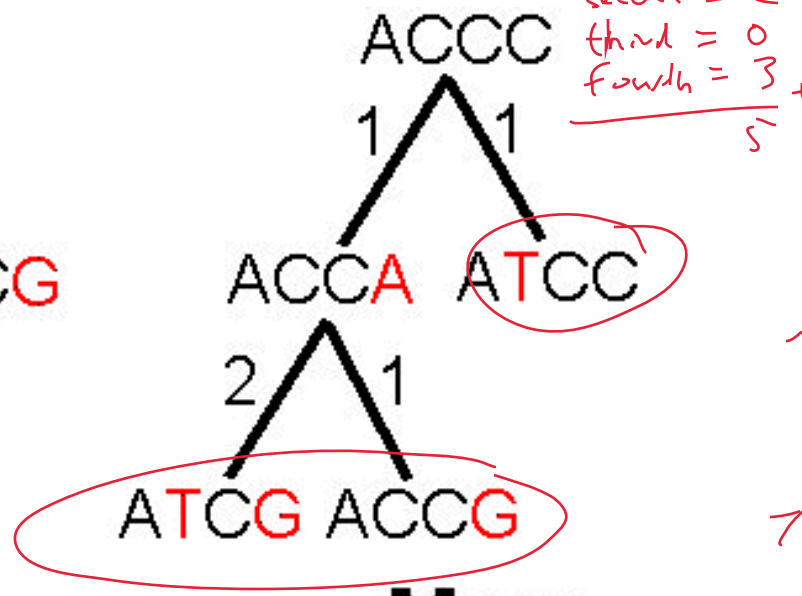
Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree T with m leaves labeled according to A and an assignment of character states to each internal vertex of T with minimum parsimony score.

Question: Are both problems easy (i.e. in P)?

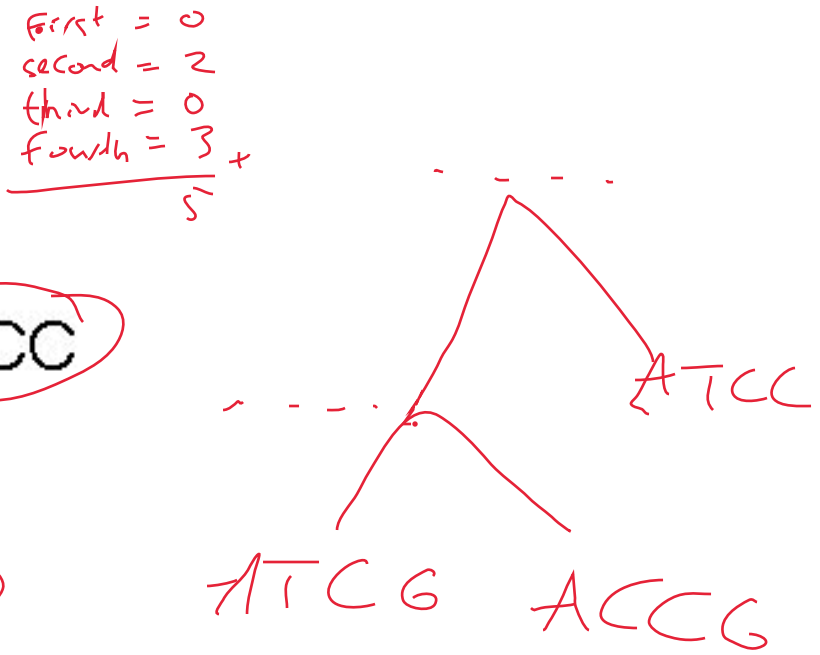
Small Maximum Parsimony Phylogeny Problem



Less Parsimonious
Score: 6



More Parsimonious
Score: 5



Question: There are $n = 4$ characters in the $m = 2$ taxa (leaves).
Can we solve each character separately?

Recurrence

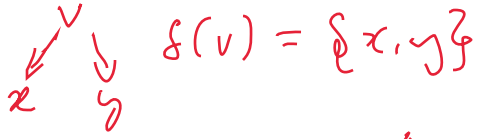
Σ set of states

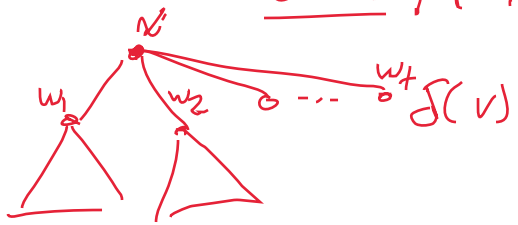
Given tree T w.th $|L(T)|$ leaves

$\sigma : L(T) \to \Sigma$

$\delta(v)$ is the set of children of v

$r(T)$ root of T





(1) $\mu : V(T) \times \Sigma \to \mathbb{N}$

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

(2) Min score is $\min_{s \in \Sigma} \{\mu(r(T), s)\}$

$\mu(r(T), s)$ is min. number of mutations when assigning s to $r(T)$ in the subtree rooted at $r(T)$

Base:

$$\mu(v, s) = \begin{cases} 0, & \text{if } v \in L(T) \text{ and } s = \sigma(v), \\ \infty, & \text{if } v \in L(T) \text{ and } s \neq \sigma(v). \end{cases}$$

Step $v \notin L(T)$

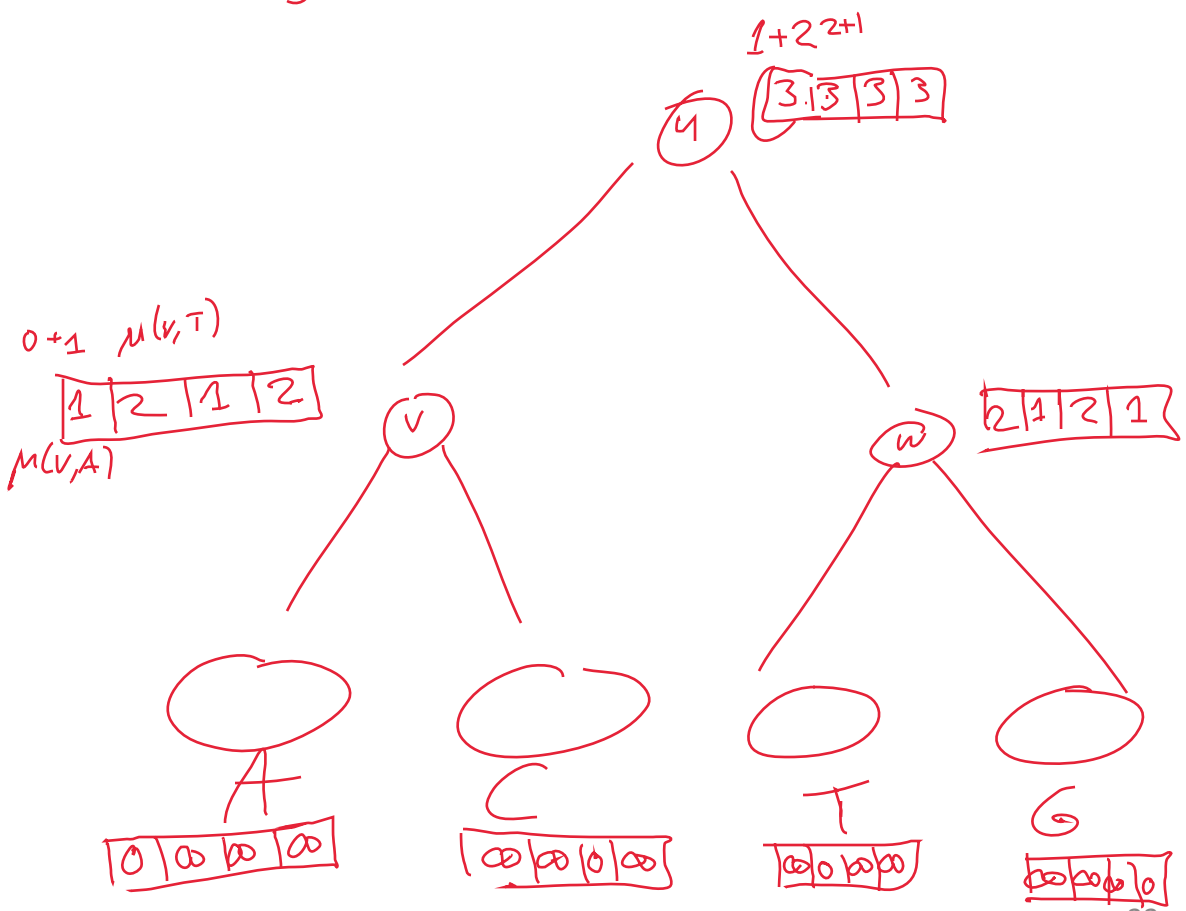
$$\mu(v, s) = \sum_{w \in \delta(v)} \min_{t \in \Sigma} \{c(s,t) + \mu(w,t)\}$$

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

Solving the Recurrence

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

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



~~Example~~ Pseudocode

$|L(T)| = m$

Filling out μ

Fill$(T, r(T), \sigma, \Sigma)$ $O(m|\Sigma|^2)$

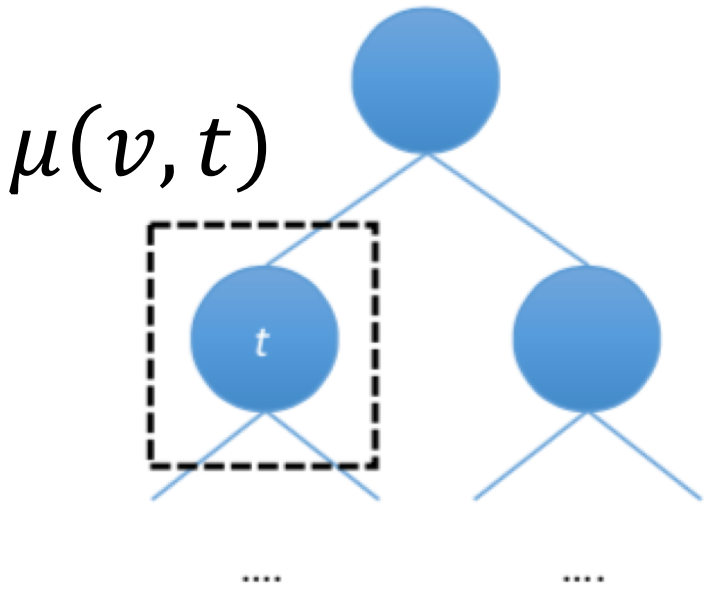
```
Fill(T, v, σ, Σ)
if v ∈ L(T) then
    For s ∈ Σ
        if s = σ(v) then
            μ(v,s) = 0
        else μ(v,s) = ∞
else
    For w ∈ δ(v)                    // children
        Fill(T, w, σ, Σ)
    μ(v,s) = 0
    For w ∈ δ(v)
        μ(v,s) += min_{t∈Σ} {c(s,t) + μ(w,t)}
```

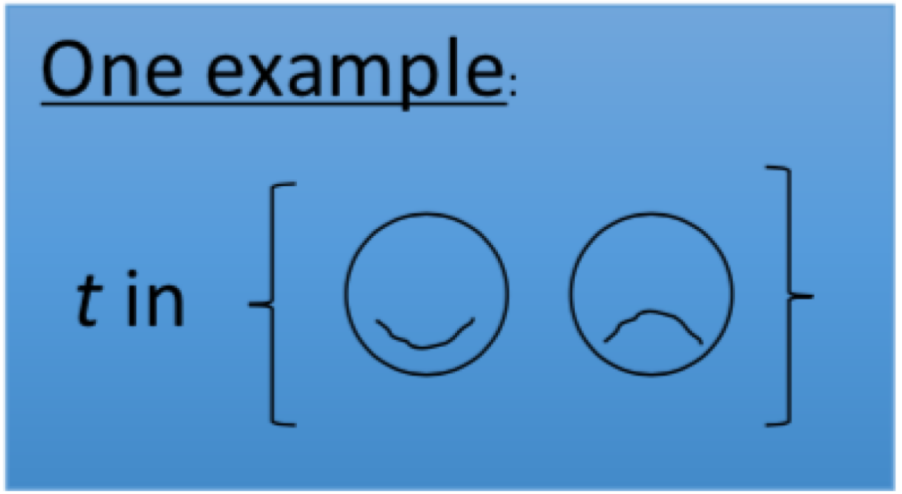
Backtrace(T, v, μ)

```
if v = r(T)
    σ(r(T)) = arg min_{s∈Σ} {μ(r(T), s)}      = σ(v)
else let u be the parent of v and let s be the state of u
    σ(v) = arg min_{t∈Σ} {c(s,t) + μ(v,t)}
For w ∈ δ(v)
    Backtrace(T, w, μ)
```

Sankoff Algorithm (Sankoff 1975)

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.





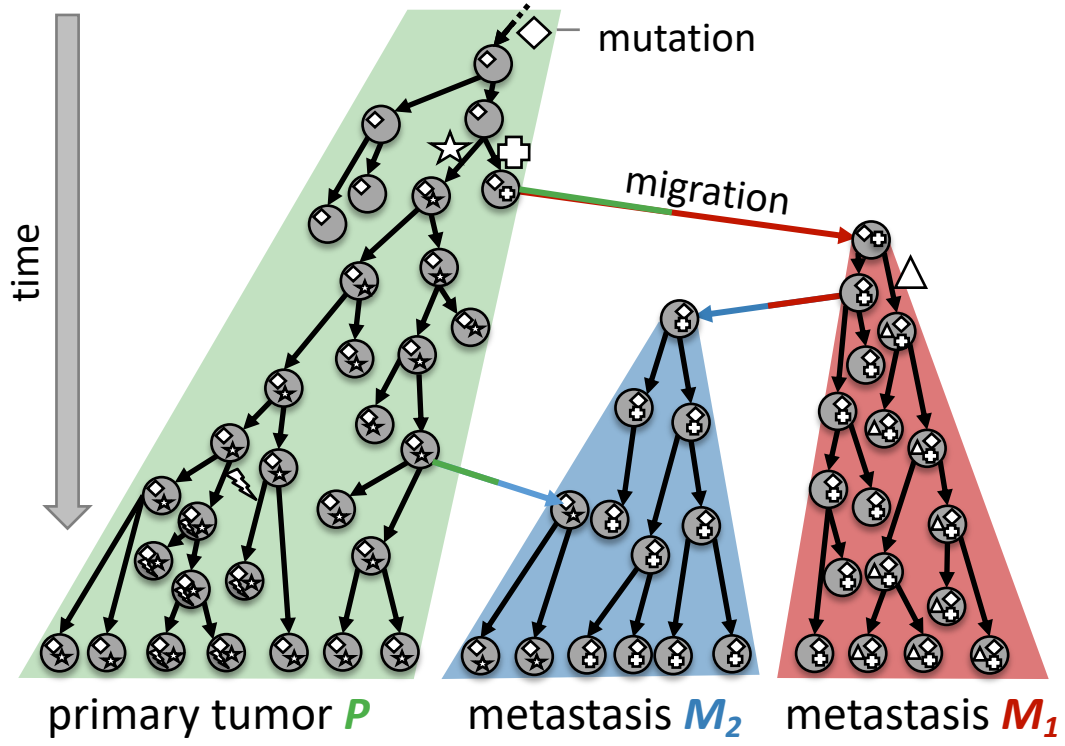
Outline

- Recap additive distance
- Neighbor joining
- Character-based phylogeny (small)
- Application to cancer

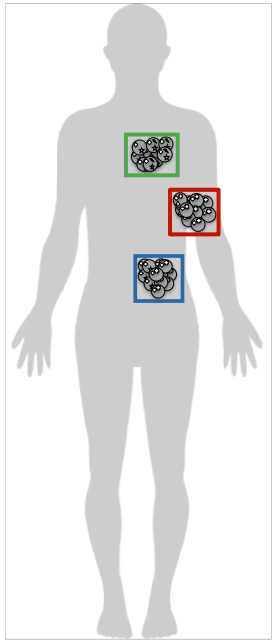
Reading:

- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner

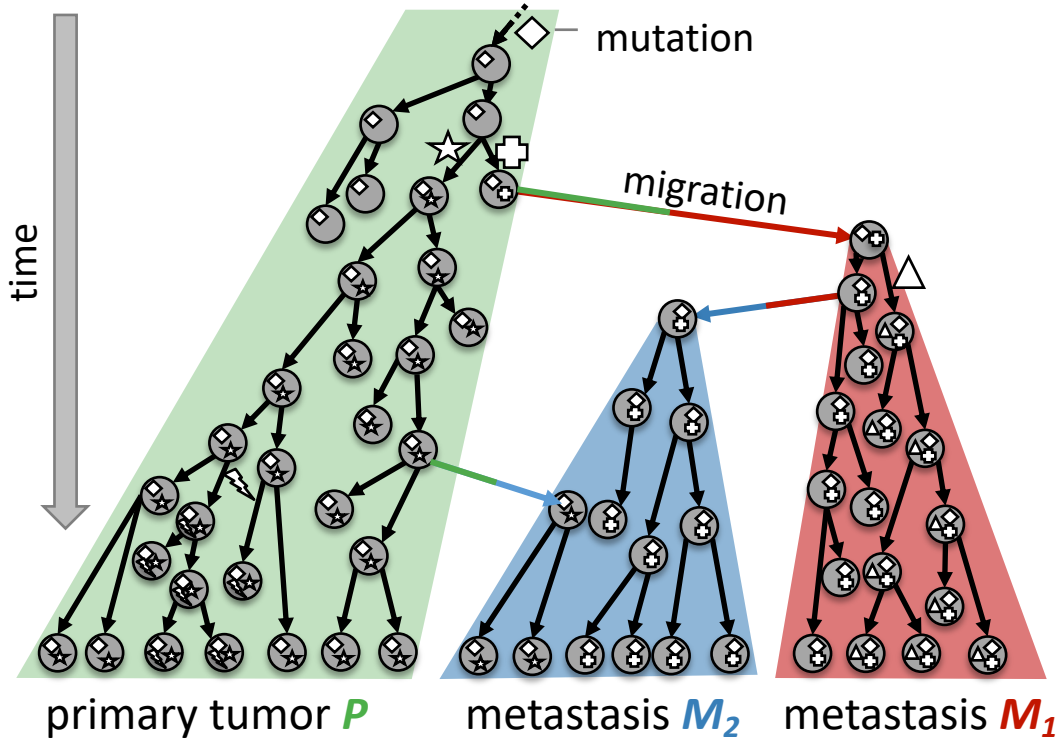
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



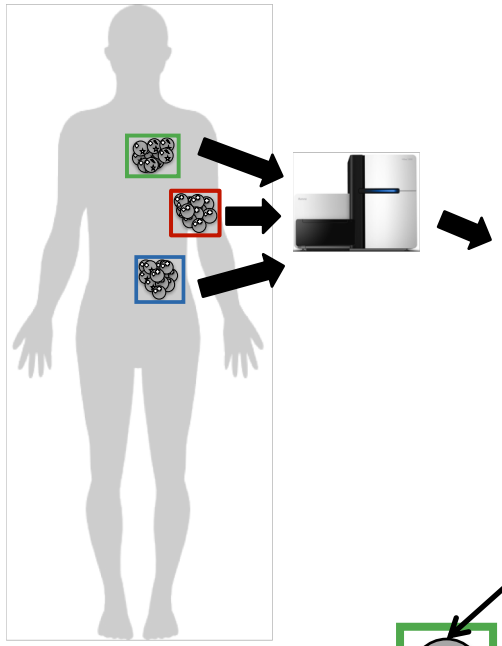
Cell Tree

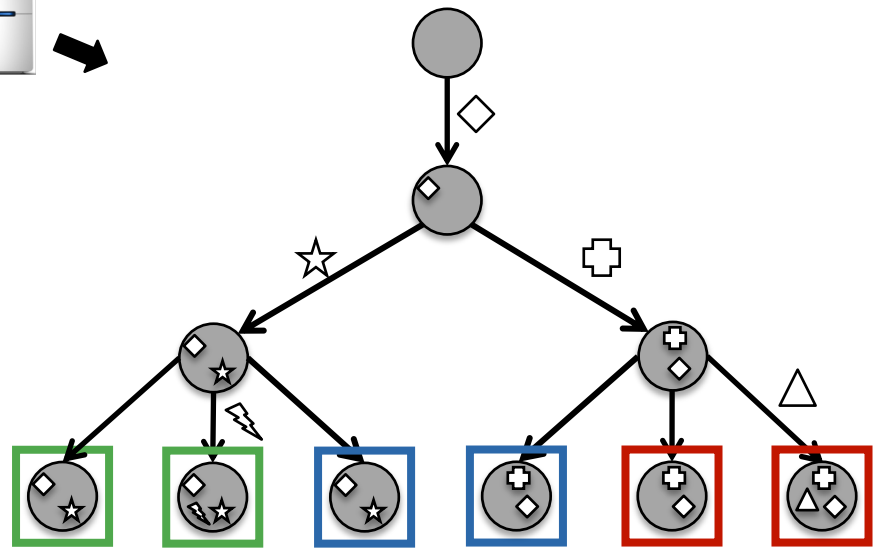


Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



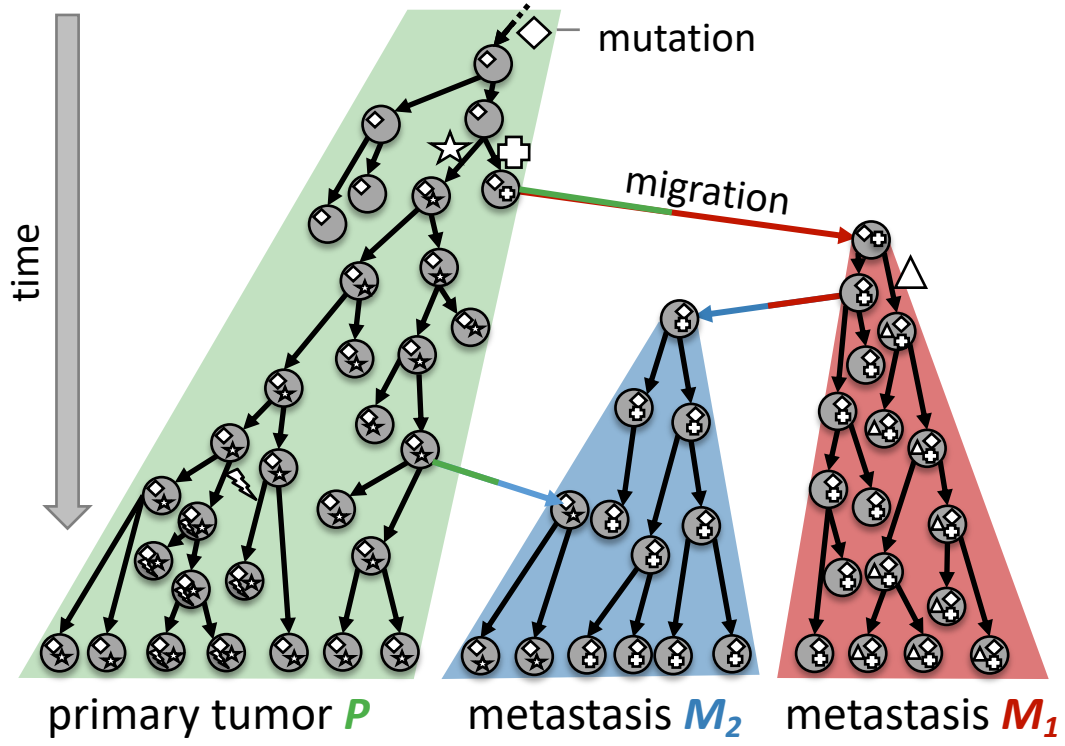
Cell Tree



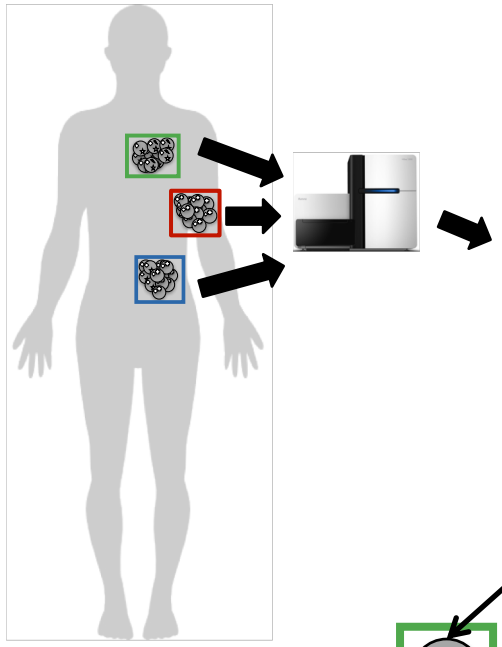


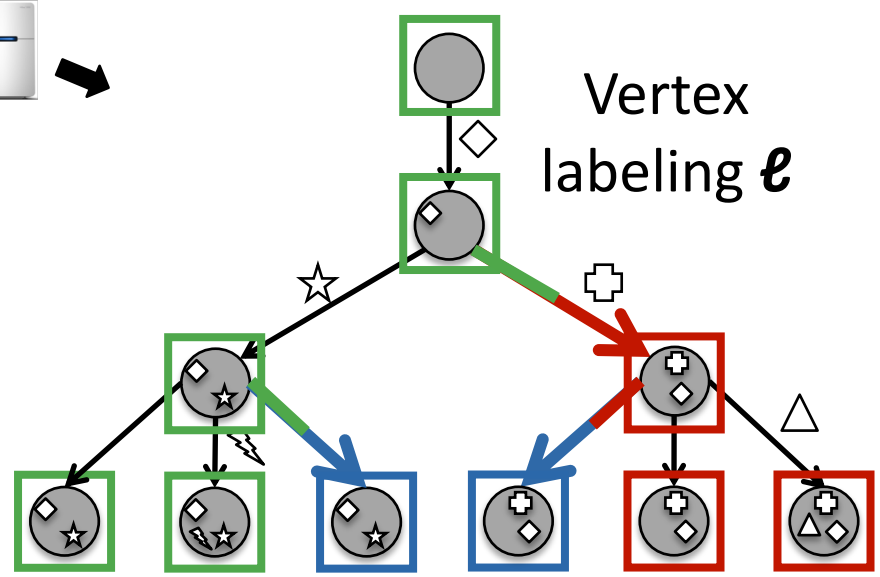
Phylogenetic Tree *T*

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



Cell Tree





Phylogenetic Tree *T*

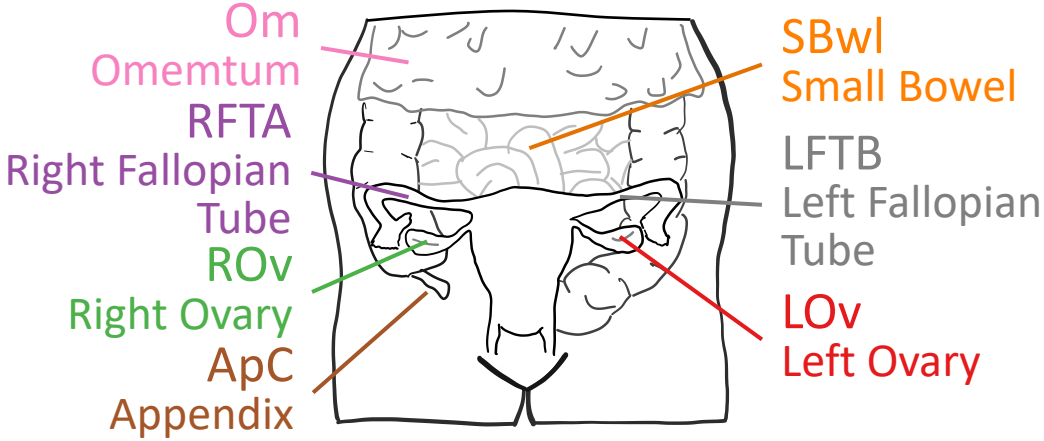
Goal: Given phylogenetic tree ***T***, find ***parsimonious*** vertex labeling ℓ with fewest migrations

Slatkin, M. and Maddison, W. P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, *123*(3), 603–613.

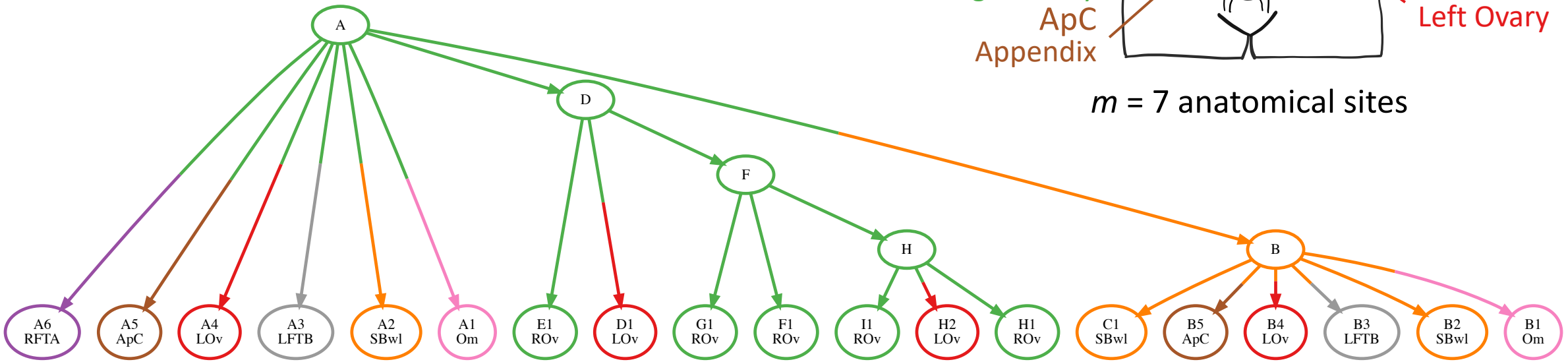
Minimum Migration Analysis in Ovarian Cancer

McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

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



m = 7 anatomical sites

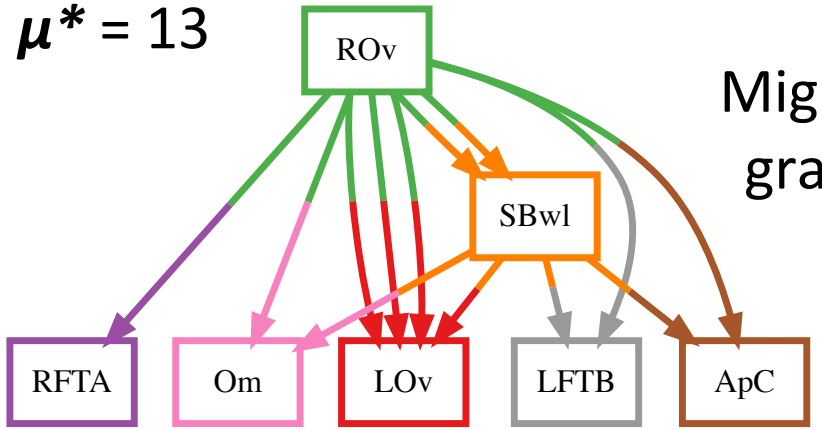


Minimum Migration Analysis in Ovarian Cancer

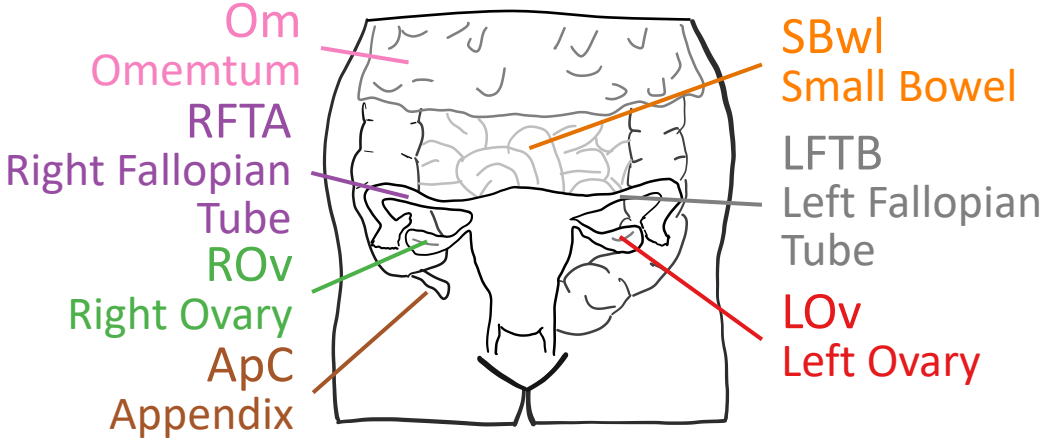
McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

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

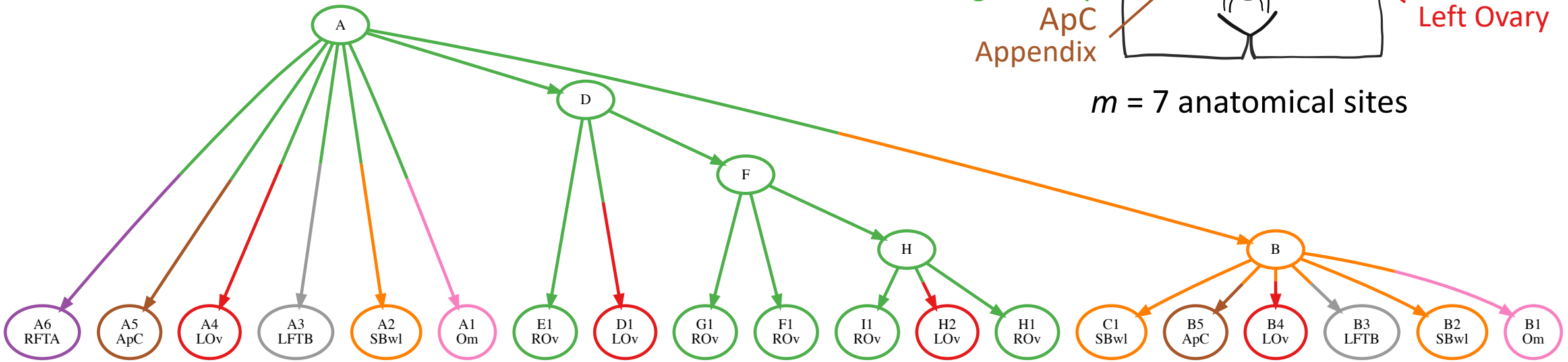
$\mu^* = 13$



Migration graph **G**



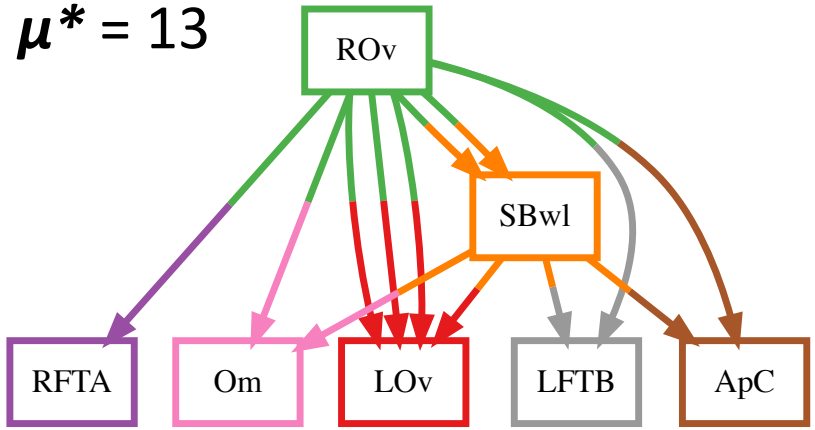
m = 7 anatomical sites



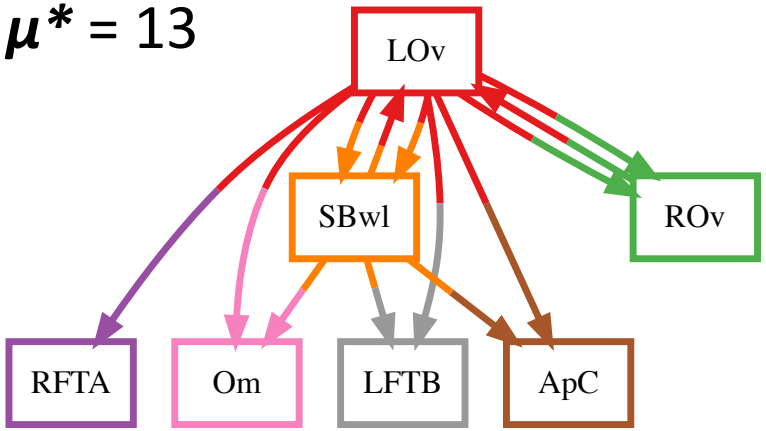
Minimum Migration History is ***Not*** Unique

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

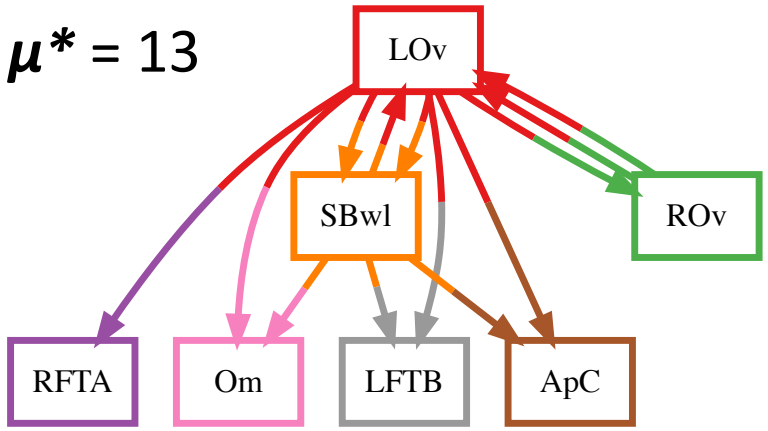
$\mu^* = 13$



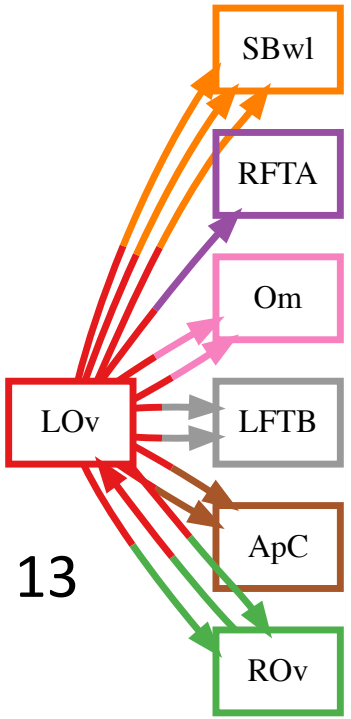
$\mu^* = 13$



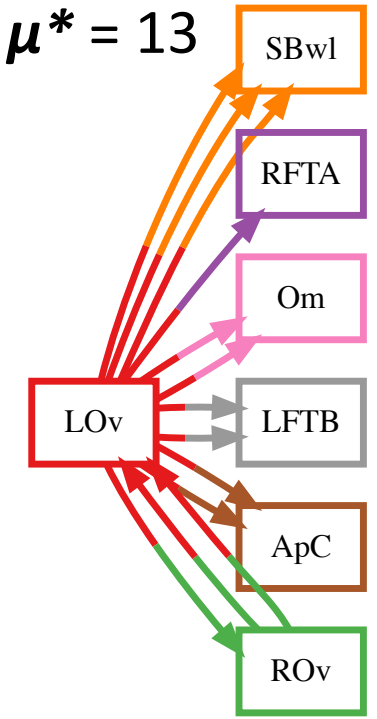
$\mu^* = 13$



$\mu^* = 13$



$\mu^* = 13$



ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

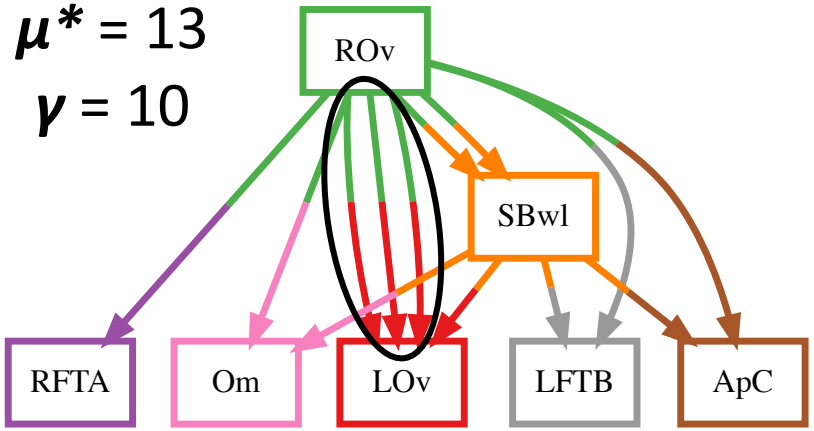
Comigrations: Simultaneous Migrations of Multiple Clones

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

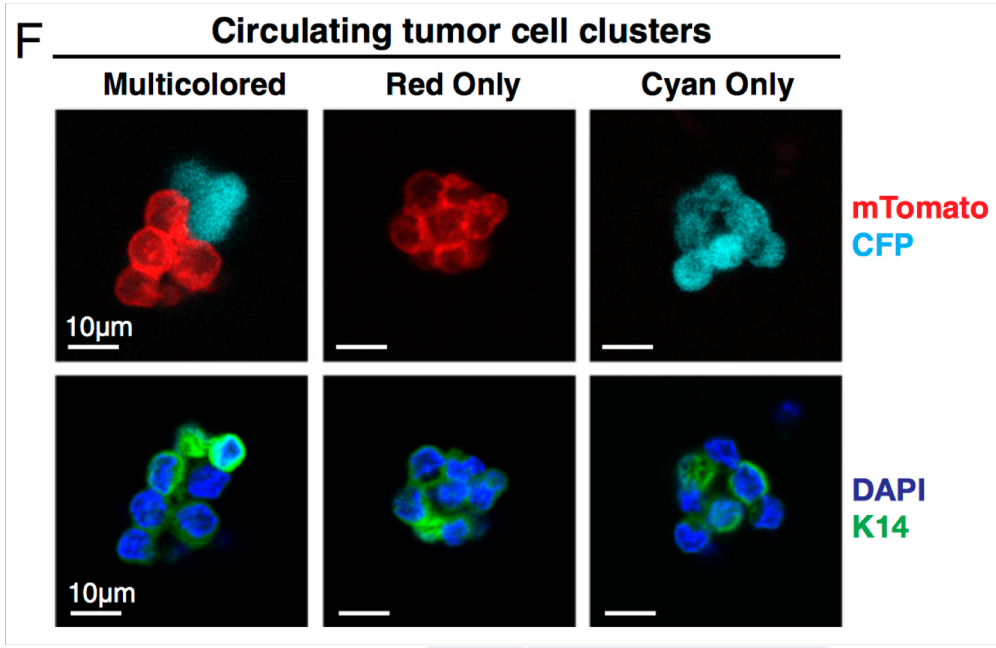
† Not necessarily true in the case of directed cycles

$\mu^* = 13$

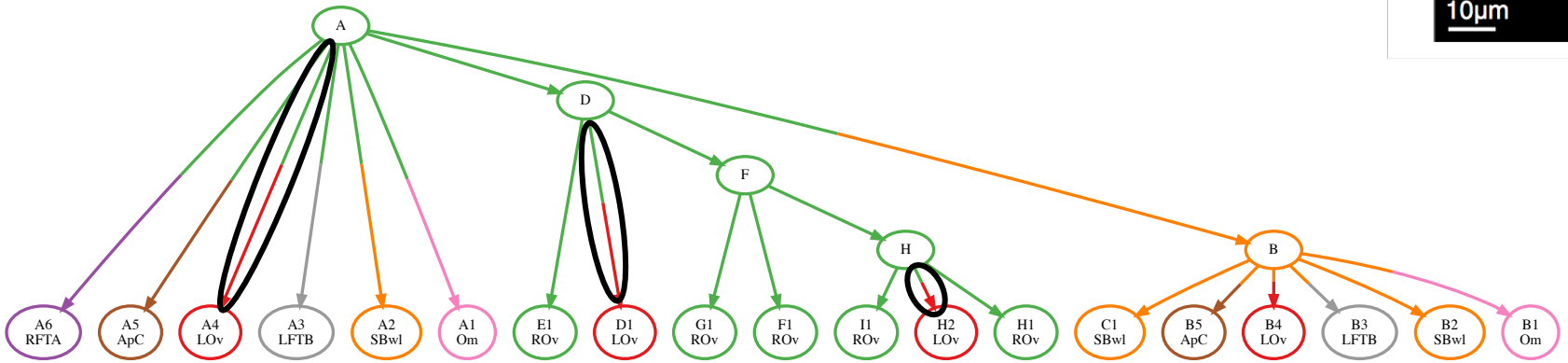
$\gamma = 10$



Migration Graph G



Clone Tree T



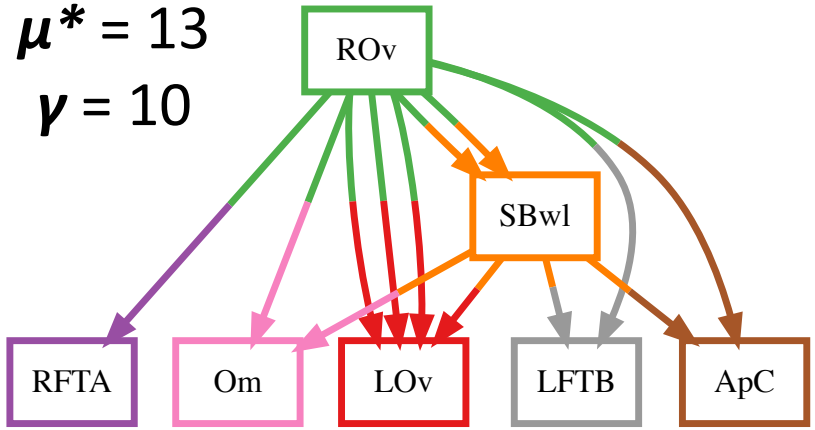
ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

Comigrations: Simultaneous Migrations of Multiple Clones

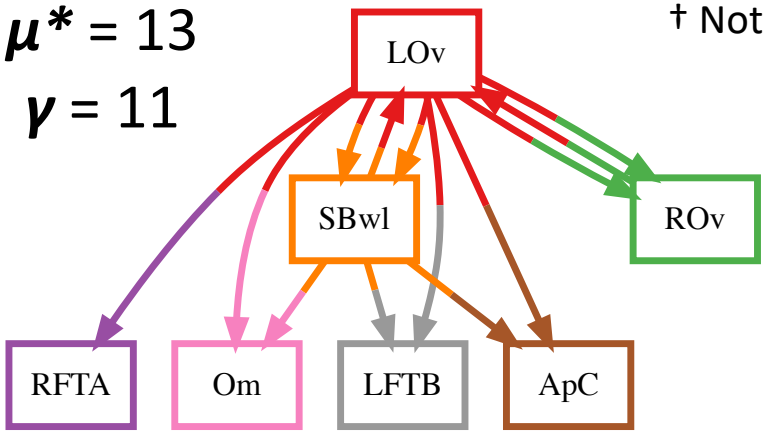
- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number γ of comigrations is the number of multi-edges in migration graph G †

† Not necessarily true in the case of directed cycles

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

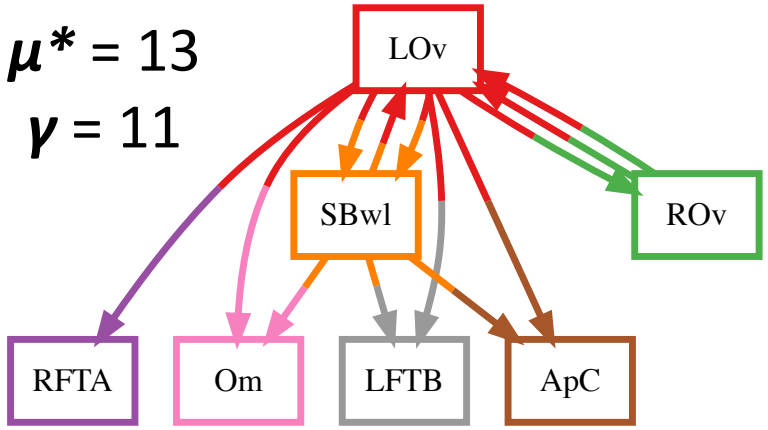


$\mu^* = 13$
$\gamma = 11$

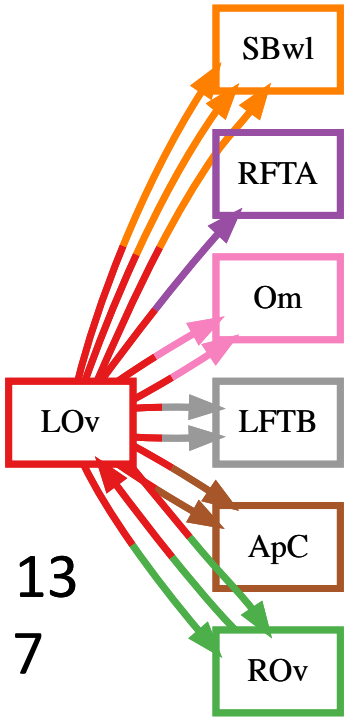


ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

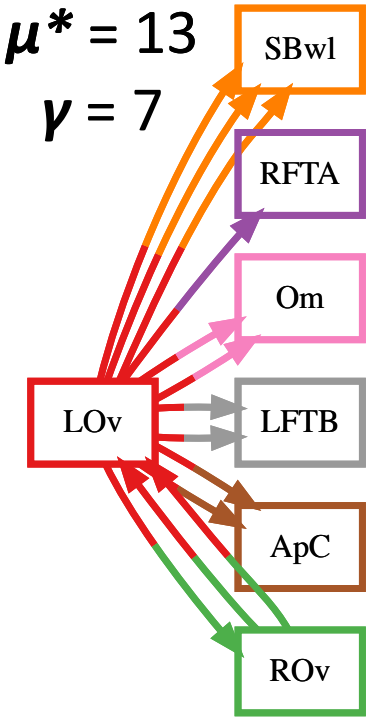
$\mu^* = 13$
$\gamma = 11$



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

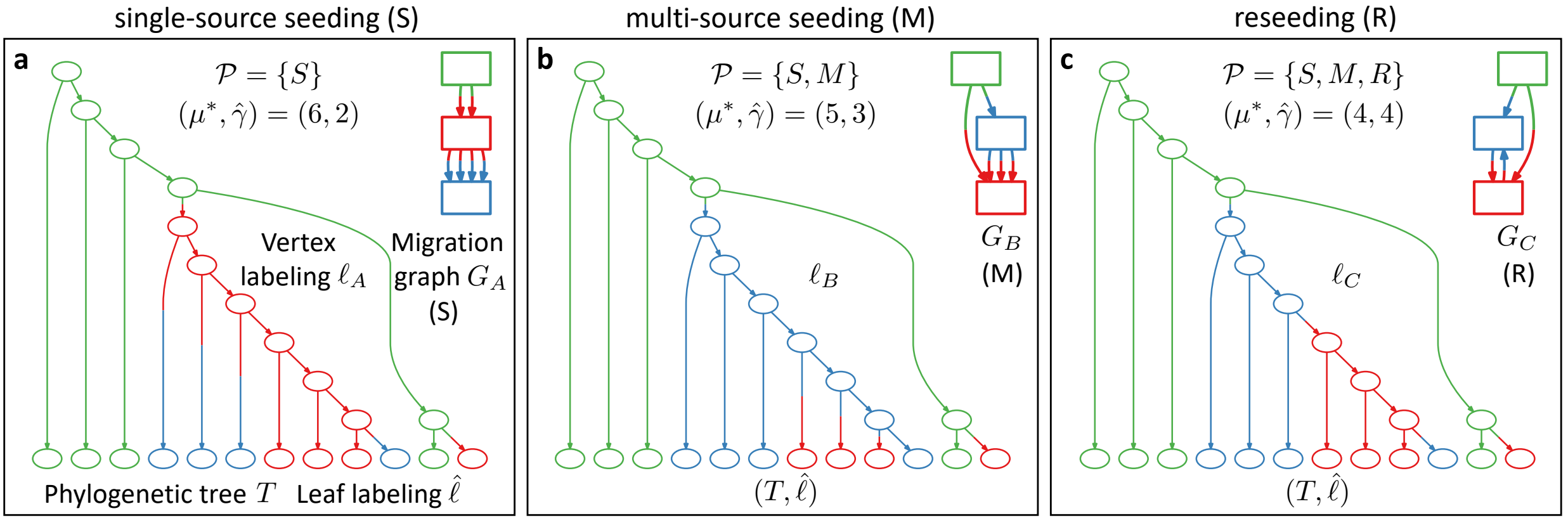


$\mu^* = 13$
$\gamma = 7$



Constrained Multi-objective Optimization Problem

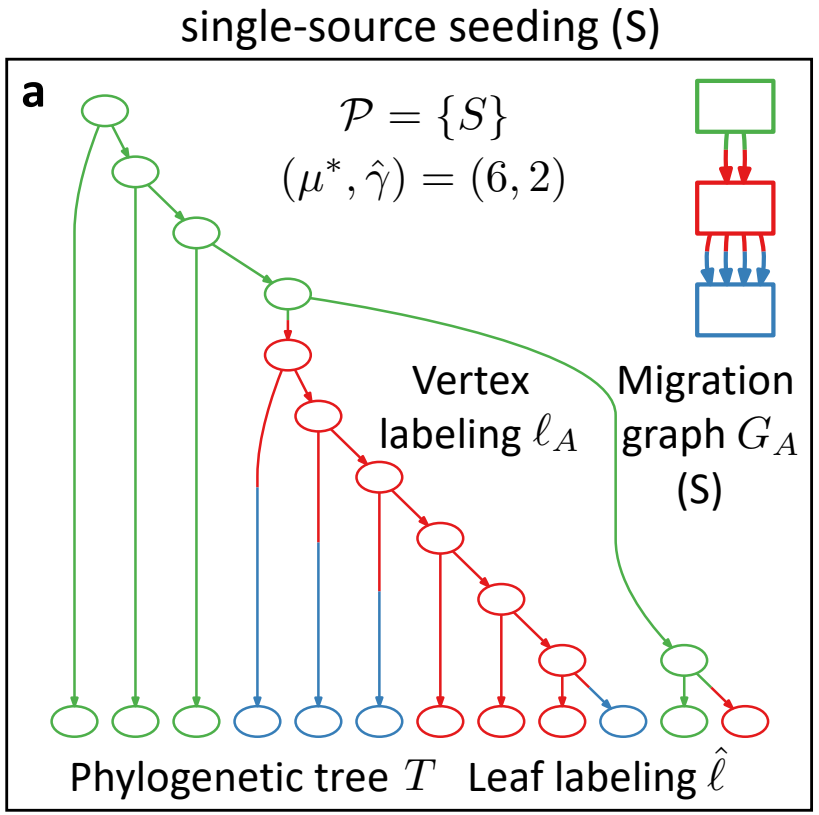
Parsimonious Migration History (PMH): Given a phylogenetic tree T and a set $\mathcal{P} \subseteq \{\mathrm{S}, \mathrm{M}, \mathrm{R}\}$ of allowed migration patterns, find vertex labeling ℓ with minimum migration number $\mu^*(T)$ and smallest comigration number $\hat{\gamma}(T)$.



El-Kebir, M., Satas, G., & Raphael, B. J. (2018). Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, *50*(5), 718–726.

Results [El-Kebir, WABI 2018]

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

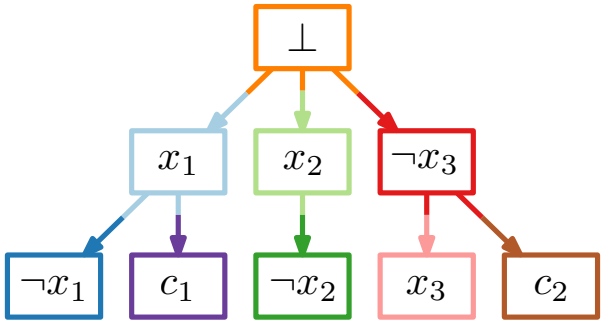


Theorem 1: PMH is NP-hard when $\mathcal{P} = \{\mathrm{S}\}$

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

PMH is NP-hard when $\mathcal{P} = \{S\}$

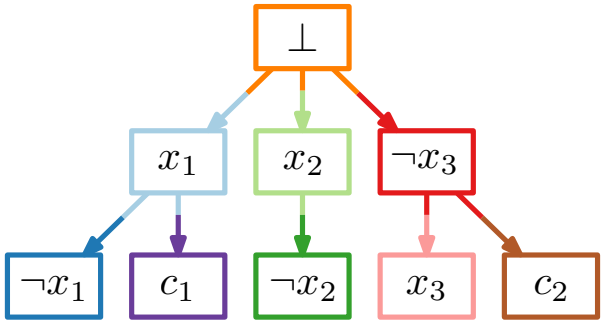
3-SAT: Given $\varphi = \Lambda_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

PMH is NP-hard when $\mathcal{P} = \{S\}$

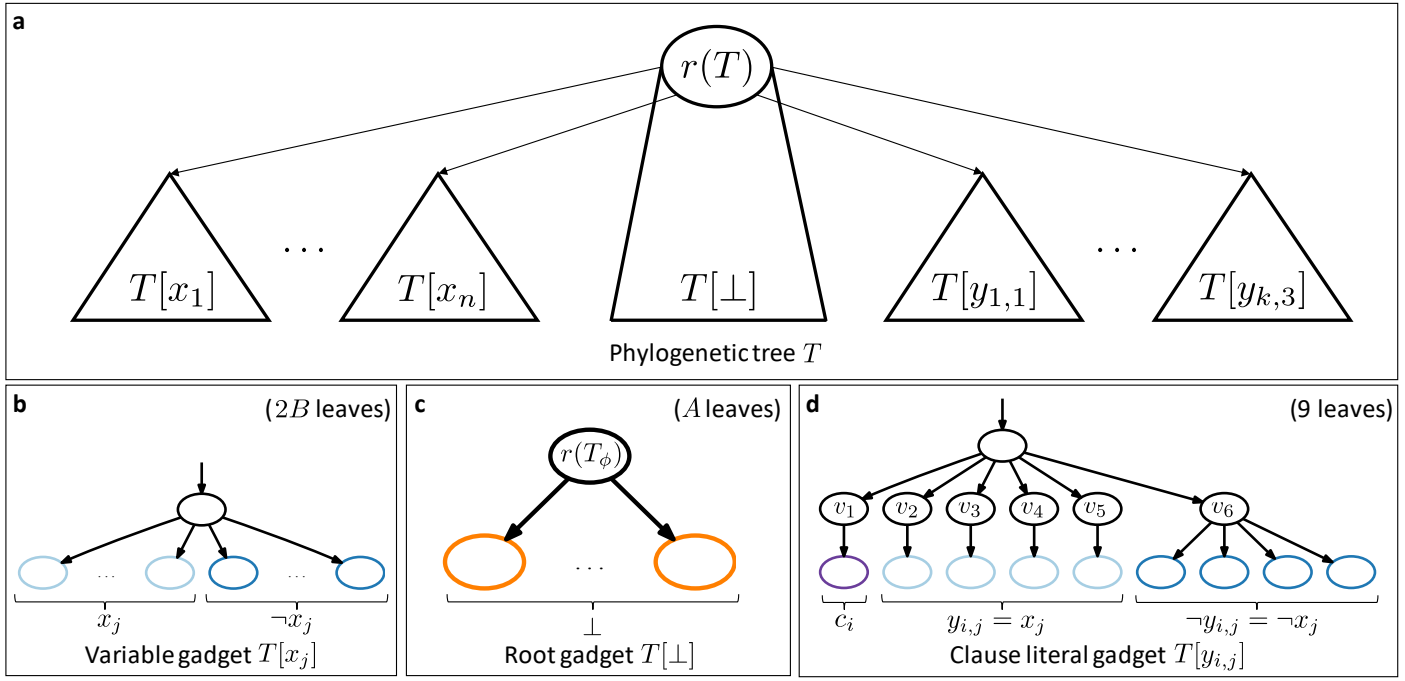
3-SAT: Given $\varphi = \Lambda_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

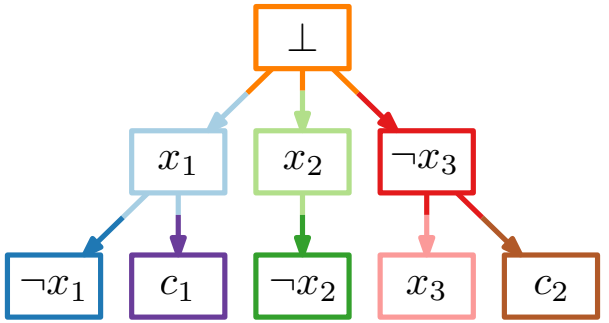
Three ideas:

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



PMH is NP-hard when $\mathcal{P} = \{S\}$

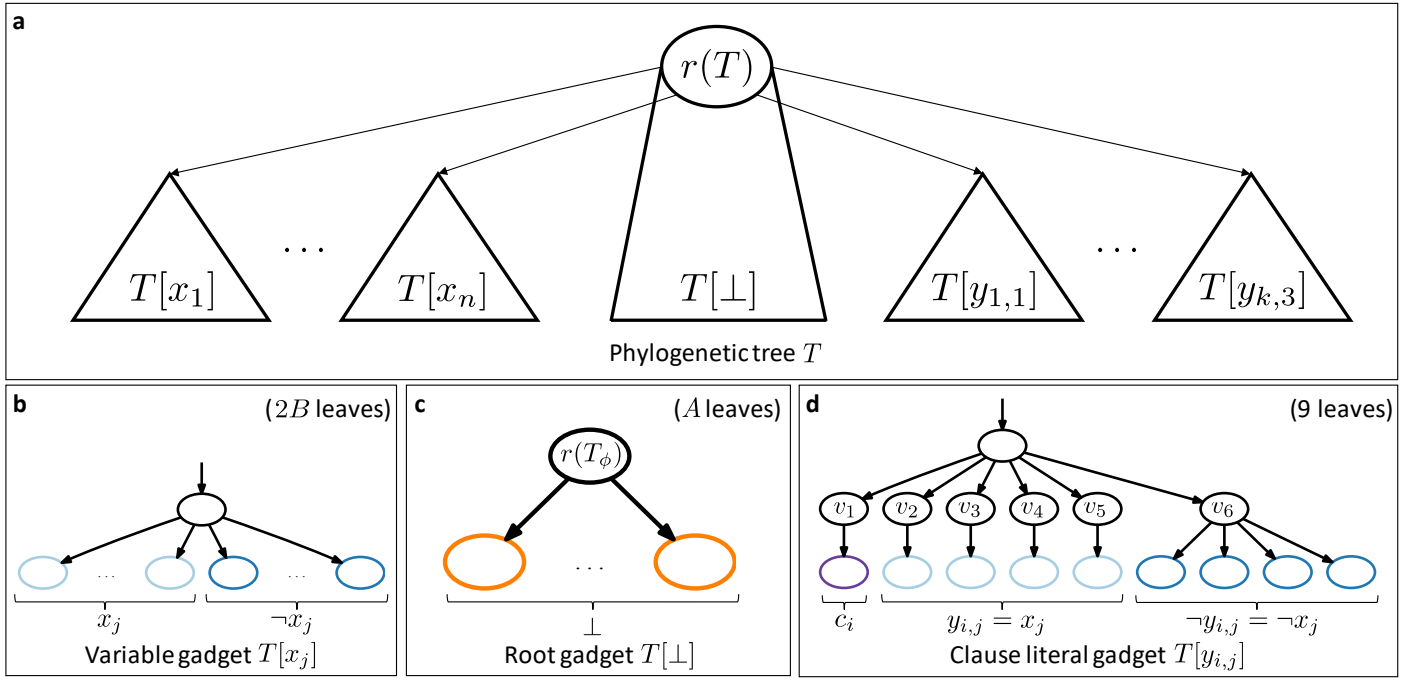
3-SAT: Given $\varphi = \Lambda_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

Three ideas:

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



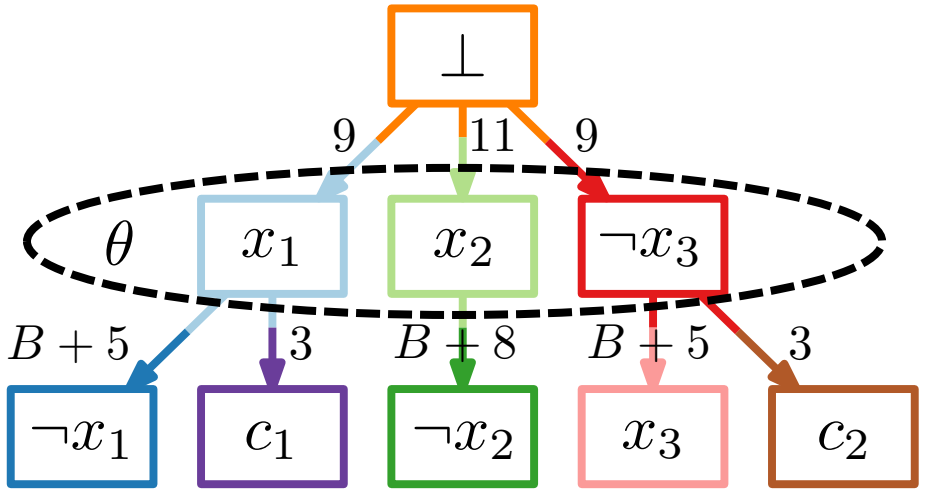
Lemma: Let $B > 10k + 1$ and $A > 2Bn + 27k$.
Then, φ is satisfiable if and only if $\mu^*(T) = (B+1)n + 25k$

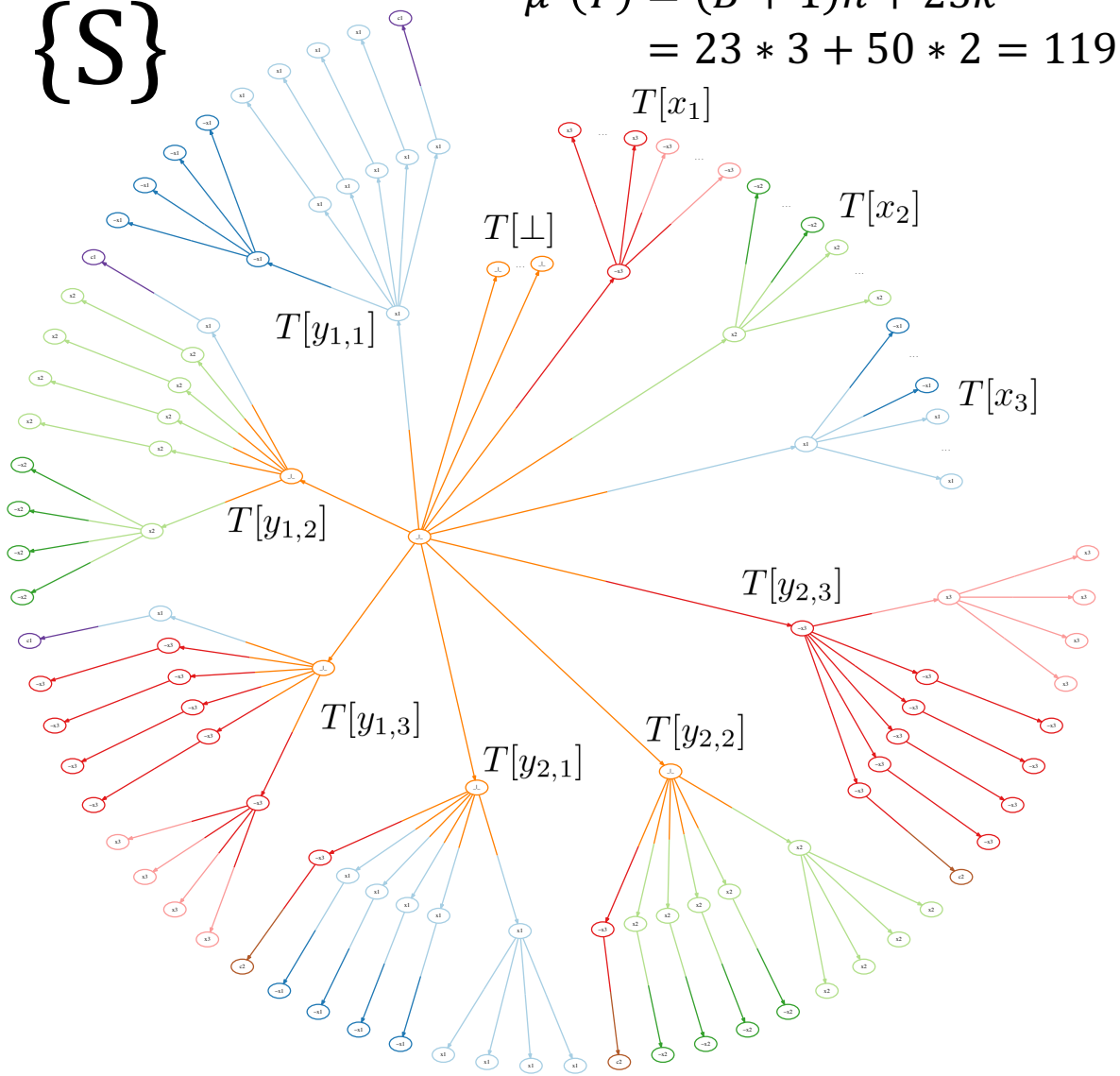
PMH is NP-hard when $\mathcal{P} = \{S\}$

$$\mu^*(T) = (B+1)n + 25k$$
$$= 23 * 3 + 50 * 2 = 119$$

$$\varphi = (x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1, \neg x_2, \neg x_3)$$

$$k = 2, n = 3$$
$$B = 10k + 2 = 22$$
$$A = 2Bn + 27\mathrm{k} + 1 = 187$$

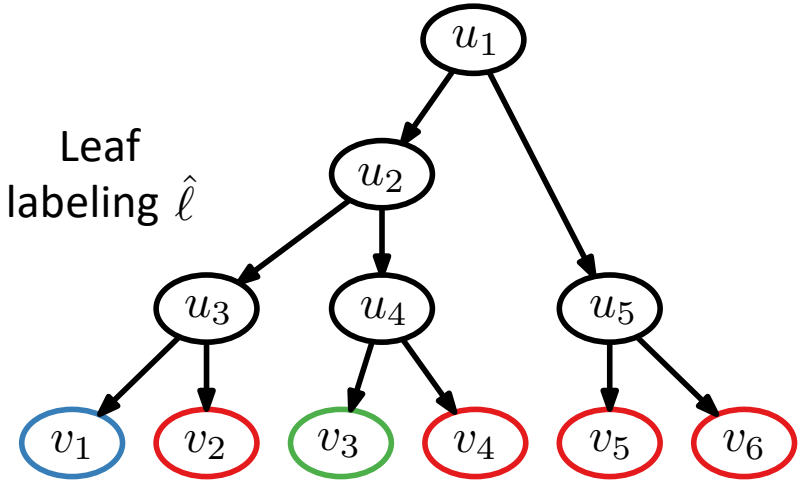




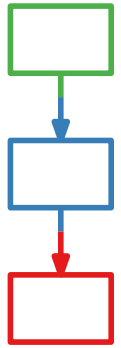
$$\Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \bot\}$$

Lemma: Let $B > 10k + 1$ and $A > 2Bn + 27k$.
Then, φ is satisfiable if and only if $\mu^*(T) = (B+1)n + 25k$

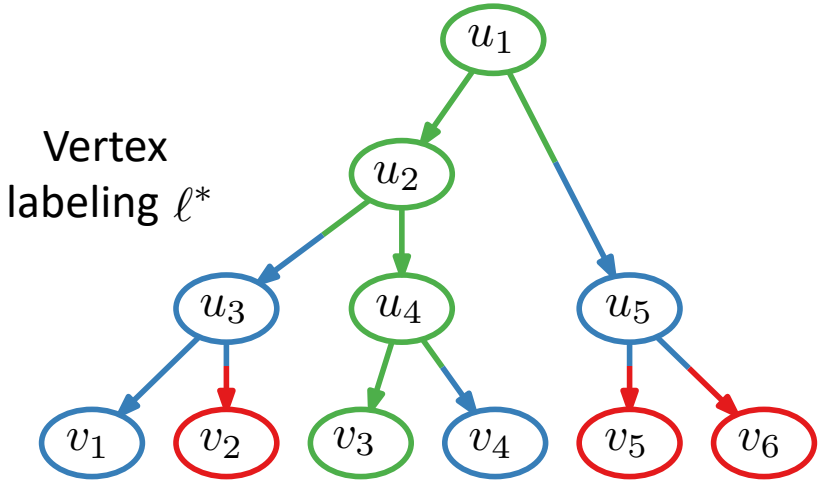
PMH is FPT in number m of locations when $\mathcal{P} = \{S\}$



Phylogenetic tree T



Migration tree $\hat{G}$



Phylogenetic tree T

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

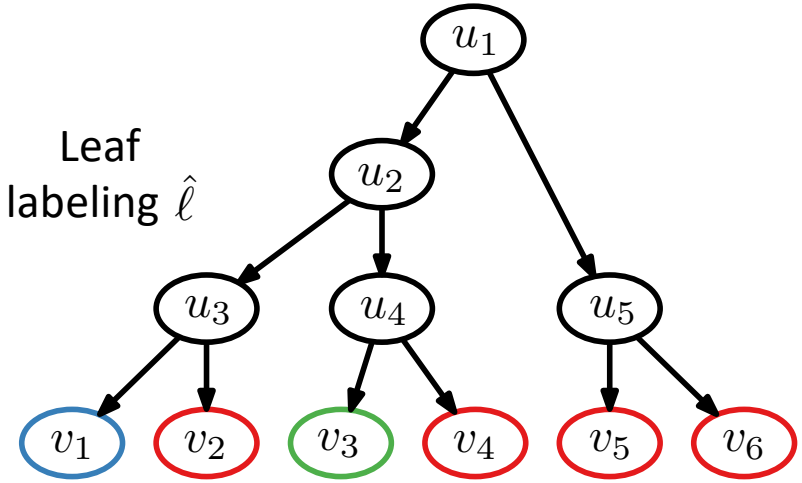
$$d_T(u,v) \geq d_{\hat{G}}(\mathrm{lca}_{\hat{G}}(u), \hat{\ell}(v)) \qquad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

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

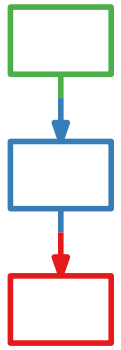
where $\sigma(s,t) = s$ if $s = t$ and otherwise $\sigma(s,t)$ is the unique child of s that lies on the path from s to t in $\hat{G}$.

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

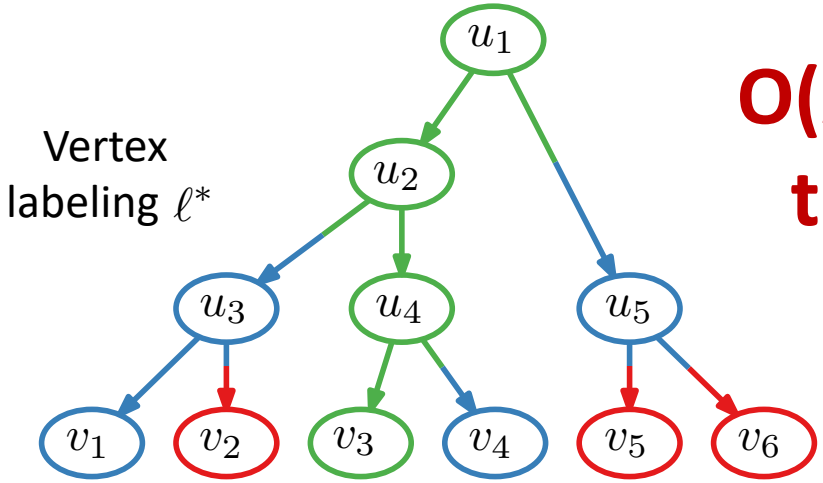
PMH is FPT in number m of locations when $\mathcal{P} = \{S\}$



Phylogenetic tree T



Migration tree $\hat{G}$



Phylogenetic tree T

O(nm^m) time

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

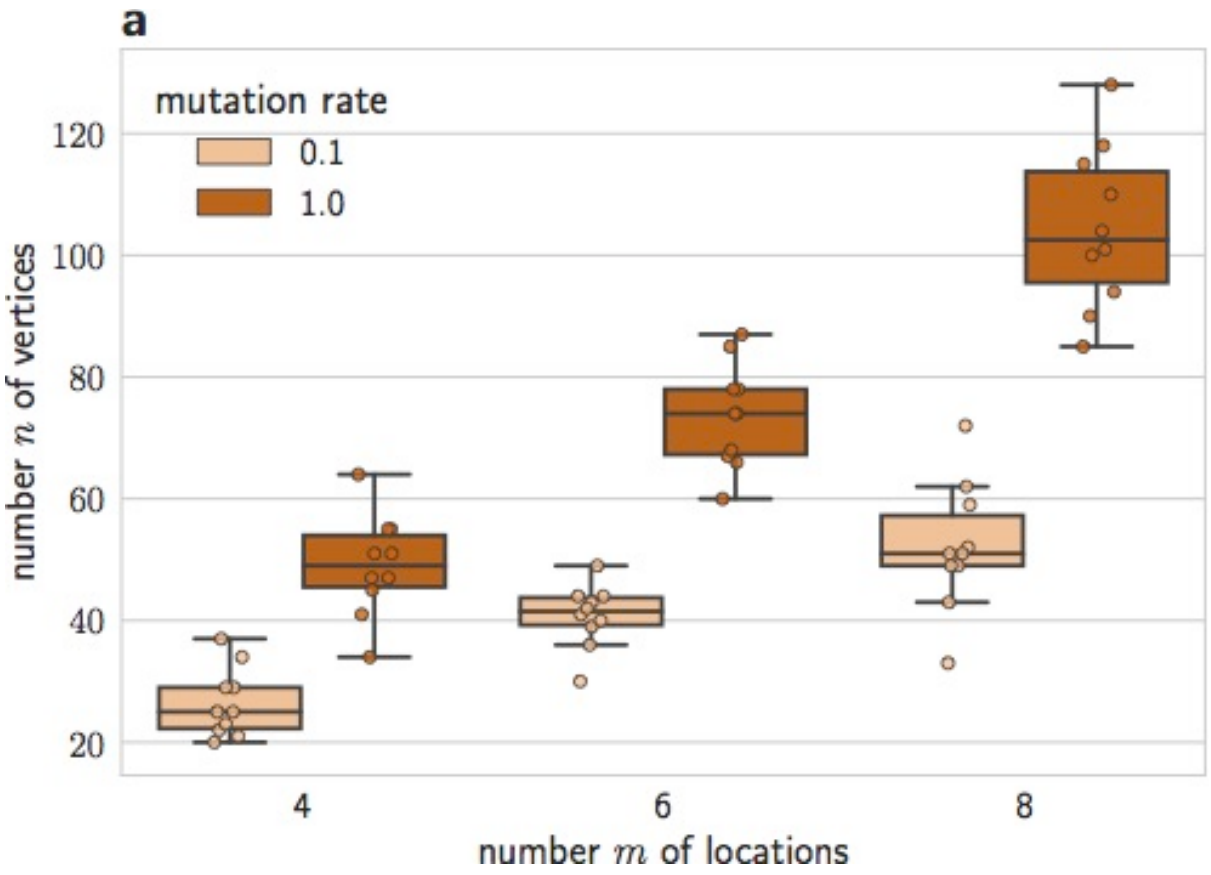
$$d_T(u,v) \geq d_{\hat{G}}(\mathrm{lca}_{\hat{G}}(u), \hat{\ell}(v)) \qquad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

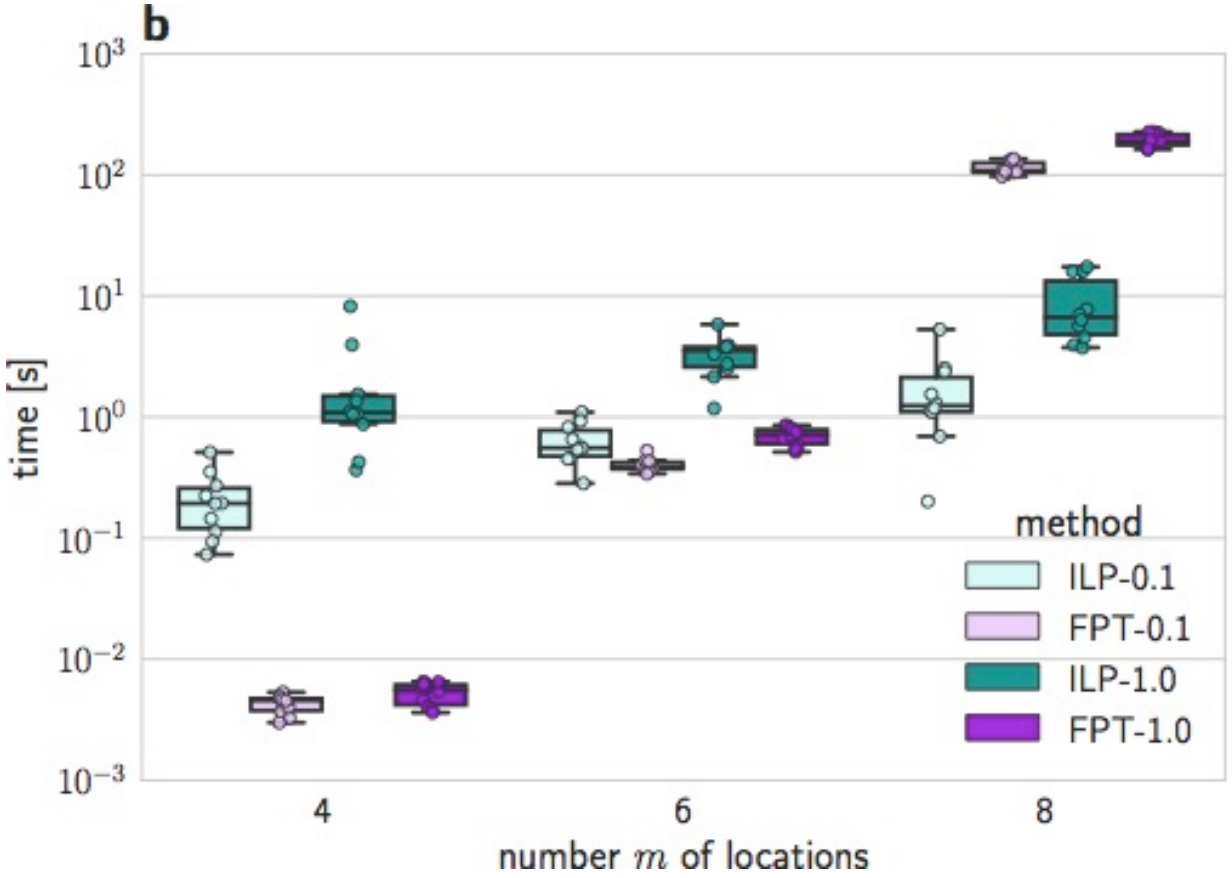
$$\ell^*(v) = \begin{cases} \mathrm{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \mathrm{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

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 ℓ^* is a minimum migration labeling consistent with $\hat{G}$.

Simulations





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

Outline

- Recap additive distance
- Neighbor joining
- Character-based phylogeny (small)
- Application to cancer

Reading:

- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner