

CS 598MEB

Computational Cancer Genomics

Lecture 2

Mohammed El-Kebir

January 28, 2021



Lecture Outline

- Recap
- Maximum Parsimony
- Two-state Perfect Phylogeny
- Two-state Perfect Phylogeny Mixtures

Reading

- Lecture notes

Hallmarks of Cancer



Inter-tumor heterogeneity:
Every tumor is different!

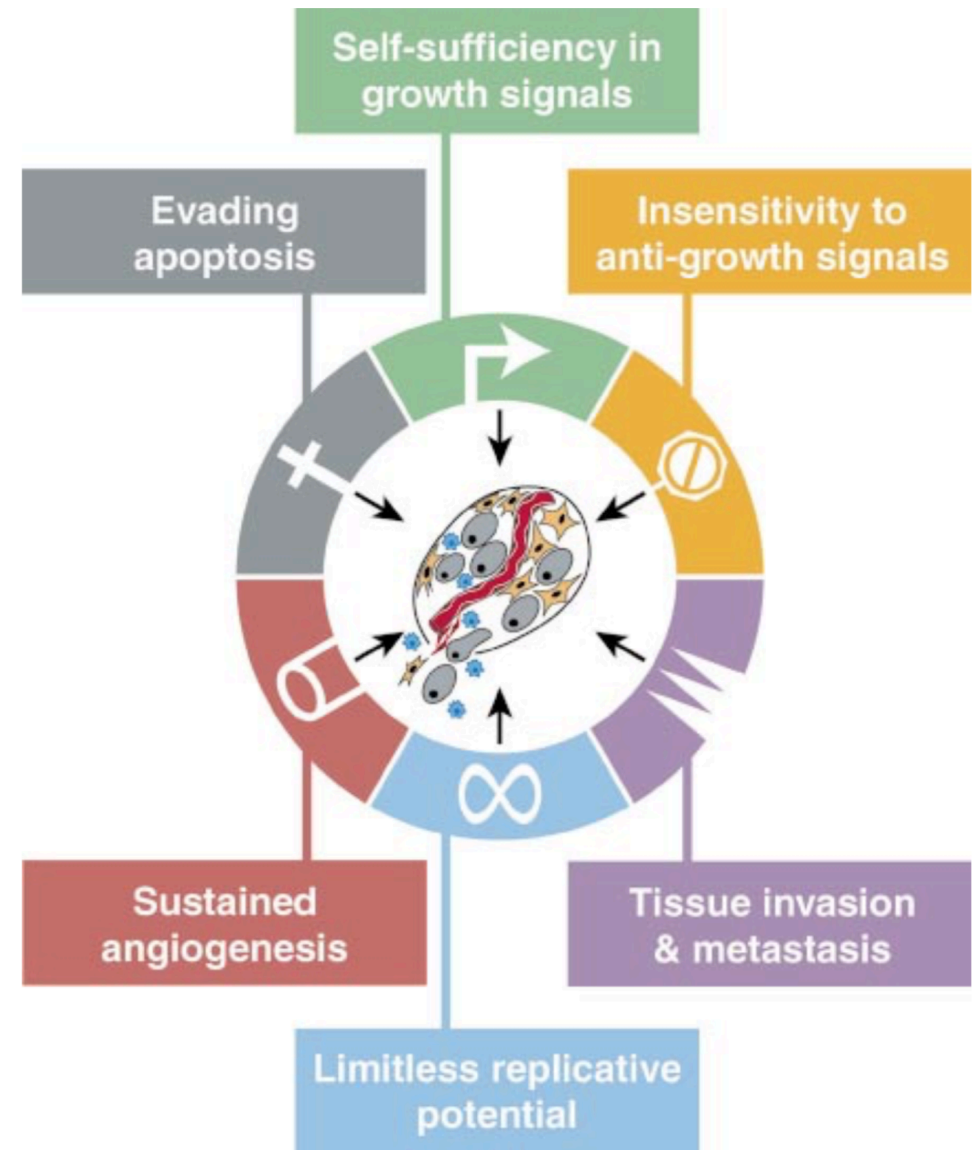
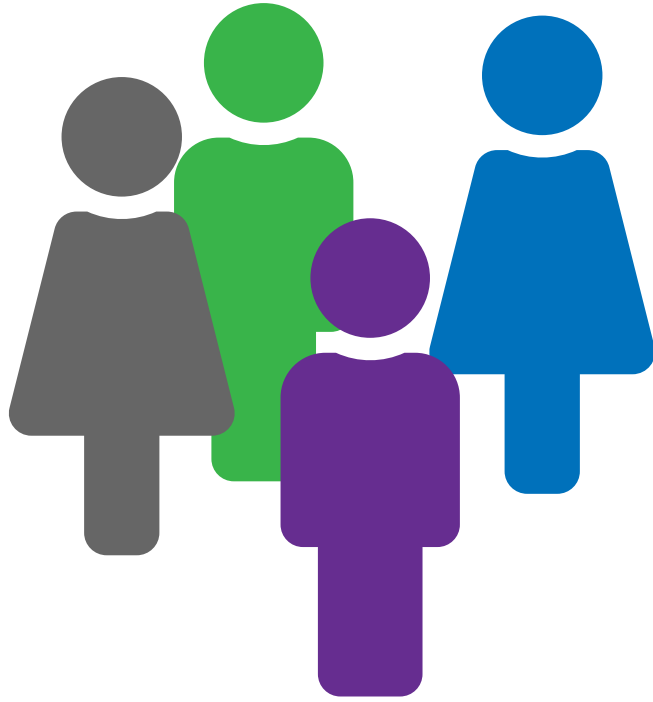


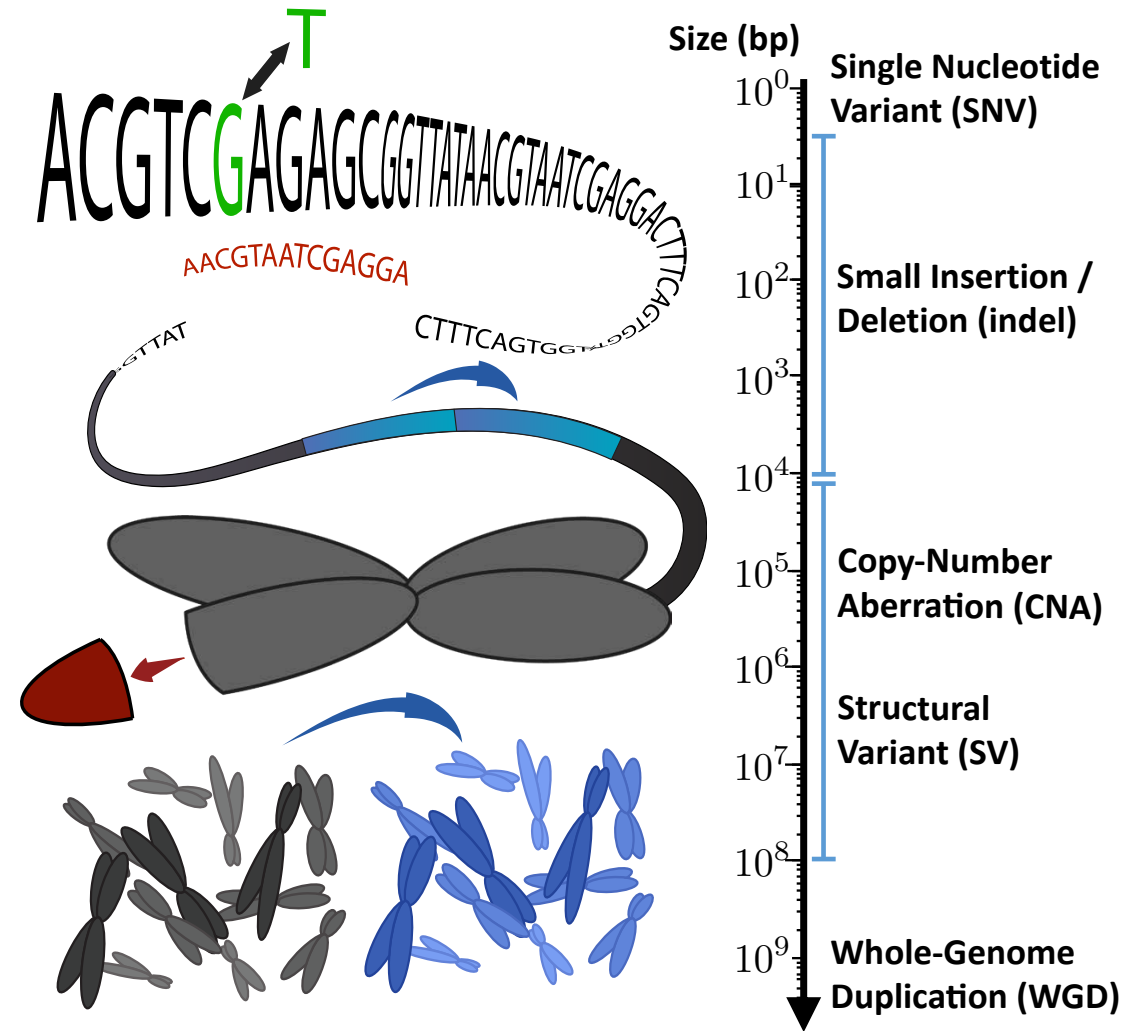
Figure 1. Acquired Capabilities of Cancer

We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.

Cancer is Caused by Somatic Mutations

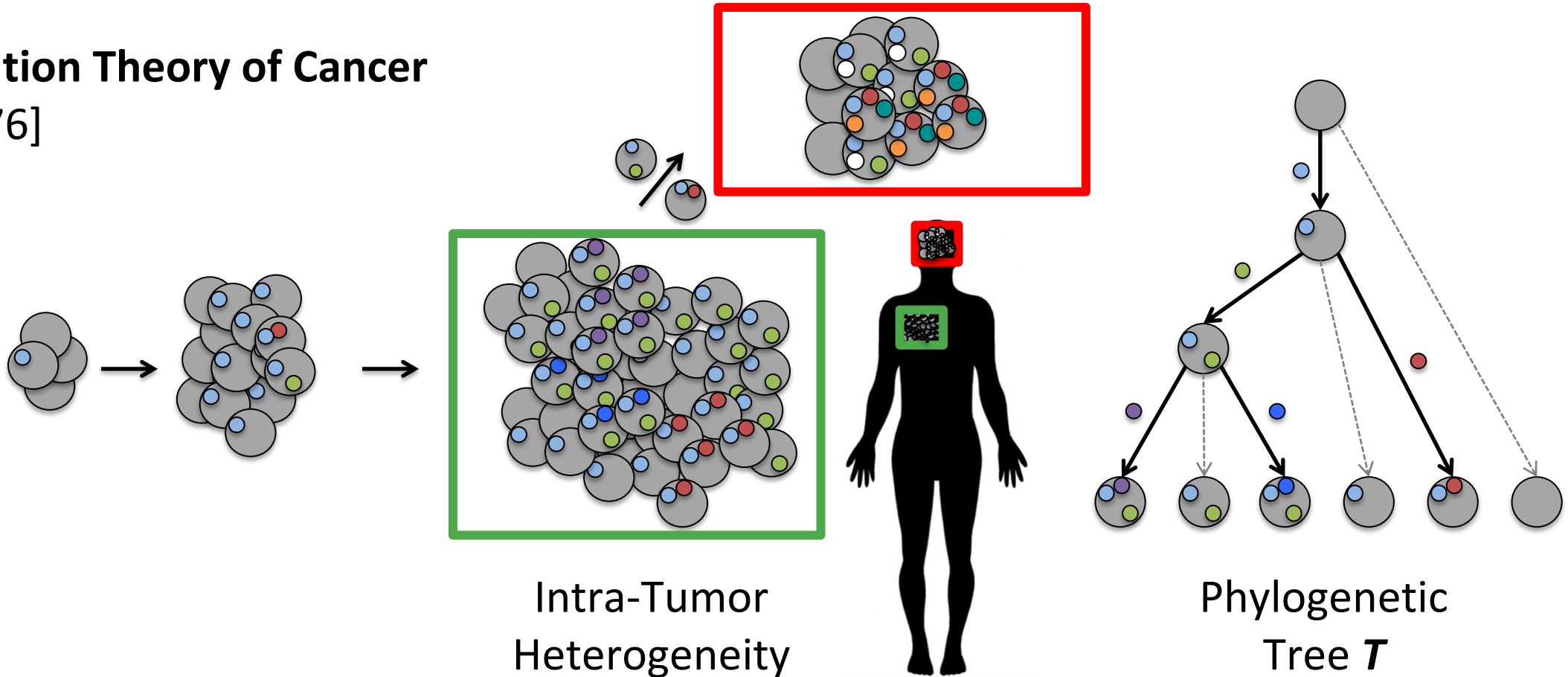


Inter-tumor heterogeneity:
Every tumor is different!



Tumorigenesis: Cell Mutation, Division & Migration

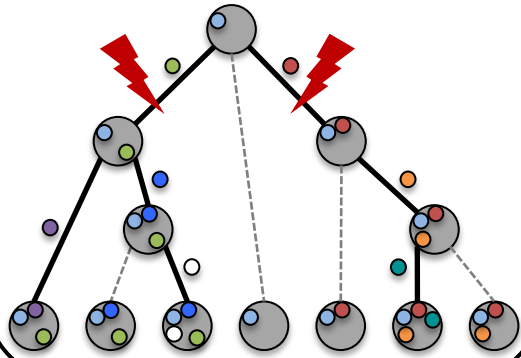
Clonal Evolution Theory of Cancer [Nowell, 1976]



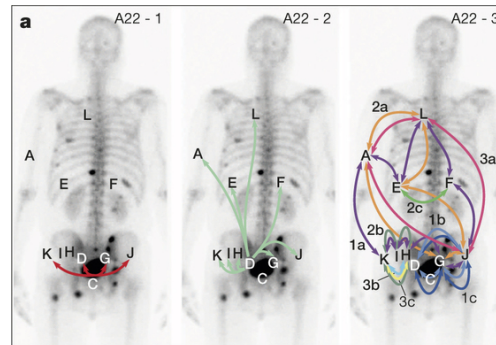
Intra-tumor heterogeneity: Every tumor cell is different

Phylogenies are Key to Understanding Cancer

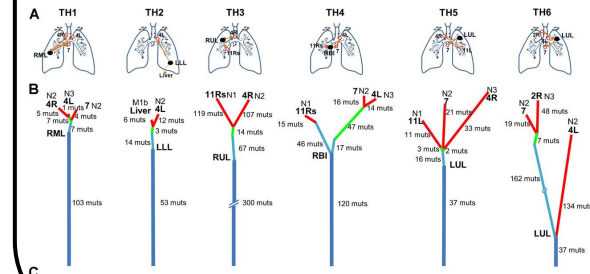
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients



These downstream analyses **critically rely** on accurate tumor phylogeny inference

Key challenge in phylogenetics:
Accurate phylogeny inference from data at present time

Lecture Outline

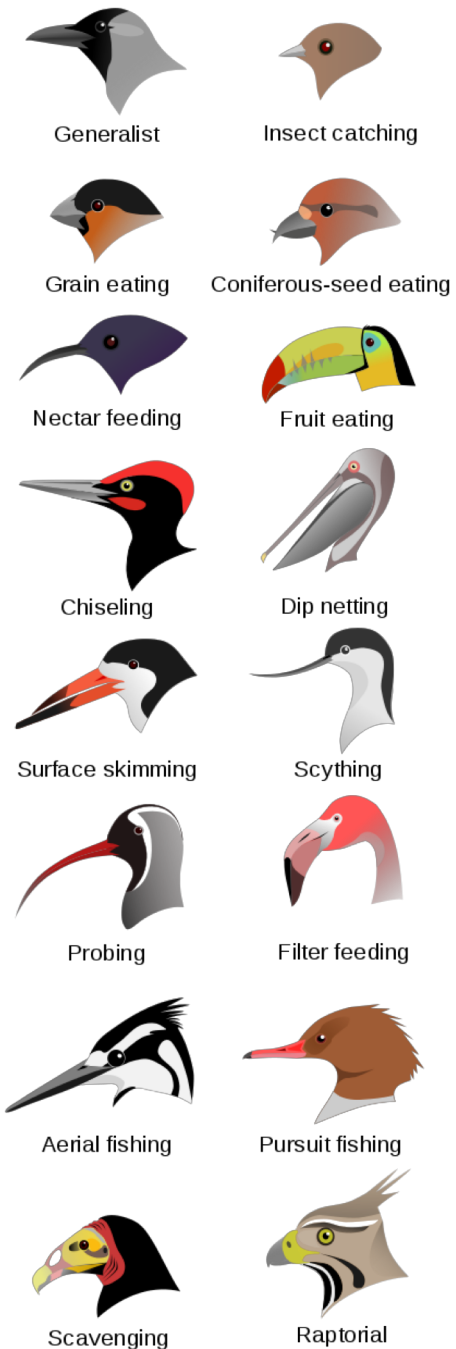
- Recap
- Maximum Parsimony
- Two-state Perfect Phylogeny
- Two-state Perfect Phylogeny Mixtures

Reading

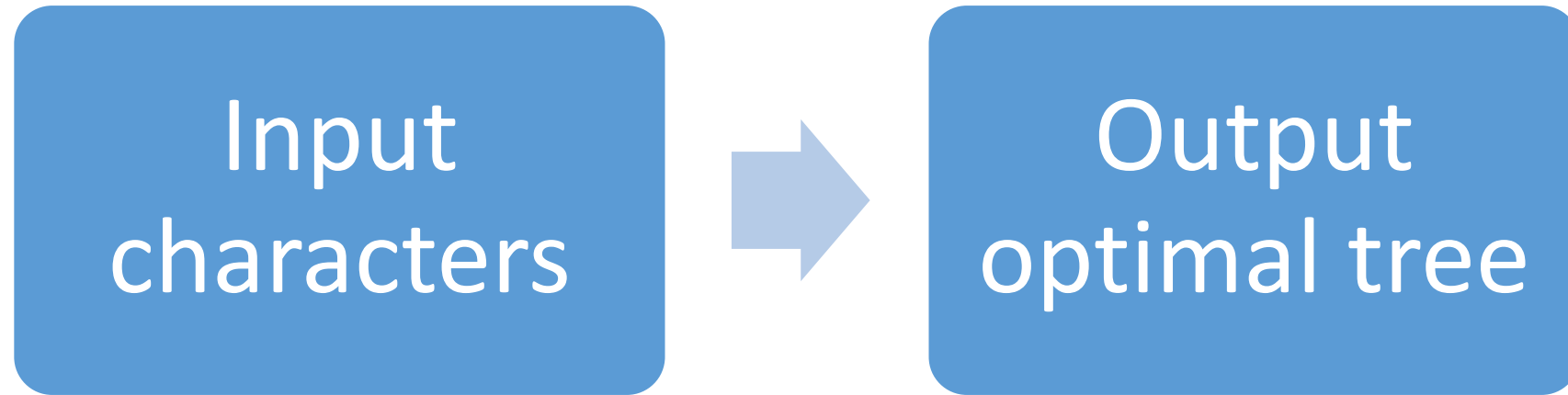
- Lecture notes

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

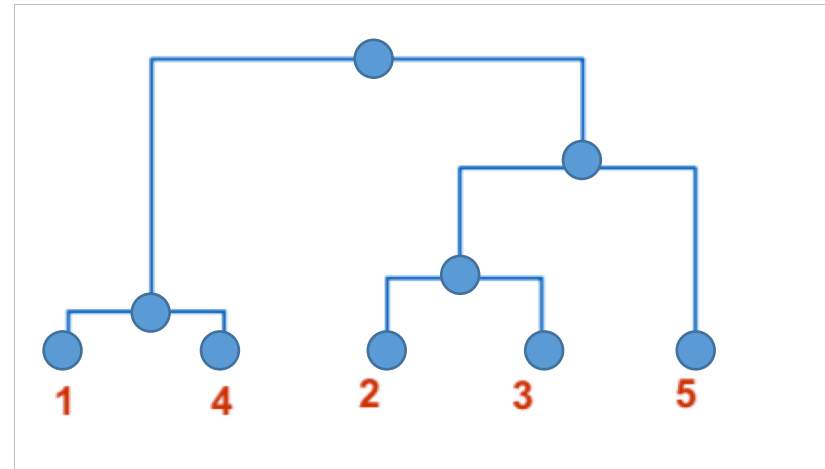


Character-Based Phylogeny Reconstruction



Question: What is optimal?

Want: Optimization criterion



Character-Based Phylogeny Reconstruction

Input
characters



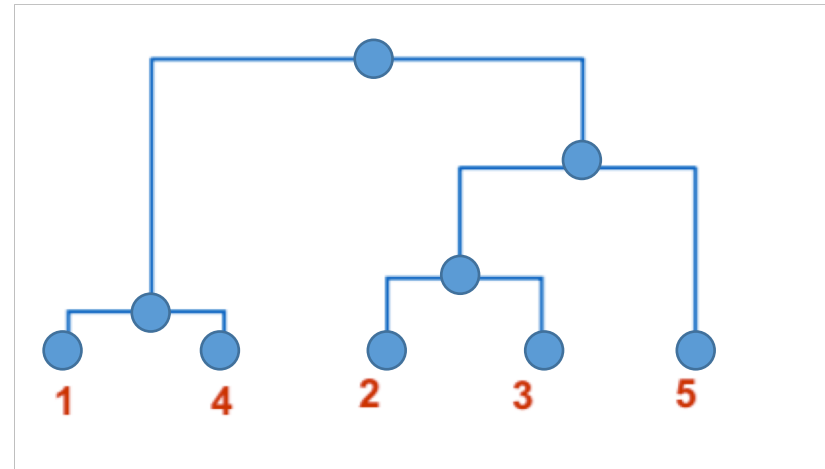
Output
optimal tree

Question: What is optimal?

Want: Optimization criterion

Question: How to optimize this criterion?

Want: Algorithm

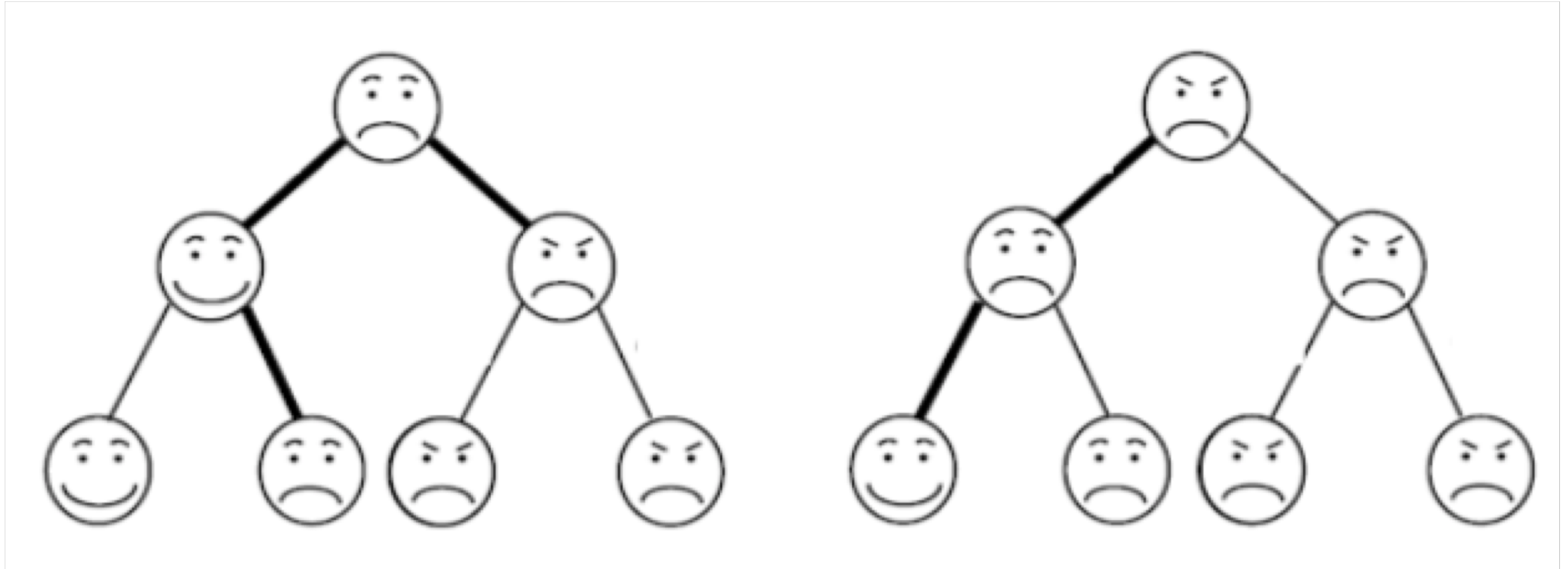


Character-Based Phylogeny Reconstruction: Input

Characters / states	State 1	State 2
Mouth	Smile	Frown
Eyebrows	Normal	Pointed

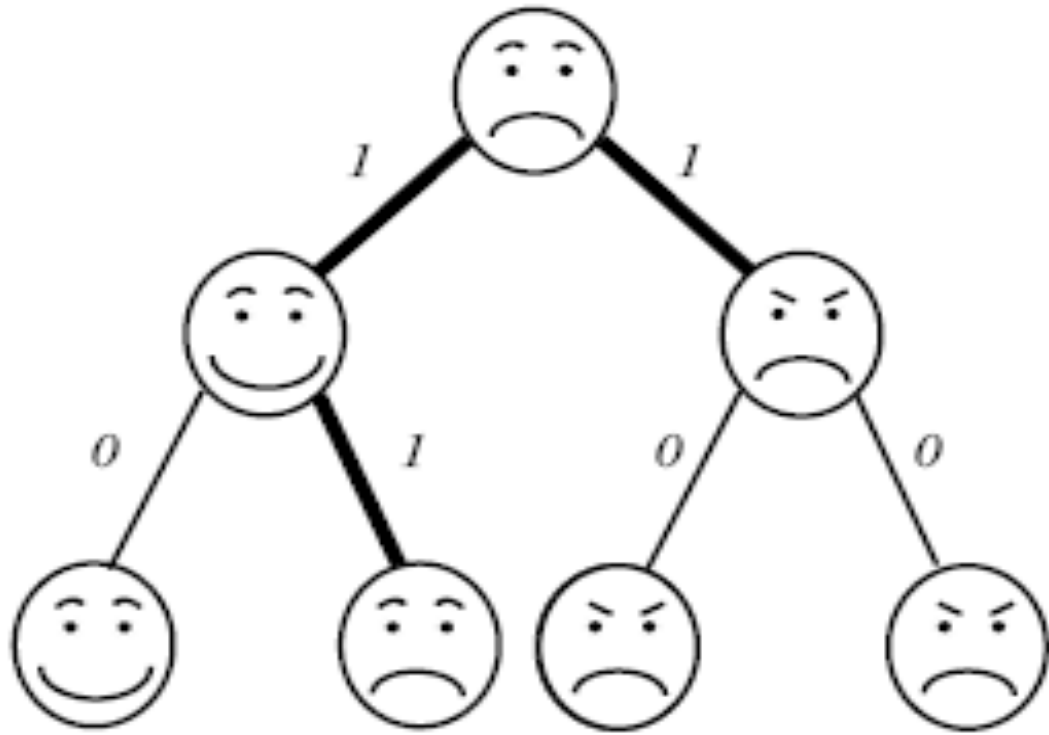


Character-Based Phylogeny Reconstruction: Criterion

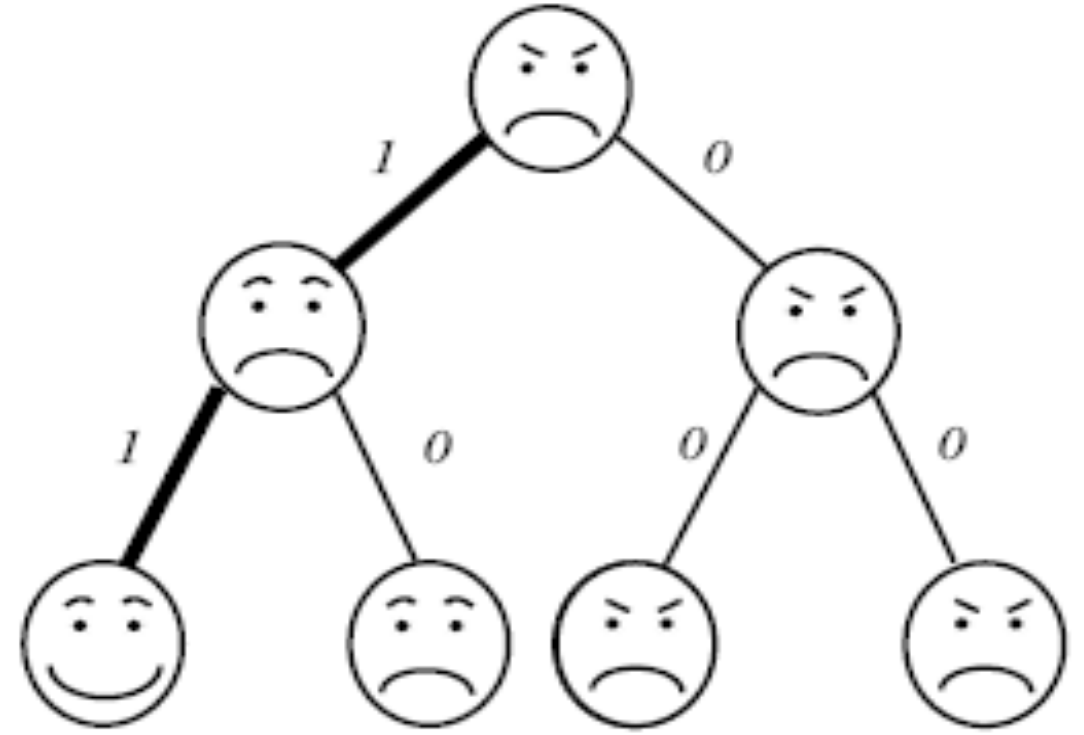


Question: Which tree is better?

Character-Based Phylogeny Reconstruction: Criterion



(a) *Parsimony Score=3*

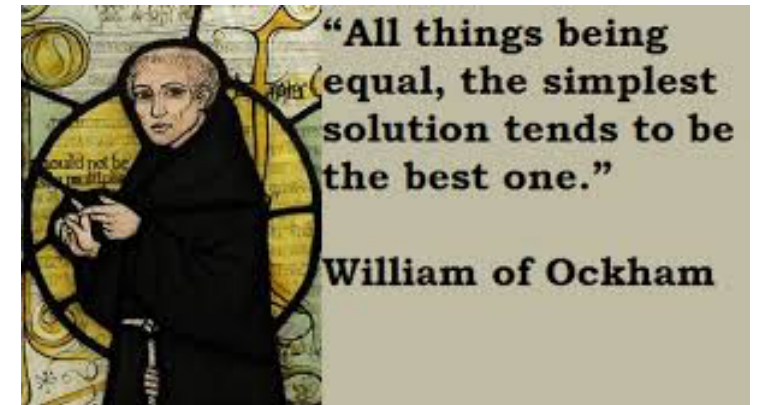


(b) *Parsimony Score=2*

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.



Binary Characters

		Characters				
		1	2	3	4	5
Species	A	0	1	1	0	0
	B	0	0	1	1	0
	C	1	1	1	1	0
	D	1	1	0	1	1

Characters only have
two possible states

Possible Encoding:
0 : not-mutated
1 : mutated

Possible Encoding:
0 : no wings
1 : wings

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.

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:

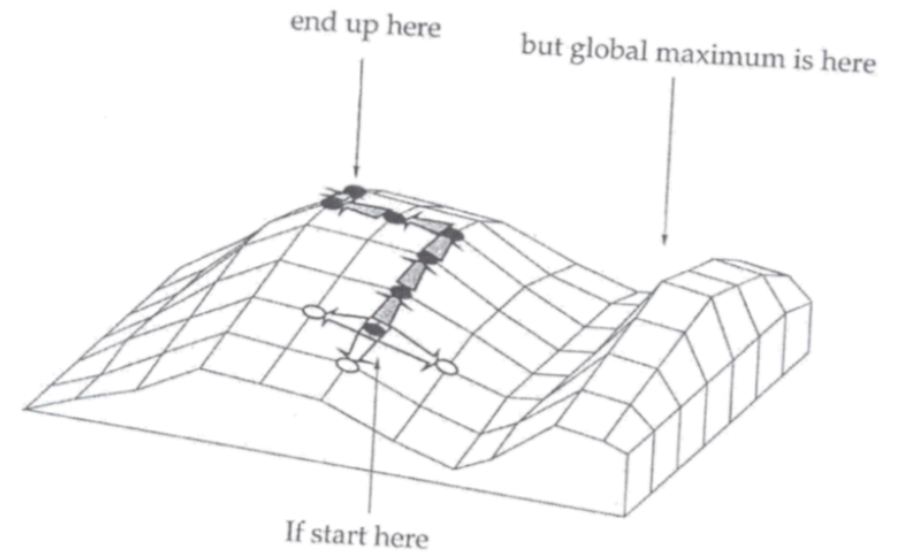
Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree T with m leaves labeled according to A and an assignment of character states to each internal vertex of T with minimum parsimony score.

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

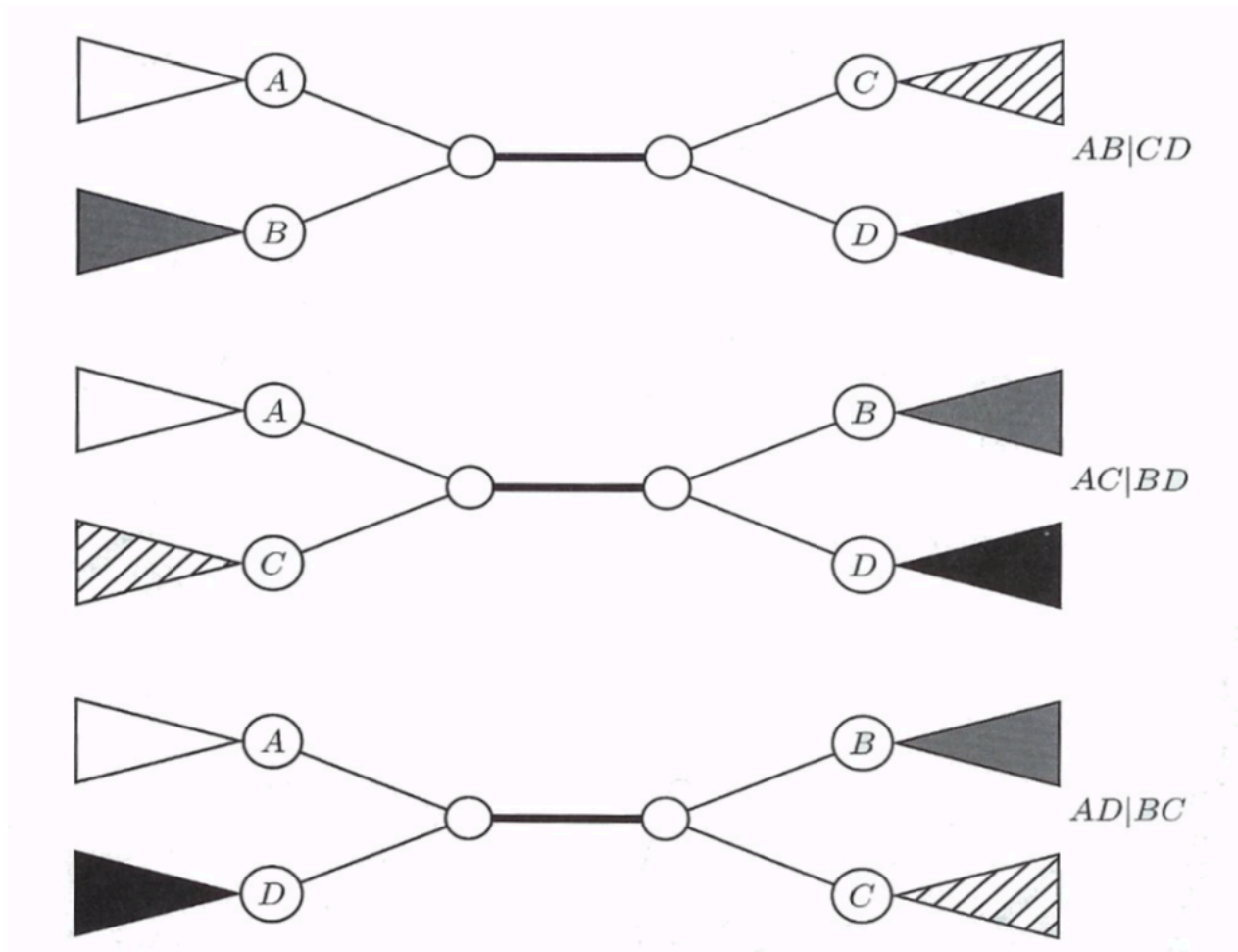
Large Maximum Parsimony Phylogeny

- This problem is NP-hard
- Heuristics using local search (tree moves)
 1. Start with an arbitrary tree T .
 2. Check “neighbors” of T .
 3. Move to a neighbor if it provides the best improvement in parsimony/likelihood score.

Caveats:
Could be stuck in **local** optimum, and not achieve global optimum



Local Search: Nearest-Neighbor Interchange (NNI)

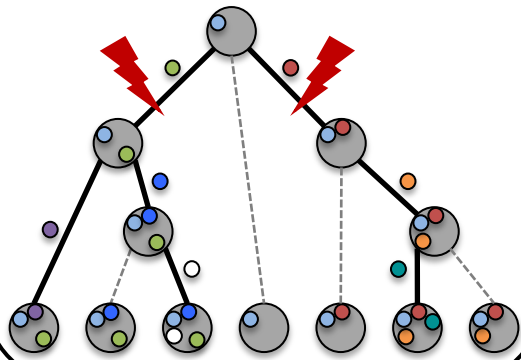


Rearrange four subtrees
defined by one
internal edge

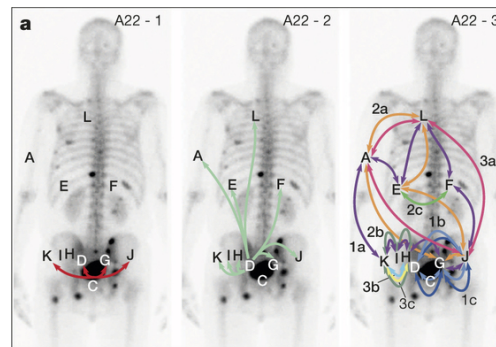
Figure: Jones and Pevzner

Phylogenies are Key to Understanding Cancer

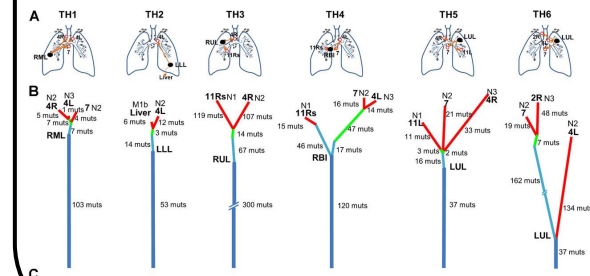
Identify targets for treatment



Understand metastatic development



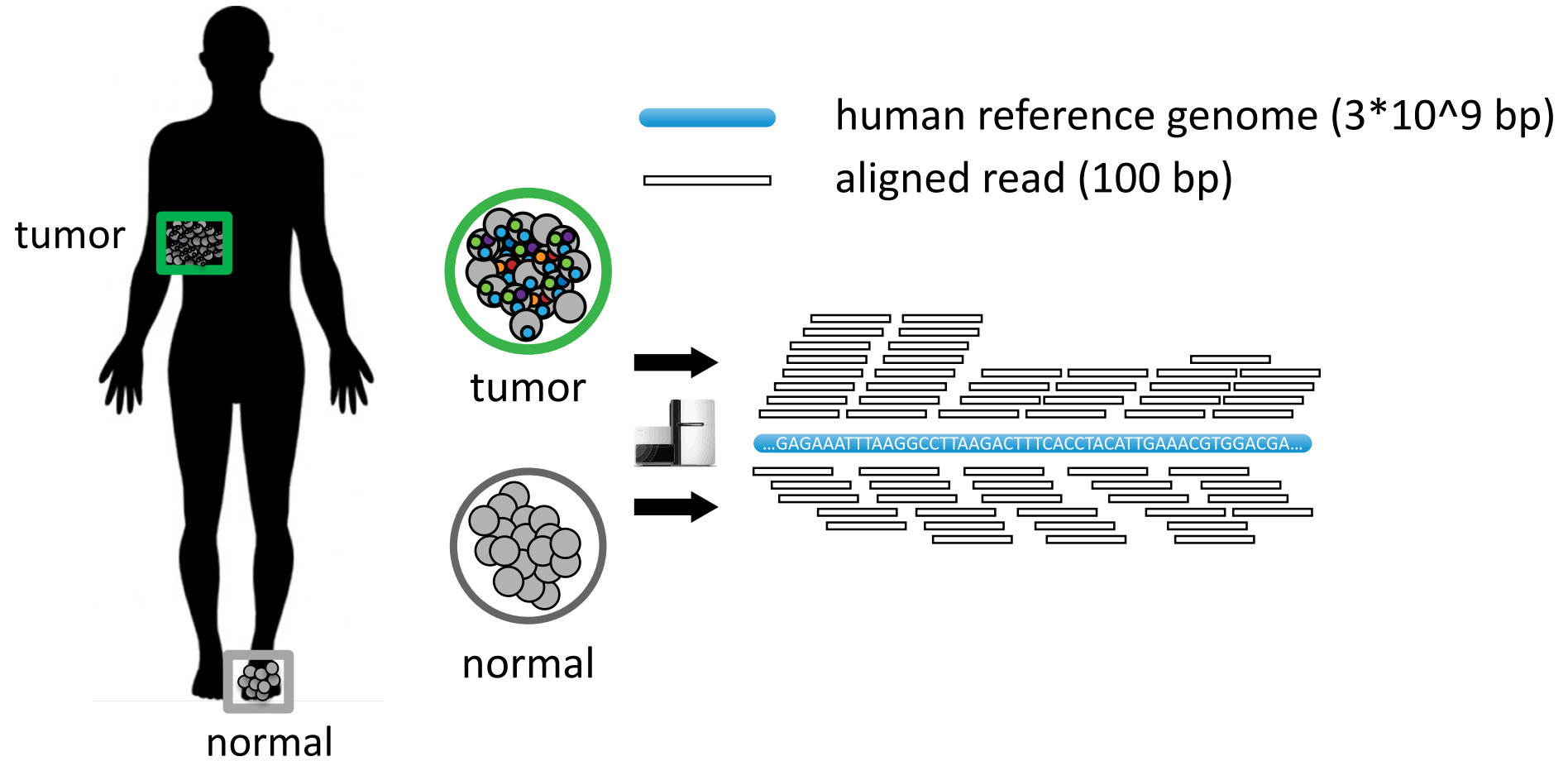
Recognize common patterns of tumor evolution across patients



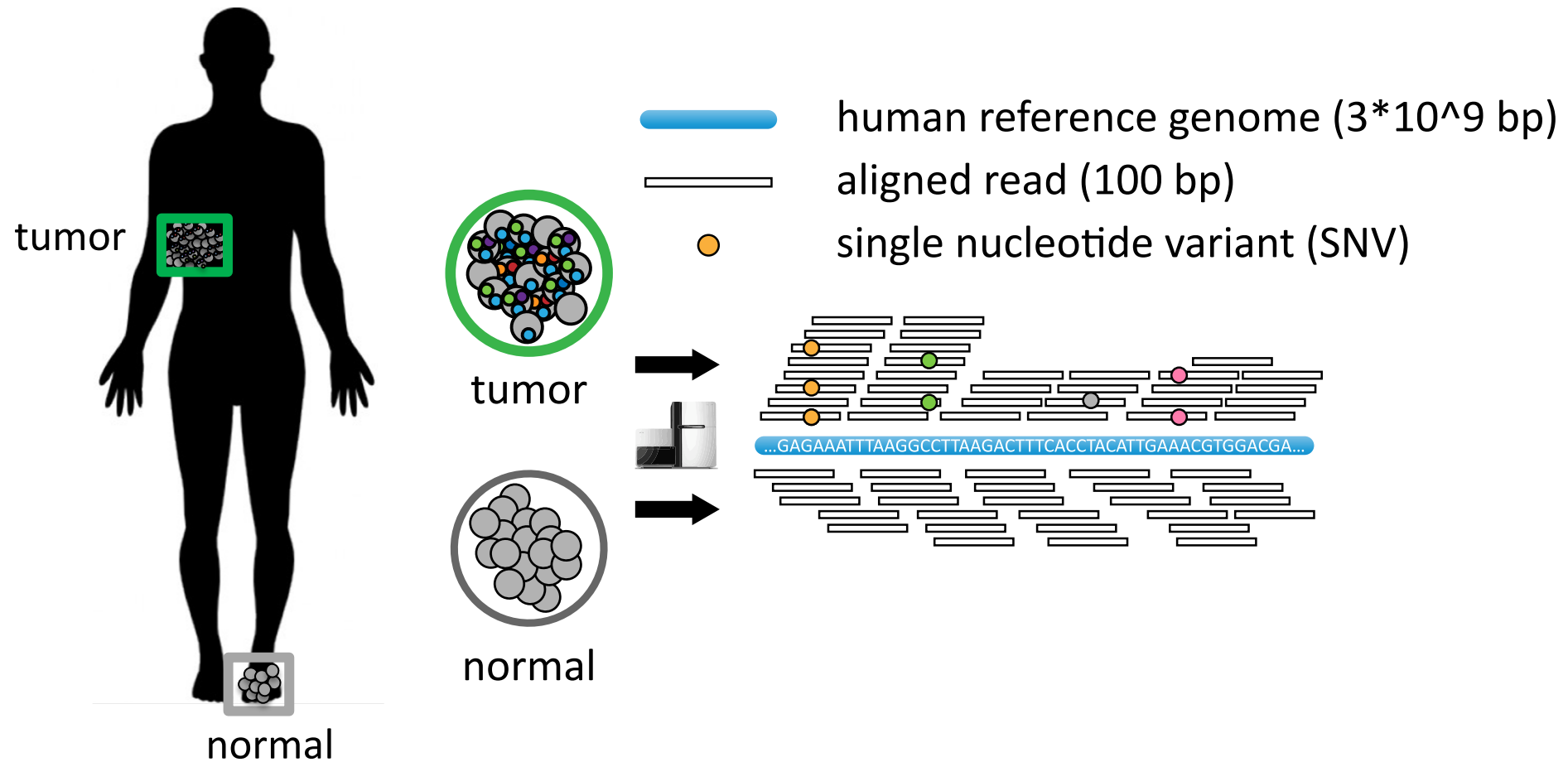
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Key challenge in phylogenetics:
Accurate phylogeny inference from data at present time

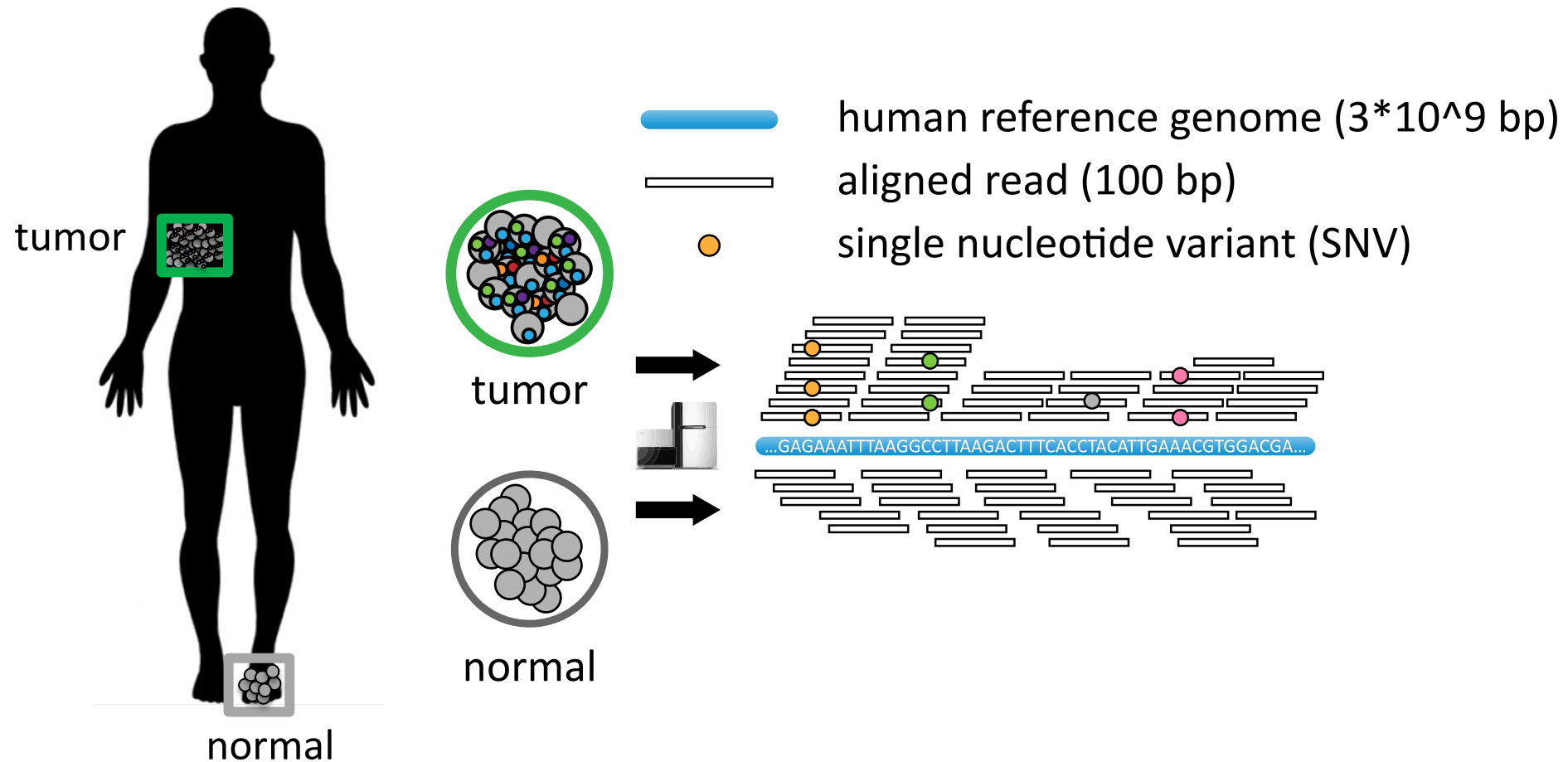
Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics









Additional challenge in cancer phylogenetics:
Phylogeny inference from **mixed bulk samples** at present time

Tumor Phylogeny Inference

Metastatic Colorectal Cancer (Patient CRC2)

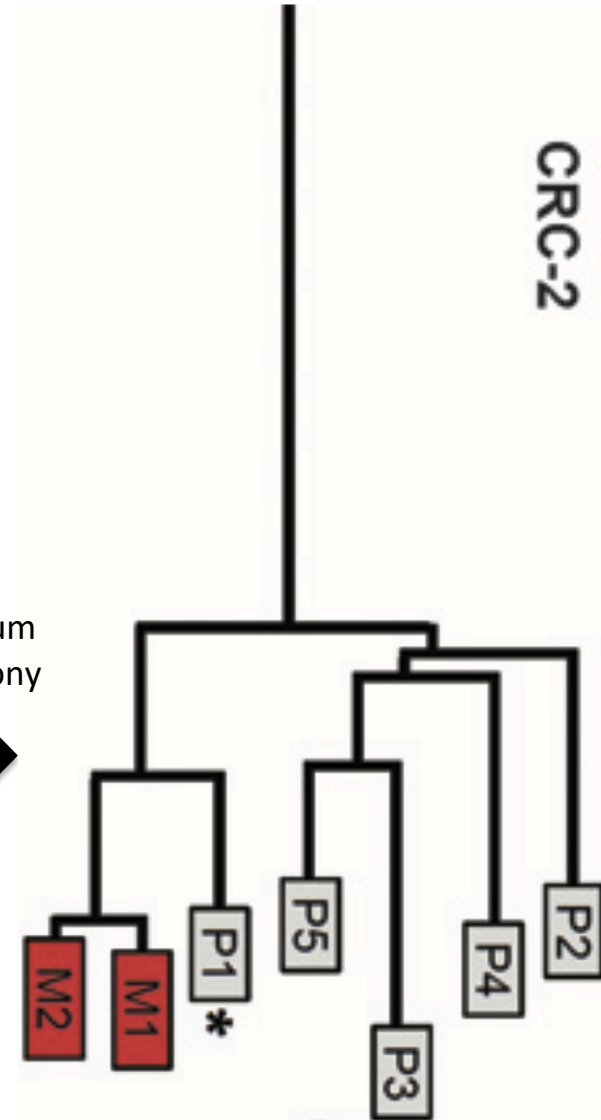
[Kim et al., *Clin Cancer Res* 21(19), 2015]:

- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)

	<i>n</i> mutations					
						
<i>m</i> samples						
P1	1	1	1	0	0	0
P2	1	1	0	1	0	0
P3	1	0	1	0	1	1
P4	0	1	1	0	0	0
P5	0	1	0	1	0	1
M1	1	1	0	0	1	0
M2	0	1	1	1	1	1

Binary Matrix **B**

Maximum
Parsimony



Tumor Phylogeny Inference







Metastatic Colorectal Cancer (Patient CRC2)

[Kim et al., *Clin Cancer Res* 21(19), 2015]:

- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (**homoplasy**)

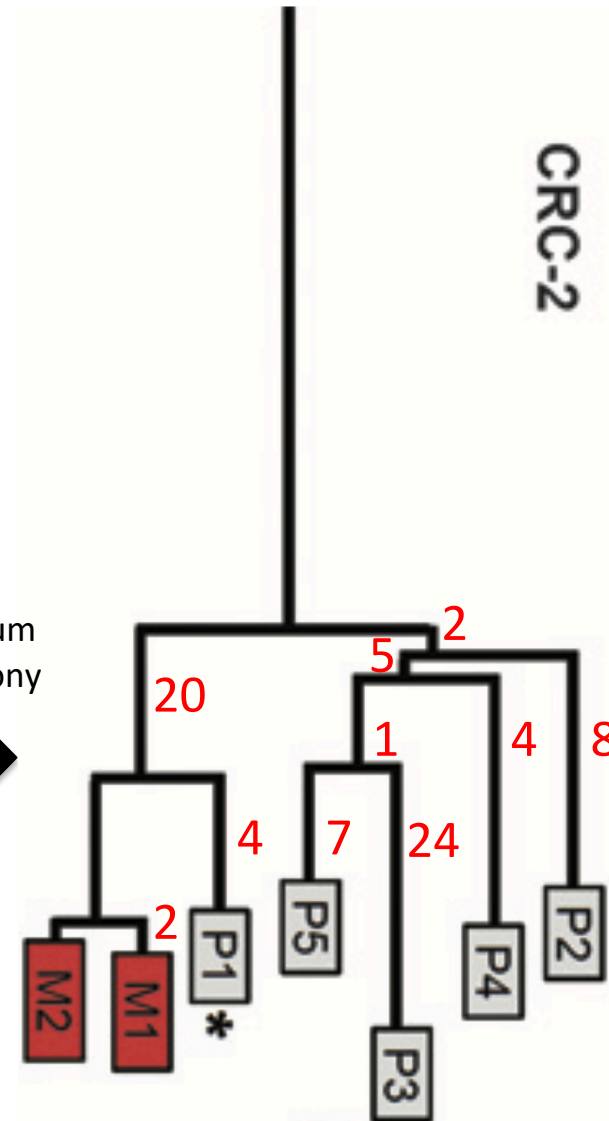
m samples

n mutations

						
P1	1	1	1	0	0	0
P2	1	1	0	1	0	0
P3	1	0	1	0	1	1
P4	0	1	1	0	0	0
P5	0	1	0	1	0	1
M1	1	1	0	0	1	0
M2	0	1	1	1	1	1

Binary Matrix B

Maximum
Parsimony

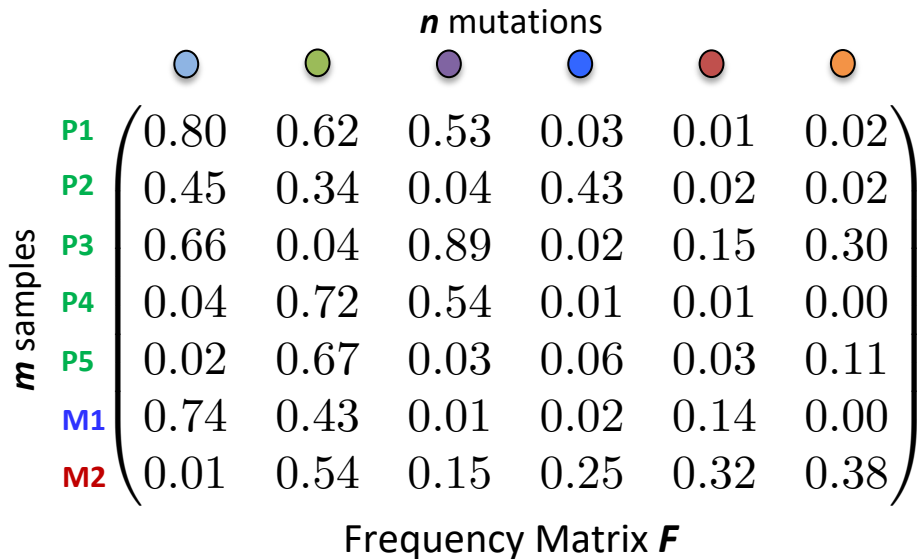


Heuristic for Tumor Phylogeny Inference

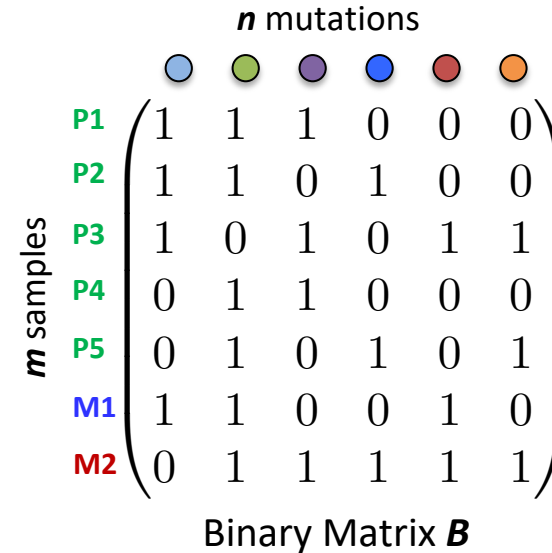
Metastatic Colorectal Cancer (Patient CRC2)

[Kim et al., *Clin Cancer Res* 21(19), 2015]:

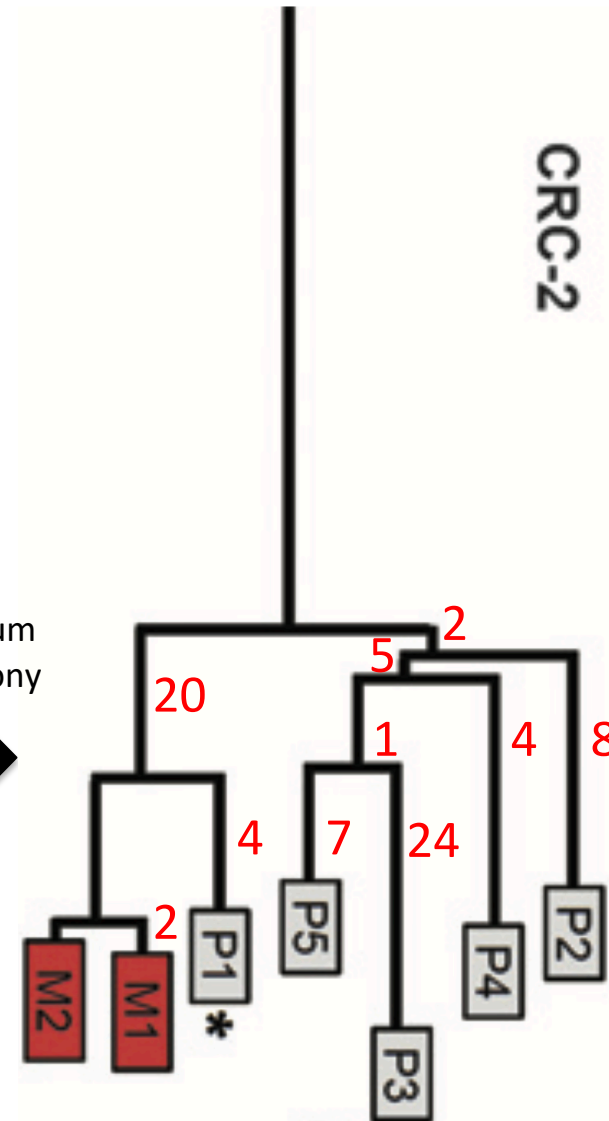
- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (**homoplasy**)



Discretize



Maximum Parsimony



Resulting **sample tree** is **not** representative of the division/mutation history or the migration history

Lecture Outline

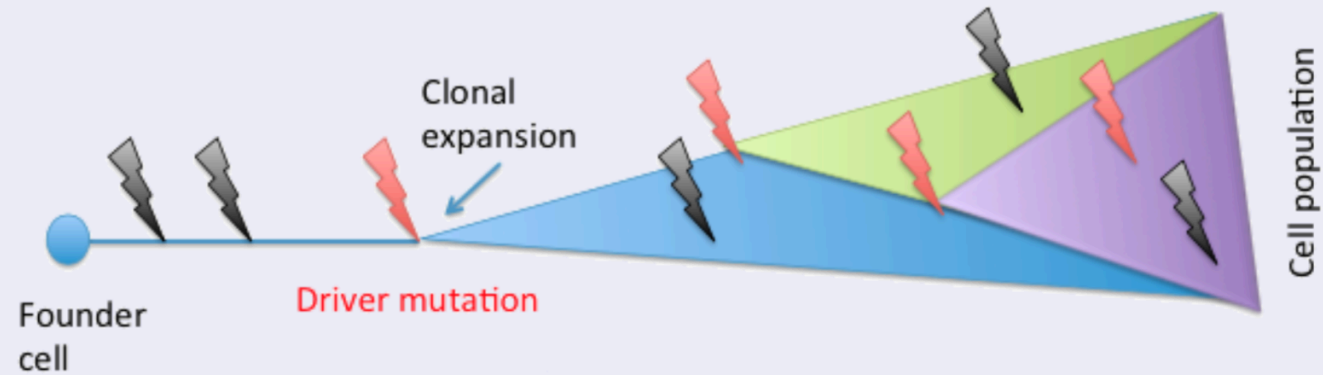
- Recap
- Maximum Parsimony
- Two-state Perfect Phylogeny
- Two-state Perfect Phylogeny Mixtures

Reading

- Lecture notes

Somatic Mutations and Cancer

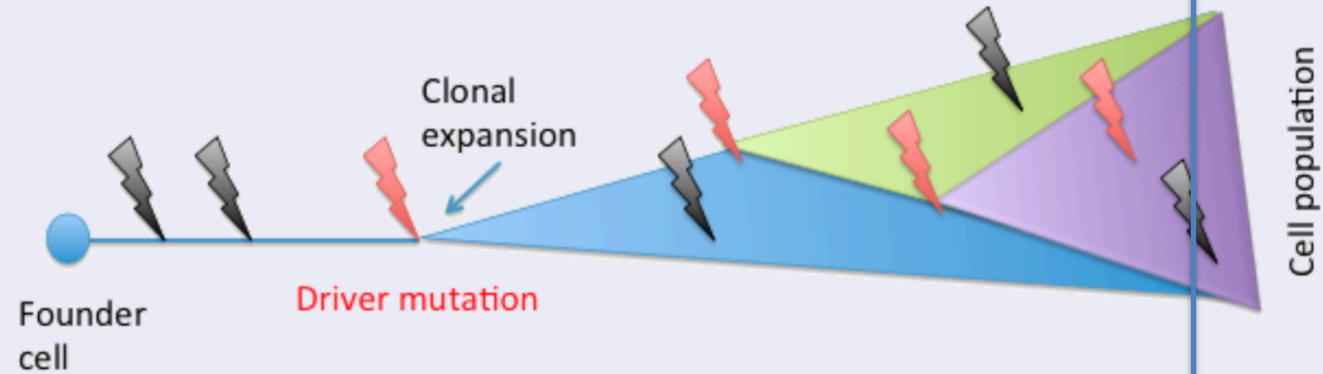
Clonal theory of cancer (Nowell, 1976)



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

Somatic Mutations and Cancer

Clonal theory of cancer (Nowell, 1976)



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



International
Cancer Genome
Consortium



Sequence genome

Progression of Somatic Mutations

Single nucleotide mutation

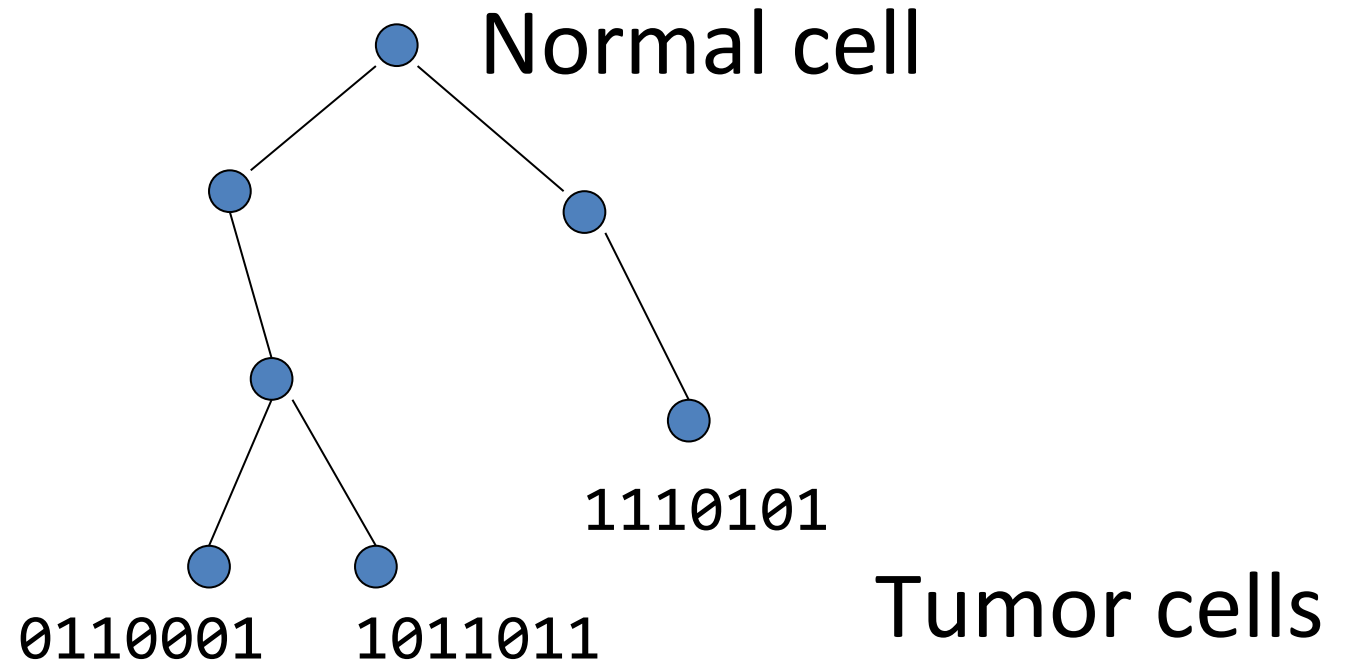
... CGT**A**TTAG ...



... CGT**C**TTAG ...

0 = normal

1 = mutated



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

Progression of Somatic Mutations

Single nucleotide mutation

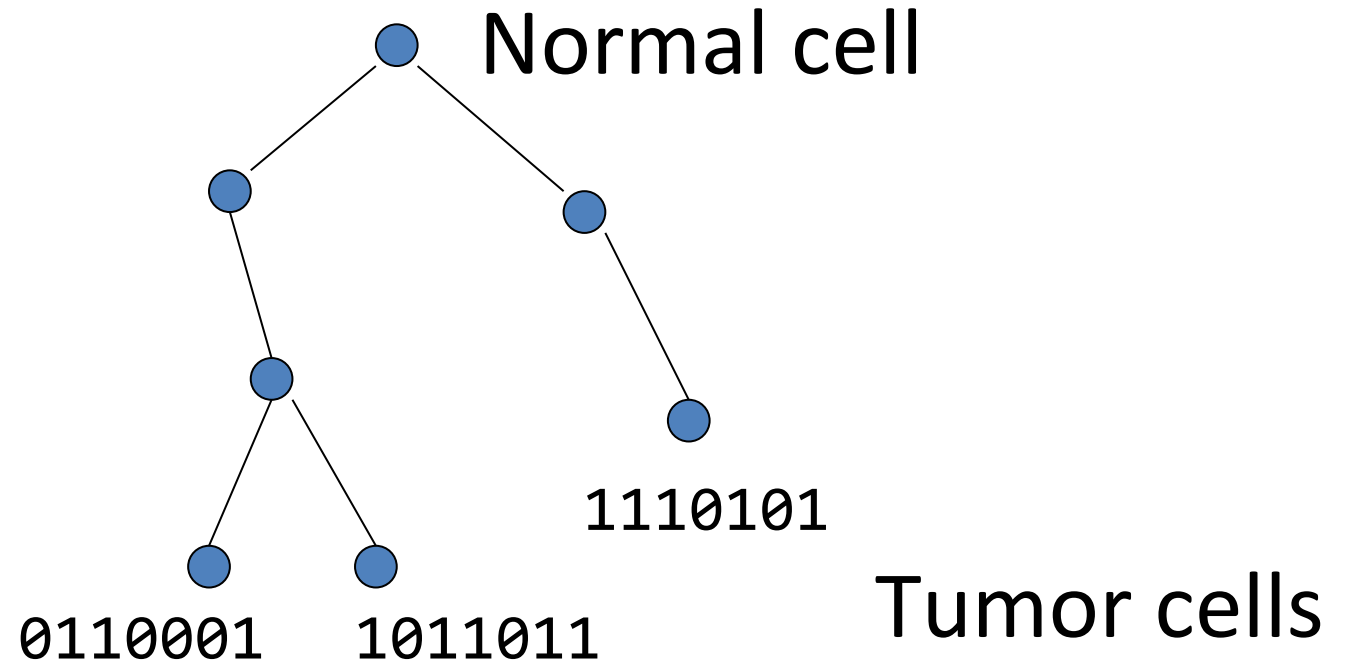
... CGT**A**TTAG ...



... CGT**C**TTAG ...

0 = normal

1 = mutated

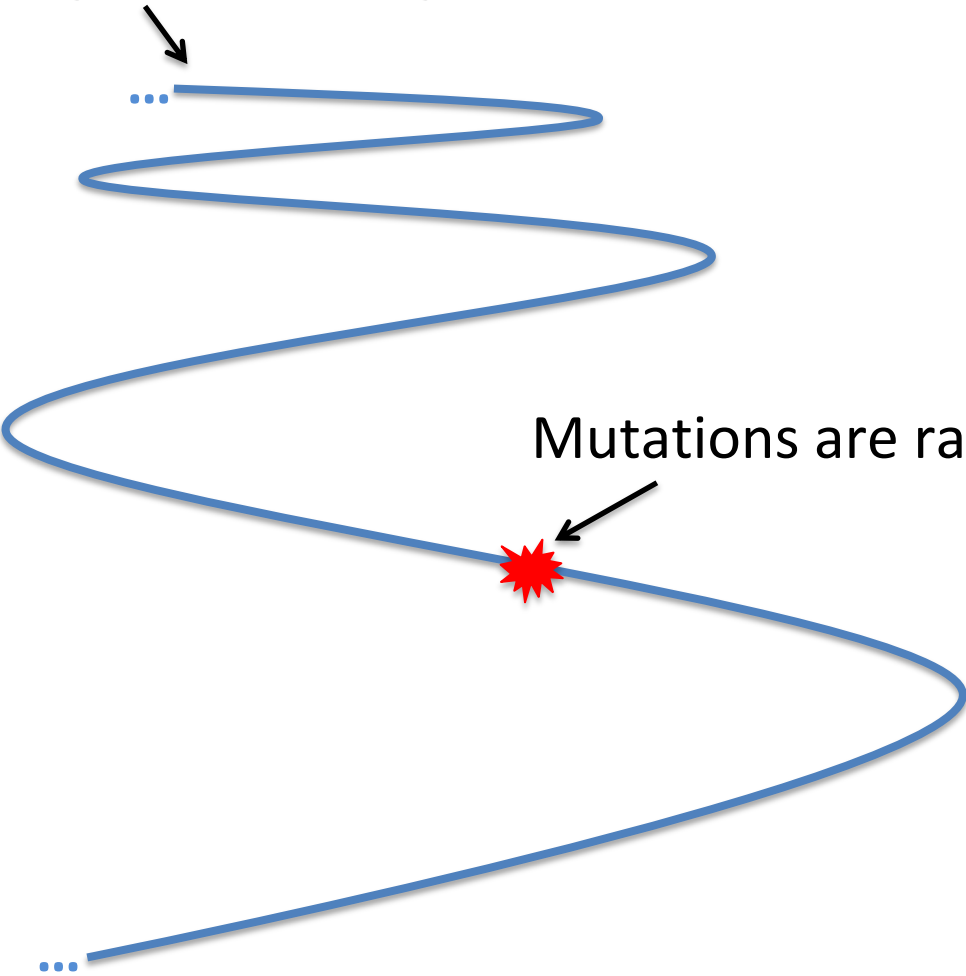


Root is the normal, founder cell and leaves are cells in tumor.

Infinite sites assumption: each locus mutates only once.







Infinite Sites Model

The genome is large



[Kimura, 1969]

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

		Mutated Loci					
							
Species (cancer cells)	A	0	0	0	0	1	1
	B	0	0	0	1	1	1
	C	0	0	1	0	1	0
	D	1	0	0	0	0	0
	E	1	1	0	0	0	0

1: mutated
0: not

All sites are bi-allelic: mutated or not.

Two-state Perfect Phylogeny

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

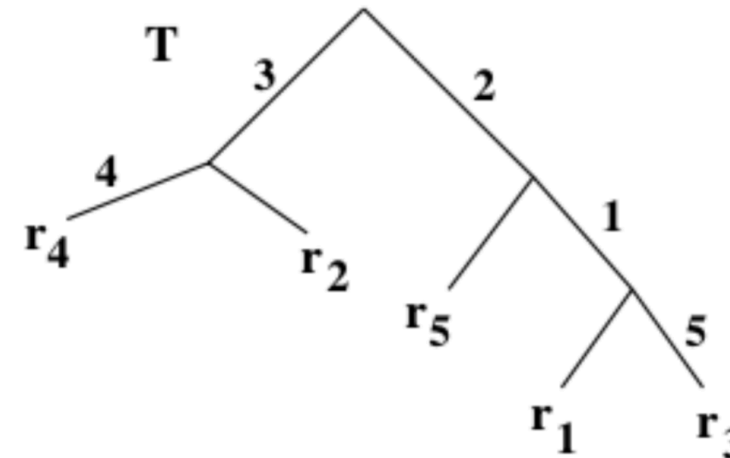
- Taxon f has state 1 for character c
 $\Leftrightarrow f$ **possesses** character c

	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

Definition

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

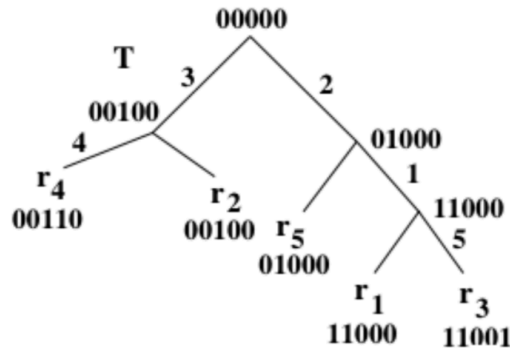
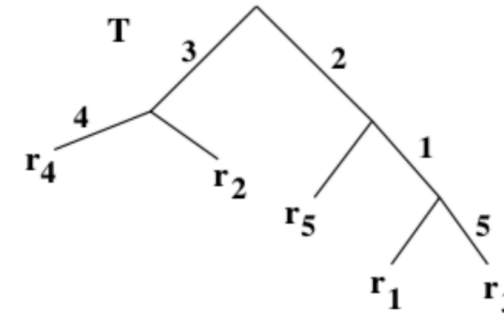
- 1 Each taxon labels only one leaf
- 2 Each character labels only one edge
- 3 Character possessed by a taxon are on unique path to root



Root node is all zero ancestor

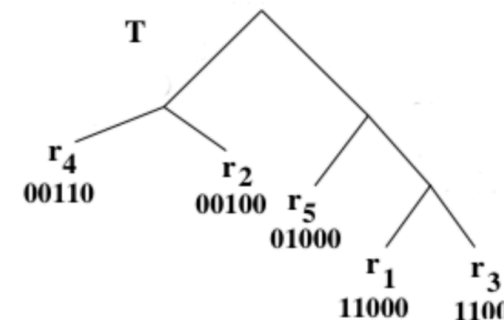
Two-state Perfect Phylogeny – Alternative Definitions

- 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
 $\Leftrightarrow f$ **possesses** character c

	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

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?

$M_1 =$

Species	Characters				
	C_1	C_2	C_3	C_4	C_5
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

$M_2 =$

Species	Characters				
	C_1	C_2	C_3	C_4	C_5
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

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 .

	c_1	c_2	c_3	c_4	c_5		c_1 (2)	c_2 (1)	c_3 (3)	c_4 (5)	c_5 (4)
r_1	1	1	0	0	0	\Rightarrow	1	1	0	0	0
r_2	0	0	1	0	0		0	0	1	0	0
r_3	1	1	0	0	1		1	1	0	1	0
r_4	0	0	1	1	0		0	0	1	0	1
r_5	0	1	0	0	0		1	0	0	0	0

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

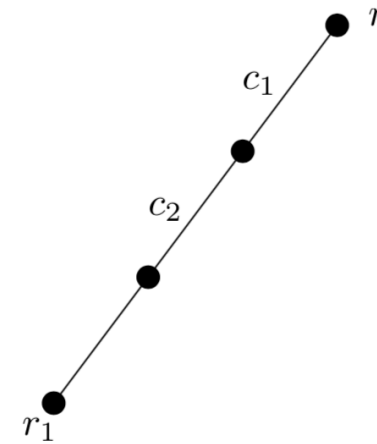
$$\sigma(r_1) = \{c_1, c_2\}$$

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 $\sigma(1)$

$$c < d \text{ iff } |I(c)| \geq |I(d)|$$

	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



- 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

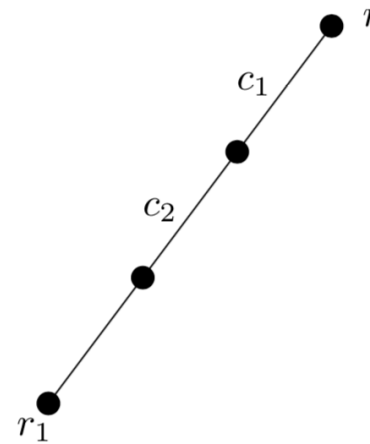
- Let v be last node on walk from r matching characters $\sigma(i+1)$

- ★ Character d is the last match

- ★ Unmatched characters $\tau(i+1)$

$$c < d \text{ iff } |I(c)| \geq |I(d)|$$

	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



- 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 $\tau(i+1)$

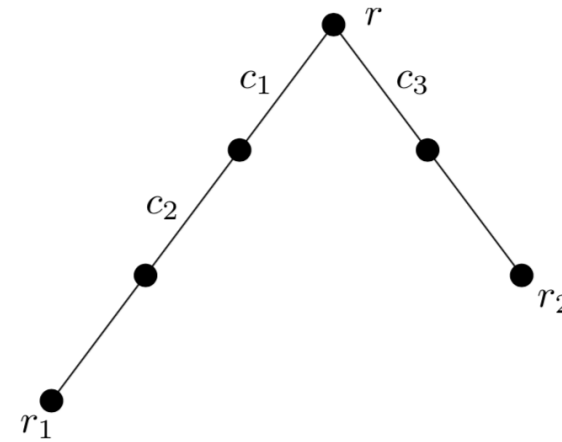
- Extend T_i with path Π

- ★ Π has terminals v and $i+1$

- ★ Π has $|\tau(i+1)| + 1$ edges labeled by $\tau(i+1)$

$$c < d \text{ iff } |I(c)| \geq |I(d)|$$

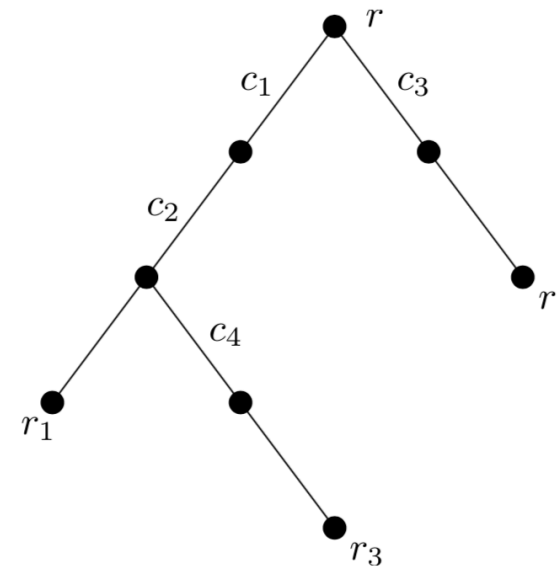
	c_1	c_2	c_3	c_4	c_5
r_1	1	1	0	0	0
r_2	0	0	1	0	0
r_3	1	1	0	1	0
r_4	0	0	1	0	1
r_5	1	0	0	0	0



- 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 Π
 - ★ Π has terminals v and $i + 1$
 - ★ Π has $|\tau(i + 1)| + 1$ edges labeled by $\tau(i + 1)$

$$c < d \text{ iff } |I(c)| \geq |I(d)|$$

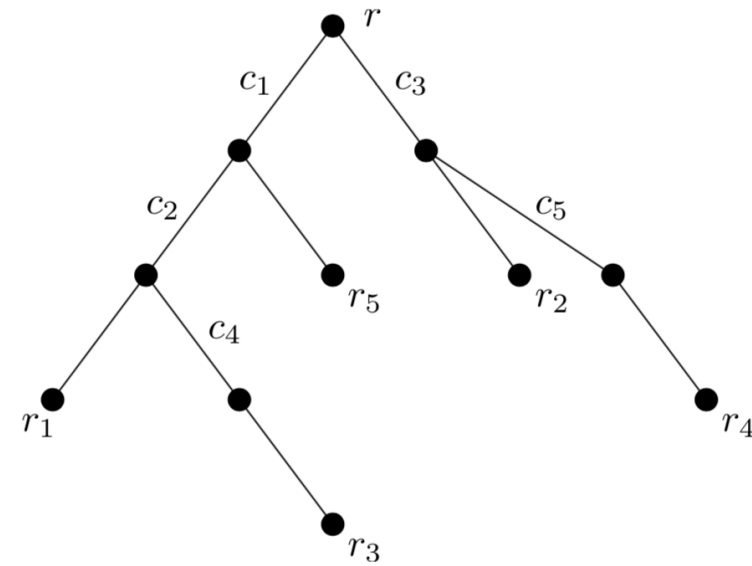
	c_1	c_2	c_3	c_4	c_5
r_1	1	1	0	0	0
r_2	0	0	1	0	0
r_3	1	1	0	1	0
r_4	0	0	1	0	1
r_5	1	0	0	0	0



- 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 Π
 - ★ Π has terminals v and $i+1$
 - ★ Π has $|\tau(i+1)| + 1$ edges labeled by $\tau(i+1)$

$$c < d \text{ iff } |I(c)| \geq |I(d)|$$

	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



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 .

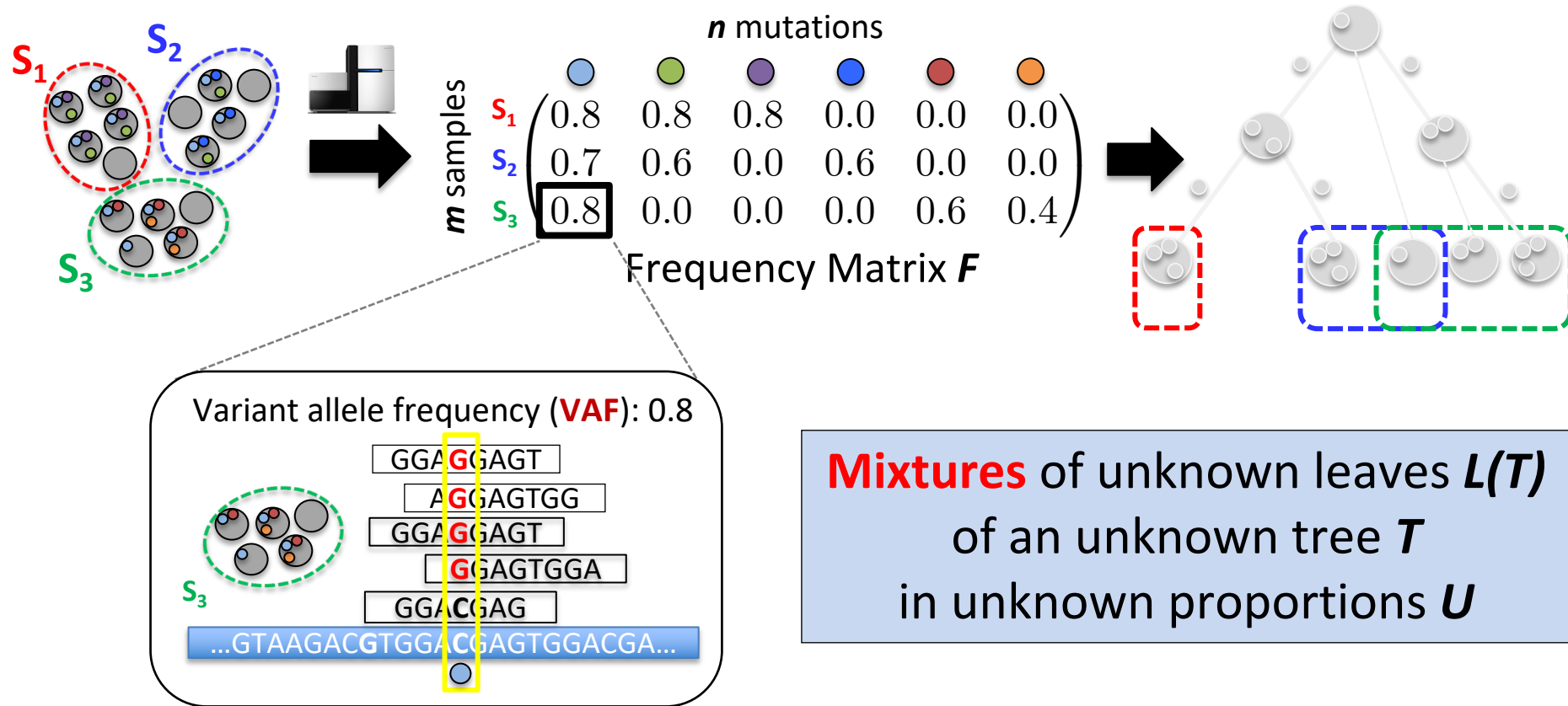
Lecture Outline

- Recap
- Maximum Parsimony
- Two-state Perfect Phylogeny
- Two-state Perfect Phylogeny Mixtures

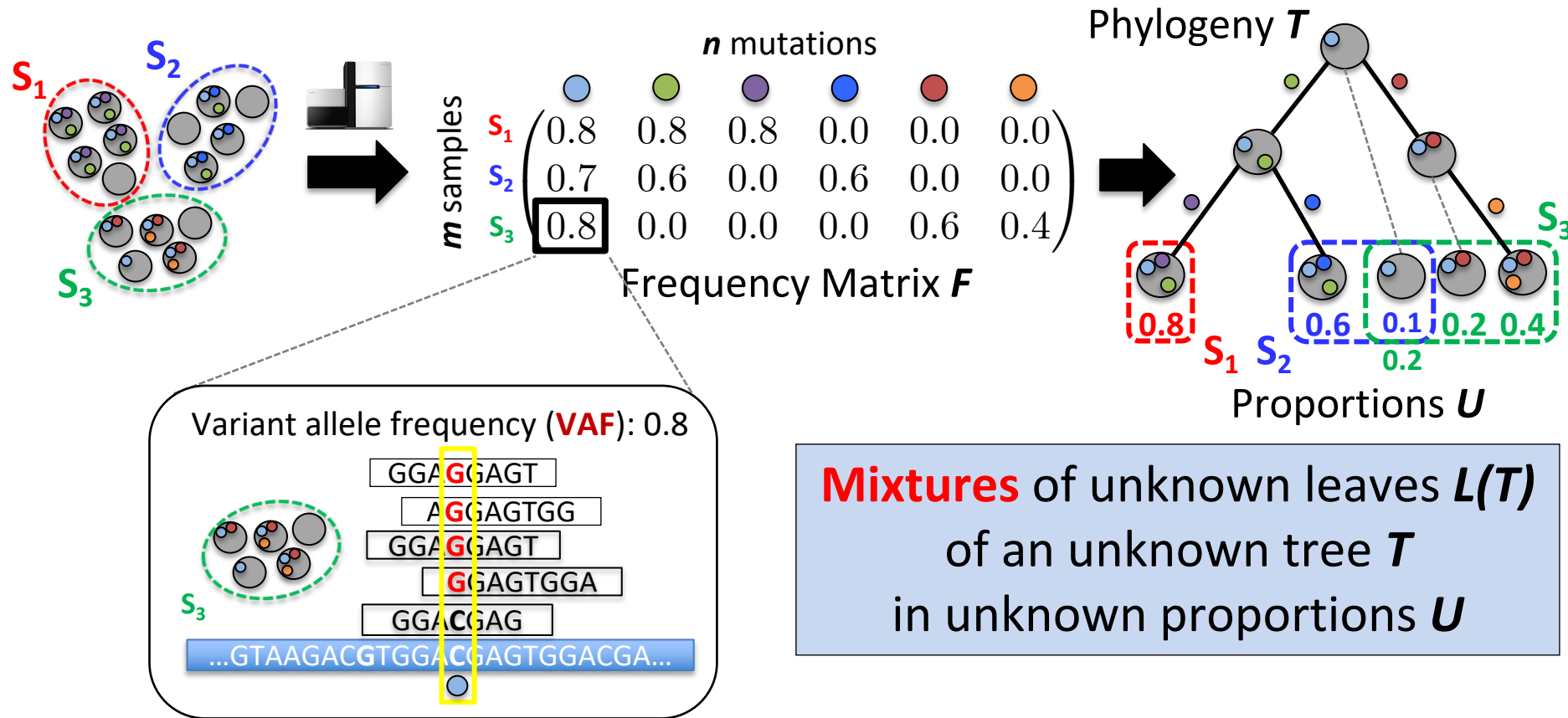
Reading

- Lecture notes

Sequencing and Tumor Phylogeny Inference

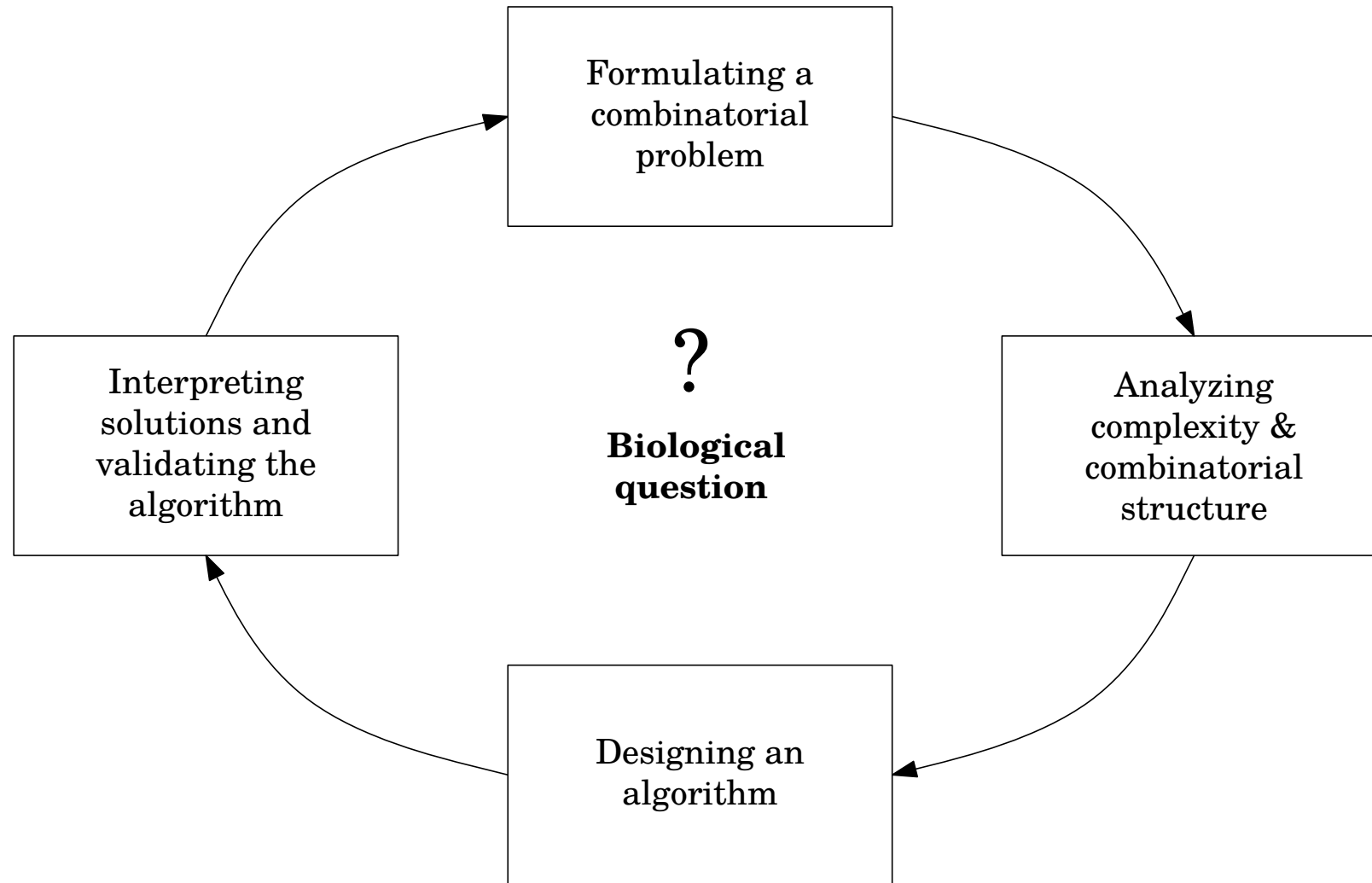


Sequencing and Tumor Phylogeny Inference



Tumor Phylogeny Inference: Given frequencies F , find phylogeny T and proportions U

Key Challenge in Computational Biology

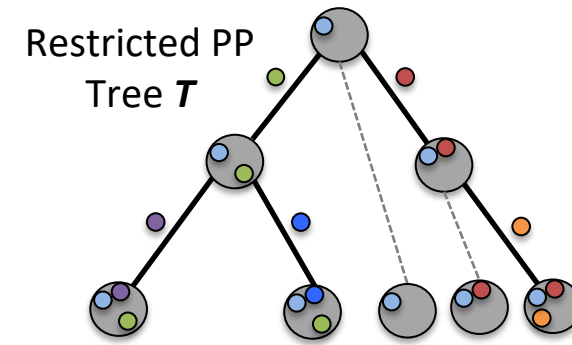


Translating a biological problem into a computational biology

Perfect Phylogeny Mixture

Assumptions:

- Infinite sites assumption: a character changes state once
- Error-free data



n mutations

m samples

Frequency Matrix F

$$F = \begin{matrix} & \begin{matrix} \text{blue} & \text{green} & \text{purple} & \text{blue} & \text{red} & \text{orange} \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix} & \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix} \end{matrix}$$

$=$

m samples

Mixture Matrix U

clones

$$U = \begin{matrix} & \begin{matrix} \text{clone 1} & \text{clone 2} & \text{clone 3} & \text{clone 4} & \text{clone 5} & \text{clone 6} \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix} & \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix} \end{matrix}$$

n mutations

Restricted PP Matrix B

clones

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

Rows of U are proportions:

$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem

[Estabrook, 1971]

[Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

Given F , find U and B such that $F = U B$

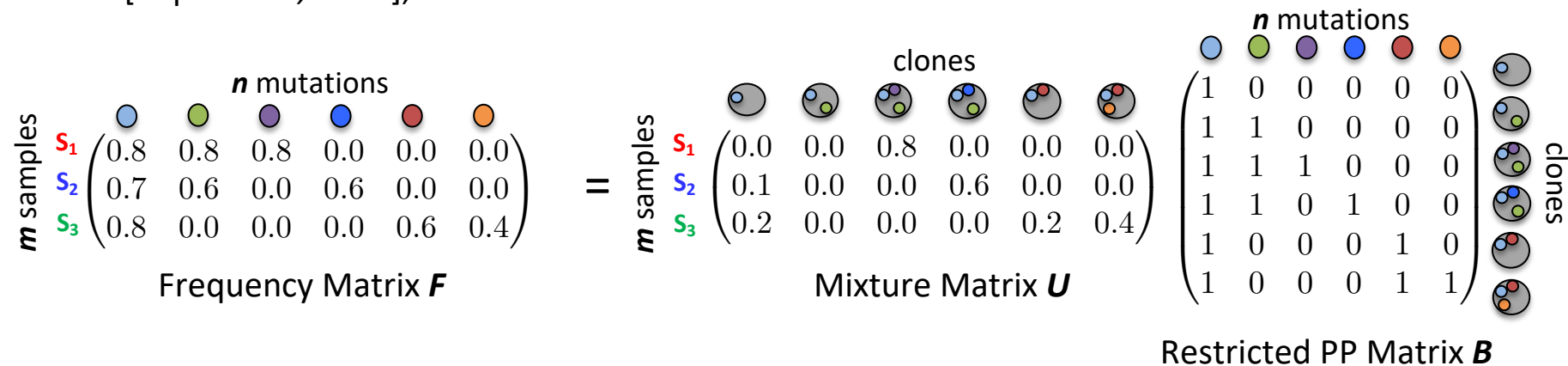
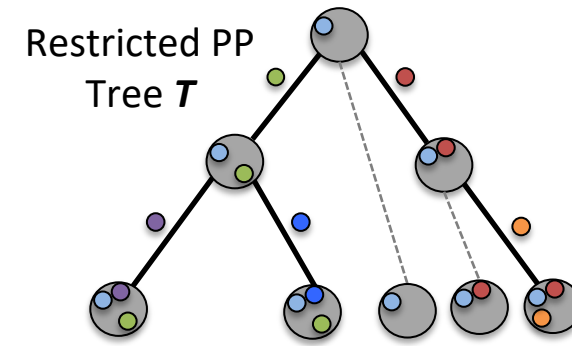
Previous Work

Variant of PPM:

TrAp [Strino *et al.*, 2013], PhyloSub [Jiao *et al.*, 2014]

CITUP [Malikic *et al.*, 2015], BitPhylogeny [Yuan *et al.*, 2015]

LICHeE [Popic *et al.*, 2015], ...



Rows of U are proportions:

$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem

[Estabrook, 1971]

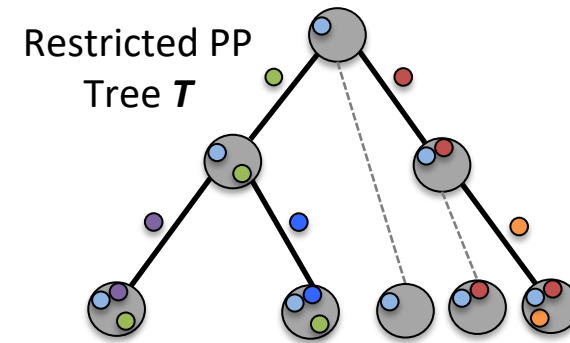
[Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

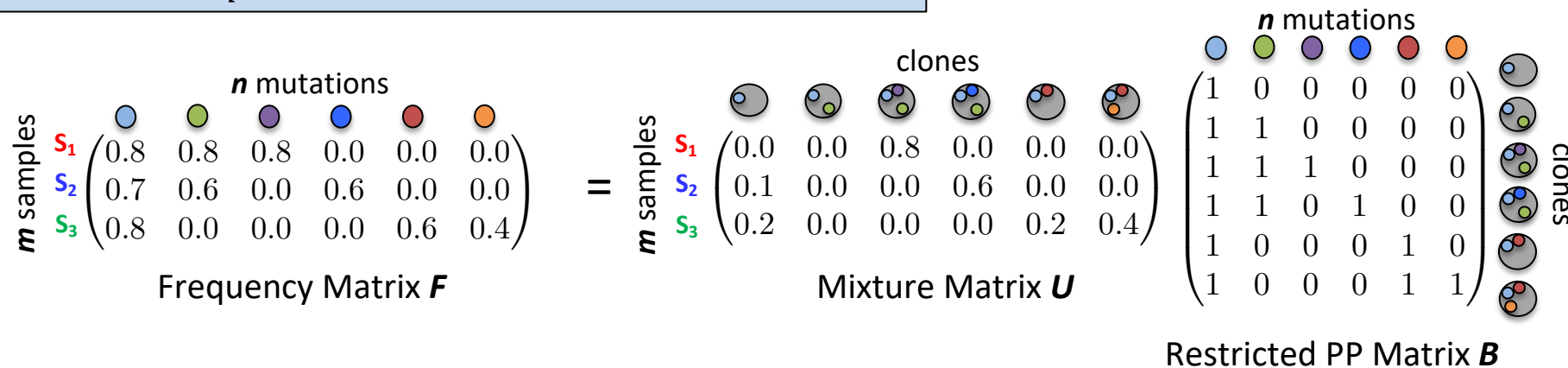
Given F , find U and B such that $F = UB$

Combinatorial Characterization

- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced i
- Usage $u_{p,i}$ is mass of node that introduced i



1-1 \updownarrow Equivalent



Rows of U are proportions:

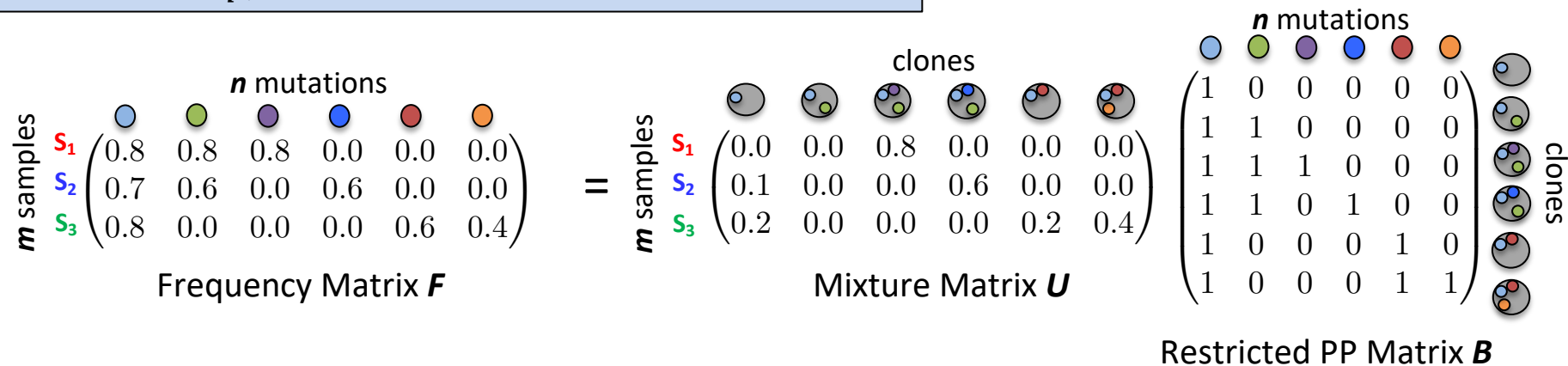
$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

Given F , find U and B such that $F = U B$

Combinatorial Characterization

- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced i
- Usage $u_{p,i}$ is mass of node that introduced i



Theorem 1:

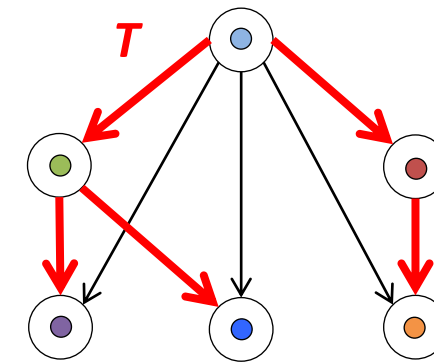
T is a solution to the PPM if and only if T is a spanning tree of G satisfying the sum condition

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

Given F , find U and B such that $F = U B$

Combinatorial Characterization

- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced i
- Usage $u_{p,i}$ is mass of node that introduced i



G

m samples $\begin{matrix} \text{S}_1 \\ \text{S}_2 \\ \text{S}_3 \end{matrix}$ n mutations

S_1	0.8	0.8	0.8	0.0	0.0	0.0
S_2	0.7	0.6	0.0	0.6	0.0	0.0
S_3	0.8	0.0	0.0	0.0	0.6	0.4

Frequency Matrix **F**

m samples $\begin{matrix} \text{S}_1 \\ \text{S}_2 \\ \text{S}_3 \end{matrix}$ clones

S_1	0.0	0.0	0.8	0.0	0.0	0.0
S_2	0.1	0.0	0.0	0.6	0.0	0.0
S_3	0.2	0.0	0.0	0.0	0.2	0.4

Mixture Matrix **U**

n mutations

S_1	1	0	0	0	0	0
S_2	1	1	0	0	0	0
S_3	1	1	1	0	0	0
S_4	1	1	0	1	0	0
S_5	1	0	0	0	1	0
S_6	1	0	0	0	1	1

Restricted PP Matrix **B**

clones

Theorem 1:

T is a solution to the PPM if and only if T is a spanning tree of **G** satisfying the sum condition

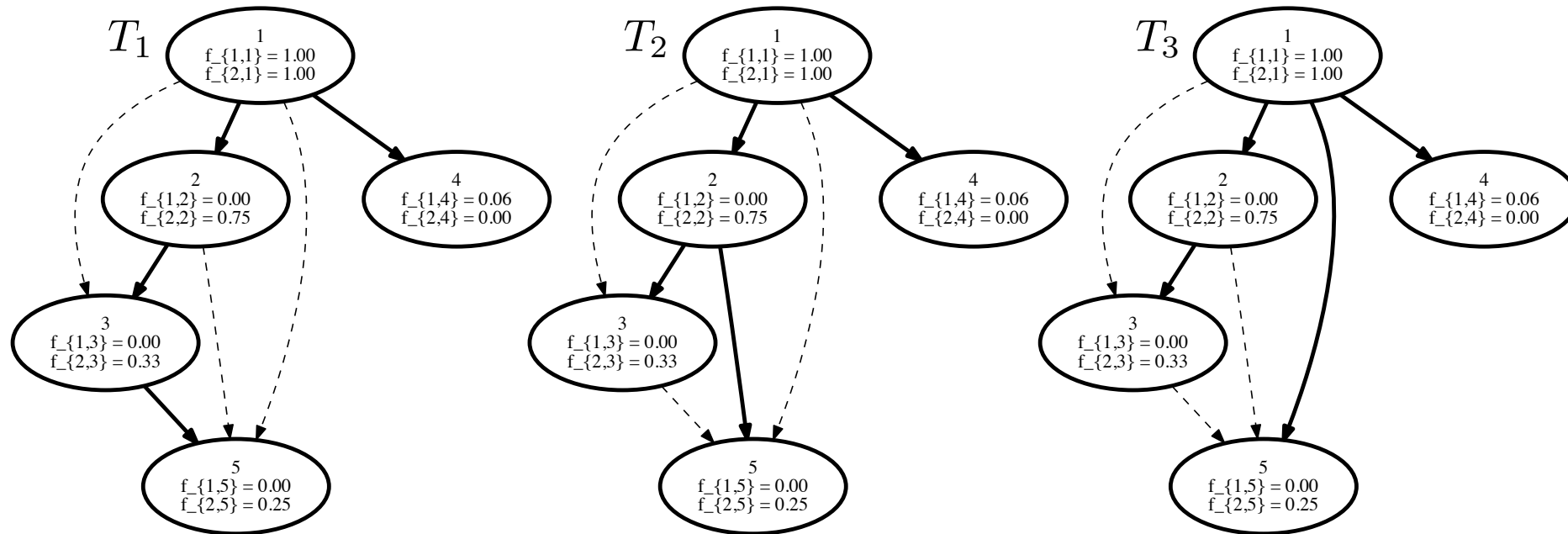
Theorem 2:

PPM is NP-complete even for $m=2$

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

Given **F**, find **U** and **B** such that **F = U B**

Non-uniqueness of Solutions to PPM



$$F = \begin{pmatrix} 1 & 0 & 0 & 0.06 & 0 \\ 1 & 0.75 & 0.33 & 0 & 0.25 \end{pmatrix}$$

Question 1: Can we determine the number of solutions?

Question 2: Can sample solutions uniformly at random?

Summary of Lectures 1 & 2

- DNA, RNA and proteins are sequences
 - Central dogma of molecular biology: DNA -> RNA -> protein
- Problem != algorithm
- Key challenge in computational biology is translating a biological problem into a computational problem
- Cancer is a genetic disease caused by somatic mutations
- Inter-tumor heterogeneity and intra-tumor heterogeneity:
 - *Not only is every tumor different, but so is every tumor cell...*
- Reading:
 - “Biology for Computer Scientists” by Lawrence Hunter
(http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)