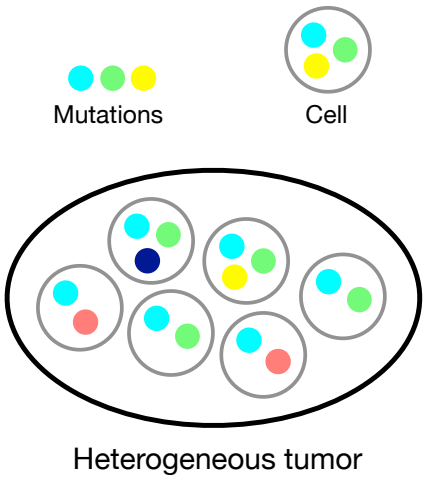
Distance Measures for Tumor Evolutionary Trees

Zach DiNardo*[1], Kiran Tomlinson*[1], Anna Ritz[2], and Layla Oesper[1]

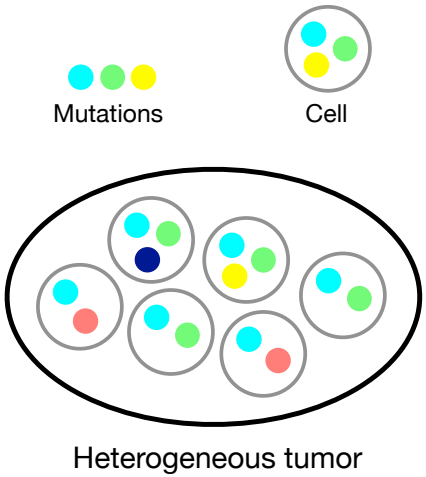
[1]Department of Computer Science, Carleton College
[2]Department of Biology, Reed College
*Joint first author

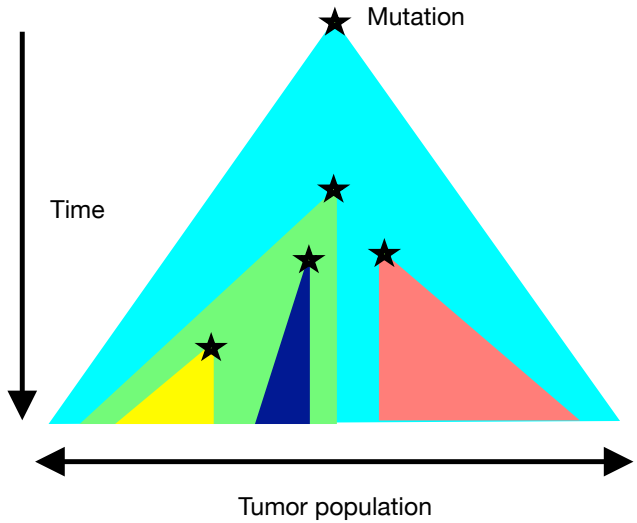
May 3, 2019

Clonal Theory (Nowell 1976)

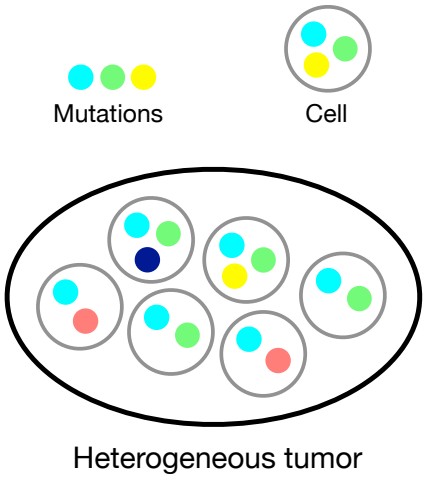


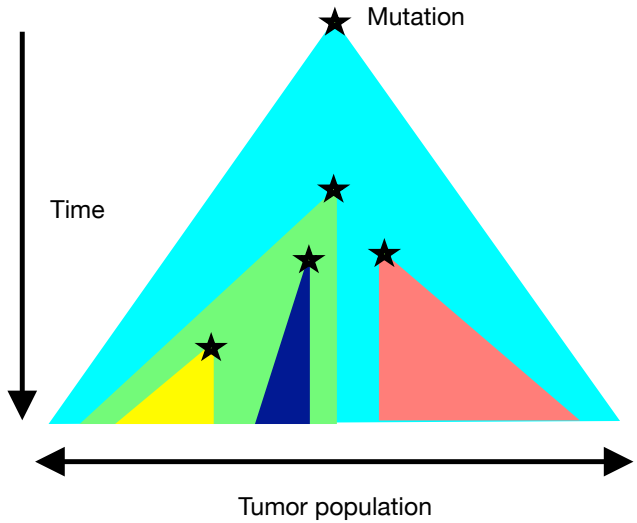
Clonal Theory (Nowell 1976)

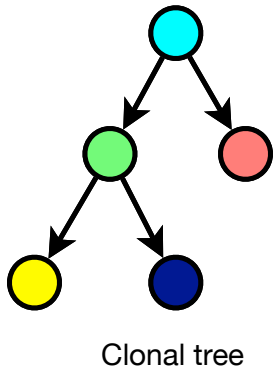




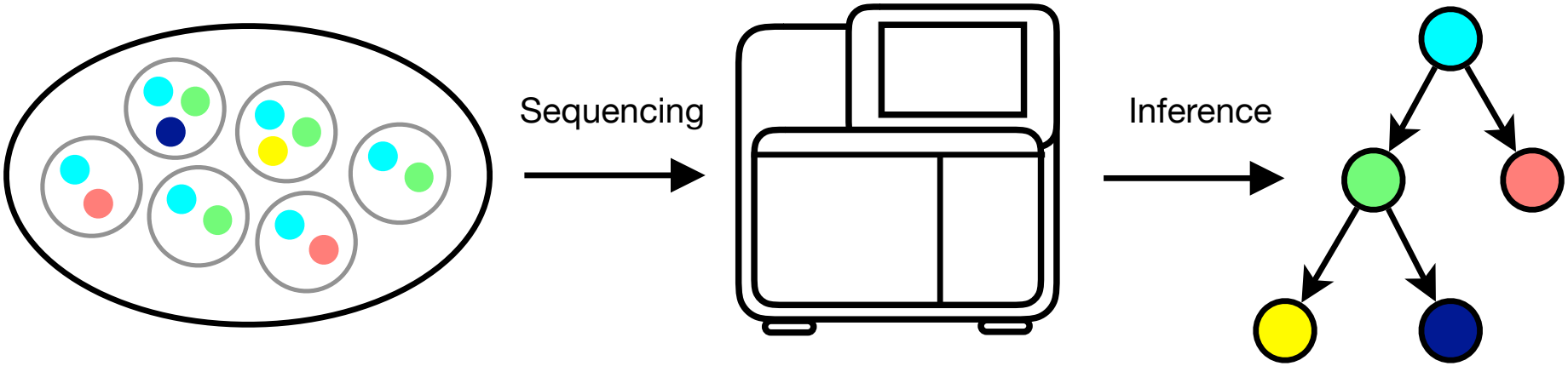
Clonal Theory (Nowell 1976)



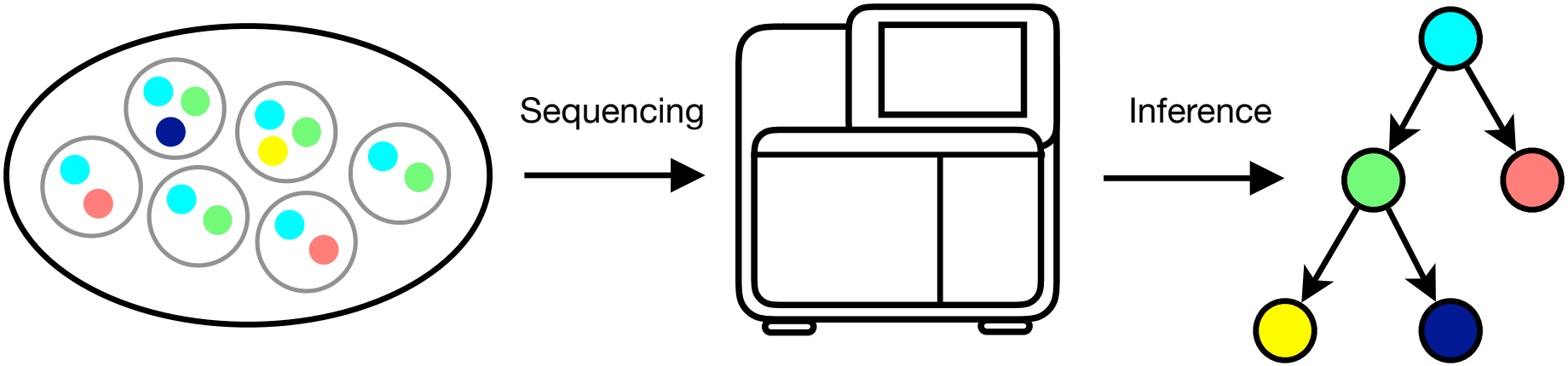




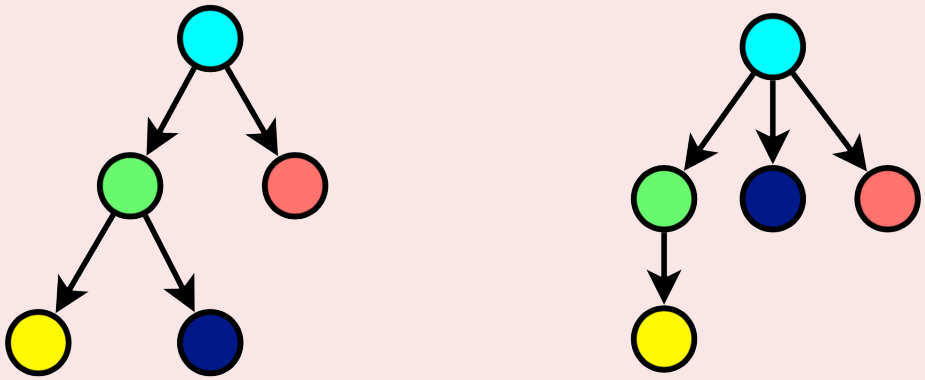
Clonal Tree Inference



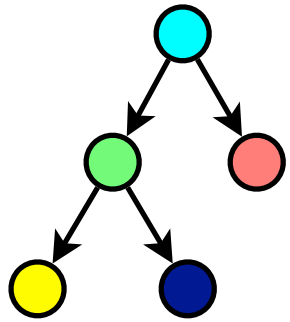
Clonal Tree Inference

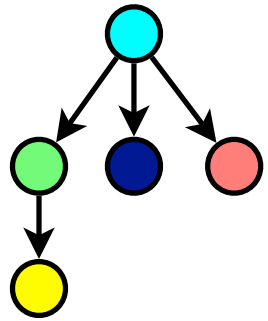


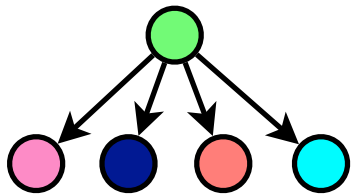
How do we compare two clonal trees?



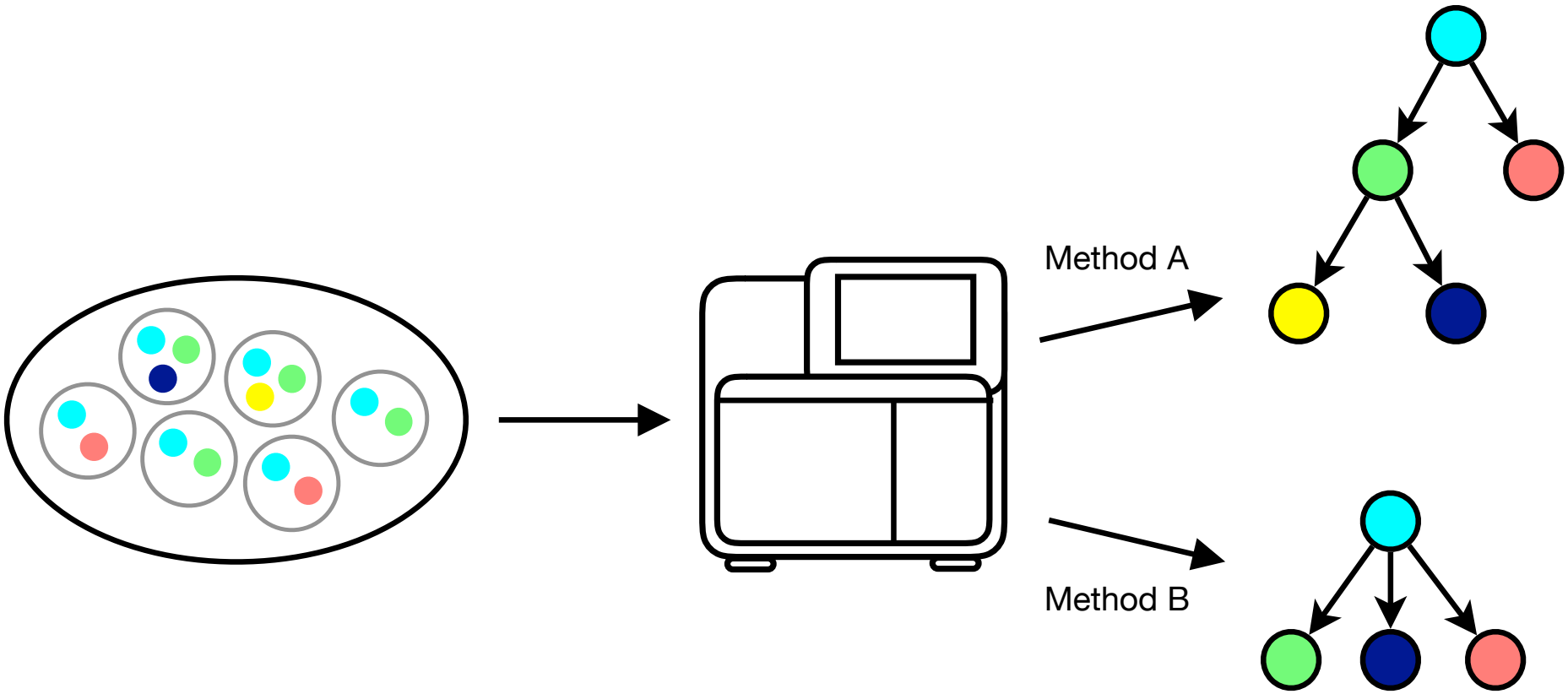
Outline



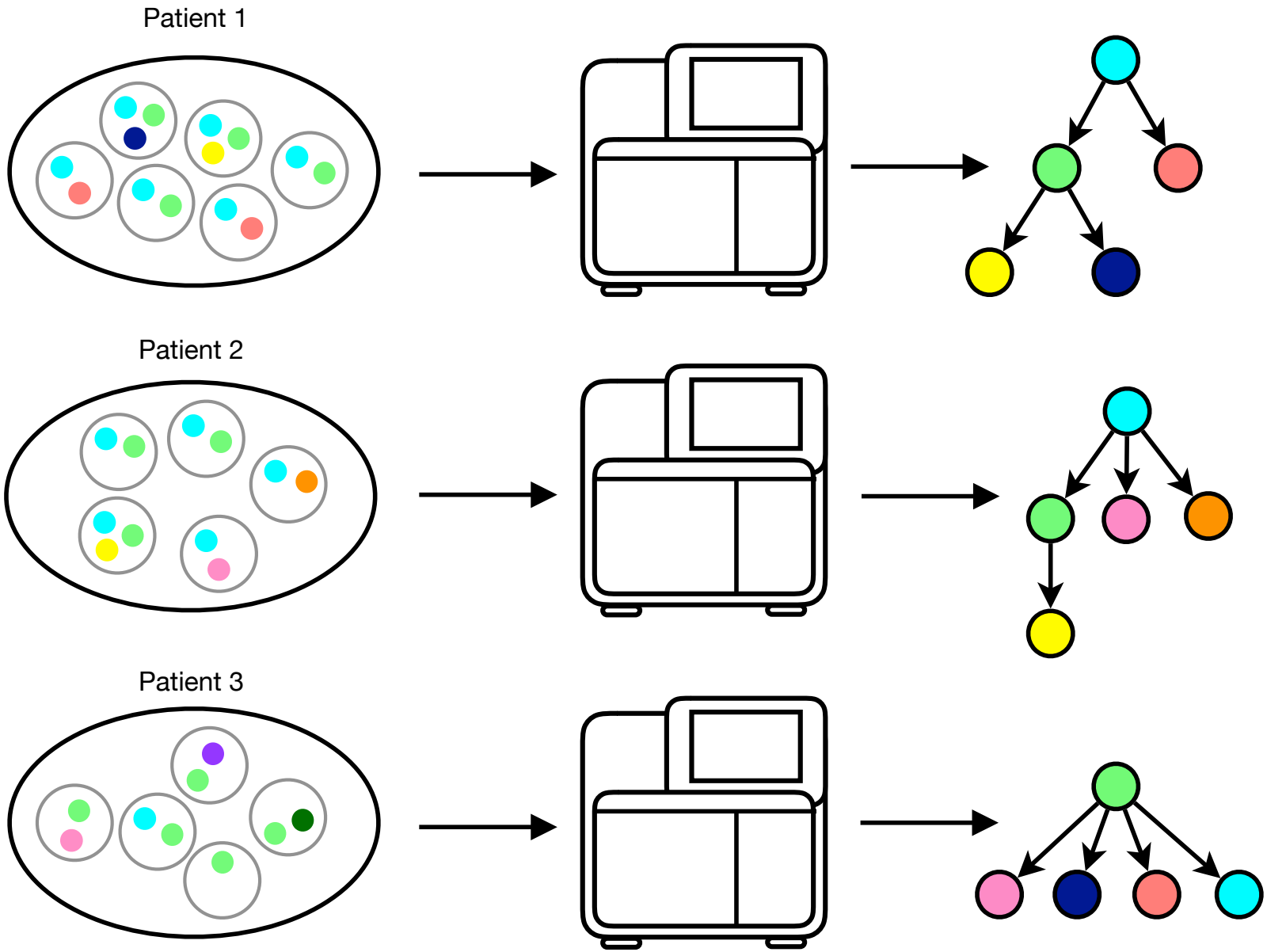




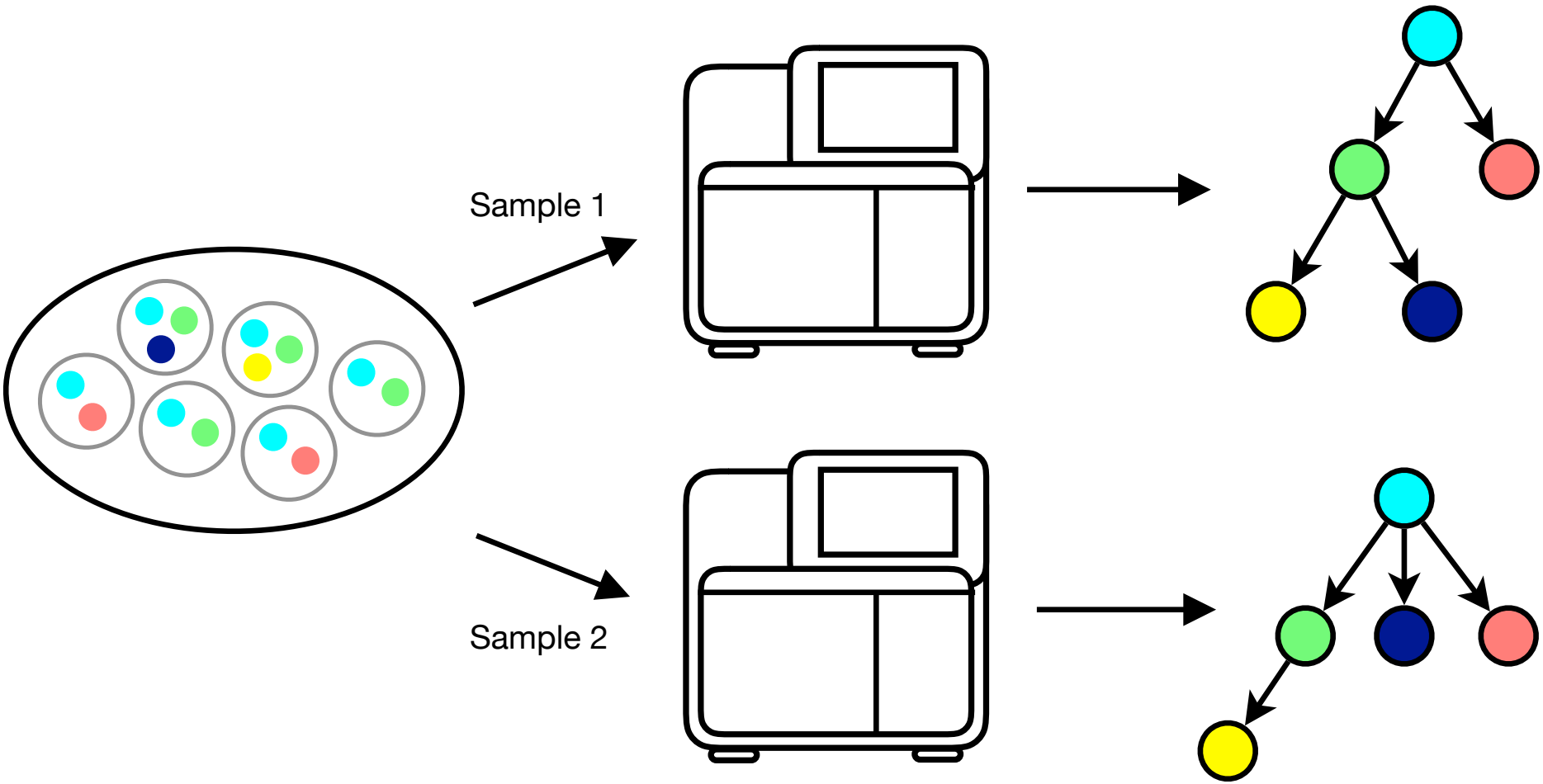
Multiple Inference Methods



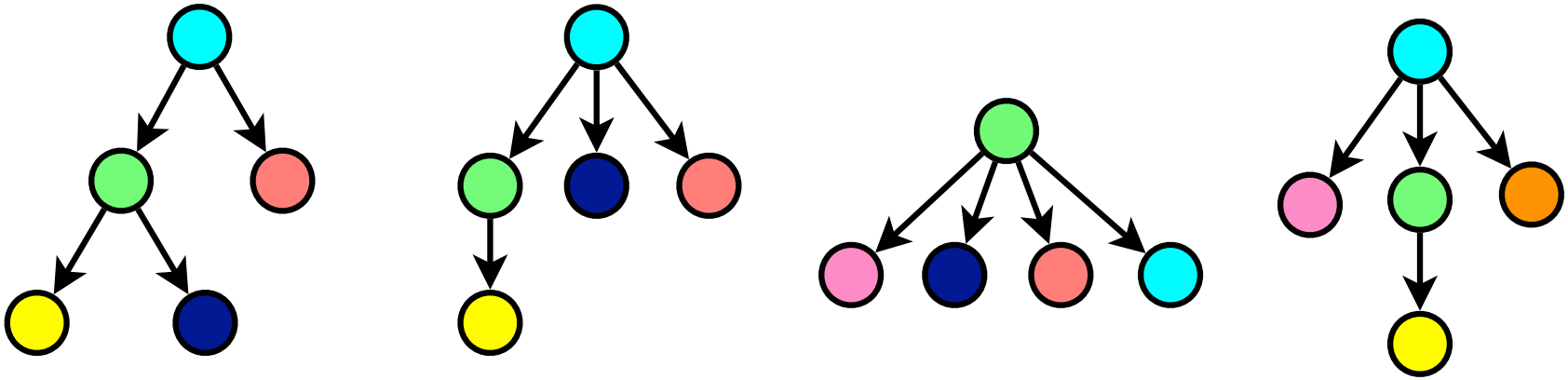
Inter-Patient Comparison



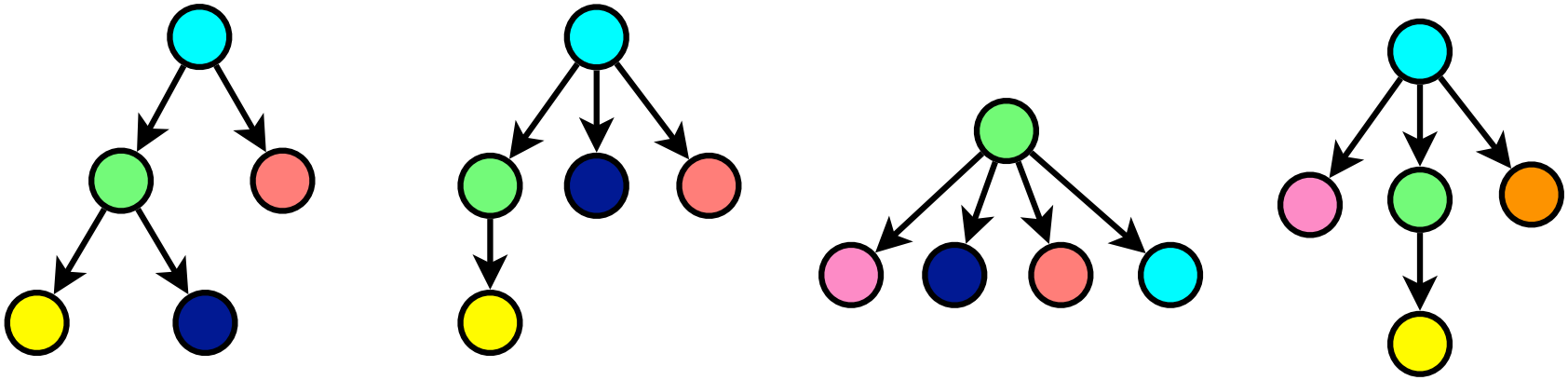
Intra-Patient Comparison



Need for Distance Measures



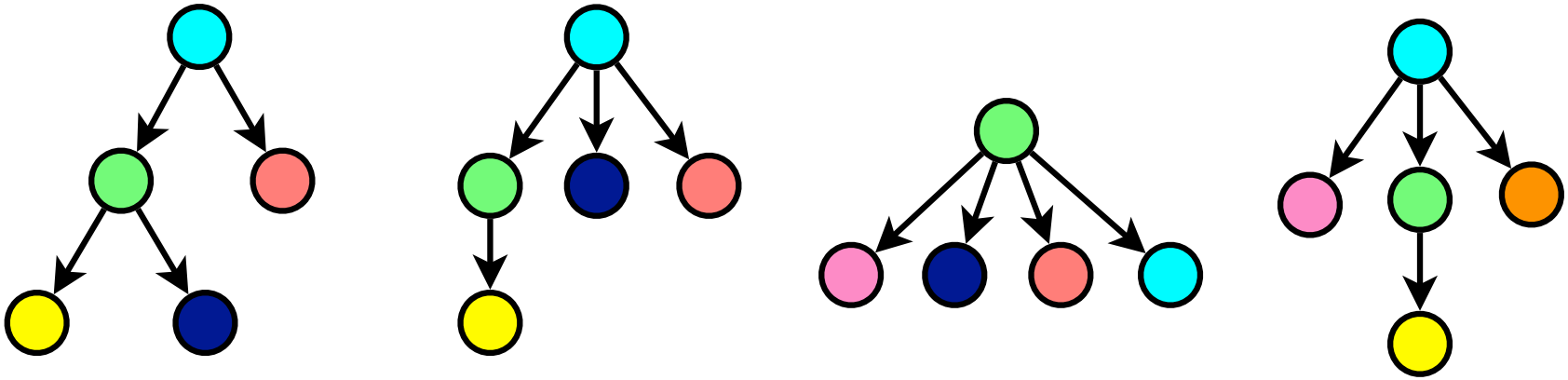
Need for Distance Measures



Uses

1. Comparing/evaluating inferred trees

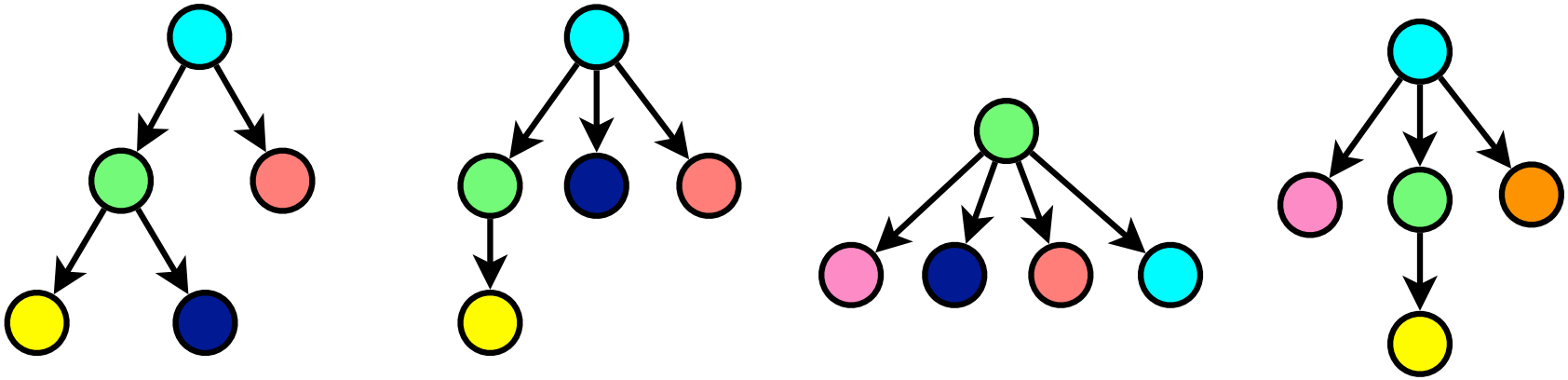
Need for Distance Measures



Uses

1. Comparing/evaluating inferred trees
2. Clustering trees

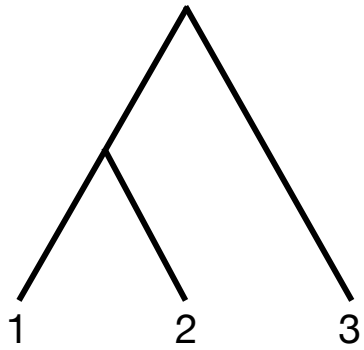
Need for Distance Measures



Uses

1. Comparing/evaluating inferred trees
2. Clustering trees
3. Inference/consensus methods (e.g., Govek et al. 2018)

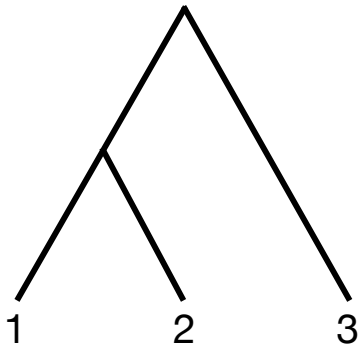
Existing Distance Measures



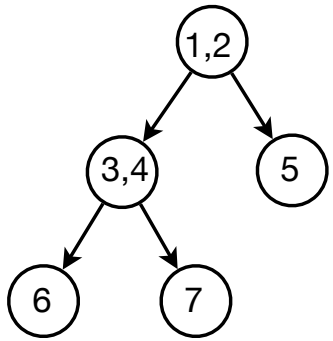
Phylogenetic trees

1. Robinson-Foulds distance (Robinson & Foulds 1981)
2. Quartet distance (Estabrook et al. 1985)
3. Triplet distance (Critchlow et al. 1996)

Existing Distance Measures



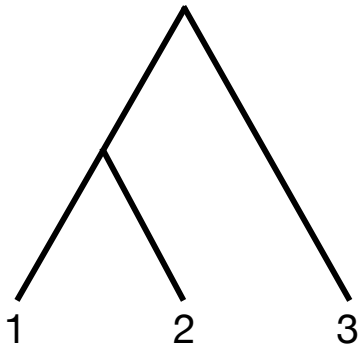
Phylogenetic trees



Clonal trees

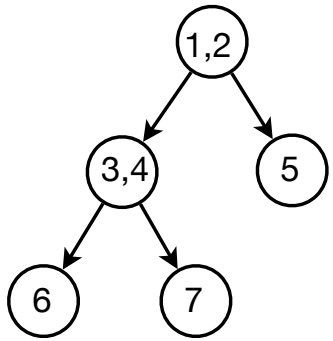
- Robinson-Foulds distance (Robinson & Foulds 1981)
- Quartet distance (Estabrook et al. 1985)
- Triplet distance (Critchlow et al. 1996)

Existing Distance Measures



Phylogenetic trees

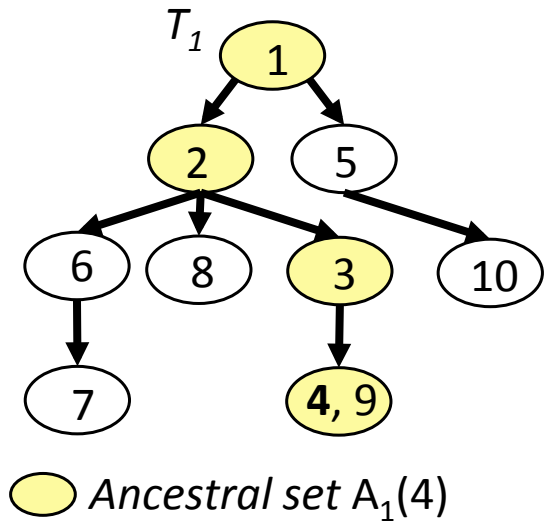
1. Robinson-Foulds distance (Robinson & Foulds 1981)
2. Quartet distance (Estabrook et al. 1985)
3. Triplet distance (Critchlow et al. 1996)



Clonal trees

1. MLTED (Karpov et al. 2018)
2. A-D distance (Govek et al. 2018)
3. Rearrangement distance (Bernardini et al. 2019)

Outline

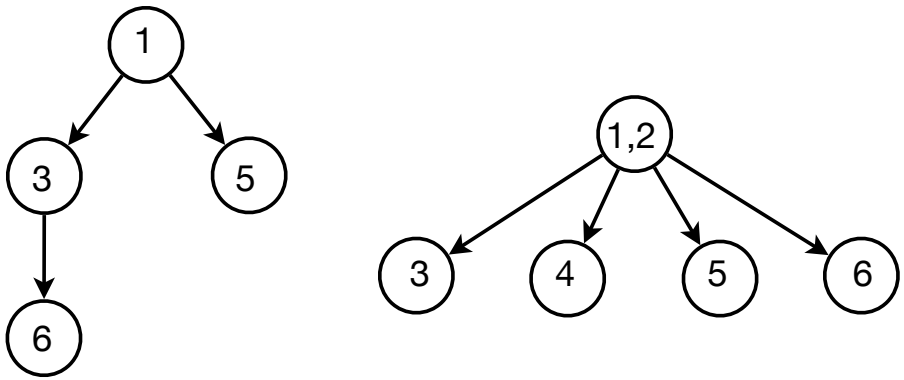


Trees

Definition

A *clonal tree* is a multi-labeled tree with unique labels.

Clonal trees:



Trees

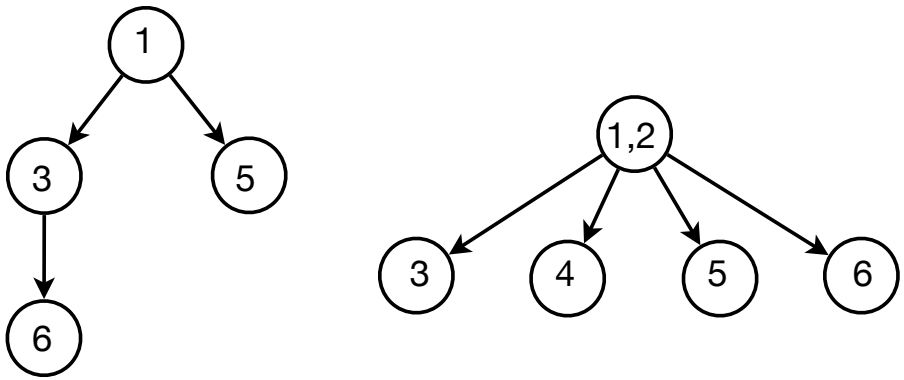
Definition

A *clonal tree* is a multi-labeled tree with unique labels.

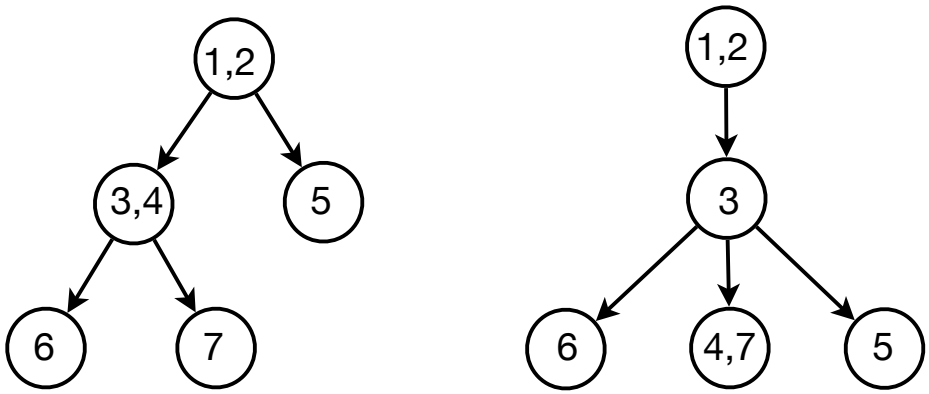
Definition

An *m-clonal tree* is a clonal tree with labels $1, \ldots, m$.

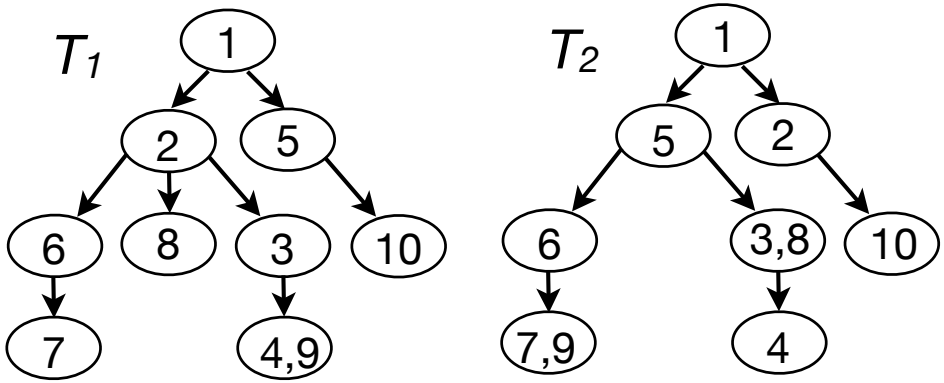
Clonal trees:



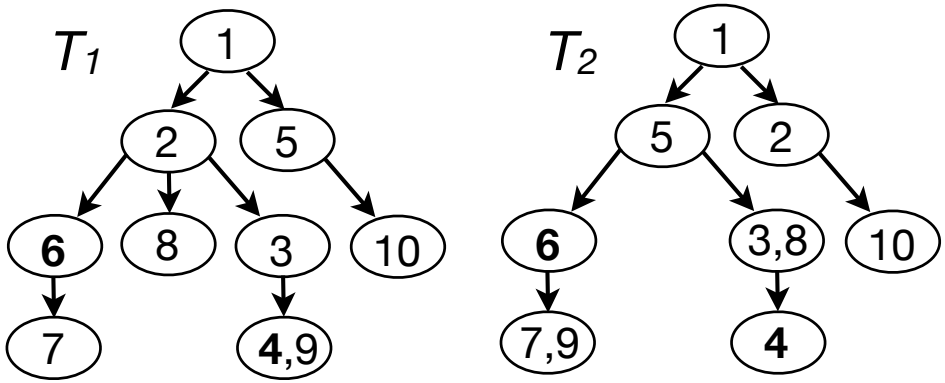
7-clonal trees:



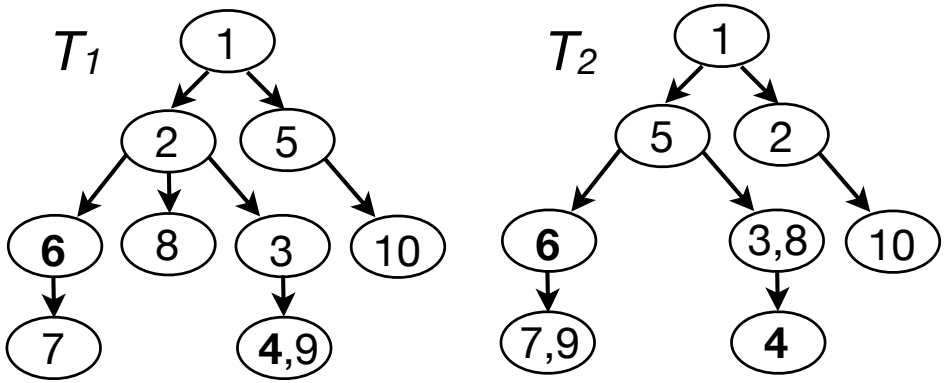
Tree Differences



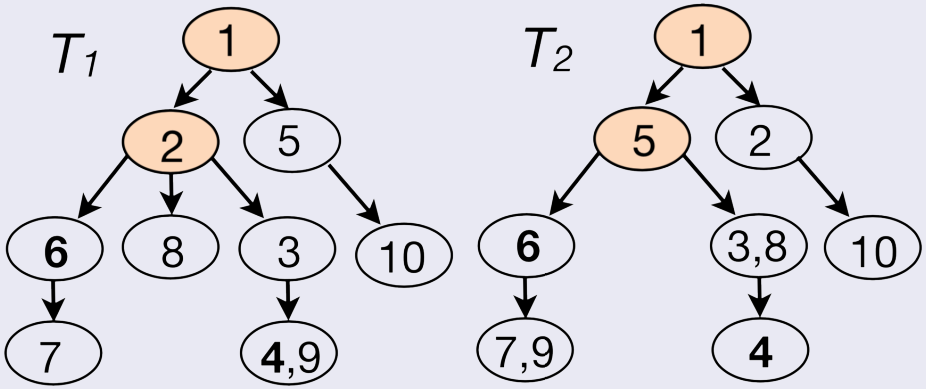
Tree Differences



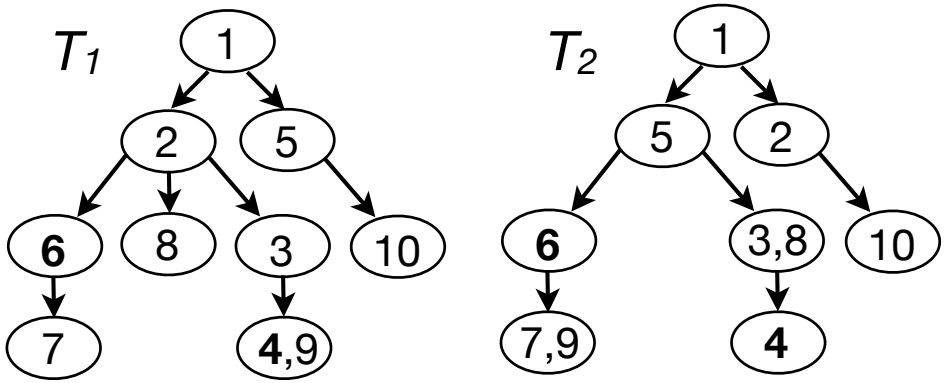
Tree Differences



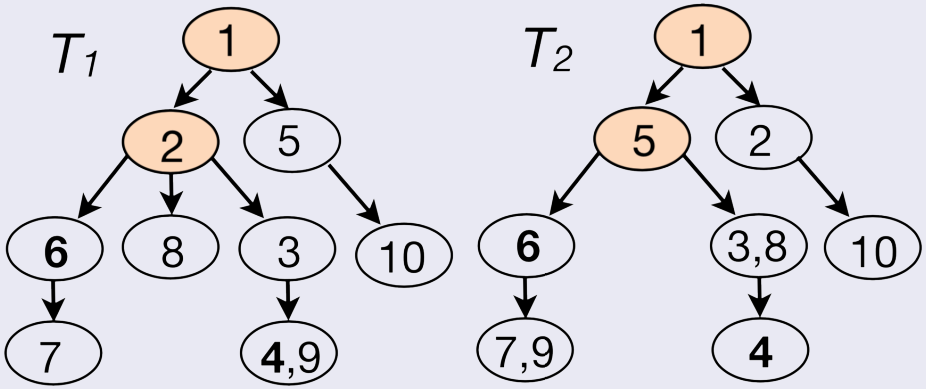
CASet



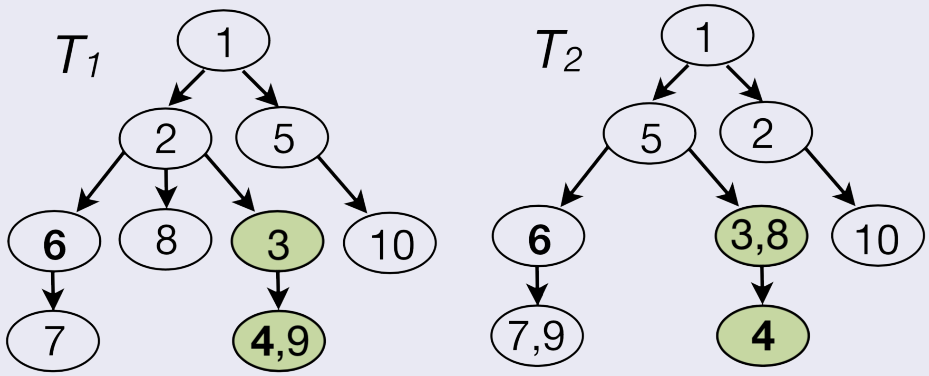
Tree Differences



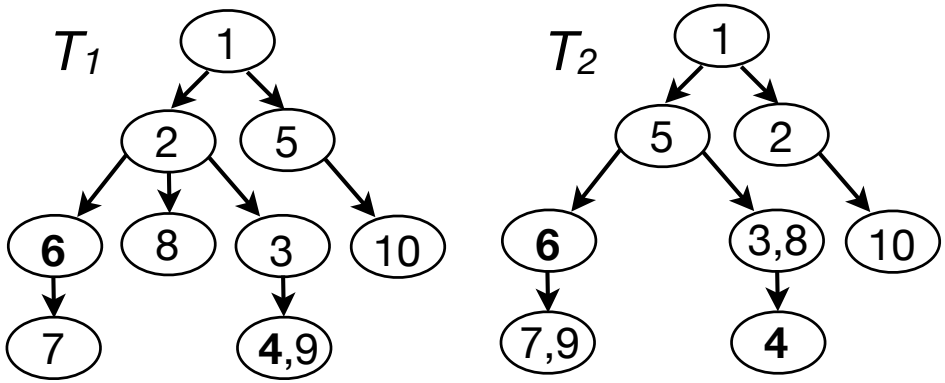
CASet



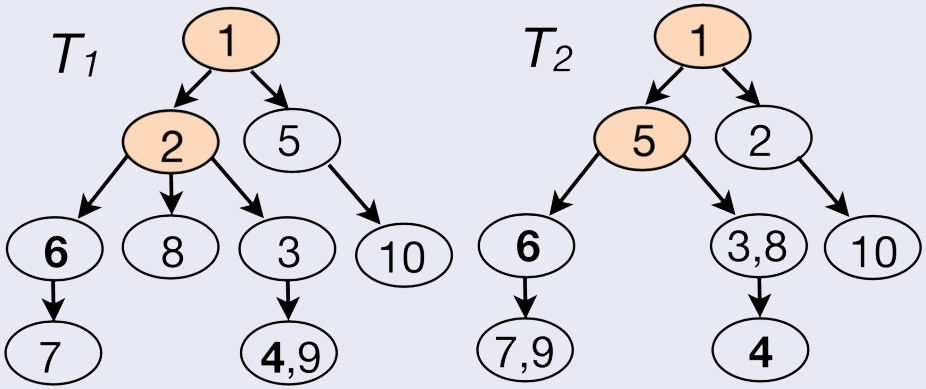
DISC



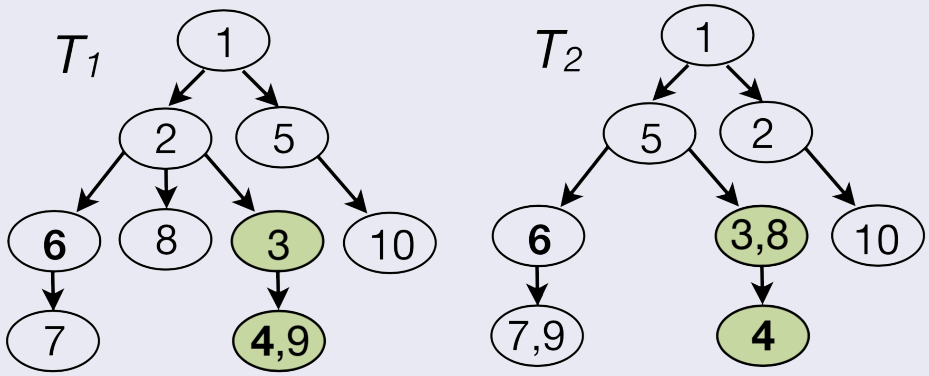
Tree Differences



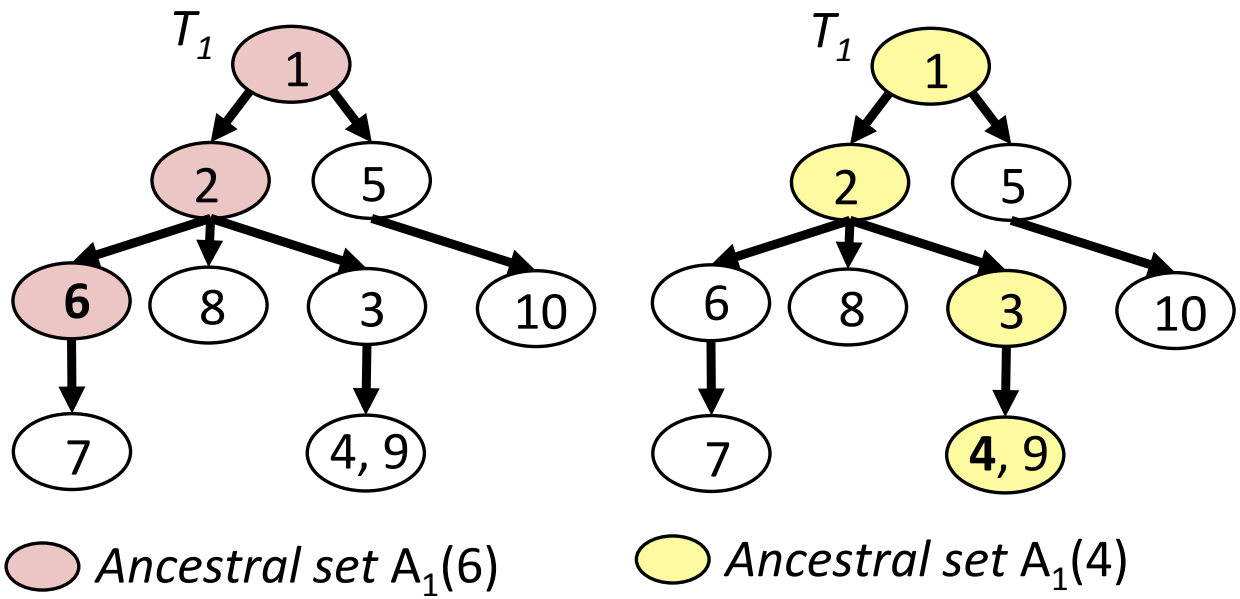
CASet



DISC



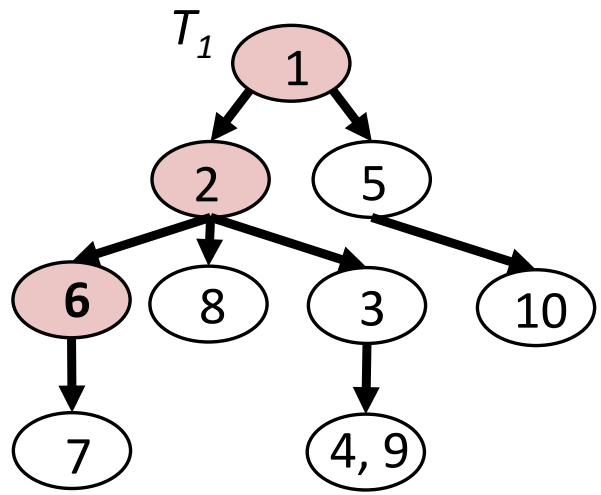
Ancestral and Common Ancestor Sets



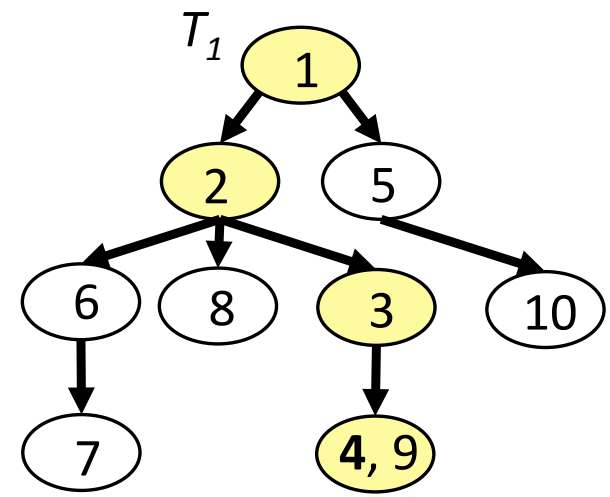
Ancestral set $A_1(6)$

Ancestral set $A_1(4)$

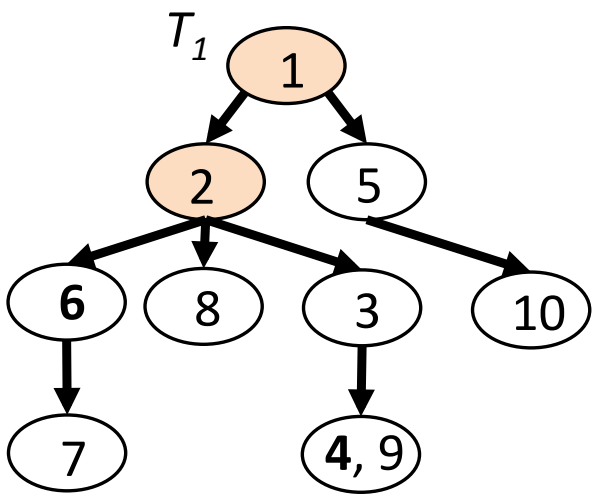
Ancestral and Common Ancestor Sets



Ancestral set $A_1(6)$

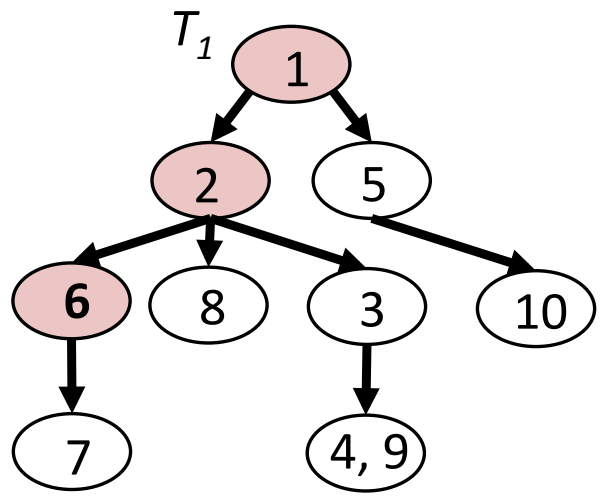


Ancestral set $A_1(4)$

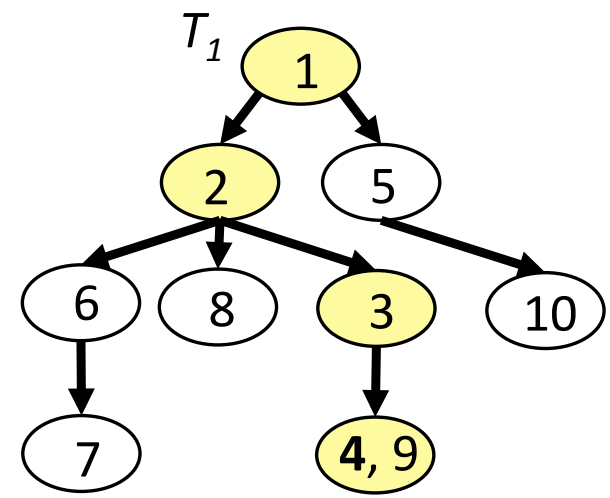


Common Ancestor set $C_1(4,6)$

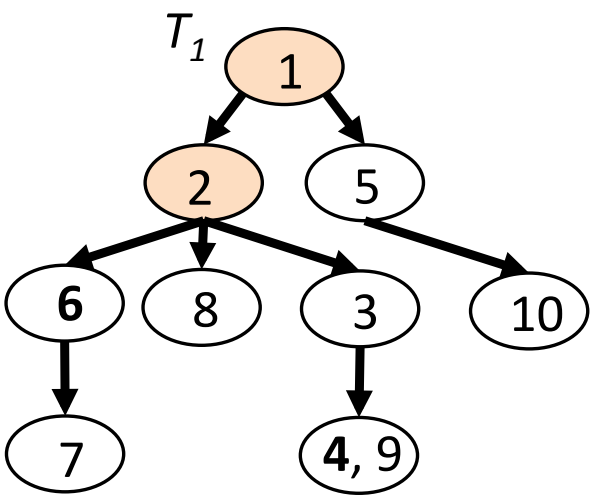
Ancestral and Common Ancestor Sets



Ancestral set $A_1(6)$



Ancestral set $A_1(4)$



Common Ancestor set $C_1(4,6)$

Common Ancestor Set

Given a clonal tree T_k and two mutations i, j,

$$C_k(i,j) = A_k(i) \cap A_k(j).$$

Comparing Sets

$$\text{Jacc}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \qquad \text{Jacc}(\emptyset, \emptyset) = 0$$

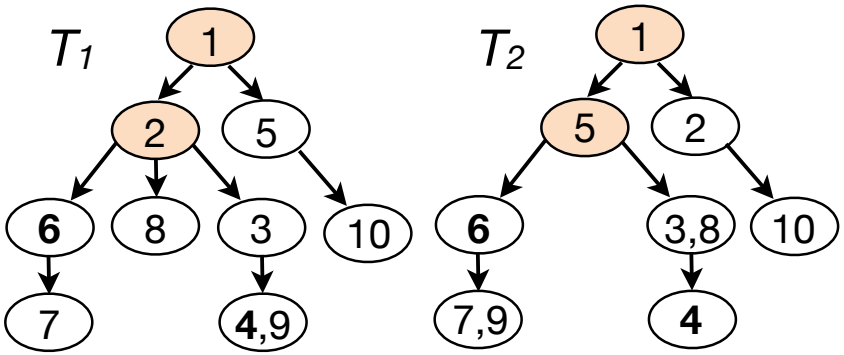
Comparing Sets

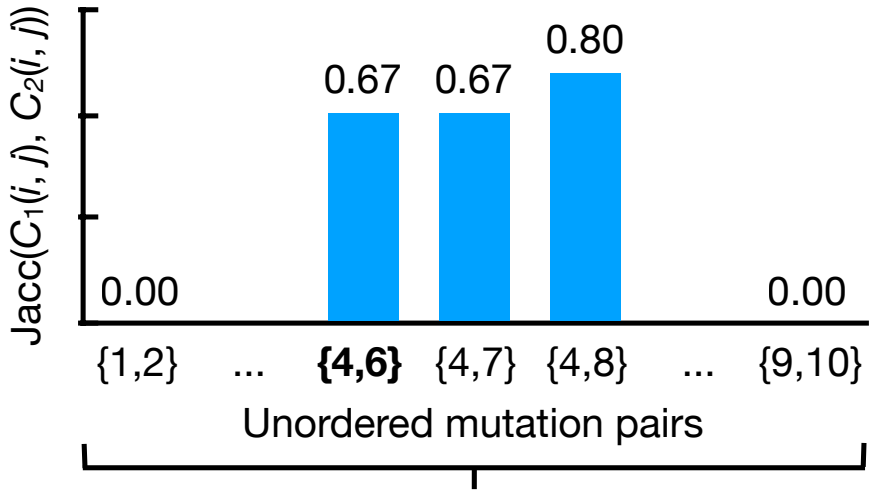
$$\text{Jacc}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \qquad \text{Jacc}(\emptyset, \emptyset) = 0$$

Theorem (e.g., Gilbert 1972)

Jaccard distance is a metric on sets.

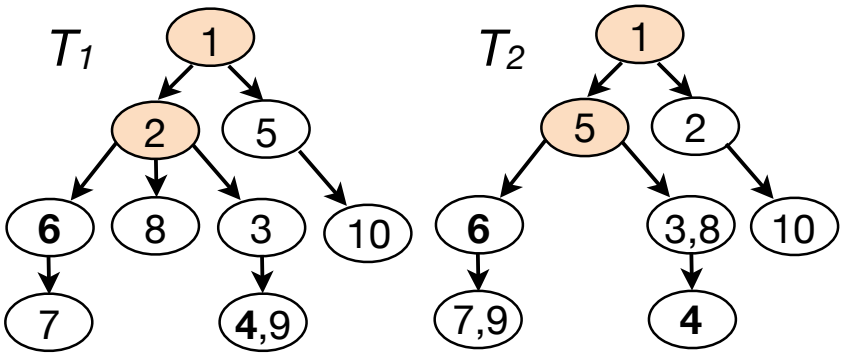
CASet

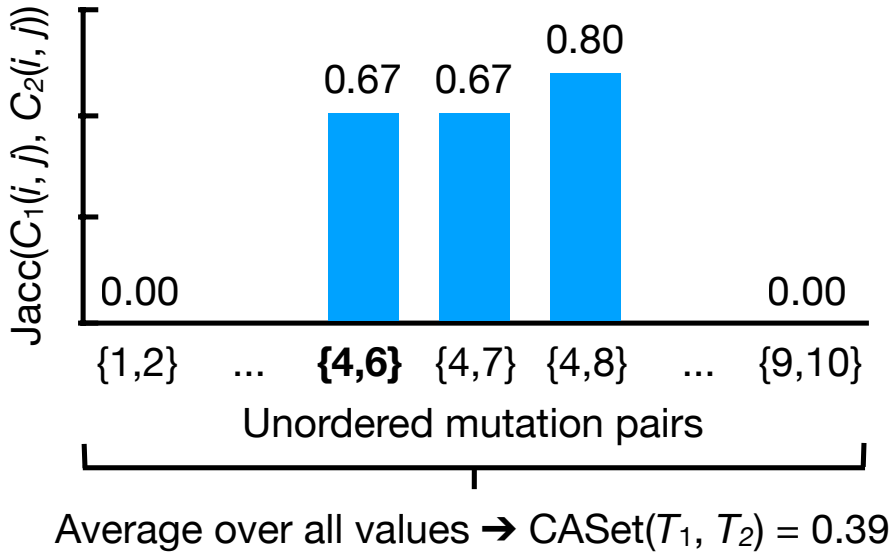




Average over all values ➔ CASet(T_1, T_2) = 0.39

CASet



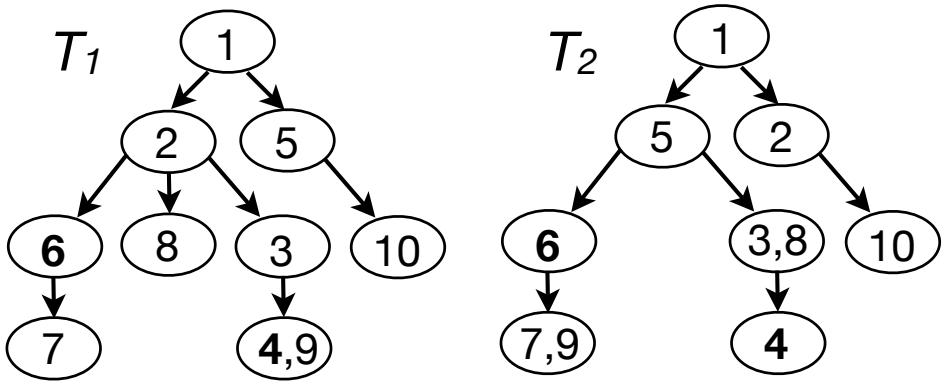


CASet Distance (**C**ommon **A**ncestor **Set**)

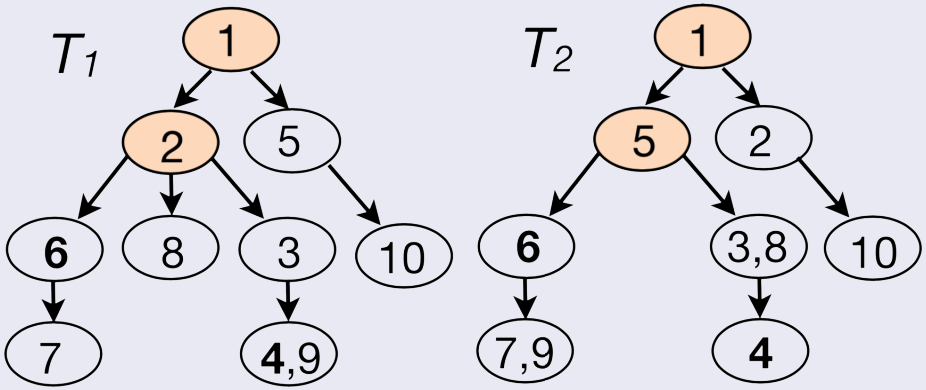
Given two m-clonal trees T_k, T_ℓ,

$$\text{CASet}(T_k, T_\ell) = \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}(C_k(i,j), C_\ell(i,j)).$$

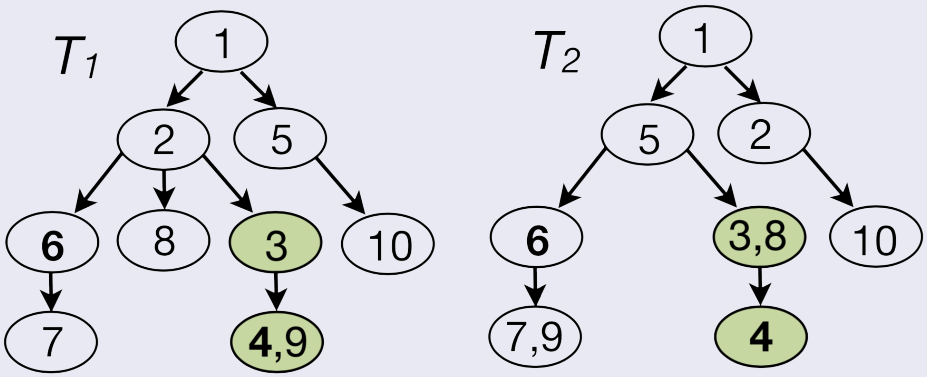
Tree Differences



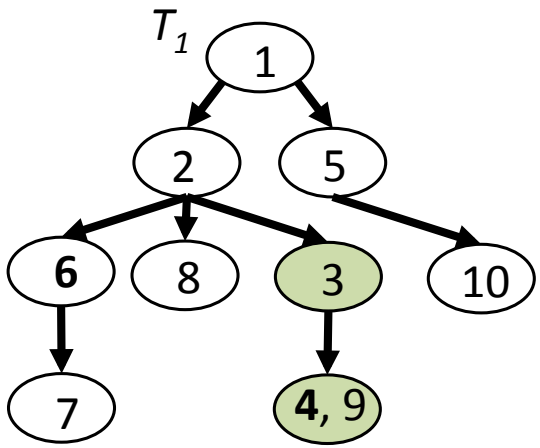
CASet

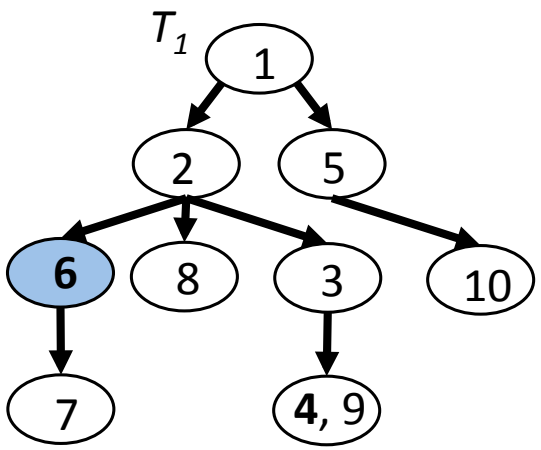


DISC

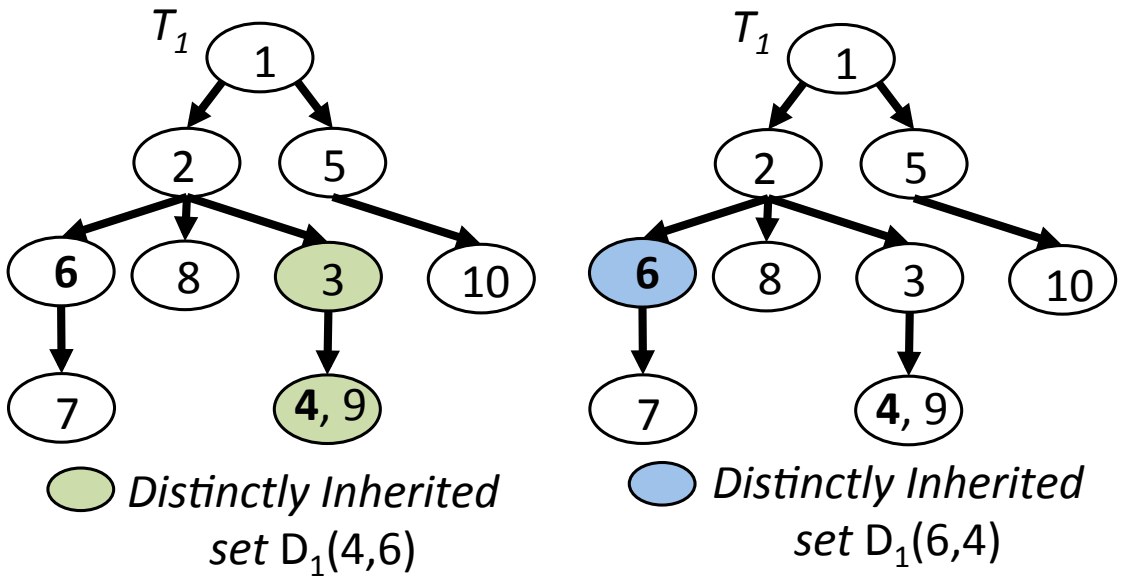


Distinctly Inherited Sets





Distinctly Inherited Sets

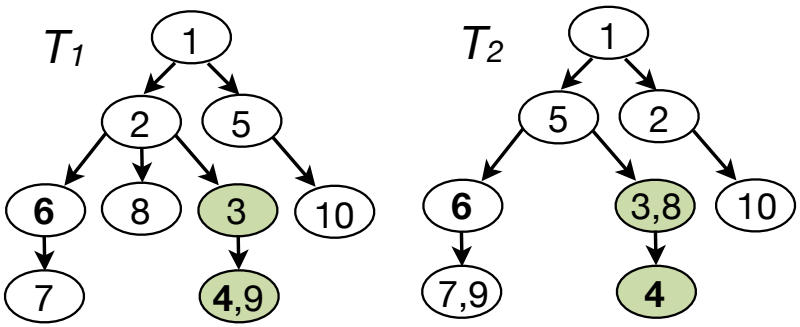


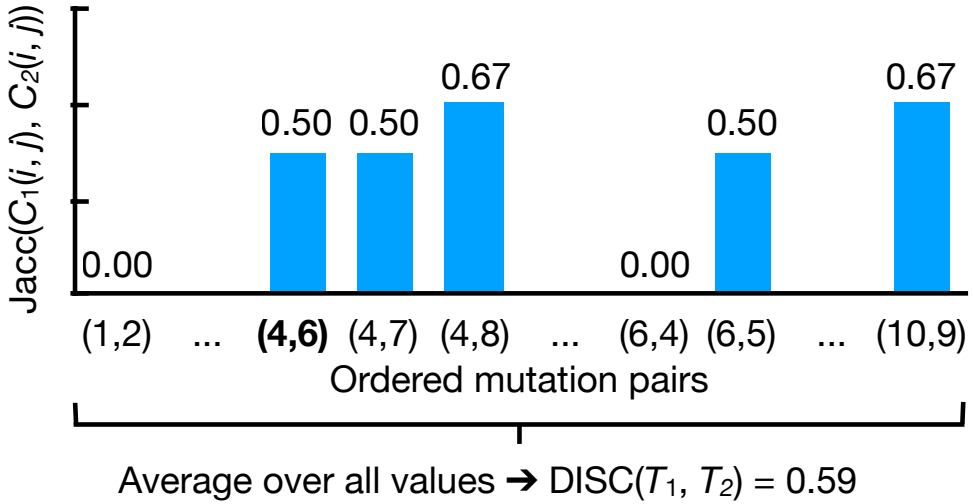
Distinctly Inherited Sets

Given a clonal tree T_k and two mutations i, j,

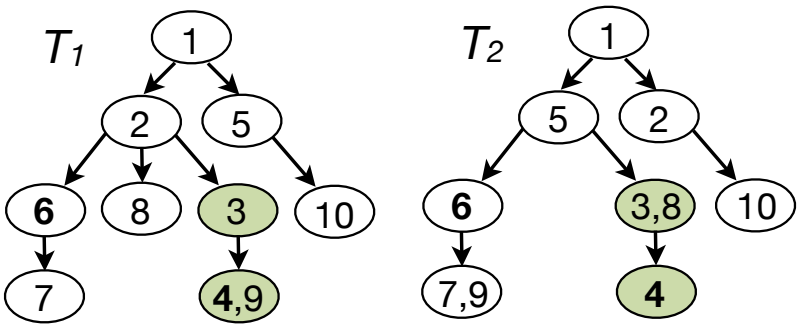
$$D_k(i,j) = A_k(i) \setminus A_k(j).$$

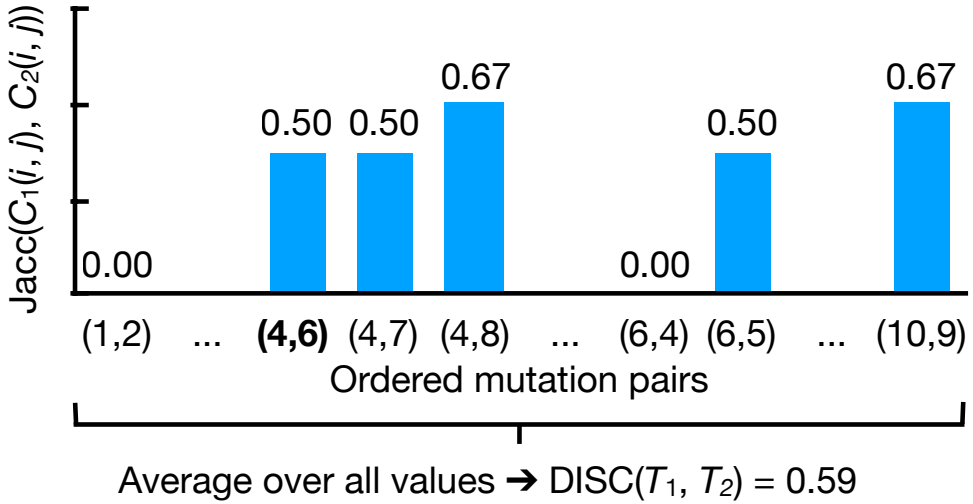
DISC


T_1
1
2
5
6
8
3
10
7
4,9
T_2
1
5
2
6
3,8
10
7,9
4


Jacc($C_1(i,j)$, $C_2(i,j)$)
0.00
0.50
0.50
0.67
0.00
0.50
0.67
(1,2) ... **(4,6)** (4,7) (4,8) ... (6,4) (6,5) ... (10,9)
Ordered mutation pairs
Average over all values ➔ DISC(T_1, T_2) = 0.59

DISC





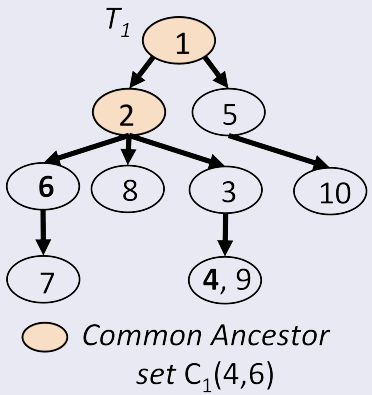
DISC Distance (**D**istinctly **I**nherited **S**et **C**omparison)

Given two m-clonal trees T_k, T_ℓ,

$$\text{DISC}(T_k, T_\ell) = \frac{1}{m(m-1)} \sum_{\substack{(i,j)\in[m]^2 \\ i\neq j}} \text{Jacc}(D_k(i,j), D_\ell(i,j)).$$

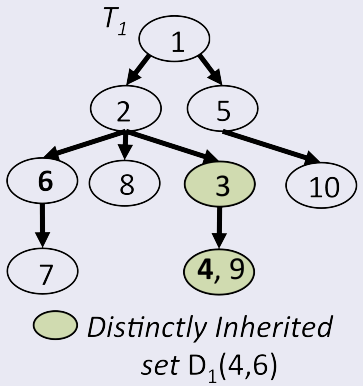
Metric Properties

CASet(T_k, T_ℓ)



$$= \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}(C_k(i,j), C_\ell(i,j))$$

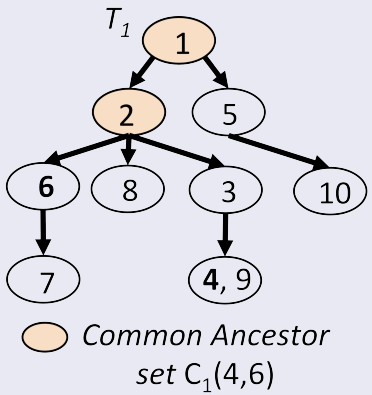
DISC(T_k, T_ℓ)



$$= \frac{1}{m(m-1)} \sum_{\substack{(i,j) \in [m]^2 \\ i \neq j}} \text{Jacc}(D_k(i,j), D_\ell(i,j))$$

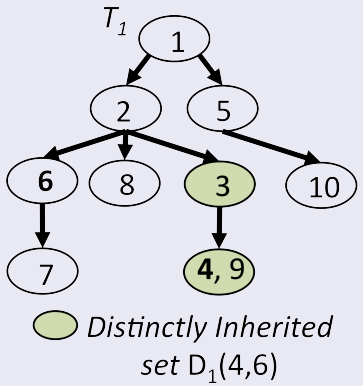
Metric Properties

CASet(T_k, T_ℓ)



$$= \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}(C_k(i,j), C_\ell(i,j))$$

DISC(T_k, T_ℓ)



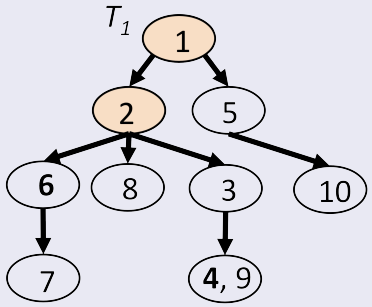
$$= \frac{1}{m(m-1)} \sum_{\substack{(i,j) \in [m]^2 \\ i \neq j}} \text{Jacc}(D_k(i,j), D_\ell(i,j))$$

Theorem

CASet and DISC are metrics on m-clonal trees.

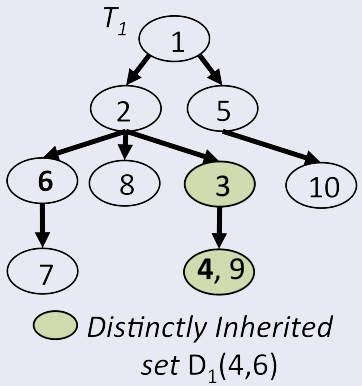
Different Label Sets: Union

$\mathrm{CASet}_{\cup}(T_k, T_\ell)$



$$= \frac{1}{\binom{|U_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq U_{k,\ell}} \mathrm{Jacc}(C_k(i,j), C_\ell(i,j))$$

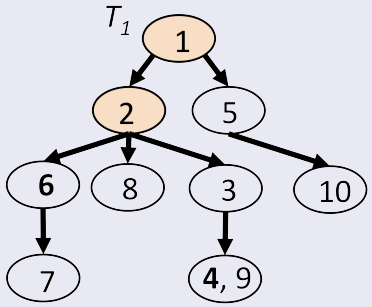
$\mathrm{DISC}_{\cup}(T_k, T_\ell)$



$$= \frac{1}{|U_{k,l}|(|U_{k,\ell}| - 1)} \sum_{\substack{(i,j) \in U^2_{k,\ell} \\ i \neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j))$$

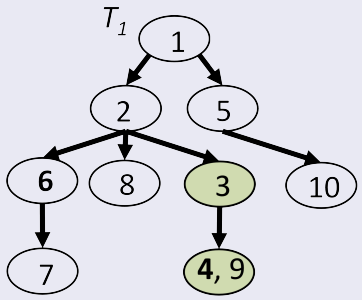
Different Label Sets: Union

$\mathrm{CASet}_{\cup}(T_k, T_\ell)$

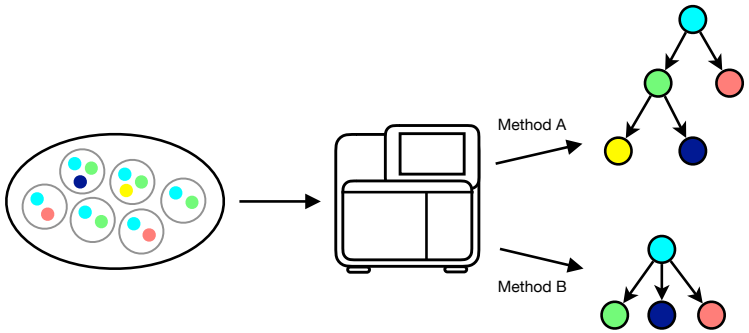


$$= \frac{1}{\binom{|U_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq U_{k,\ell}} \mathrm{Jacc}(C_k(i,j), C_\ell(i,j))$$

$\mathrm{DISC}_{\cup}(T_k, T_\ell)$

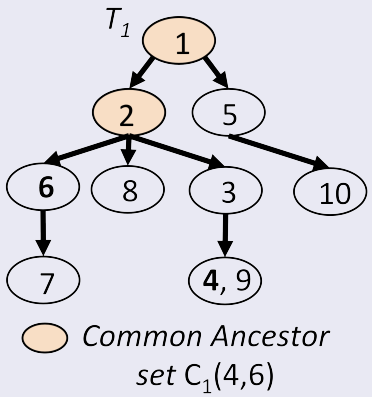


$$= \frac{1}{|U_{k,l}|(|U_{k,\ell}| - 1)} \sum_{\substack{(i,j) \in U^2_{k,\ell} \\ i \neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j))$$



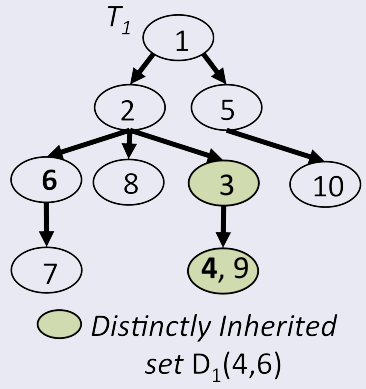
Different Label Sets: Intersection

$\mathrm{CASet}_{\cap}(T_k, T_\ell)$



$$= \frac{1}{\binom{|I_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq I_{k,\ell}} \mathrm{Jacc}(C_k(i,j), C_\ell(i,j))$$

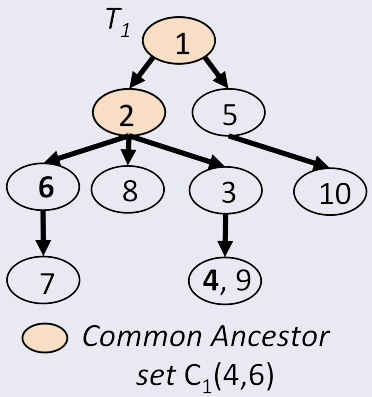
$\mathrm{DISC}_{\cap}(T_k, T_\ell)$



$$= \frac{1}{|I_{k,l}|(|I_{k,\ell}| - 1)} \sum_{\substack{(i,j) \in I^2_{k,\ell} \\ i \neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j))$$

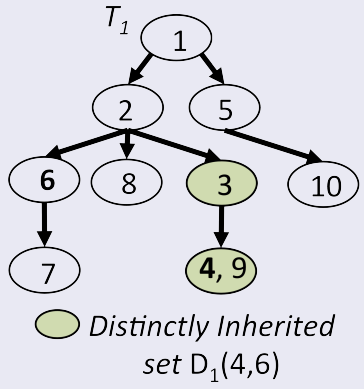
Different Label Sets: Intersection

$\mathrm{CASet}_{\cap}(T_k, T_\ell)$

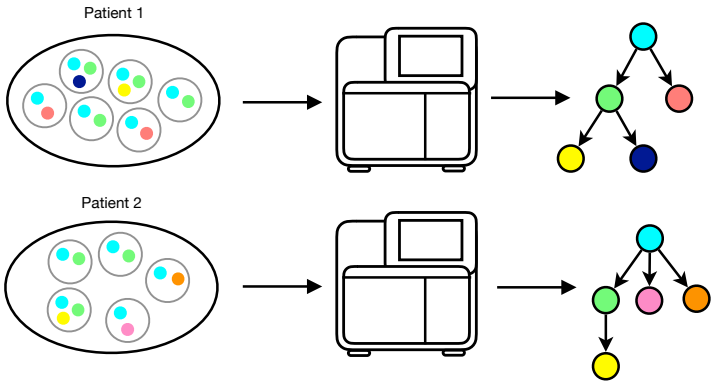


$$= \frac{1}{\binom{|I_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq I_{k,\ell}} \mathrm{Jacc}(C_k(i,j), C_\ell(i,j))$$

$\mathrm{DISC}_{\cap}(T_k, T_\ell)$

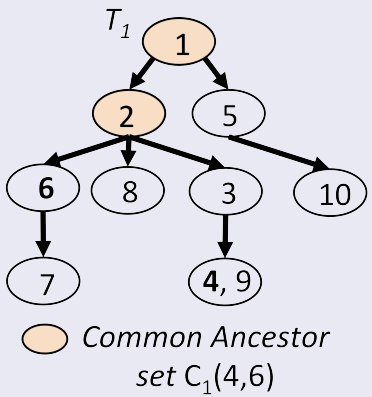


$$= \frac{1}{|I_{k,l}|(|I_{k,\ell}| - 1)} \sum_{\substack{(i,j) \in I^2_{k,\ell} \\ i \neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j))$$



Overview of Distance Measures

CASet(T_k, T_ℓ)



$$= \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}(C_k(i,j), C_\ell(i,j))$$

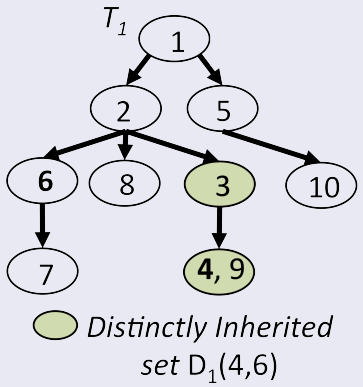
Union

$$\sum_{\{i,j\} \subseteq \boldsymbol{U_{k,\ell}}}$$

Intersection

$$\sum_{\{i,j\} \subseteq \boldsymbol{I_{k,\ell}}}$$

DISC(T_k, T_ℓ)



$$= \frac{1}{m(m-1)} \sum_{\substack{(i,j) \in [m]^2 \\ i \neq j}} \text{Jacc}(D_k(i,j), D_\ell(i,j))$$

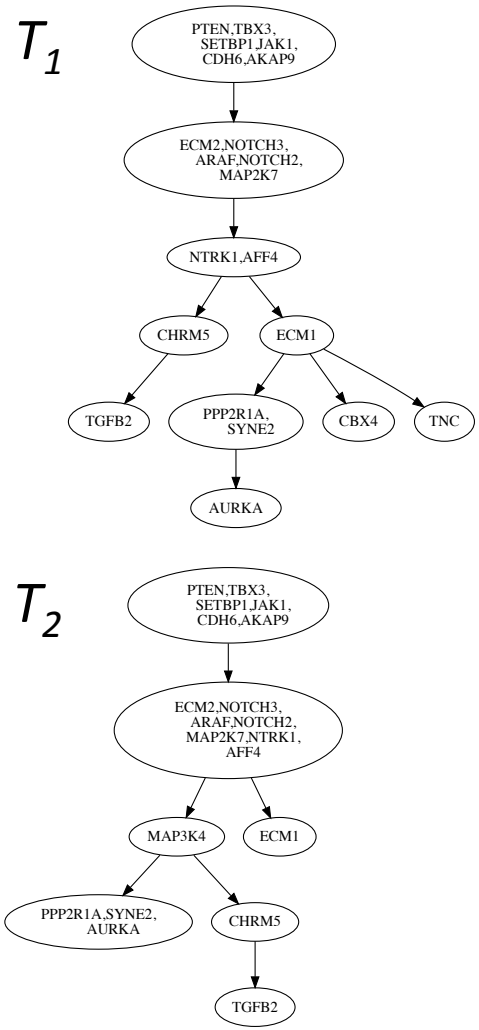
Union

$$\sum_{\substack{(i,j) \in \boldsymbol{U}^2_{\boldsymbol{k,\ell}} \\ i \neq j}}$$

Intersection

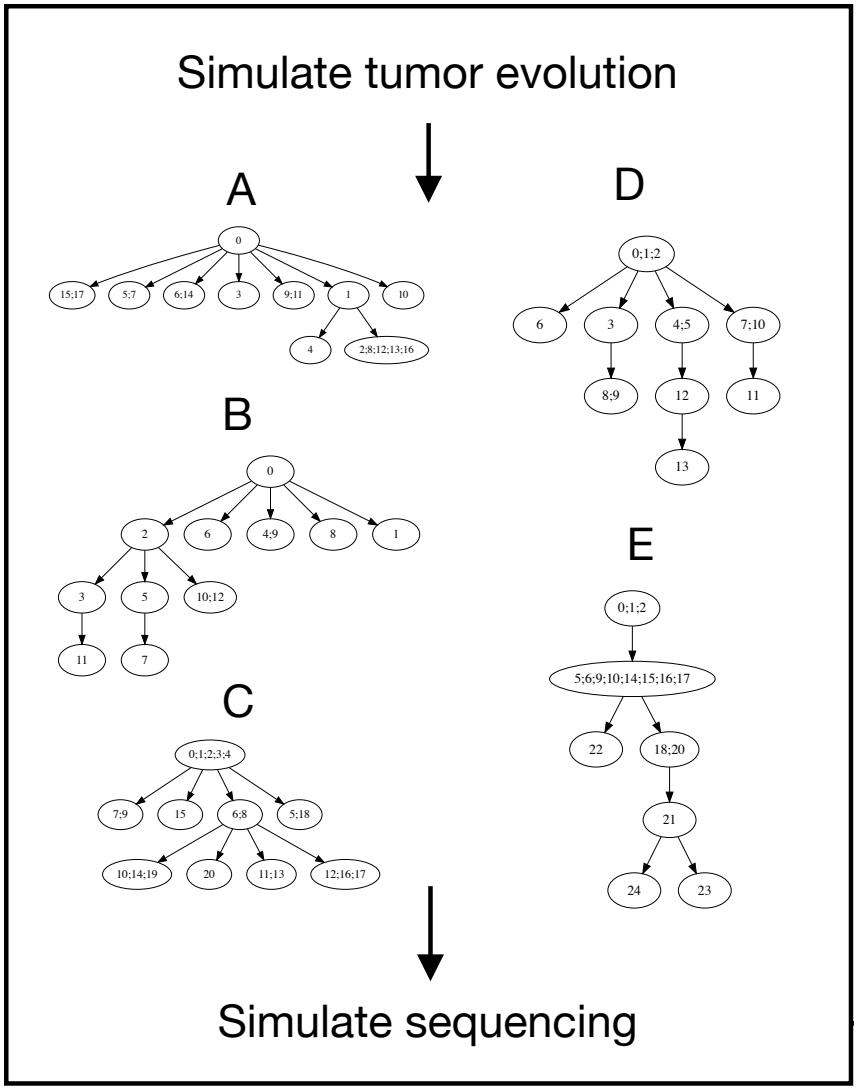
$$\sum_{\substack{(i,j) \in \boldsymbol{I}^2_{\boldsymbol{k,\ell}} \\ i \neq j}}$$

Outline



Simulated Data

OncoLib
(El-Kebir Group 2018)

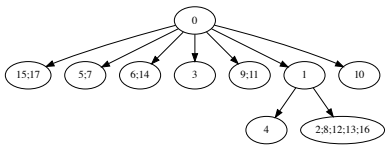


Simulated Data

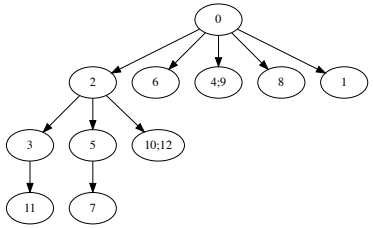
OncoLib
(El-Kebir Group 2018)

Simulate tumor evolution

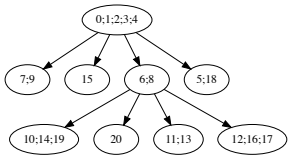
A



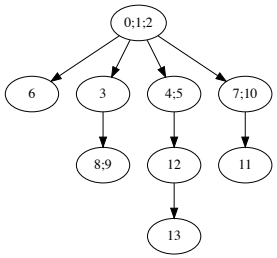
B



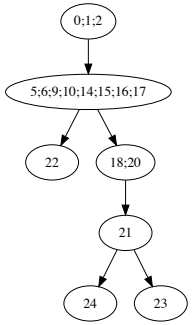
C



D



E



Simulate sequencing

Modified AncesTree
(El-Kebir et al. 2015, Tomlinson & Oesper 2018)

Simulated sequencing data

Enumerate compatible trees

Random sampling

Family A { … }

B { … }

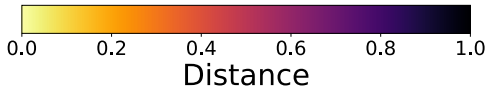
C { … }

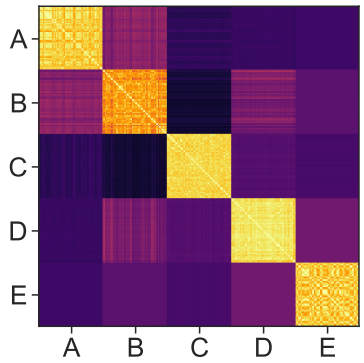
D { … }

E { … }

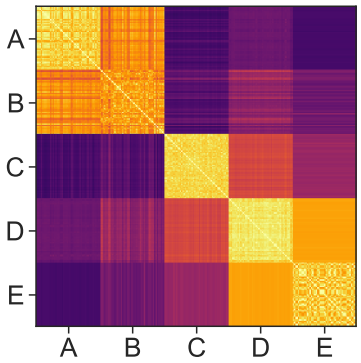
(50 trees each)

Clustering OncoLib Trees

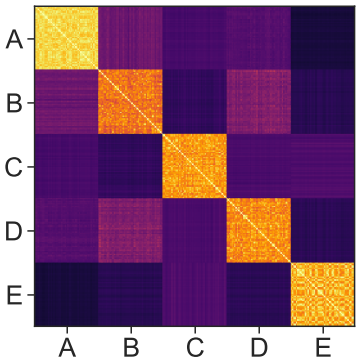




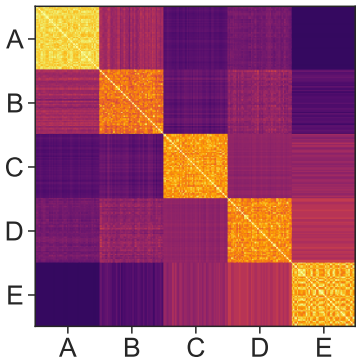
(a) $\text{CASet}_{\cup}$



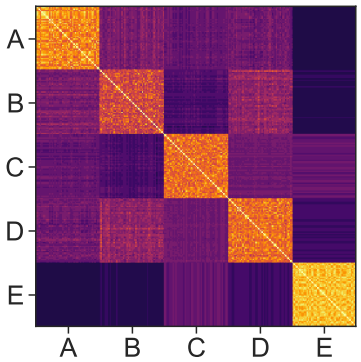
(b) $\text{CASet}_{\cap}$



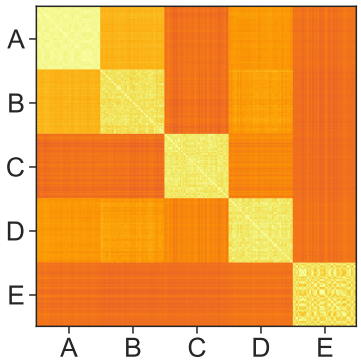
(c) $\text{DISC}_{\cup}$



(d) $\text{DISC}_{\cap}$

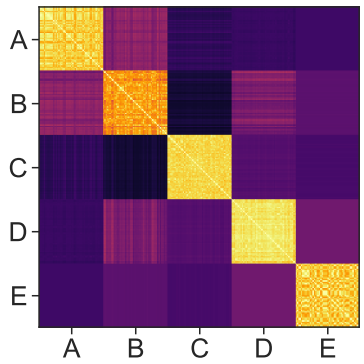


(e) MLTED

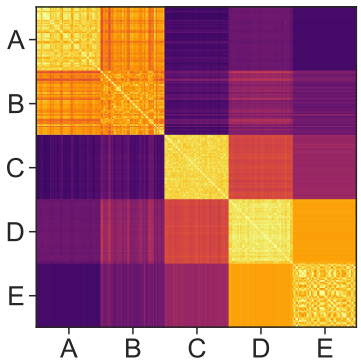


(f) A-D

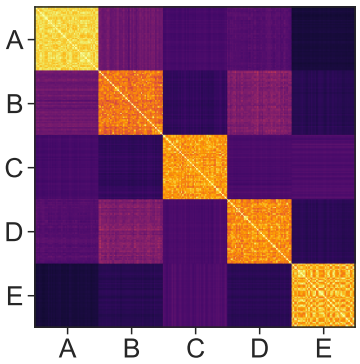
Clustering OncoLib Trees



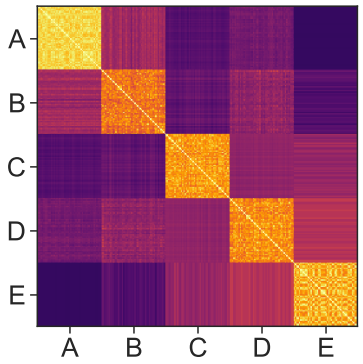
(a) $\text{CASet}_{\cup}$ (0.81)



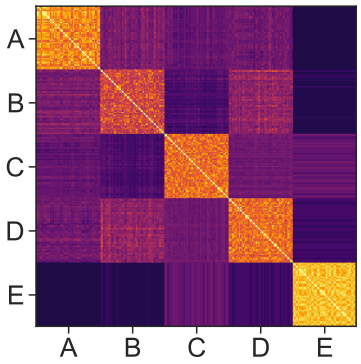
(b) $\text{CASet}_{\cap}$ (0.57)



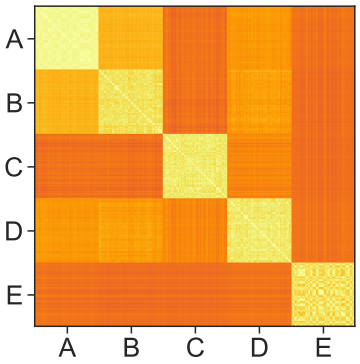
(c) $\text{DISC}_{\cup}$ (0.70)



(d) $\text{DISC}_{\cap}$ (0.62)

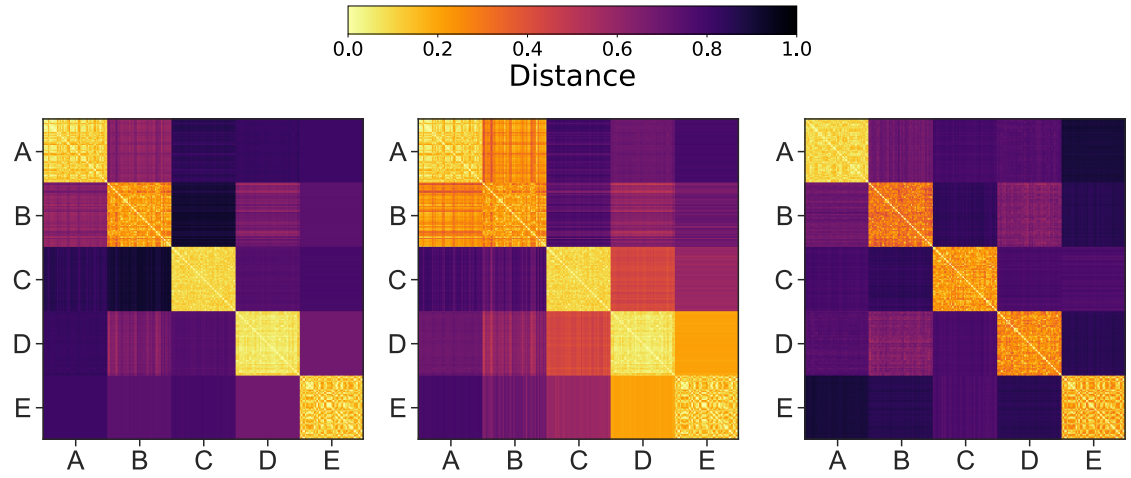


(e) MLTED (0.54)



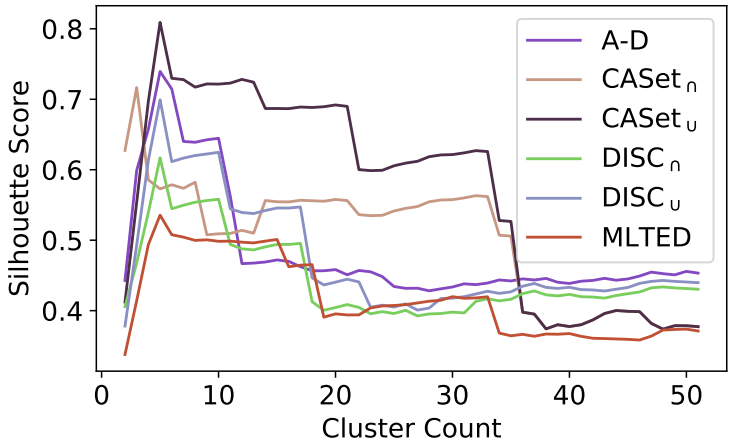
(f) A-D (0.74)

Clustering OncoLib Trees

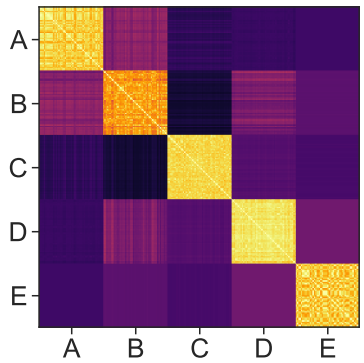


(a) $\text{CASet}_{\cup}$ (0.81) (b) $\text{CASet}_{\cap}$ (0.57) (c) $\text{DISC}_{\cup}$ (0.70)

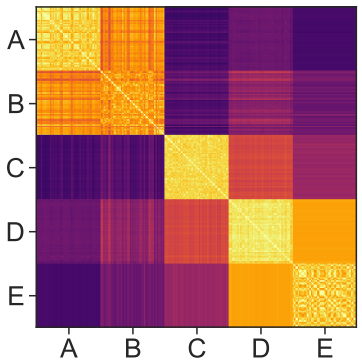
(d) $\text{DISC}_{\cap}$ (0.62) (e) MLTED (0.54) (f) A-D (0.74)



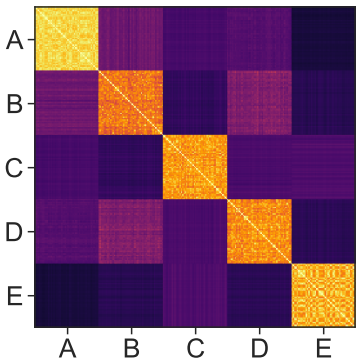
Clustering OncoLib Trees



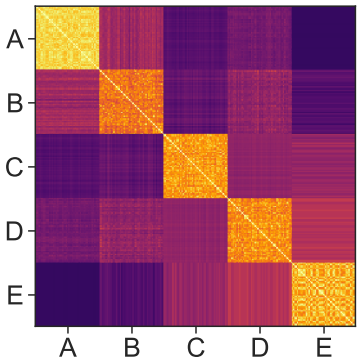
(a) $\text{CASet}_{\cup}$ (0.81)



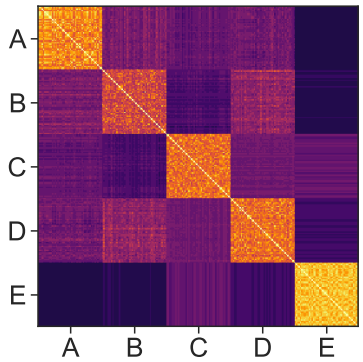
(b) $\text{CASet}_{\cap}$ (0.57)



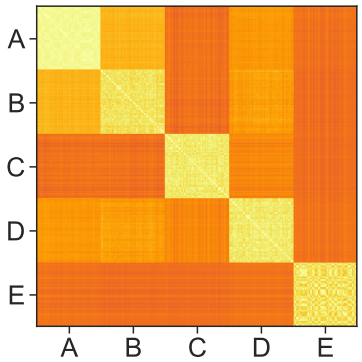
(c) $\text{DISC}_{\cup}$ (0.70)



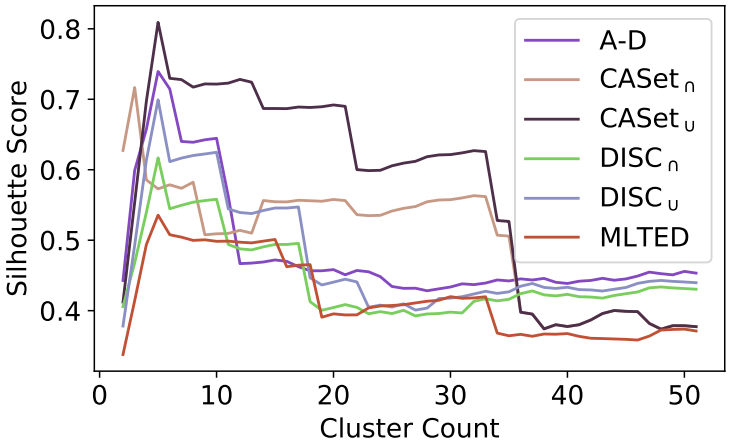
(d) $\text{DISC}_{\cap}$ (0.62)

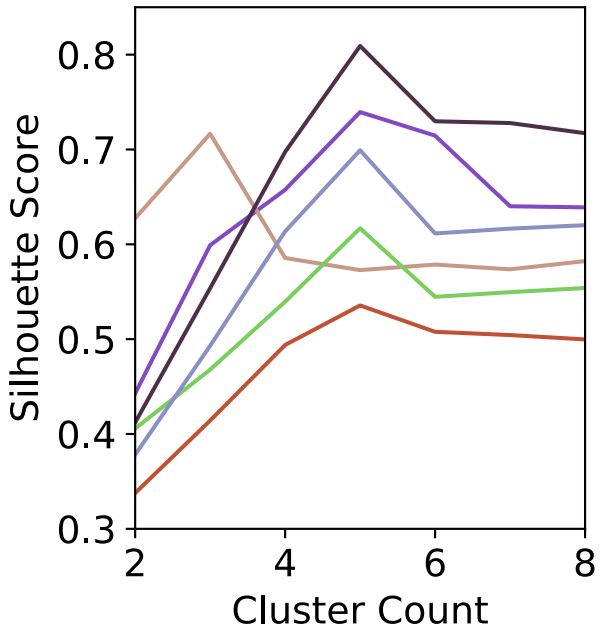


(e) MLTED (0.54)

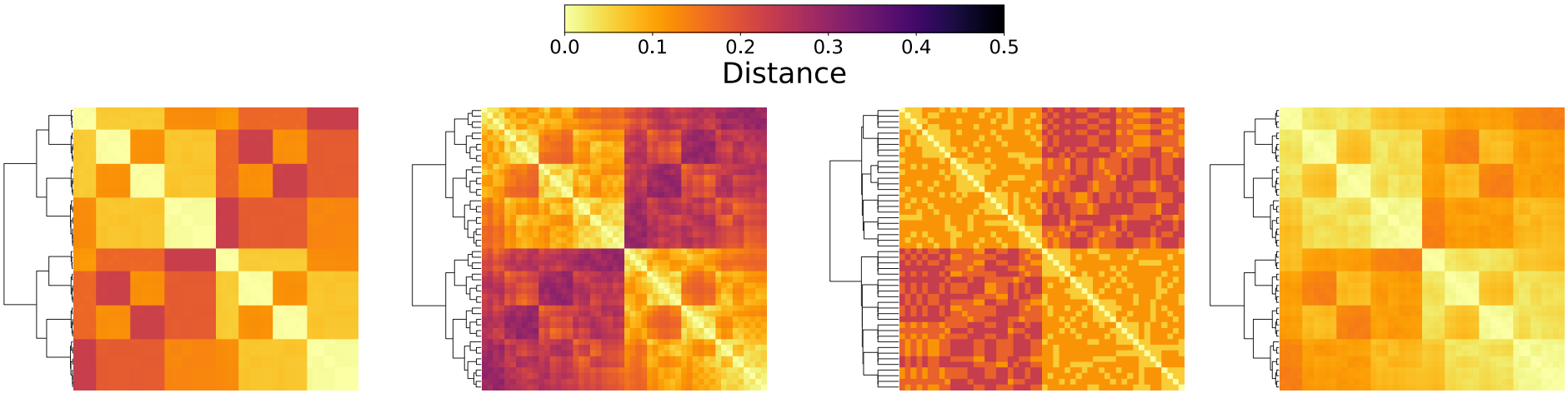


(f) A-D (0.74)





Intra-Family Structure

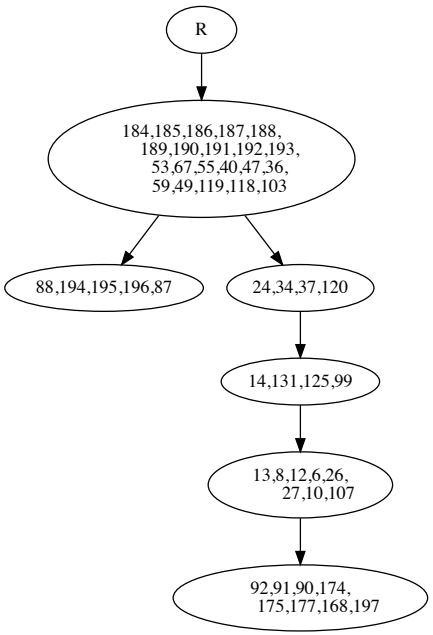


(a) CASet (b) DISC (c) MLTED (d) A-D

Family E

Real Datasets

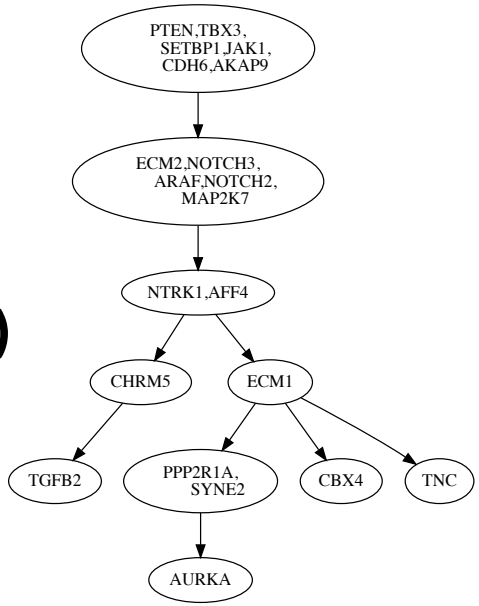
1. Triple negative breast cancer (Wang et al. 2014)
   - Single-cell seq. at 72× coverage
   - Bulk deep seq. at 118,743× coverage
2. Breast cancer xenograft (Eirew et al. 2015)
   - Whole-genome seq. at 35-72× coverage
   - MiSeq targeted deep seq.



TNBC Tree Inference Analysis

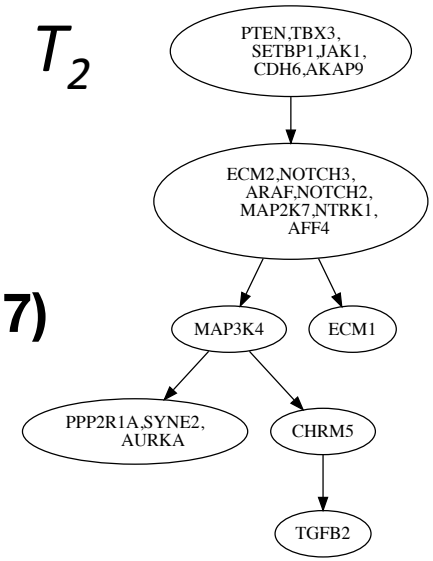
PHiSCS
(Malikic et al., 2018)

T_1



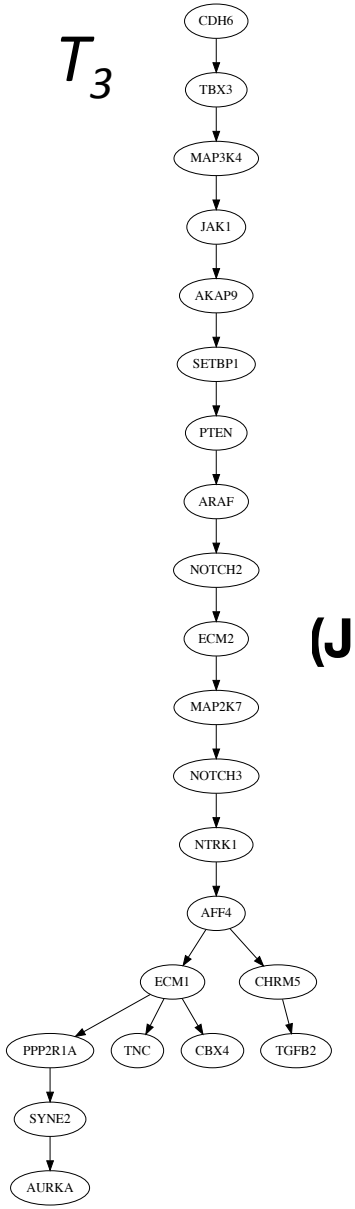
SiFit
(Zafar et al., 2017)

T_2

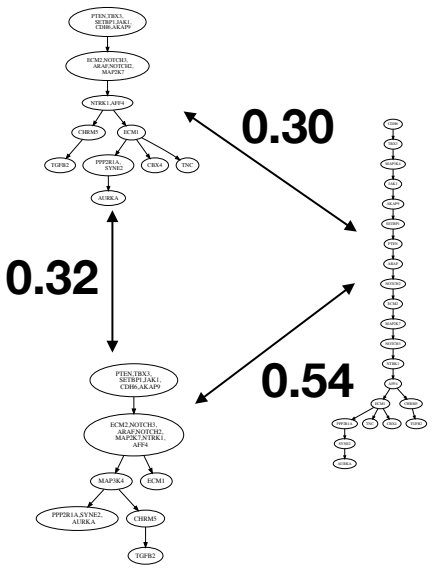


SCITE
(Jahn et al., 2016)

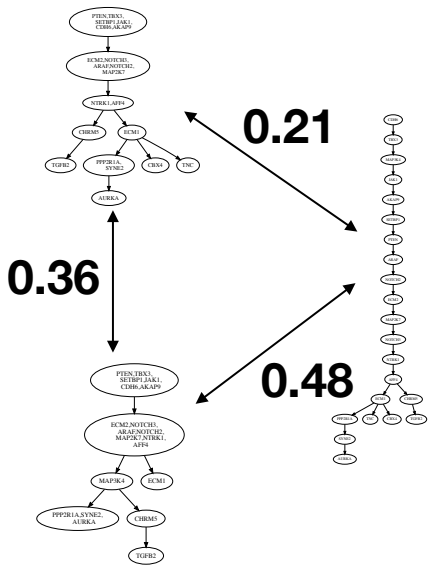
T_3



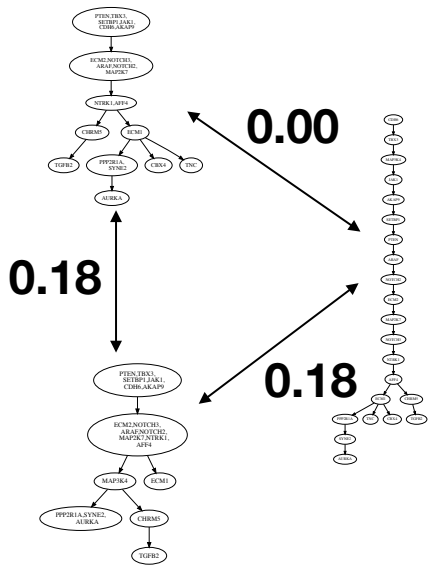
TNBC Tree Inference Analysis



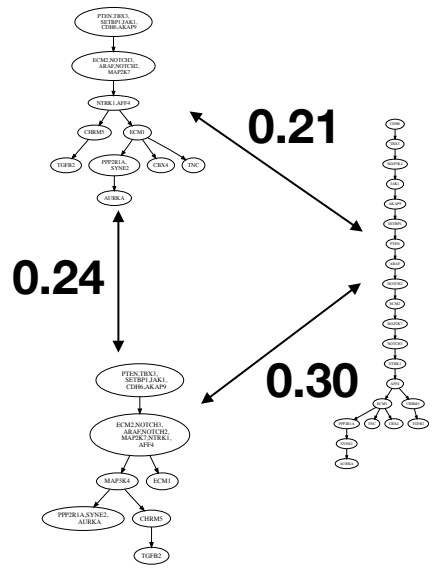
CASet$_\cup$



DISC$_\cup$

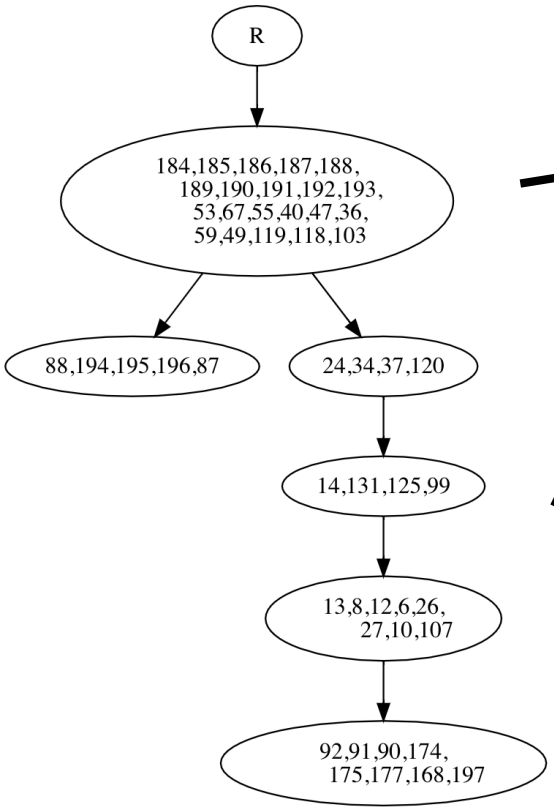


MLTED



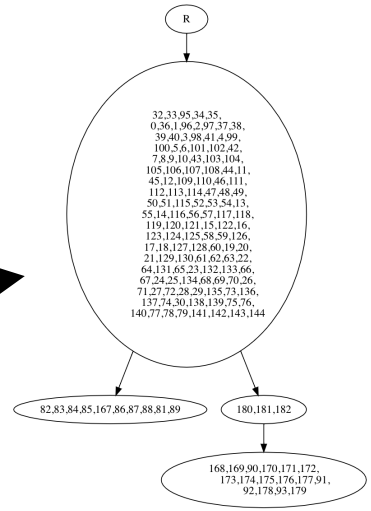
A-D

Xenoengraftment Tree Inference Analysis



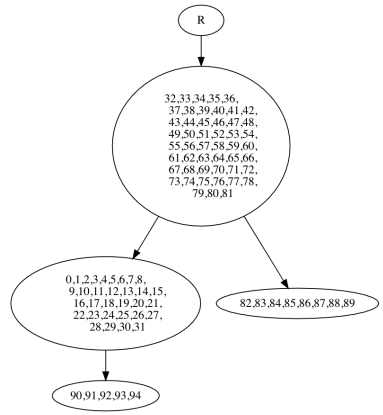
SA501 Base Tree

(Eirew et al., 2014)



MIPUP inference

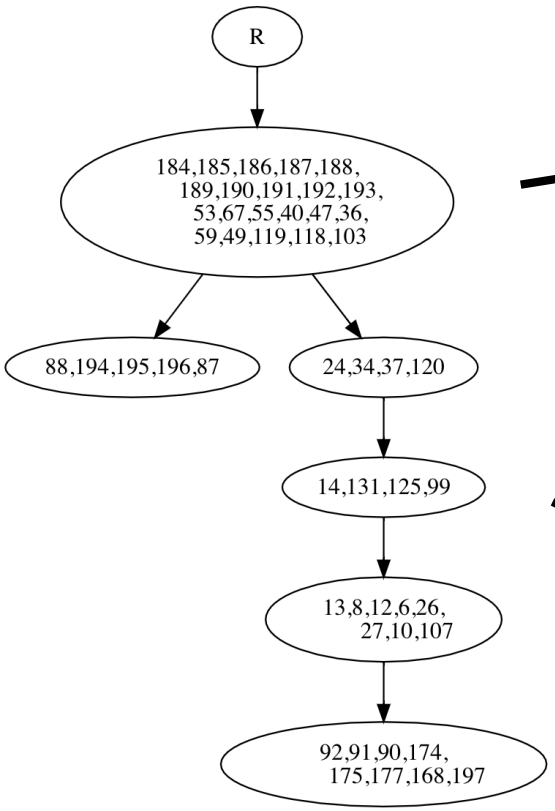
(Husić et al., 2018)



LICHeE inference

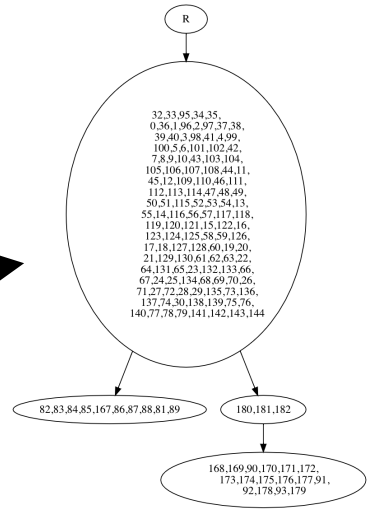
(Popic et al., 2015)

Xenoengraftment Tree Inference Analysis



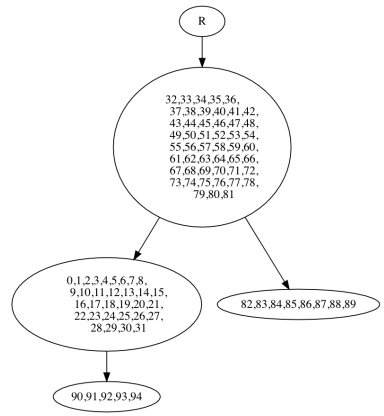
SA501 Base Tree

Measure	Distance
$CASet_{\cup}$	0.88
$CASet_{\cap}$	0.84
$DISC_{\cup}$	**0.40**
$DISC_{\cap}$	0.40
MLTED	**0.80**
A-D	0.70



MIPUP inference

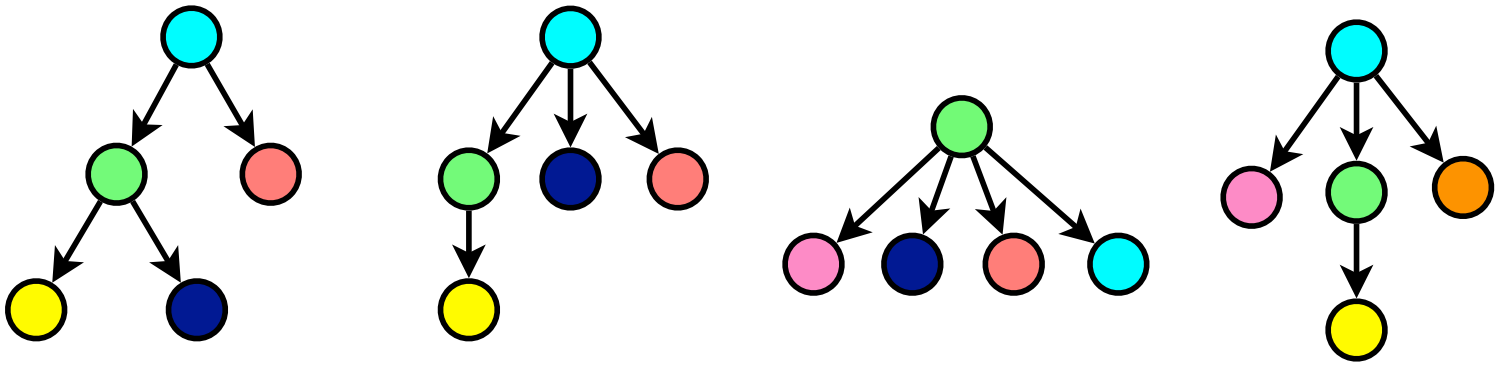
Measure	Distance
$CASet_{\cup}$	**0.74**
$CASet_{\cap}$	**0.78**
$DISC_{\cup}$	0.60
$DISC_{\cap}$	**0.38**
MLTED	0.81
A-D	**0.46**



LICHeE inference

Conclusions

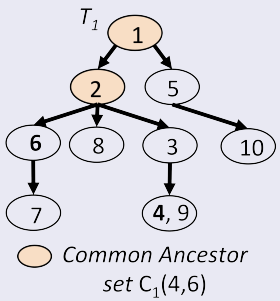
1. Distance measures are important for tumor tree analysis
2. We introduced two novel distance metrics, CASet and DISC
3. $\text{CASet}_{\cup}$ clusters trees more clearly than existing measures
4. CASet and DISC have high resolution on simulated and real data



Conclusions

1. Distance measures are important for tumor tree analysis
2. We introduced two novel distance metrics, CASet and DISC
3. $\text{CASet}_{\cup}$ clusters trees more clearly than existing measures
4. CASet and DISC have high resolution on simulated and real data

CASet(T_k, T_ℓ)



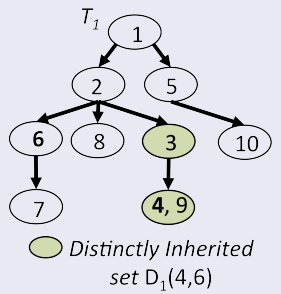
Union

$$\sum_{\{i,j\} \subseteq U_{k,\ell}}$$

Intersection

$$\sum_{\{i,j\} \subseteq I_{k,\ell}}$$

DISC(T_k, T_ℓ)



Union

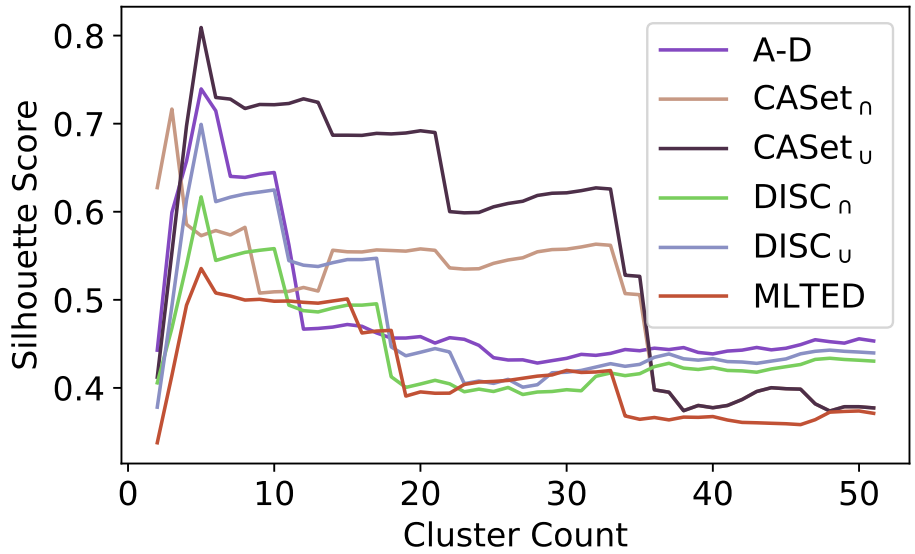
$$\sum_{\substack{(i,j) \in U^2_{k,\ell} \\ i \neq j}}$$

Intersection

$$\sum_{\substack{(i,j) \in I^2_{k,\ell} \\ i \neq j}}$$

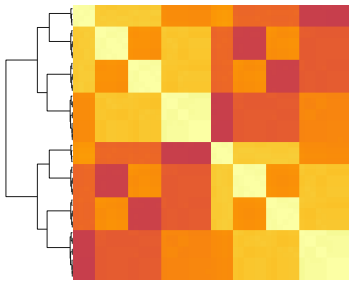
Conclusions

1. Distance measures are important for tumor tree analysis
2. We introduced two novel distance metrics, CASet and DISC
3. CASet$_\cup$ clusters trees more clearly than existing measures
4. CASet and DISC have high resolution on simulated and real data

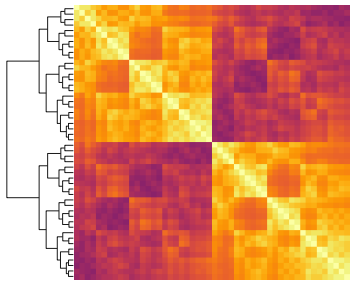


Conclusions

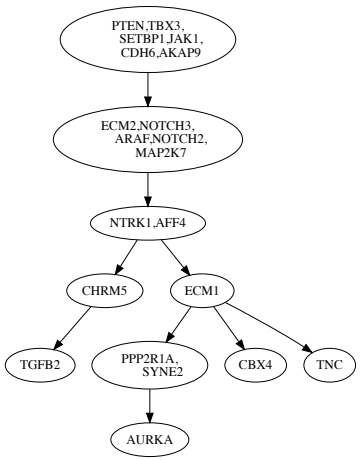
1. Distance measures are important for tumor tree analysis
2. We introduced two novel distance metrics, CASet and DISC
3. $\text{CASet}_{\cup}$ clusters trees more clearly than existing measures
4. CASet and DISC have high resolution on simulated and real data

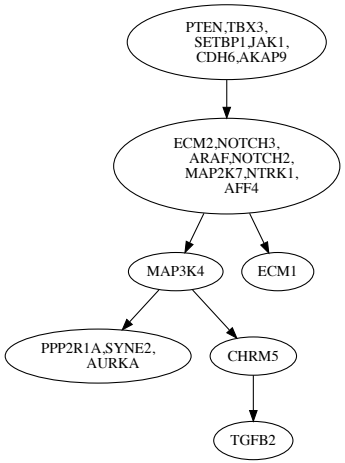


(a) CASet



(b) DISC





Acknowledgment

- This project is supported by the NSF and Elledge, Eugster, and Class of '49 Fellowships from Carleton College (to LO).
- Thank you to the Carleton College Computer Science Department, Layla Oesper, Anna Ritz, Rosa Zhou, and Thais Del Rosario Hernandez.

Availability

Implementations of CASet and DISC are available at `https://bitbucket.org/oesperlab/stereodist`.
Preprint available at `https://www.biorxiv.org/`.