Markovian Models of Genetic Inheritance

Elchanan Mossel,

U.C. Berkeley

mossel@stat.berkeley.edu,
http://www.cs.berkeley.edu/~mossel/

General plan

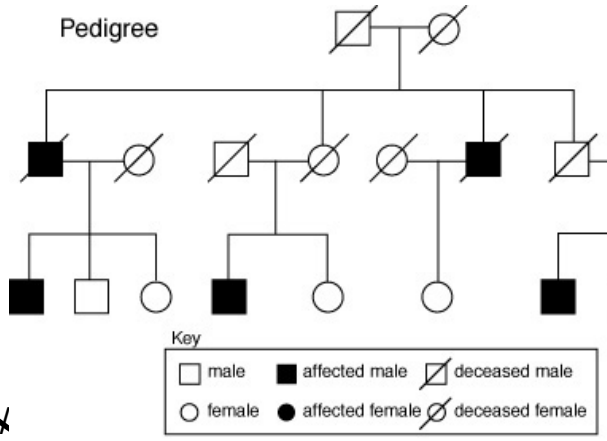
- Define a number of Markovian Inheritance Models (MIM)
- Discuss how to estimate and reconstruct from data.
- Lecture 1: Definition of Models
- Lecture 2: Reconstruction via metric estimates.
- Lecture 3: Decay of information and impossibility results.
- Lecture 4: Reconstruction.
- Lecture 5: Survey of more advanced topics.

General plan

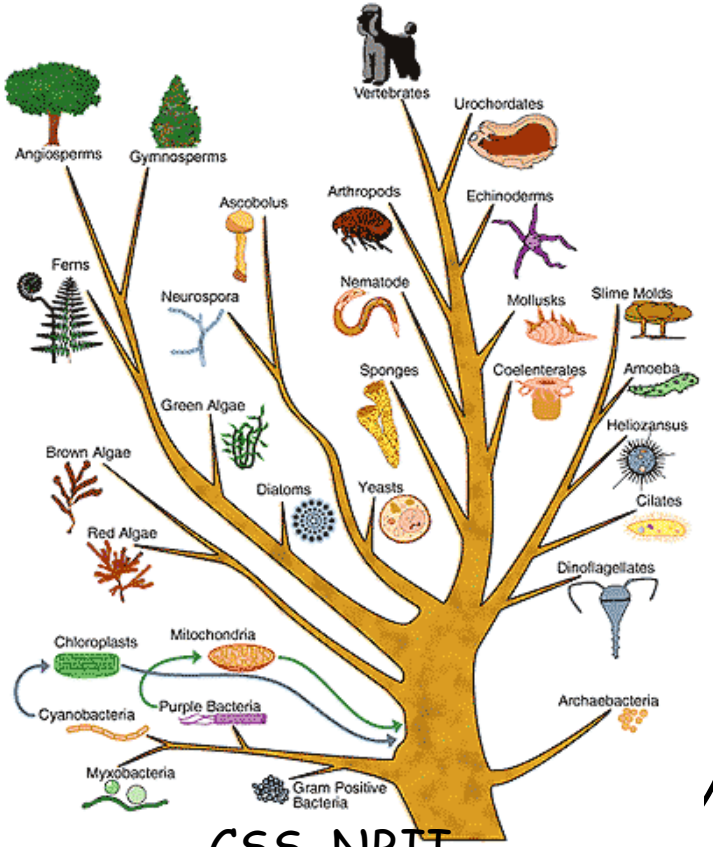
- Disclaimers:
- Won't prove anything hard.
- Many of easy facts are exercises.
- Questions!

Markovian Inheritance Models

- An inheritance graph is nothing but
- A directed acyclic graph (DAG) (V,E).
- u -> v := u is a parent of v, direct ancestor;
- Par(v) := {parents of v}.
- If u -> v_1 -> v_2 -> … v_k = v
- v is a descendant of u, etc.
- Anc(v) = {Ancestors of v}.



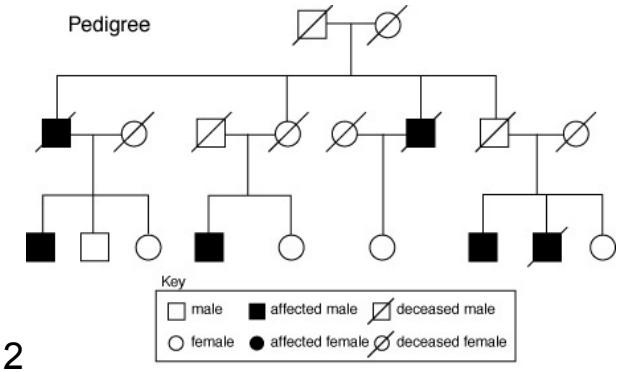
NHGIR, Darryl Lega

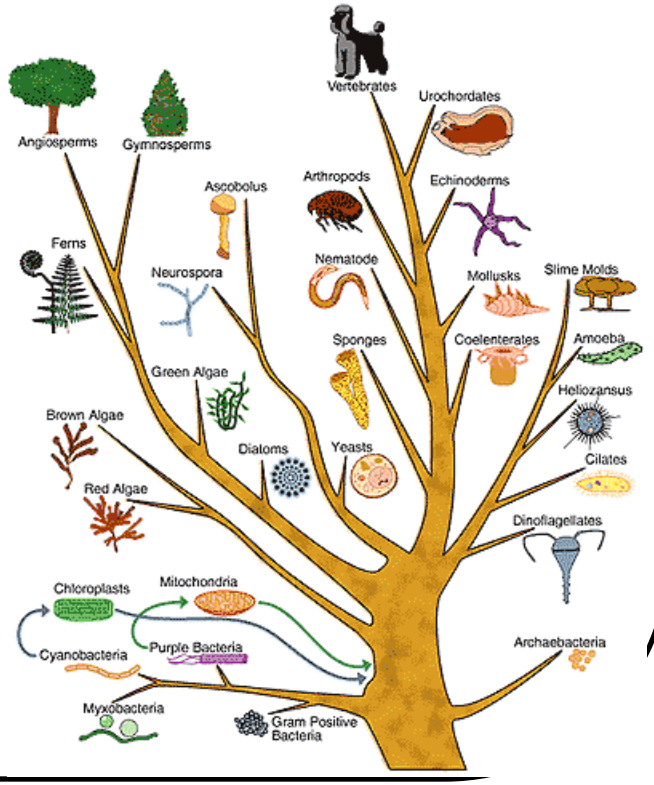


CSS, NBII

Markovian Inheritance Models

- For each $v \in V$, genetic content is given by $\sigma(v)$.
- Def: An MIM is given by 1) a DAG (V,E)
- 2) A probability distribution P on Σ^V satisfying the Markov property:
- $P(\sigma(v) = * \mid \sigma(Anc(v))) = P(\sigma(v) = * \mid \sigma(Par(v)))$
- Ex 1: Phylogeny ↔ speciation.
- Ex 2: Pedigrees ↔ H. genetics.



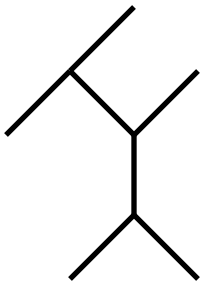


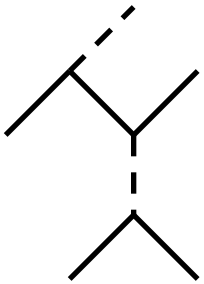
Phylogenetic product models

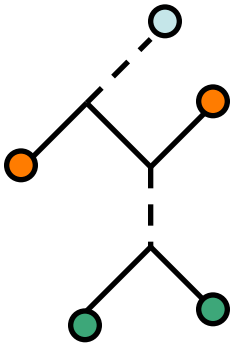
- Def: A Phylogenetic tree is an MIM where (V,E) is a tree.
- Many models are given by products of simpler models.
- Lemma: Let (P,V,E) be an MIM taking values in Σ^V. Then $(P^{\otimes k}, V, E)$ is an MIM taking values in $(\Sigma^k)^V$.
- Pf: Exercise.
- In biological terms:
- Genetic data is given in sequences of letters.
- Each letter evolves independently according to the same law (law includes the DAG (V,E)).

The "random cluster" model

- Infinite set A of colors.
  - "real life" – large $|A|$; e.g. gene order.
- Defined on an un-rooted tree $T=(V,E)$.
- Edge e has (non-mutation) probability $\theta(e)$.
- Character: Perform percolation – edge e open with probability $\theta(e)$.
- All the vertices v in the same open-cluster have the same color σ_v. Different clusters get different colors. This is the "random cluster" model (both for (P,V,E) and $(P^{\otimes k}, V, E)$







Markov models on trees

- Finite set Σ of information values.
- Tree $T=(V,E)$ rooted at r.
- Vertex $v \in V$, has information $\sigma_v \in \Sigma$.
- Edge $e=(v, u)$, where v is the parent of u, has a mutation matrix M^e of size $|\Sigma| \times |\Sigma|$:
- $M_{i,j}^{(v,u)} = P[\sigma_u = j \mid \sigma_v = i]$
- For each character σ, we are given $\sigma_{\partial T} = (\sigma_v)_{v \in \partial T}$, where ∂T is the boundary of the tree.
- Most well knows is the Ising-CFN model.

$$M^e = \begin{pmatrix} \frac{1+\theta(e)}{2} & \frac{1-\theta(e)}{2} \\ \frac{1-\theta(e)}{2} & \frac{1+\theta(e)}{2} \end{pmatrix}.$$

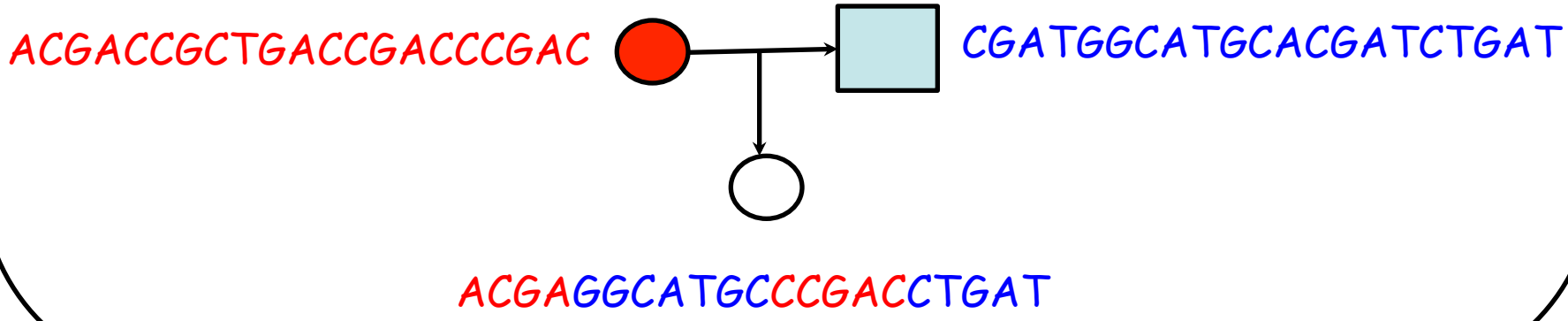
Insertions and Deletions on Trees

- Not a product model (Thorne, Kishino, Felsenstein 91-2)
- Vertex $v \in V$, has information $\sigma_v \in \Sigma^*$ .Then:
- Apply Markov model (e.g. CFN) to each site independently.
- Delete each letter indep. With prob $p_d(e)$.
- There also exist variants with insertions.

ACGACCGCTGACCGACCCGACGTTGTAAACCGT	Original Sequence
ACGACCGTTGACCGACCCGACATTGTAAACTGT	Mutations
ACGACCGTTGACCGACCCGACATTGTAAACTGT	Deletions
ACGCCGTTGACCGCCCGACTTGTAACTGT	Mutated Sequence

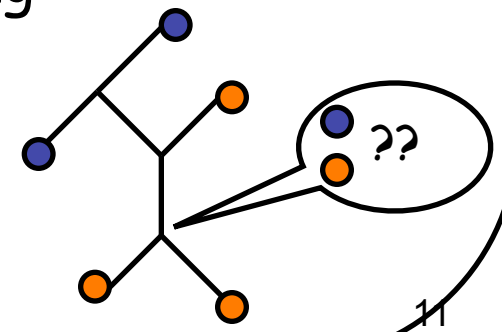
A simple model of recombination on pedigrees

- Vertex $v \in V$, has information $\sigma_v \in \Sigma^k$.
- Let π be a probability distribution over subsets of $[k]$.
- Let u,w be the father and mother of v.
- Let S be drawn from π and let:
- $\sigma_v(S) = \sigma_u(S), \quad \sigma_v(S^c) = \sigma_w(S^c)$.
- Example: i.i.d. "Hot spot" process on $[k]$: $\{X_1,\ldots X_r\}$
  Let $S = [1,X_1] \cup [X_2,X_3] \cup \ldots$



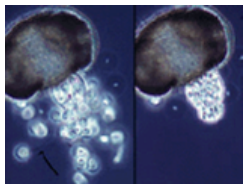
The reconstruction problem

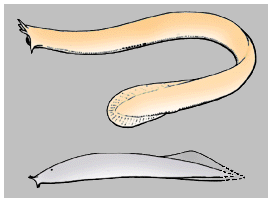
- We discuss two related problems.
- In both, want to reconstruct/estimate unknown parameters from observations.
- The first is the "reconstruction problem".
- Here we are given the tree/DAG and
- the values of the random variables at a subset of the vertices.
- Want to reconstruct the value of the random variable at a specific vertex ("root").
- For trees this is algorithmically easy using Dynamic programs / recursion.



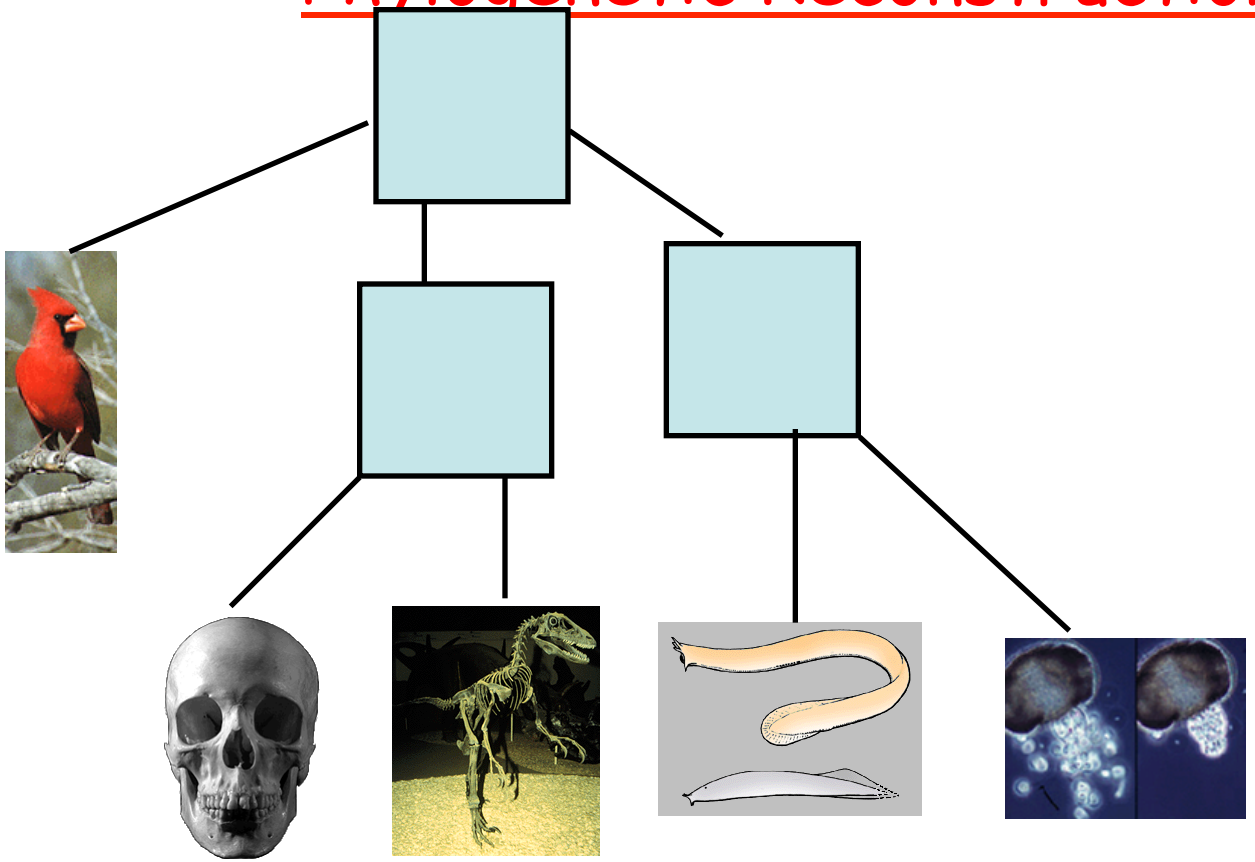
Phylogenetic Reconstruction

- Here the tree/DAG etc. is unknown.
- Given a sequence of collections of random variables at the leaves ("species").
- Want to reconstruct the tree (un-rooted).





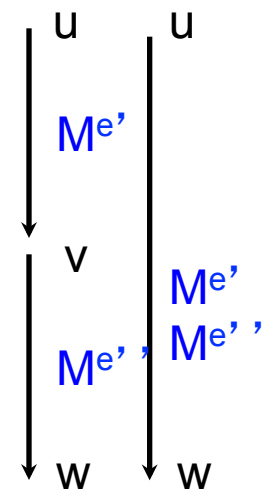
Phylogenetic Reconstruction

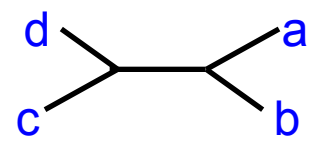


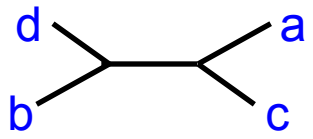
- Algorithmically "hard". Many heuristics based on Maximum-Likelihood, Bayesian Statistics used in practice.

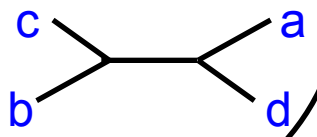
Trees

- In biology, all internal degrees ≥ 3.
- Given a set of species (labeled vertices) X, an X-tree is a tree which has X as the set of leaves.
- Two X-trees T_1 and T_2 are identical if there's a graph isomorphism between T_1 and T_2 that is the identity map on X.









Highlights for next lectures

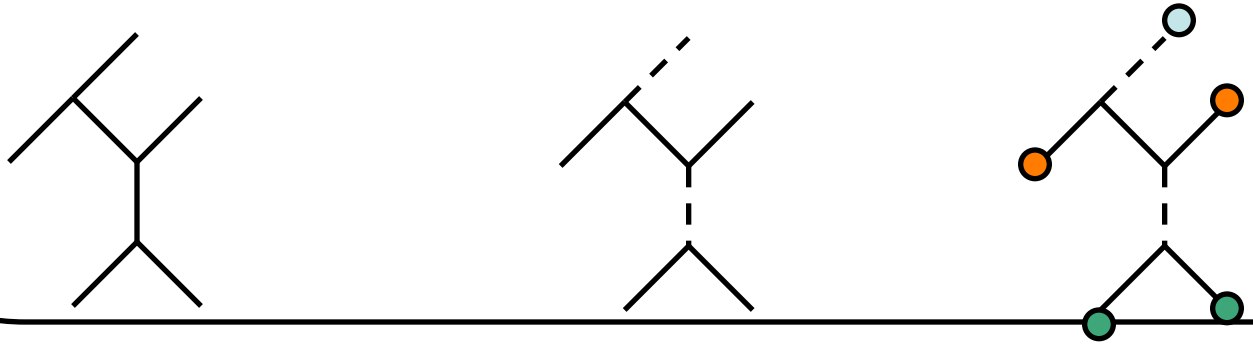
- Develop methods to reconstruct Phylogenies with the following guarantees.
- Consider large trees (# of leaves n -> ∞)
- Show that for all trees with high probability (over randomness of inheritance) recover the true tree.
- Upper and lower bounds on amount of information needed.
- Surprising connections with phase transitions in statistical physics.
- Briefly discuss why non-tree models are much harder.

Lecture plan

- Lecture 2: Reconstruction via metric estimates.
- Metrics from stochastic models.
- Tree Metrics determine trees.
- Approximate Tree Metrics determine trees.
- Some tree reconstruction algorithms.
- Metric and geometric ideas for tree mixtures.
- Metrics and pedigrees.

The "random cluster" model

- Infinite set A of colors.
  - "real life" – large $|A|$; e.g. gene order.
- Defined on an un-rooted tree $T=(V,E)$.
- Edge e has (non-mutation) probability $\theta(e)$.
- Character: Perform percolation – edge e open with probability $\theta(e)$.
- All the vertices v in the same open-cluster have the same color σ_v. Different clusters get different colors. This is the "random cluster" model (both for (P,V,E) and $(P^{\otimes k}, V, E)$



An additive metric for the RC model

- Claim: For all u,v: $P(\sigma_u = \sigma_v) = \prod_{e \in path(u,v)} \theta(e)$, where the product is over all e in the path connecting u to v.
- Def: Let $d(e) = -\log \theta(e)$, and $d(u,v) = \sum_{e \in path(u,v)} d(e) = -\log P(\sigma_u = \sigma_v)$
- Claim: d(u,v) is a metric
  - Pf: Exercise

Markov models on trees

- Finite set Σ of information values.
- Tree $T=(V,E)$ rooted at r.
- Vertex $v \in V$, has information $\sigma_v \in \Sigma$.
- Edge $e=(v, u)$, where v is the parent of u, has a mutation matrix M^e of size $|\Sigma| \times |\Sigma|$:
- $M_{i,j}^{(v,u)} = P[\sigma_u = j \mid \sigma_v = i]$
- For each character σ, we are given $\sigma_{\partial T} = (\sigma_v)_{v \in \partial T}$, where ∂T is the boundary of the tree.
- Most well knows is the Ising-CFN model.

$$M^e = \begin{pmatrix} \frac{1+\theta(e)}{2} & \frac{1-\theta(e)}{2} \\ \frac{1-\theta(e)}{2} & \frac{1+\theta(e)}{2} \end{pmatrix}.$$

Markov models on trees

- Most well knows is the Ising-CFN model on {-1,1}:

$$M^e = \begin{pmatrix} \frac{1+\theta(e)}{2} & \frac{1-\theta(e)}{2} \\ \frac{1-\theta(e)}{2} & \frac{1+\theta(e)}{2} \end{pmatrix}.$$

- Claim: For all u,v: $E[\sigma_u \sigma_v] = \prod_{e \in \text{path}(u,v)} \theta(e)$.
- Pf: Exercise.
- Claim: $d(u,v) = -\log E[\sigma_u \sigma_v]$ is a metric and $d(u,v) = \sum_{e \in \text{path}(u,v)} d(e)$
- This a special case of the log-det distance for General Markov models on trees (Steel 94)

  $d(u,v) \sim -\log |\det \prod_{e \in \text{path}(u,v)} M^e|$

Insertions and Deletions on Trees

- Not a product model (Thorne, Kishino, Felsenstein 91-2)
- Vertex $v \in V$, has information $\sigma_v \in \Sigma^*$ .Then:
- Delete each letter indep. With prob $p_d(e)$.

ACGACCGTTGACCGACCCGACATTGTAAACTGT Original Sequence

ACGACCGTTGACCGACCCGACATTGTAAACTGT Deletions

ACGCCGTTGACCGCCCGACTTGTAACTGT Mutated Sequence

- Define $d(u,v) = -\log E[Avg(\sigma_u)\, Avg(\sigma_v)]$
- This is a metric (Ex ; Daskalakis-Roch 10).
- Same also works if also insertions and mutations allowed.

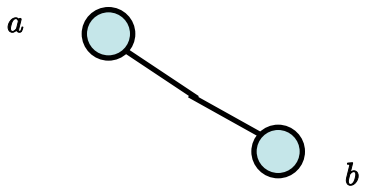
From metrics to trees

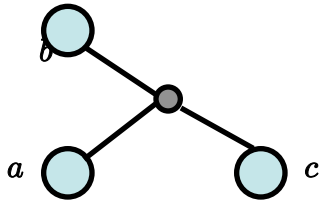
- Def: Given a tree $T=(V,E)$ a tree metric is defined by a collection of positive numbers $\{ d(e) : e \in E\}$ by:
  letting: $d(u,v) = \sum_{e \in \text{ path}(u,v)} d(e)$ all $u,v \in V$.

- Claim: Let $T=(V,E)$ a tree with all internal degrees at least 3, let d be a tree metric on T and let L be the set of leaves of T. Then $\{ d(u,v) : u,v \in L \}$ determines the tree T uniquely.

Think small: trees on 2 and 3 leaves

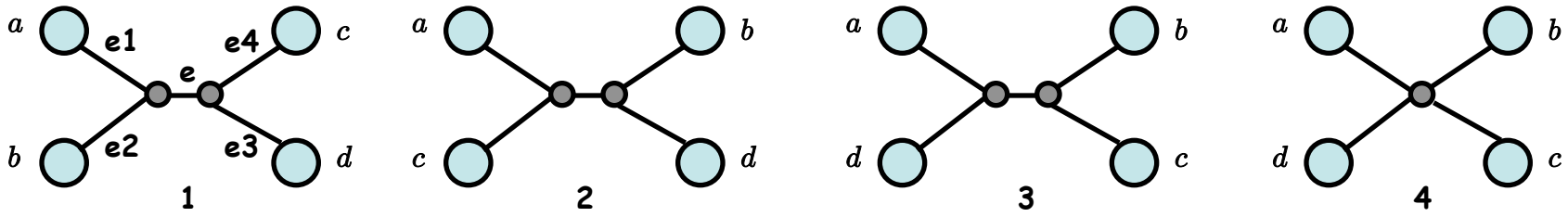
- Q: What are the possible trees on 2 / 3 leaves a,b,c?
- A: Only one tree if we assume all int. deg > 2.





Think small: trees on 4 leaves

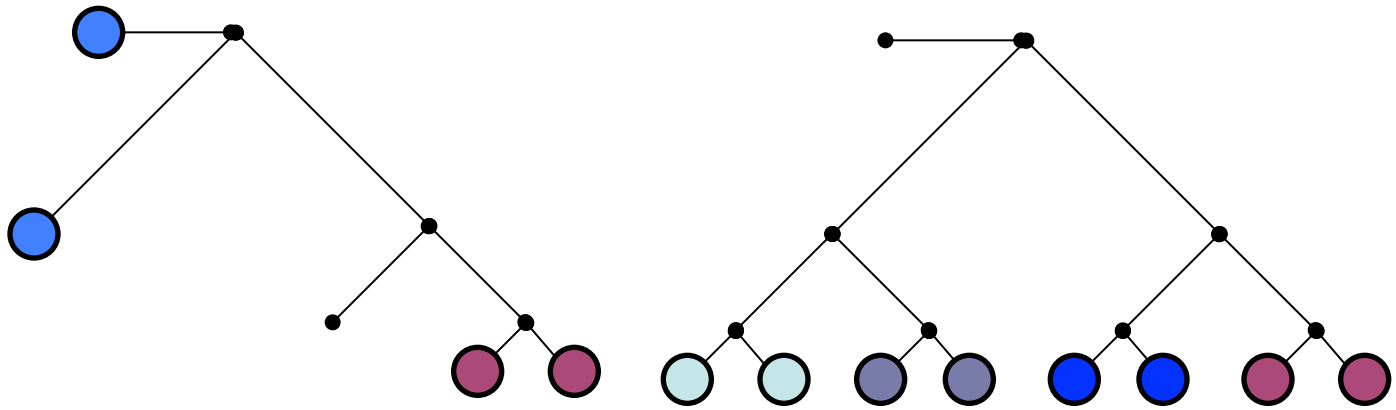
- Q: What are the possible trees on 4 leaves a,b,c,d?
- A: ab|cd, ac|bd, ad|bc or abcd



- Q: How to distinguish between them , given the leaves' pairwise distances of the leaves?
- A: Look at partition xy, zw minimizing d(x,y) + d(z,w)
  - Case 1-3 : The partition corresponding to the tree will give the optimum distance – d(e1)+d(e2)+d(e3)+d(e4), while all other partitions will give distance bigger by 2d(e) (go through the middle edge twice).
  - Case 4 (star) : All partitions will give the same result.
  - Note: Approximate distances (+/- d(e)/8) suffice!

From Small Tree to Big Trees

- Claim: In order to recover tree topology suffice to know for each set of 4 leaves what is the induced tree.
- Pf: By induction on size of tree using Cherries.
- Definition: A cherry is a pair of leaves at graph distance 2.
- Claim1 : vertices x,y make a cherry in the tree T iff they are a cherry in all trees created of 4 of the it's leaves.
- Claim2 : Every tree with all internal degrees ≥ 3 has a cherry
- Proof : Pick a root, take u to be the leaf farthest away from the root. The sibling of u (must exist one as the degree ≥ 3) must be a leaf as well.



From leaf pairwise distances to trees

- Algorithm to build tree from quartets :
  - Find cherries (pairs of vertices which are coupled in all 4-leaves combinations).
  - For each cherry <x,y> replace it by a single leaf x (remove all quartets involving both x,y; each quartet including only y – replace the y by x)
  - Repeat (until # leaves ≤ 4)

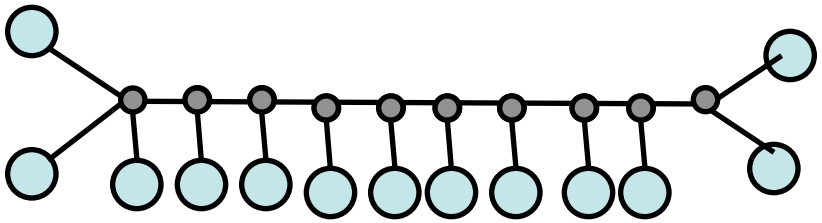
- A statistical Q: How many samples k are needed?
- In other words: what is the seq length needed?
- A: We would like to have enough samples so we can estimate d(u,v) with accuracy $\min_e\{d(e)/8\}$
- Define $f = \min_e d(e)$, $g = \max_e d(e)$, $D = \max_{\{u,v \text{ leaves}\}} d(u,v)$.

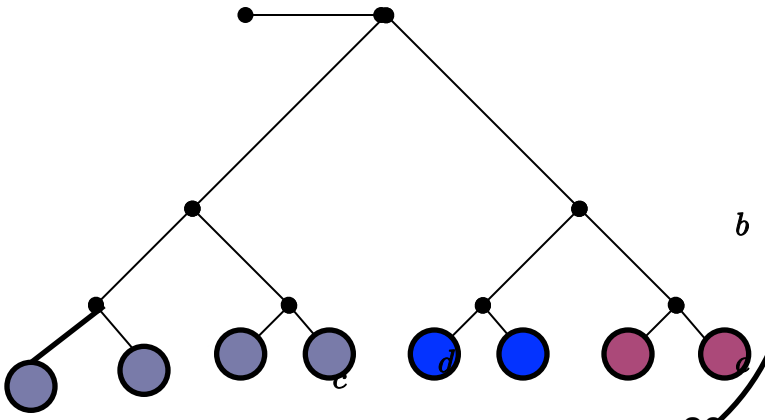
From leaf pairwise distances to trees

- A statistical Q: How many samples are actually needed?
- A: We would like to have enough samples so we can estimate $d(u,v)$ with accuracy $\min_e\{d(e)/8\}$
- Define $f = \min_e d(e)$, $g = \max_e d(e)$, $D = \max_{\{u,v \text{ leaves}\}} d(u,v)$.
- In RC-model: e^{-D} vs. $e^{-D-f/8}$ agreement.
- In CFN: e^{-D} vs. $e^{-D-f/8}$ correlation.
- Etc.
- Claim: In both models need at least $O(e^D/g^2)$ samples to estimate all distances within required accuracy.
- Claim: In both models $O(\log n\, e^D/g^2)$ suffice to estimate all distances with required accuracy with good probability.
- Exercises!

From leaf pairwise distances to trees

- Claim: In both models need at least $O(e^D/g)$ samples to estimate all distances within required accuracy.
- Claim: In both models $O(\log n\, e^D/g^2)$ suffice to estimate all distances with required accuracy with good probability.
- Q: Is this bad? How large can D be? Let n = # leaves.
- D can be as small as $O(\log n)$ and as large as $O(n)$.
- If $D = f\, n$ need $O(e^{f\, n}/g^2)$ samples!
- Can we do better?

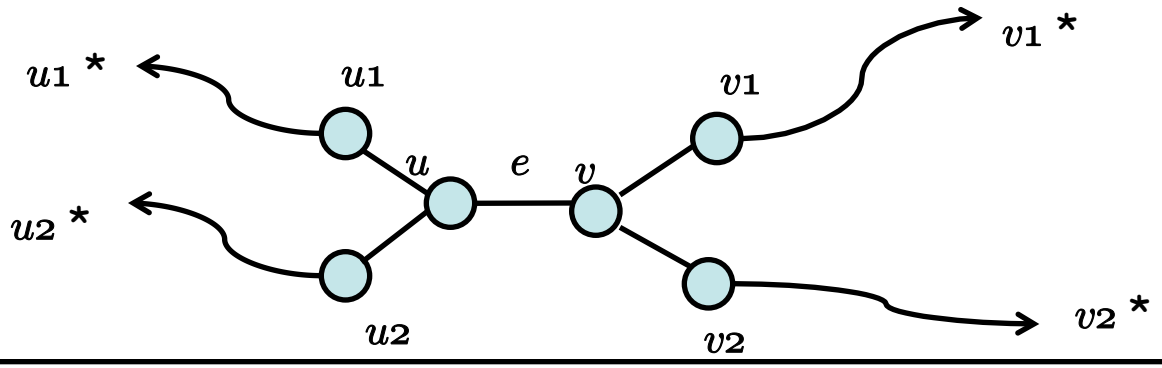




From leaf pairwise distances to trees

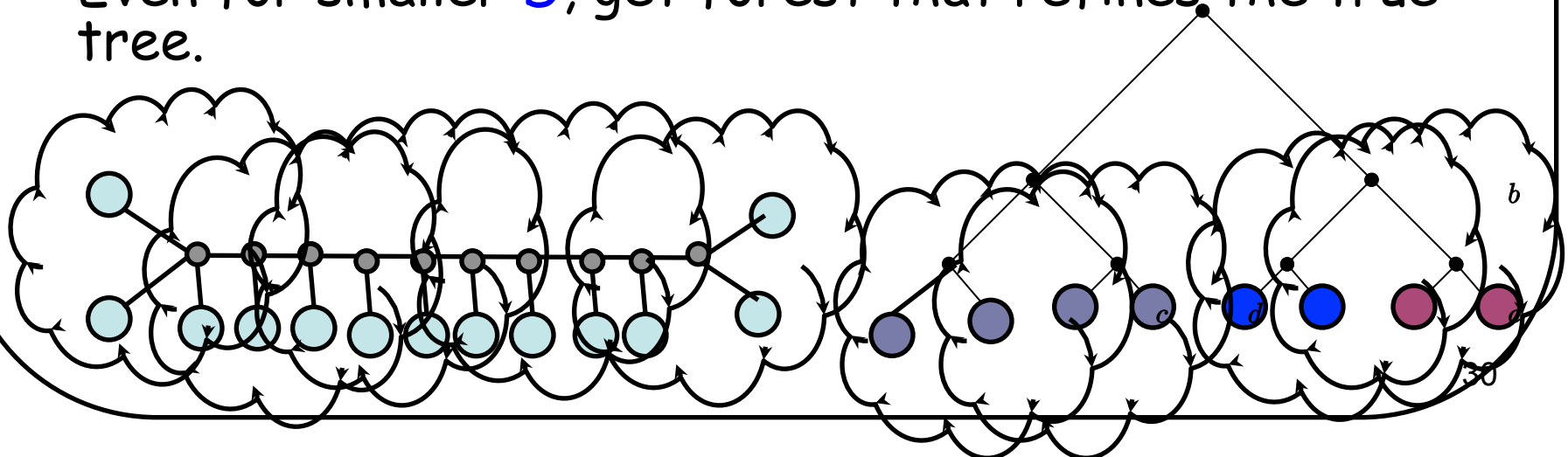
- Can we do better?
- Do we actually need *all* pairwise distances?
- Do we actually need *all* quartets?

- In fact: Need only "short quartets" so actual # of samples needed is $O(e^{8 f \log n} / g^2)$ (Erods-Steel-Szekeley-Warnow-96).
- An alternative approach is in Mossel-09:



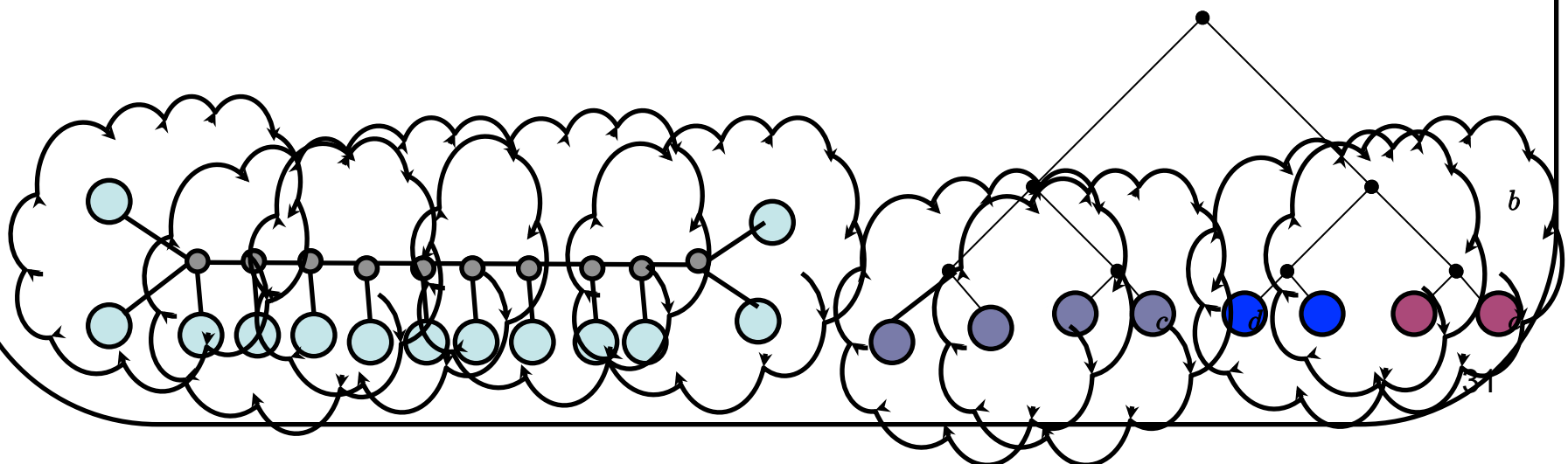
Distorted metrics idea sketch

- Construction: given a radius D:
- For each leaf u look at C(u,D) = all leaves v whose estimated distance to u is at most D.
- Construct the tree T(u,D) on C(u,D).
- Algorithm to stitch T(u,D)'s (main combinatorial argument)
- Sequence length needed is $O(e^{2D}/g^2)$
- Lemma: if $D > 2\, g \log n$, will cover the tree.
- Even for smaller D, get forest that refines the true tree.



Short and long edges

- Gronau, Moran, Snir 2008: dealing with short edges (sometimes need to contract)
- Daskalakis, Mossel, Roch 09: dealing with both short and long edges: "contracting the short, pruning the deep".



Can we do better?

- Consider e.g. the CFN model with sequence length k.
- Results so far $\Rightarrow$ model can be reconstruct when $k = O(n^{\alpha})$ where $\alpha = \alpha(f,g)$.
- Can we do better?
- Can we prove lower bounds?

Can we do better?

- Can we prove lower bounds?

- Trivial lower bound:
- Claim 1: T_n = set of leaf labeled trees on n leaves (and all degrees at least 3). Then $|T_n| = \exp(\Theta(n \log n))$.
- Pf: Exercise.

- Claim 2: # of possible sequences at the leaves is $2^{k\,n}$.

- Conclusion: To have good prob. of reconstruction need
- $2^{n\,k} > \exp(\Theta(n \log n))$. $\Rightarrow k \geq \Omega(\log n)$

Can we do better?

- More formally:

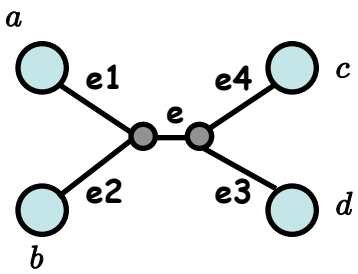
- Claim: Consider a uniform prior over trees μ.
- Then for all possible estimators Est
- E_μ P[Est is correct] $\leq 2^{n\,k} / |T_n|$.
- Pf sketch:
- The optimal estimator is deterministic:
- Est : $\{0,1\}^{n\,k}$ -> T_n.
- E_μ P[Est is correct] $\leq$ |Image(Est)| / $|T_n| \leq 2^{n\,k} / |T_n|$

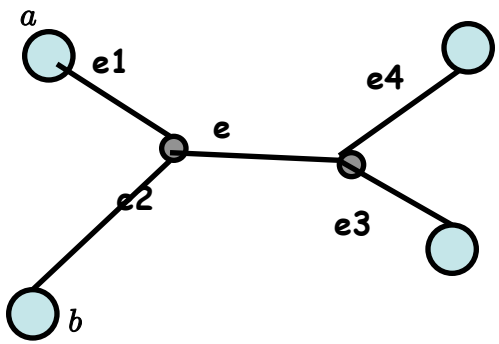
- Conclusion: Impossible to reconstruct if $k \leq 0.5 \log n$ and possible if $k \geq n^\alpha$. What is the truth?
- Next lecture …

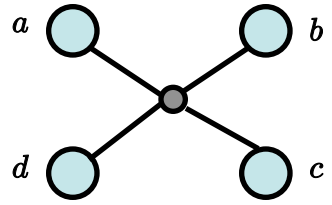
Metric ideas for tree mixtures

- Def: Let $T_1=(V_1,E_1,P_1)$ and $T_2 = (V_2, E_2, P_2)$ be two phylogenetic models on the same leaf set L.
- The $(\alpha,1-\alpha)$ mixture of the two models is the probability distribution $\alpha P_1 + (1-\alpha) P_2$

- Construction (Matsen Steel 2009):
- There exist 3 phylogenies T_1, T_2, T_3 for the CFN model with $(V_1,E_1) = (V_2,E_2) \neq (V_3, E_3)$ and $T_3= 0.5(T_1 + T_2)$
- $\Rightarrow$ Mixtures are not identifiable!







Metric ideas for tree mixtures

- Construction (Matsen Steel 2009):
- There exist 3 phylogenies T_1, T_2, T_3 for the CFN model with $(V_1,E_1) = (V_2,E_2) \neq (V_3, E_3)$ and $T_3 = 0.5(T_1 + T_2)$
- $\Rightarrow$ Mixtures are not identifiable!

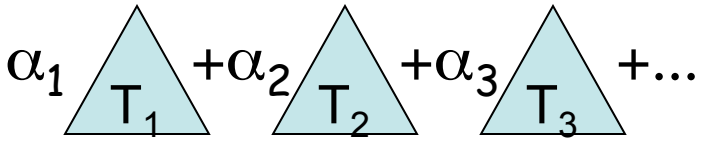
- On the other hand, using metric idea in a recent work with Roch we show that when n is large and the trees T_1 and T_2 are generic it is possible to find both of them with high probability.

Metric ideas for tree mixtures

- Proof sketch: Fix a radius $D \geq 10g$.
- Let $S_1 = \{ u, v \in \text{Leaves}: d_1(u,v) \leq D\}$
- Easy to show that $|S_2|, |S_1| \geq \Omega(n)$
- For "generic trees" we have $|S_2 \cap S_1| = o(n)$
- By looking for high correlation between leaves we can approximately recover $S_1 \cup S_2$.
- Note: Pairs in S_1 will tend to be correlated in samples from T_1 and pairs in S_2 will be correlated in samples from T_2.
- By checking co-occurrence of correlation can approximately recover both S_1 and S_2.
- Using S_1 and S_2 can determine for each sample if it comes from T_1 or from T_2
- Same ideas can be used for different rates …

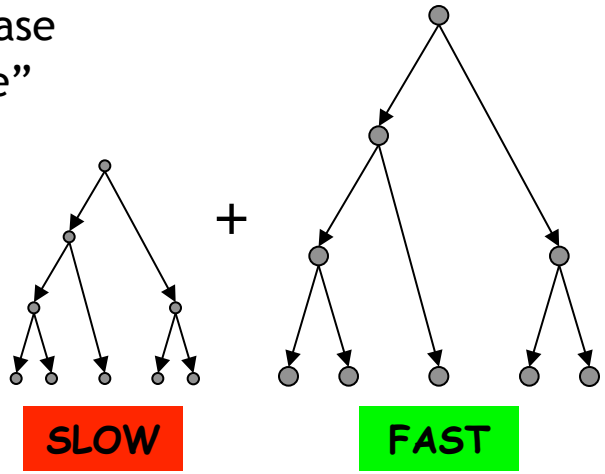
heterogeneous data

- **phylogenetic mixtures** – definition by picture:

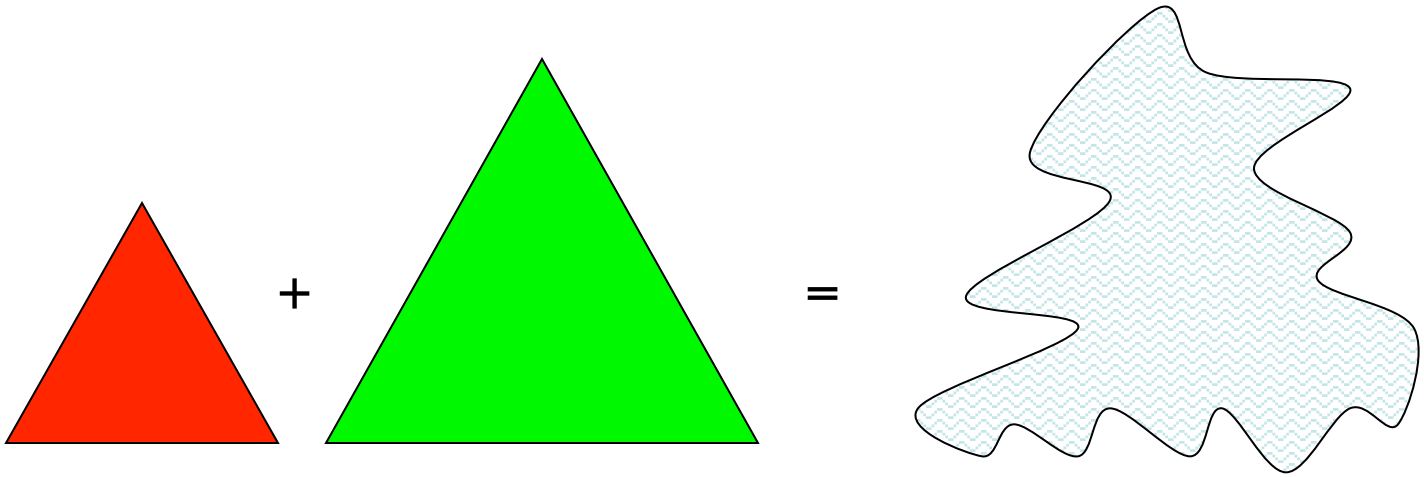


- **special case** – "rates-across-sites"
  - trees are the same up to **random scaling**
  - in this talk, will focus on two-scaling case
  - can think of scaling as "hidden variable"

- **biological motivation**
  - heterogeneous mutation rates
  - inconsistent lineage histories
  - hybrid speciation, gene transfer
  - corrupted data

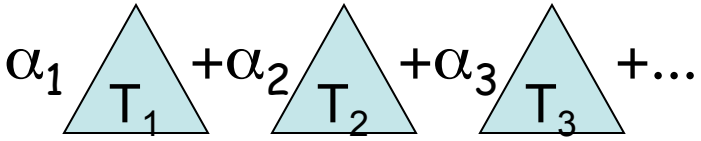


but, on a mixture…



why are mixtures problematic?

- **identifiability** – does the distribution at the leaves determine the α's and T's?
  - negative results: e.g. [Steel et al.'94], [Stefankovic-Vigoda'07], [Matsen-Steel'07], etc.
  - positive results: e.g. [Allman, Rhodes'06,'08], [Allman, Ane, Rhodes'08], [Chai-Housworth'10], etc.



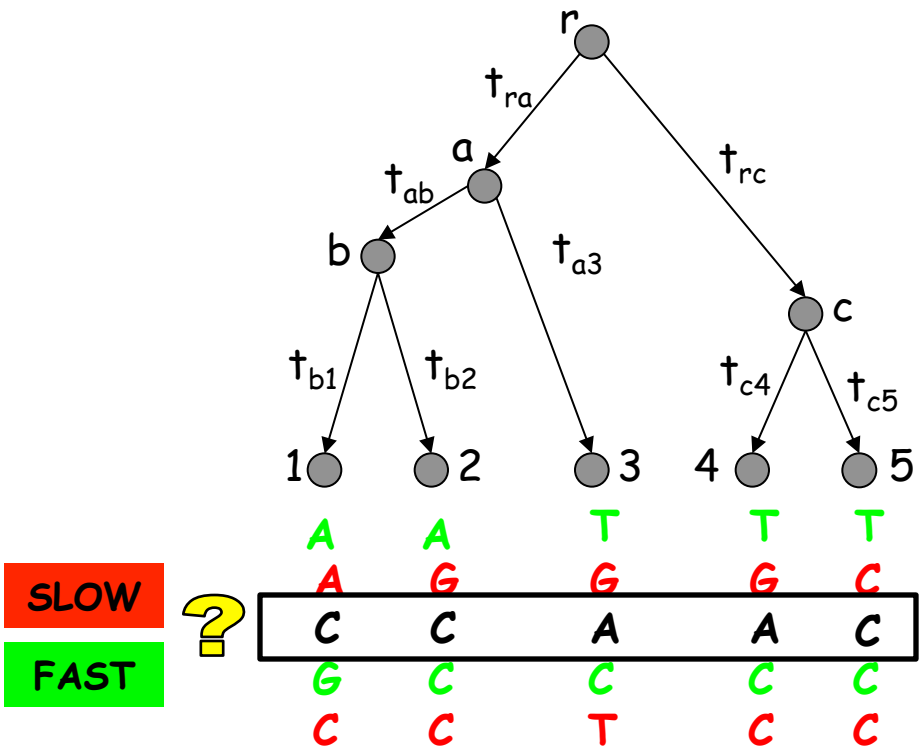
- **algorithmic** – assuming identifiability, can we reconstruct the topologies efficiently?
  - can mislead standard methods;
  - ML under the full model is consistent in identifiable cases; BUT ML is already NP-hard for pure case [Chor,Tuller'06, R.'06]

a new site clustering approach

- **new results** [M-Roch, 2011] – we give a simple way to **determine** which sites come from which component
  - based on concentration of measure in large-tree limit

site clustering

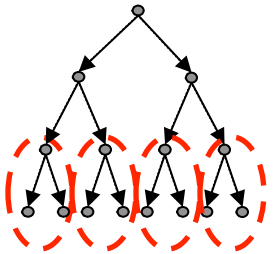
- ideally, guess **which sites** were produced by **each component**
  - scaling is “hidden” but we can try to infer it
  - to be useful, a test should work with high confidence



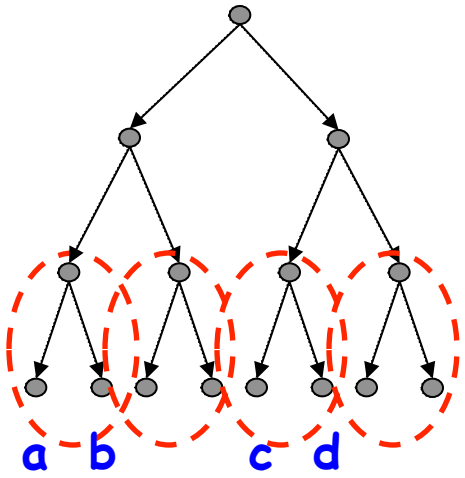
leaf agreement

- **a natural place to start** - impact of scaling on **leaf agreement**
  - one pair of leaves is not very informative
  - we can look at many pairs

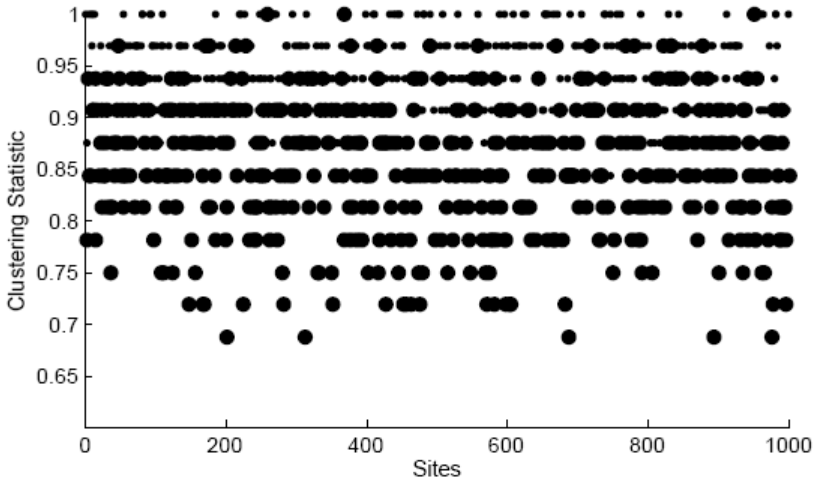
$$C = \sum_{(a,b)\in R \subseteq L^2} \mathrm{I}\{s_a = s_b\}$$

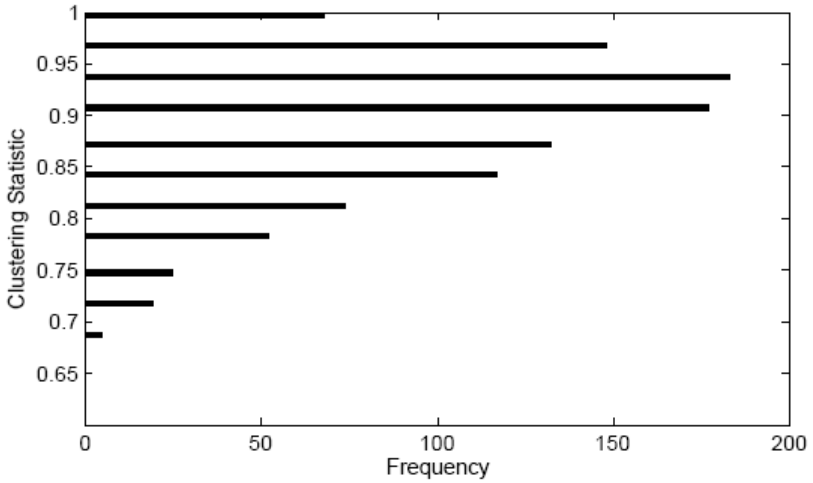


- we would like C to be **concentrated**:
  - large number of pairs
  - each pair has a small contribution
  - independent (or almost independent) pairs
  - nice separation between SLOW and FAST

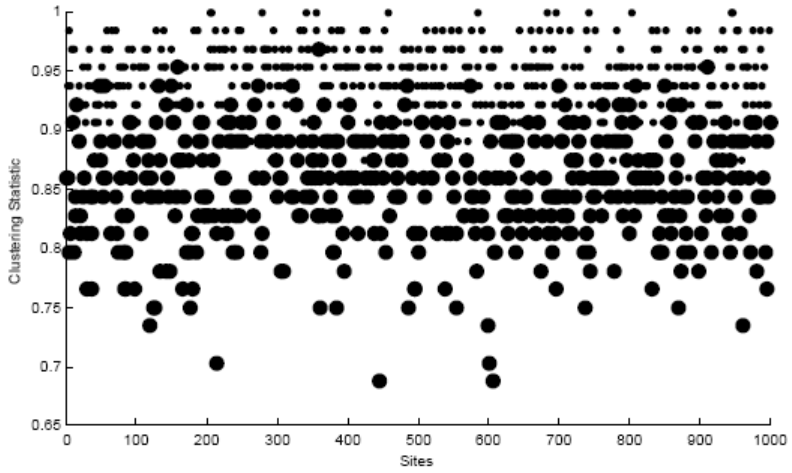


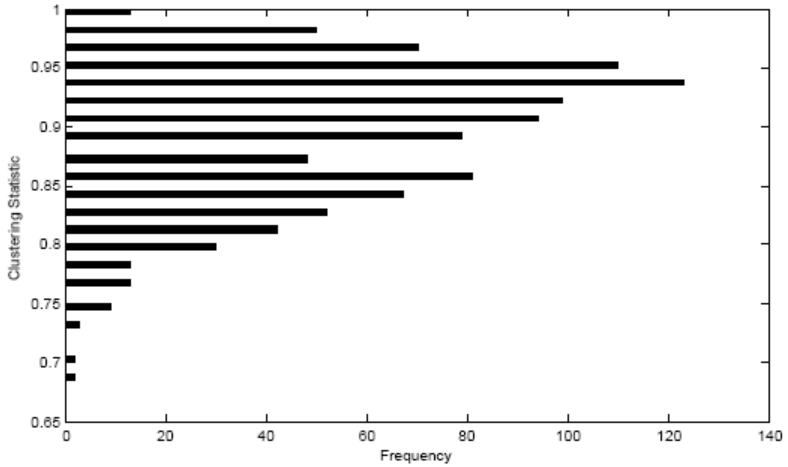
64 leaves


Clustering Statistic
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
200
400
600
800
1000
Sites

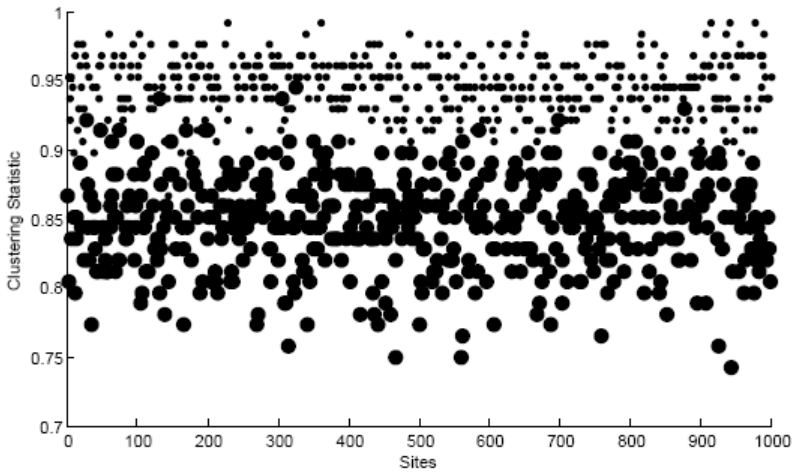

Clustering Statistic
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
50
100
150
200
Frequency

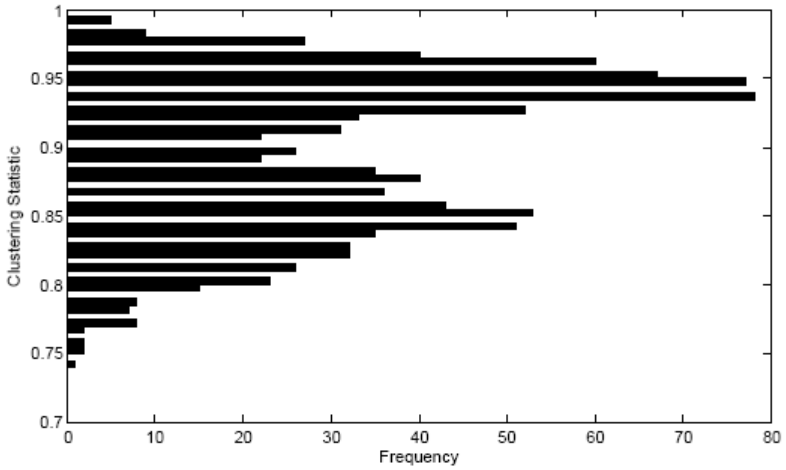
128 leaves


Clustering Statistic
Sites
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
100
200
300
400
500
600
700
800
900
1000

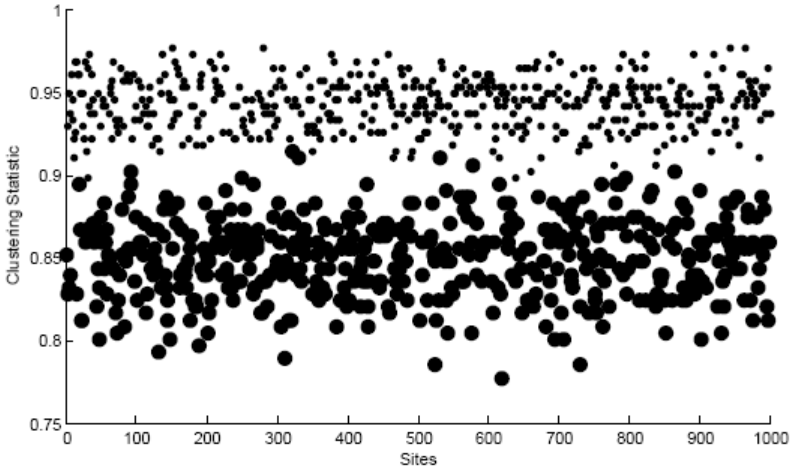

Clustering Statistic
Frequency
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
20
40
60
80
100
120
140

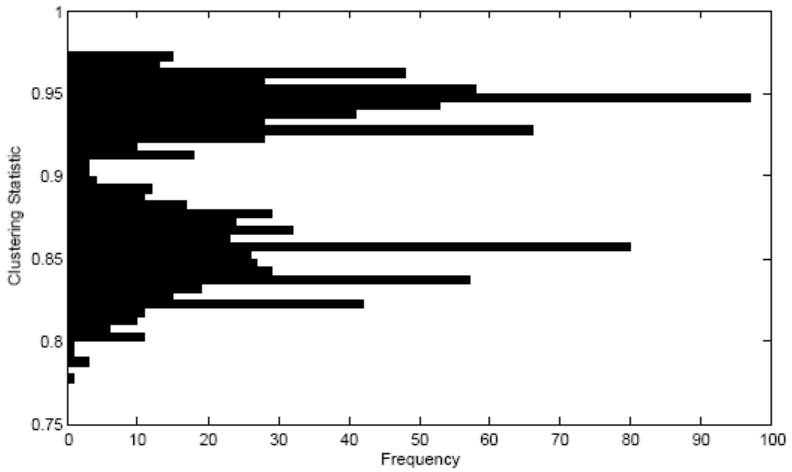
256 leaves


Clustering Statistic
Sites
1
0.95
0.9
0.85
0.8
0.75
0.7
0
100
200
300
400
500
600
700
800
900
1000


Clustering Statistic
Frequency
1
0.95
0.9
0.85
0.8
0.75
0.7
0
10
20
30
40
50
60
70
80

512 leaves

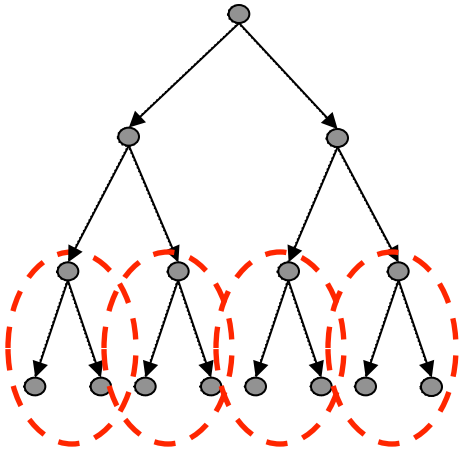

Clustering Statistic
1
0.95
0.9
0.85
0.8
0.75
0
100
200
300
400
500
600
700
800
900
1000
Sites


Clustering Statistic
1
0.95
0.9
0.85
0.8
0.75
0
10
20
30
40
50
60
70
80
90
100
Frequency

but the tree is not complete…

- **lemma 1** – on a general binary tree, the set of all pairs of leaves at distance at most 10 is linear in n
  - proof: count the number of leaves with no other leaves at distance 5
- **lemma 2** – in fact, can find a linear set of leaf pairs that are non-intersecting
  - proof: sparsify above
- this is enough to build a concentrated statistic

$$\hat{C} = \sum_{(a,b)\in\hat{R}\subseteq L^2} \mathrm{I}\{s_a = s_b\}$$

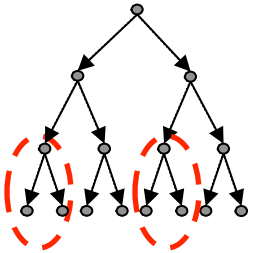


but we don't know the tree…

- **a simple algorithm** – cannot compute exact distances but can tell which pairs are more or less correlated

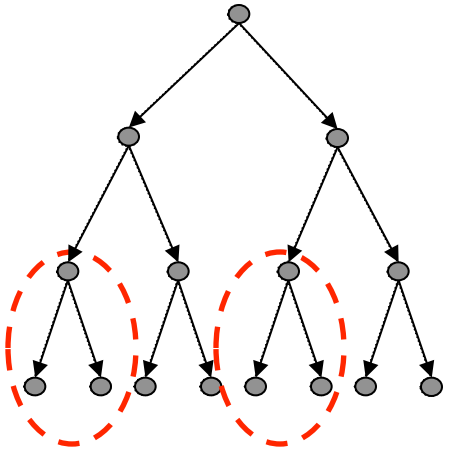
  - find "close" pairs
  - starting with one pair, remove all pairs that are too close
  - pick one of the remaining pairs and repeat

$$\hat{C} = \sum_{(a,b)\in\hat{R}\subseteq L^2} \mathrm{I}\{s_a = s_b\}$$

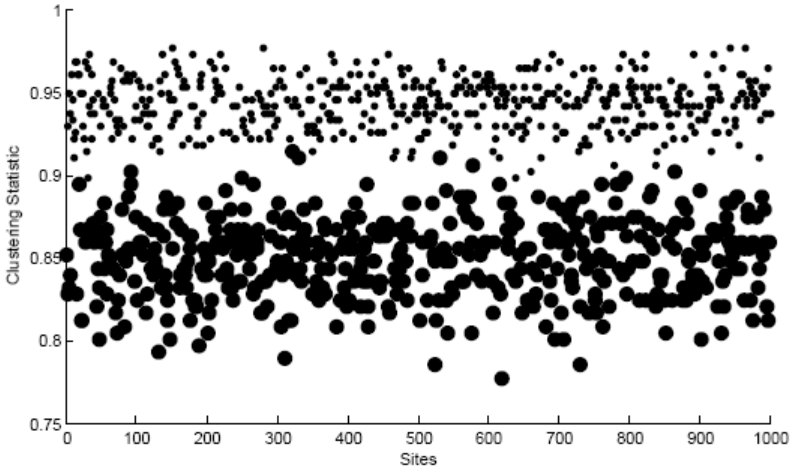


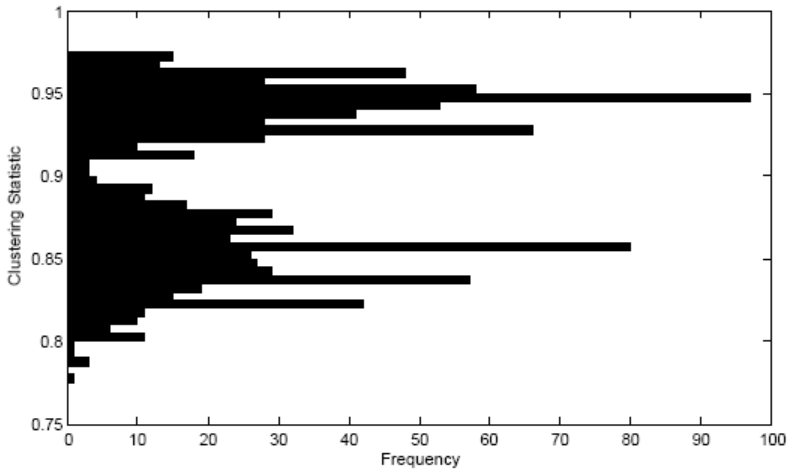
- **claim** – this gives a nicely concentrated variable (for large enough trees)

  - large number of pairs
  - independent (or almost independent) pairs
  - nice separation between SLOW and FAST



site clustering + reconstruction





summary

Proposition 4 (Site Clustering: RAS-JC Model) *Under the assumptions stated in Section 2 on the RAS-JC model, for any given tolerance on the mutation and mixture parameters, there exists a high-confidence site clustering algorithm.*

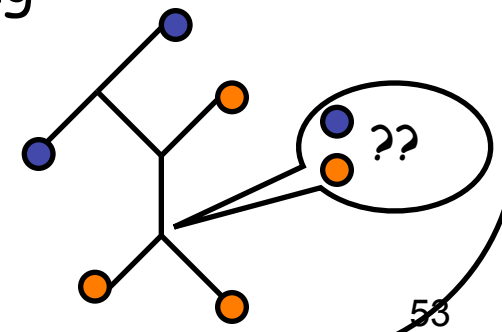
Proposition 5 (Full Reconstruction: RAS-JC Model) *Under the assumptions stated in Section 2 on the RAS-JC model, for any given tolerance on the mutation and mixture parameters, there exists a high-probability reconstruction algorithm using polynomial-length sequences and running in polynomial time.*

Metric ideas for pedigrees

- Correlation measure = inheritance by decent
- Doesn't really measure distance but something more complicated …

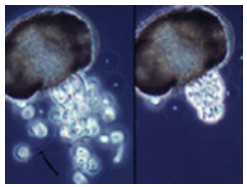
The reconstruction problem

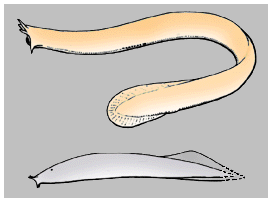
- We discuss two related problems.
- In both, want to reconstruct/estimate unknown parameters from observations.
- The first is the "reconstruction problem".
- Here we are given the tree/DAG and
- the values of the random variables at a subset of the vertices.
- Want to reconstruct the value of the random variable at a specific vertex ("root").
- For trees this is algorithmically easy using Dynamic programs / recursion.



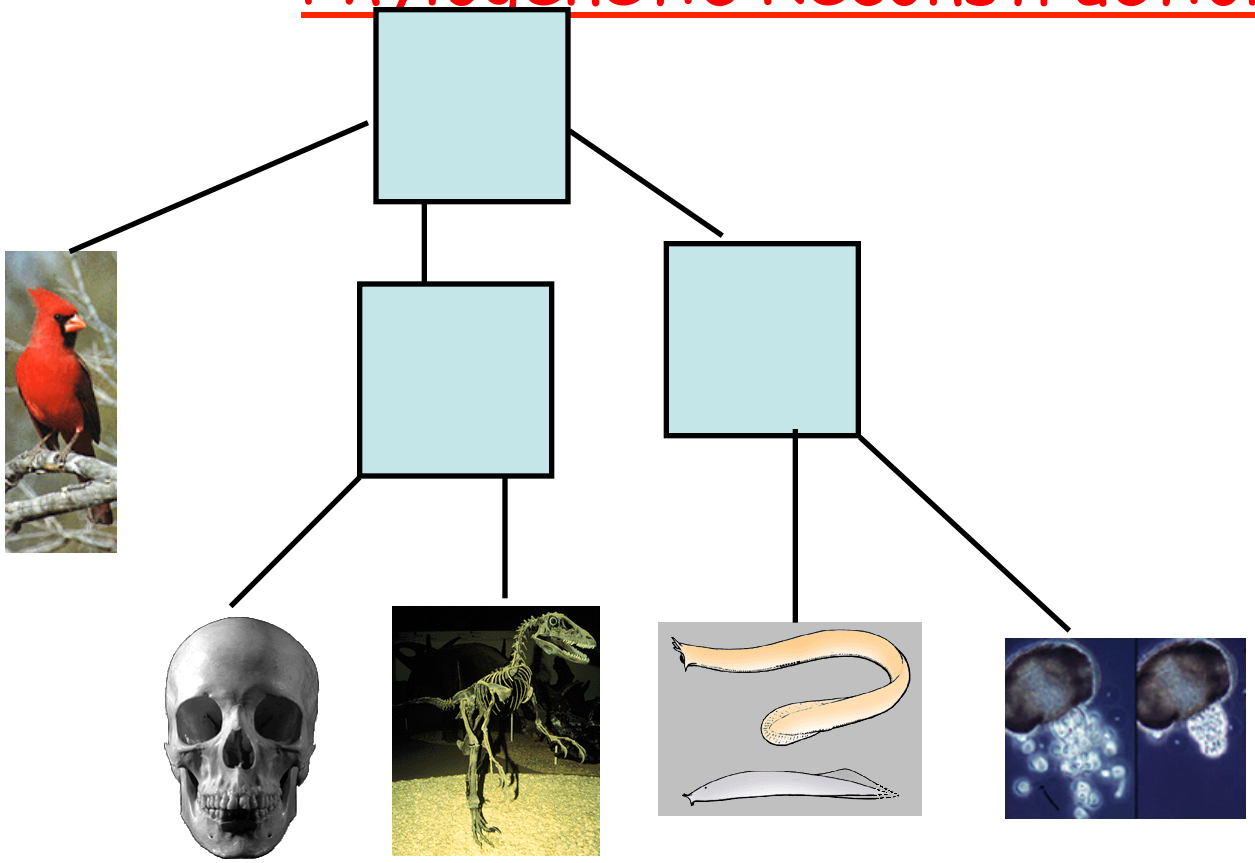
Phylogenetic Reconstruction

- Here the tree/DAG etc. is unknown.
- Given a sequence of collections of random variables at the leaves ("species").
- Want to reconstruct the tree (un-rooted).





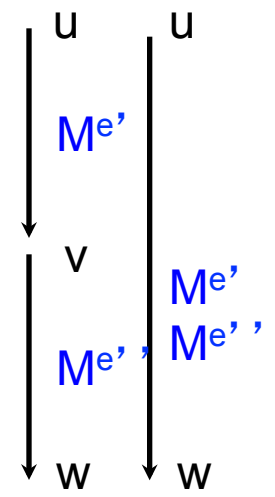
Phylogenetic Reconstruction

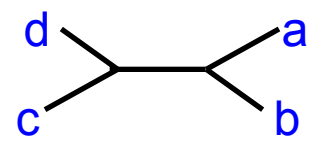


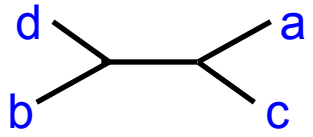
- Algorithmically "hard". Many heuristics based on Maximum-Likelihood, Bayesian Statistics used in practice.

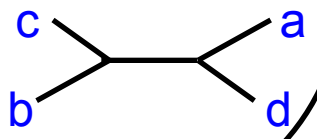
Trees

- In biology, all internal degrees ≥ 3.
- Given a set of species (labeled vertices) X, an X-tree is a tree which has X as the set of leaves.
- Two X-trees T_1 and T_2 are identical if there's a graph isomorphism between T_1 and T_2 that is the identity map on X.









Highlights for next lectures

- Develop methods to reconstruct Phylogenies with the following guarantees.
- Consider large trees (# of leaves n -> ∞)
- Show that for all trees with high probability (over randomness of inheritance) recover the true tree.
- Upper and lower bounds on amount of information needed.
- Surprising connections with phase transitions in statistical physics.
- Briefly discuss why non-tree models are much harder.