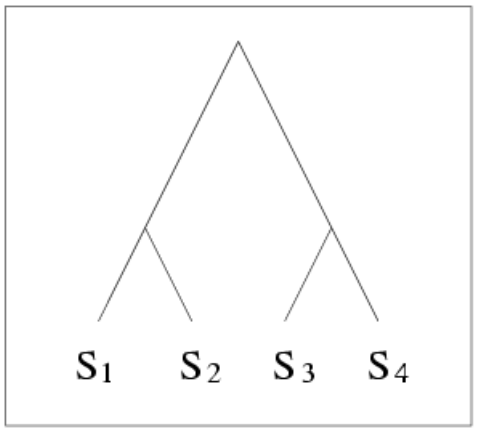
Distance-based phylogeny estimation

CS 598 AGB

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

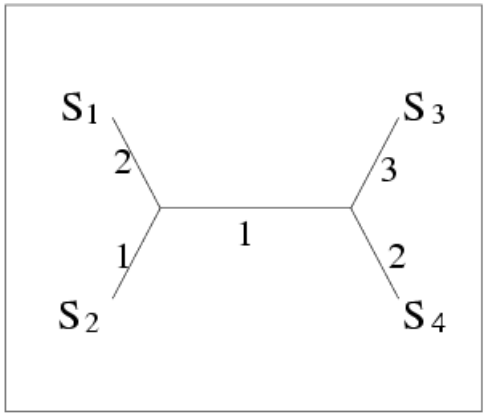
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Today's Class

- Phylogeny as statistical estimation problem
- Stochastic models of evolution
- Distance-based estimation
- Statistical consistency of distance-based methods

Note overlap with material from January 21, 2016

Simplest model of binary character evolution: **Cavender-Farris**

- For each edge **e**, there is a probability **p(e)** of the property "changing state" (going from 0 to 1, or vice-versa), with $0<p(e)<0.5$ (to ensure that CF trees are identifiable).
- Every position evolves under the same process, independently of the others.

CF tree estimation

- Instead of directly estimating the tree, we try to estimate the process itself.
- For example, we try to estimate the probability that two leaves will have different states for a random character.

Cavender-Farris pattern probabilities

- Let x and y denote nodes in the tree, and p_{xy} denote the probability that x and y exhibit different states.
- In other words,

$$p_{xy} = \Pr[(x=0 \ \& \ y=1) \text{ OR } (x=1 \ \& \ y=0)]$$

Cavender-Farris pattern probabilities

- Theorem: Let p_i be the substitution probability for edge e_i, and let x and y be connected by path $e_1e_2e_3\ldots e_k$. Then

  $1-2p_{xy} = (1-2p_1)(1-2p_2)\ldots(1-2p_k)$

- Proof: by induction on k.

And then take logarithms

- The theorem gave us:

$$1-2p_{xy} = (1-2p_1)(1-2p_2)\ldots(1-2p_k)$$

- If we take logarithms, we obtain

$$\ln(1-2p_{xy}) = \ln(1-2p_1) + \ln(1-2p_2)+\ldots+\ln(1-2p_k)$$

- Since $0<p_i<0.5$, $0<1-2p_i<1$, and $\ln(1-2p_i) < 0$.
- Multiply by -1 to get positive numbers.

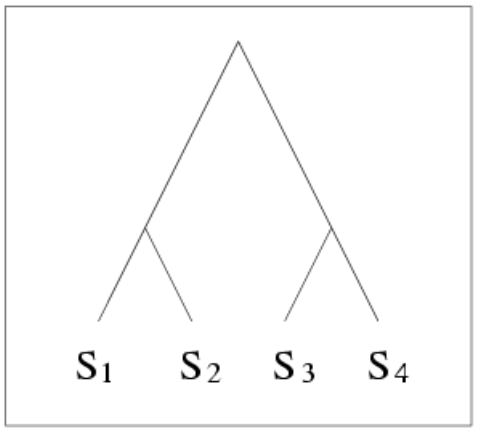
Additive Distance Matrix

- Consider $D_{xy} = -\ln(1-2p_{xy})$
- Let $w(e_i) = -\ln(1-2p_i)$. Note that $w(e_i) > 0$.
- We have shown $D_{xy} = w(e_1) + w(e_2)+\ldots+w(e_k)$, where $e_1,e_2,\ldots,e_k$ is the path in the true tree between leaves x and y.

In other words, the distance matrix D is **additive on the true tree T.** Furthermore, given D we can reconstruct the tree T and its edge weights.

We write this as D->(T,w).

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

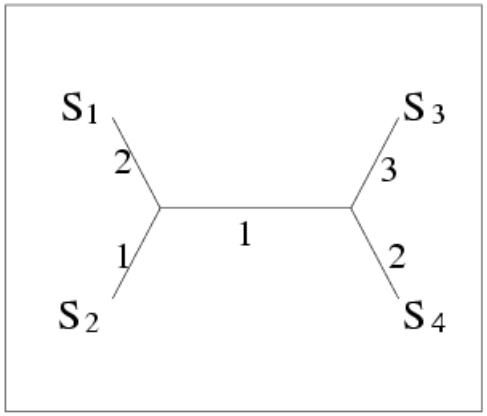
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

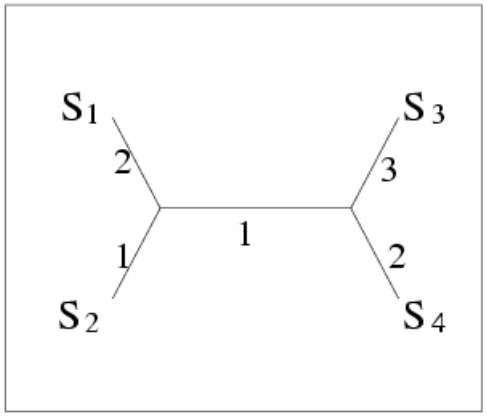
DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Additive distance matrix



INFERRED TREE

METHODS SUCH AS NEIGHBOR JOINING

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

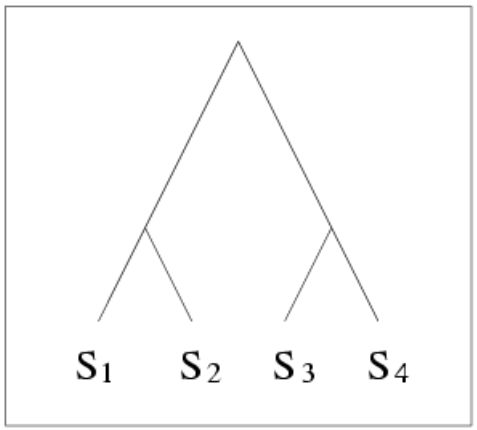
DISTANCE MATRIX

What now?

- The matrix D defined by $D_{ij} = -\ln(1-2p_{ij})$ is additive, and if we have the matrix, we can compute the true tree in polynomial time.
- But we don't know these p_{ij} values.
- Question: can we estimate p_{ij} in a statistically consistent manner?

Recall how to estimate the probability of a head…

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

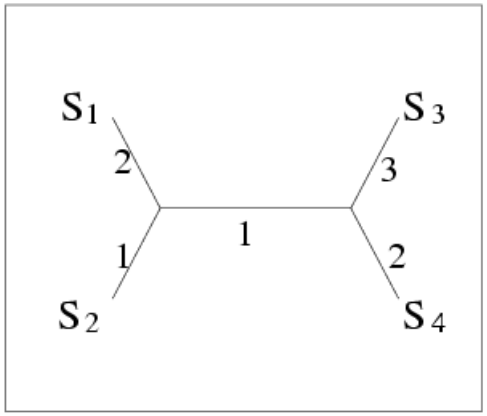
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Estimating CF distances

Consider H_{ij}/k, where k is the number of characters, and H_{ij} is the Hamming distance between sequences s_i and s_j.

Theorem: as k increases, $H_{ij}/k \rightarrow p_{ij}$

Estimating CF distances

Consider H_{ij}/k, where k is the number of characters, and H_{ij} is the Hamming distance between sequences s_i and s_j.

Theorem: as k increases, $H_{ij}/k \rightarrow p_{ij}$

And therefore as k increases,

$$d_{ij} = -\ln(1-2H_{ij}/k) \rightarrow -\ln(1-2p_{ij}) = D_{ij}$$

Estimating CF distances

Consider H_{ij}/k, where k is the number of characters, and H_{ij} is the Hamming distance between sequences s_i and s_j.

Theorem: as k increases, $H_{ij}/k \rightarrow p_{ij}$

And therefore as k increases,

$$d_{ij} = -\ln(1-2H_{ij}/k) \rightarrow -\ln(1-2p_{ij}) = D_{ij}$$

Therefore, the matrix d of estimated distances converges to the matrix D of model distances, which is additive on the true tree T.

Distance-based CF tree estimation

- Step 1: Compute Hamming distances
- Step 2: Correct the Hamming distances, using the CF distance calculation
- Step 3: Use distance-based method (neighbor joining, naïve quartet method, etc.)

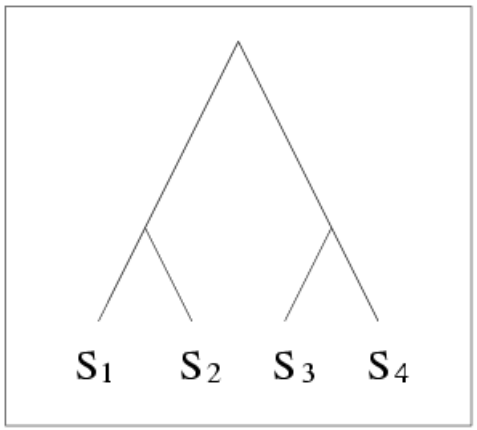
In this scenario, the only thing we are trying to estimate is the tree T, and not necessarily the lengths of the edges.

Distance-based CF tree estimation

- Step 1: Compute Hamming distances
- Step 2: Correct the Hamming distances, using the CF distance calculation
- Step 3: Use distance-based method (neighbor joining, naïve quartet method, etc.)

In this scenario, the only thing we are trying to estimate is the tree T, and not necessarily the lengths of the edges.

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

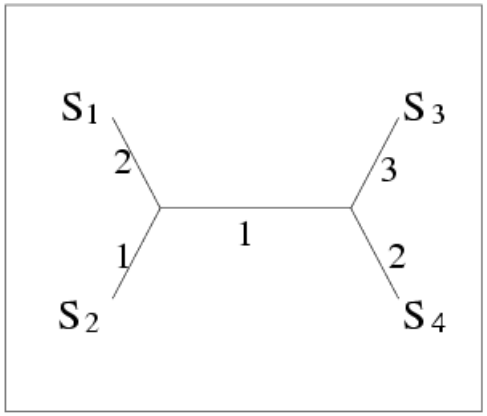
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Statistical consistency

To prove that a method M for estimating CF model tree topologies is statistically consistent, we need to show that

for all model CF trees (T,w) and for all $\varepsilon>0$,

there is some sequence length L such that

$Pr[M(S)=T]>1-\varepsilon$, given sequences S of length L generated on (T,w).

Statistical consistency

- In other words, it's not enough to show that the method M returns T given additive distance matrix for T.

Statistical consistency of distance-based methods

- Let D be defined by $D_{ij}=-\ln(1-2p_{ij})$. Recall that D is additive on T, and that $d_{ij}->D_{ij}$ for all i,j, (specifically, d converges to D in probability).
- Suppose there exists $\delta > 0$ such that

  $M(d)=M(D)=T$ whenever $\max_{ij}|d_{ij}-D_{ij}| < \delta$
- Then we will have a proof of statistical consistency for M!

This is called “error tolerance”

Four Point Method

- Input: Given 4x4 dissimilarity matrix A and four leaves i,j,k,l
- Output: a tree on i,j,k,l
- Step 1: Compute the three pairwise sums,
- Step 2: Return quartet tree ij|kl if

$$A_{ij}+A_{kl} < \min\{A_{ik}+A_{jl}, A_{il}+A_{jk}\}$$

Error tolerance for FPM

Theorem: Suppose D->(T,w) where T is the true tree, and $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$. Then the Four Point Method will return the correct tree on any four leaves in T.

Error tolerance for FPM

Theorem: Suppose D->(T,w) where T is the true tree, and $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$. Then the Four Point Method will return the correct tree on any four leaves in T.

Proof: Let i,j,k,l be any four leaves and suppose ij|kl is the true tree in T.

Error tolerance for FPM

Theorem: Suppose D->(T,w) where T is the true tree, and $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$. Then the Four Point Method will return the correct tree on any four leaves in T.

Proof: Let i,j,k,l be any four leaves and suppose ij|kl is the true tree in T.

Hence $D_{ij}+D_{kl} <= \min\{D_{ik}+D_{jl}, D_{il}+D_{jk}\}-2f$.

Error tolerance for FPM

Theorem: Suppose D->(T,w) where T is the true tree, and $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$. Then the Four Point Method will return the correct tree on any four leaves in T.

Proof: Let i,j,k,l be any four leaves and suppose ij|kl is the true tree in T.

Hence $D_{ij}+D_{kl} <= \min\{D_{ik}+D_{jl}, D_{il}+D_{jk}\}-2f$.

Suppose $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$.

Error tolerance for FPM

Theorem: Suppose D->(T,w) where T is the true tree, and $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$. Then the Four Point Method will return the correct tree on any four leaves in T.

Proof: Let i,j,k,l be any four leaves and suppose ij|kl is the true tree in T.

Hence $D_{ij}+D_{kl} <= \min\{D_{ik}+D_{jl}, D_{il}+D_{jk}\}-2f$.

Suppose $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$.

Then $d_{ij}+d_{kl} < \min\{d_{ik}+d_{jl}, d_{il}+d_{jk}\}$.

Error tolerance for FPM

Theorem: Suppose D->(T,w) where T is the true tree, and $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$. Then the Four Point Method will return the correct tree on any four leaves in T.

Proof: Let i,j,k,l be any four leaves and suppose ij|kl is the true tree in T.

Hence $D_{ij}+D_{kl} <= \min\{D_{ik}+D_{jl}, D_{il}+D_{jk}\}-2f$.

Suppose $\max_{ij}|d_{ij}-D_{ij}| < f/2$ where $f=\min_e w(e)$.

Then $d_{ij}+d_{kl} < \min\{d_{ik}+d_{jl}, d_{il}+d_{jk}\}$.

Hence, the FPM will return ij|kl given matrix d.

Naïve Quartet Method

- Compute the tree on each quartet using the four-point condition
- Merge them into a tree on the entire set if they are compatible using the All Quartets Method:
  - Find a sibling pair A,B
  - Recurse on S-{A}
  - If S-{A} has a tree T, insert A into T by making A a sibling to B, and return the tree

Error tolerance for NQM

- Suppose every pairwise distance is estimated well enough (within f/2, for f the minimum length of any edge).
- Then the Four Point Method returns the correct tree on every quartet.
- And so all quartet trees are compatible, and NQM returns the true tree.

In other words:

- The NQM method is statistically consistent methods for estimating Cavender-Farris trees!

- Plus it is polynomial time!

DNA substitution models

- Every edge has a substitution probability
- The model also allows 4x4 substitution matrices on the edges:
  - Simplest model: Jukes-Cantor (JC) assumes that all substitutions are equiprobable
  - General Time Reversible (GTR) Model: one 4x4 substitution matrix for all edges
  - General Markov (GM) model: different 4x4 matrices allowed on each edge

Jukes-Cantor DNA model

- Character states are A,C,T,G (nucleotides).
- All substitutions have equal probability.
- On each edge **e**, there is a value **p(e)** indicating the probability of change from one nucleotide to another on the edge, with $0<p(e)<0.75$ (to ensure that JC trees are identifiable).
- The state (nucleotide) at the root is random (all nucleotides occur with equal probability).
- All the positions in the sequence evolve identically and independently.

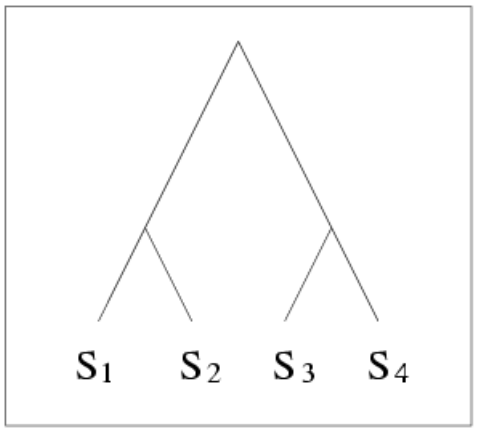
Jukes-Cantor distances

- D_{ij} = -3/4 ln(1-4/3 H(i,j)/k)) where k is the sequence length
- These distances converge to an additive matrix, just like with
  Cavender-Farris distances
- Hence, the NQM is statistically consistent under the JC model.

The NQM is statistically consistent under other models

- Note that the proof that the NQM is statistically consistent depends on d_{ij}->D_{ij} (i.e., we can estimate the model distances in a statistically consistent manner).
- This property is true for these other stochastic models of evolution (but not for arbitrary models).

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

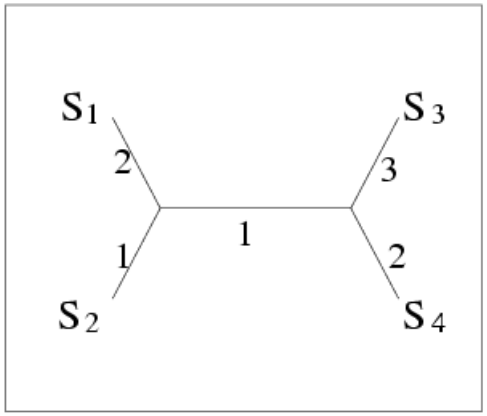
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Other distance-based methods

- Neighbor Joining (Saitou and Nei):
  - Kevin Atteson proved NJ has error tolerance of f/2
- 3-approximation to L-infinity nearest tree (Agarwala et al.)
  - Easy to see has error tolerance of f/8
  - Exact algorithm has error tolerance of f/4
- FastME, BioNJ, Weighbor, NINJA, FastTree, and many others (most statistically consistent but not analyzed for error tolerance)
- UPGMA – not consistent

UPGMA

While |S|>2:

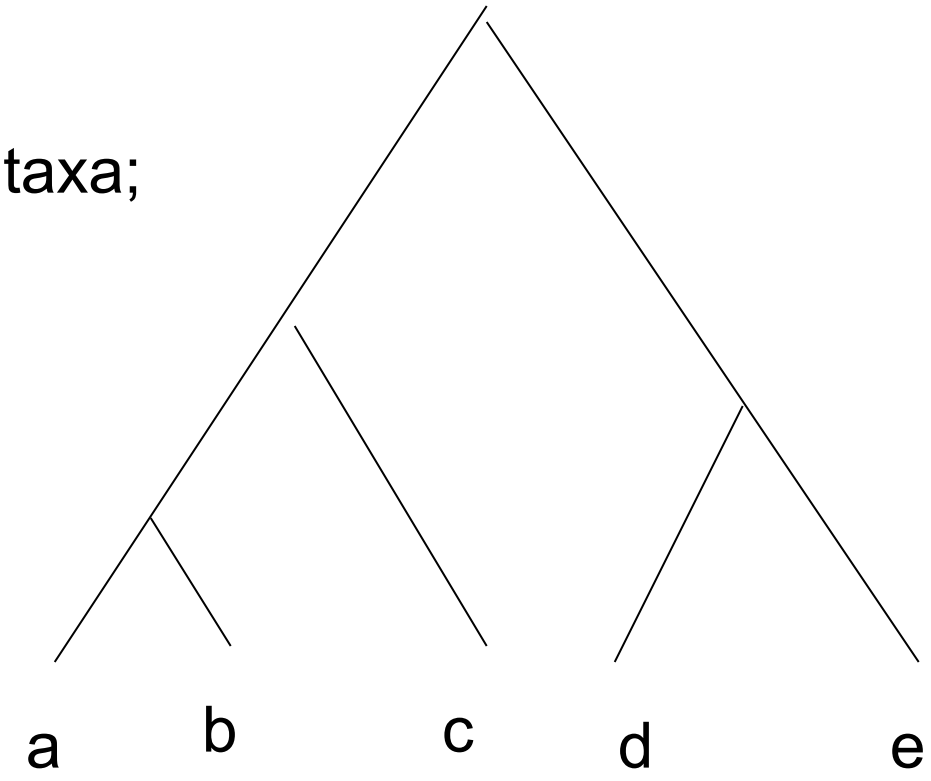
find pair x,y of closest taxa;

delete x

Recurse on S-{x}

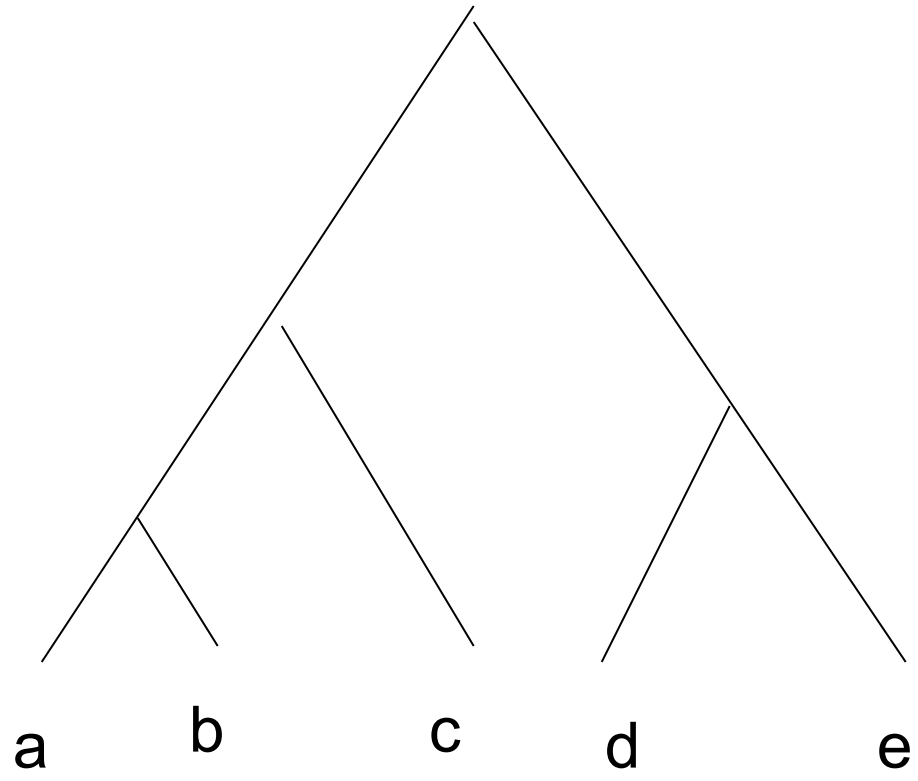
Insert y as sibling to x

Return tree



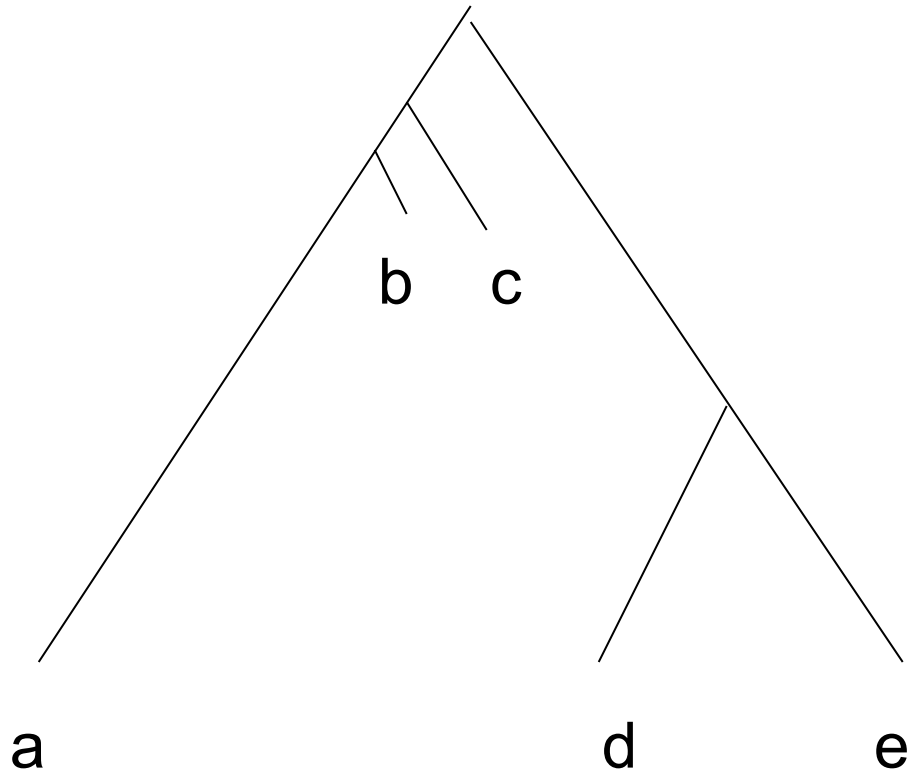
UPGMA

Works when evolution is "clocklike"



UPGMA

Fails to produce true tree if evolution deviates too much from a clock!



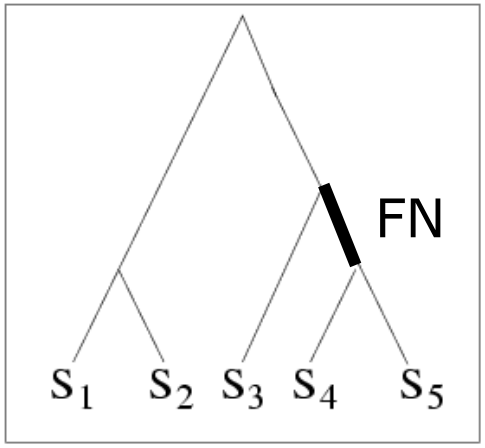
Summary (so far)

- Distance-based phylogeny estimation is popular because it is computationally fast.
- Many distance-based methods are statistically consistent under standard stochastic models. The proofs require
  - (a) that an additive distance matrix for the true tree can be estimated in a statistically consistent manner, and
  - (b) the method M return the true tree T given an additive matrix for (T,w), for any positive edge weighting w, and
  - (c) the method M has error tolerance.
- The ability to estimate additive distances for the true tree depends on the stochastic model, and does not apply to all models.
- The ability to compute the tree T from a noisy distance matrix depends on the method M, and does not apply to all methods.

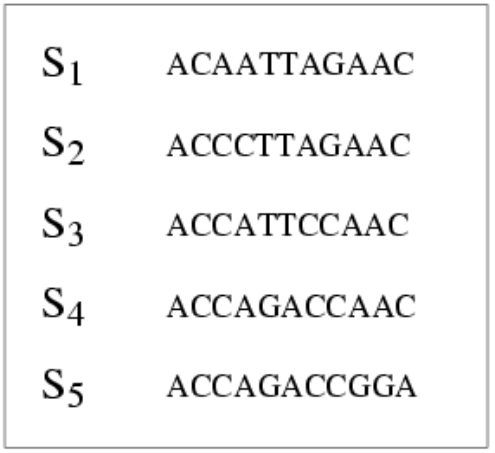
Other statistically consistent methods

- Maximum Likelihood
- Bayesian MCMC methods
- Absolute fast converging methods, such as the Dyadic Closure Method (Erdos et al., 1997) and DCM_{NJ} (Warnow et al. 2001).

Quantifying Error



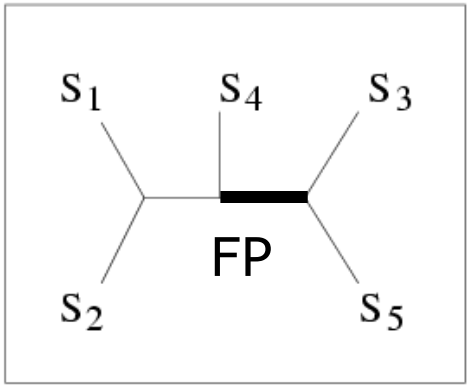
TRUE TREE



DNA SEQUENCES

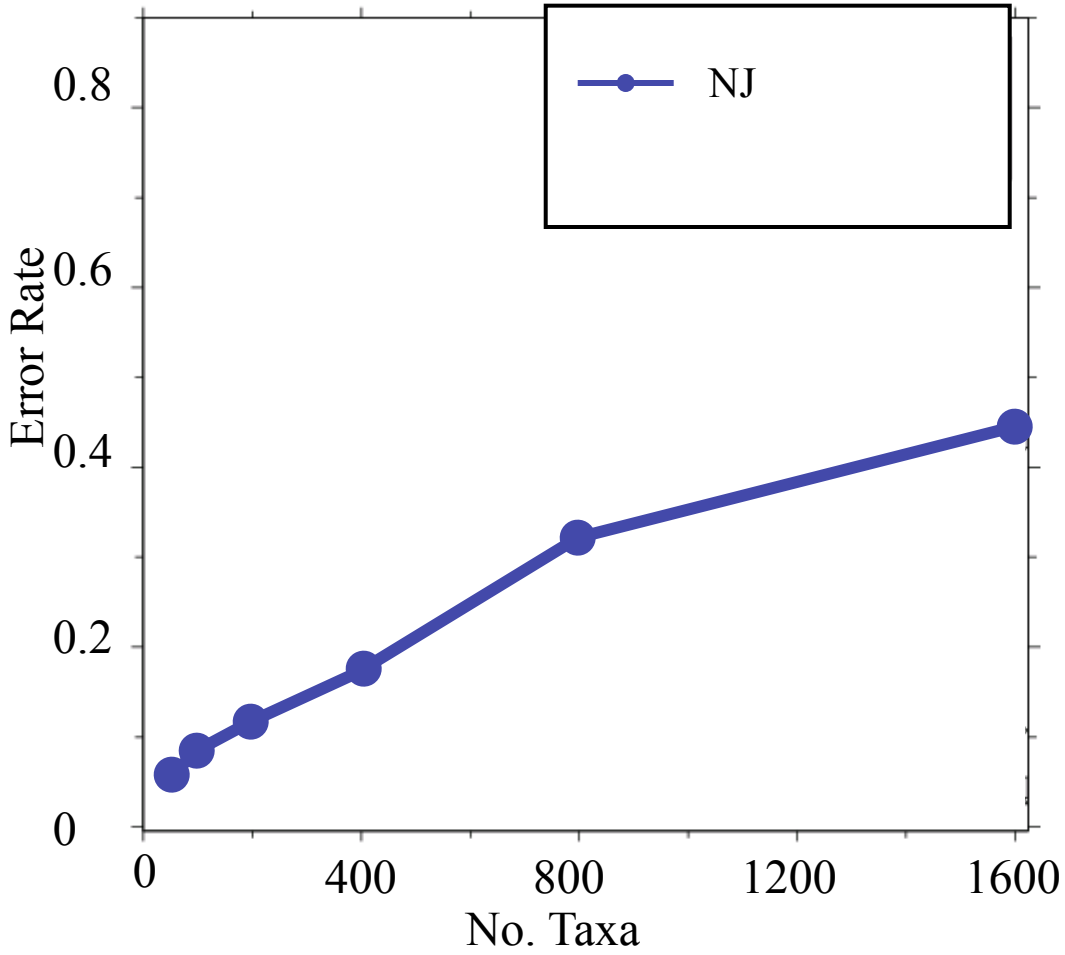
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Neighbor joining has poor performance on large diameter trees *[Nakhleh et al. ISMB 2001]*

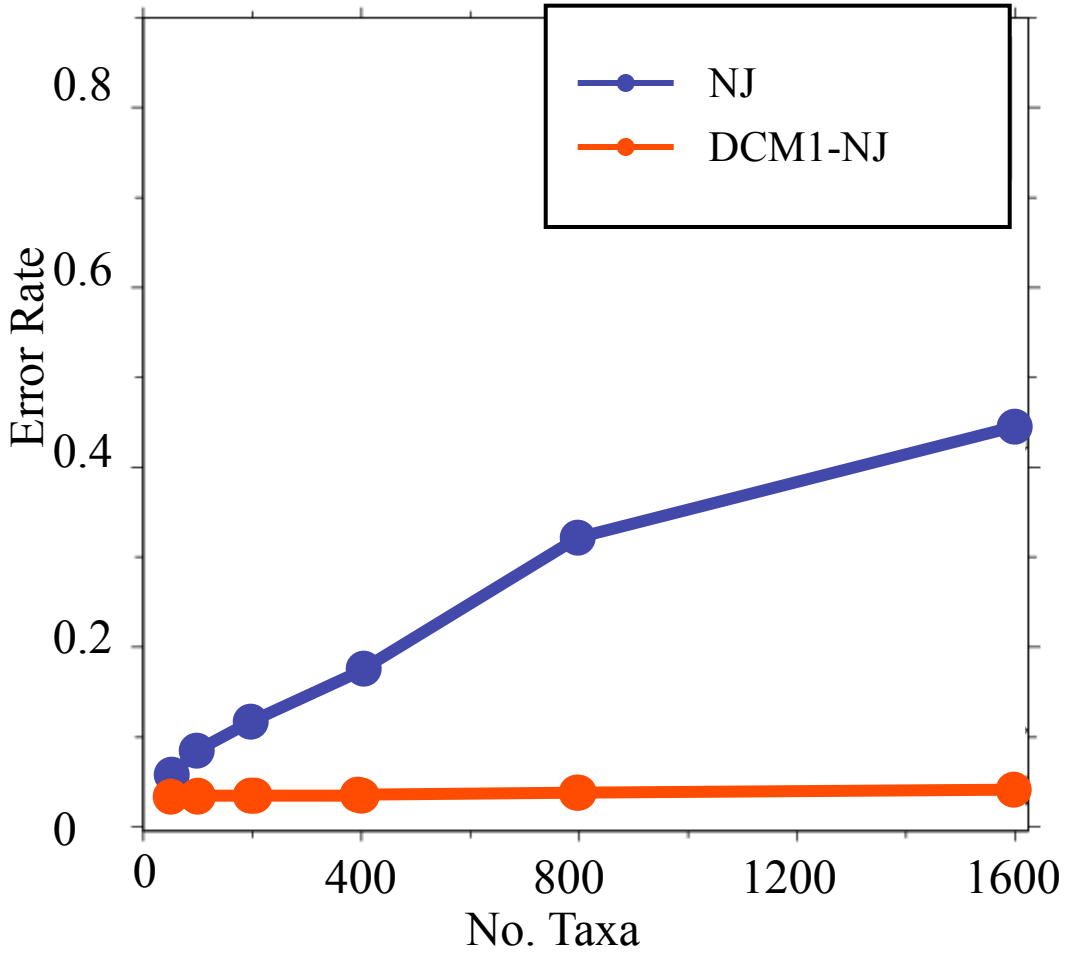


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



Theorem (Warnow et al., SODA 2001): DCM1-NJ converges to the true tree from polynomial length sequences

Advanced material

- Absolute Fast Converging Methods
- Will be covered at the end of the course