CS 581

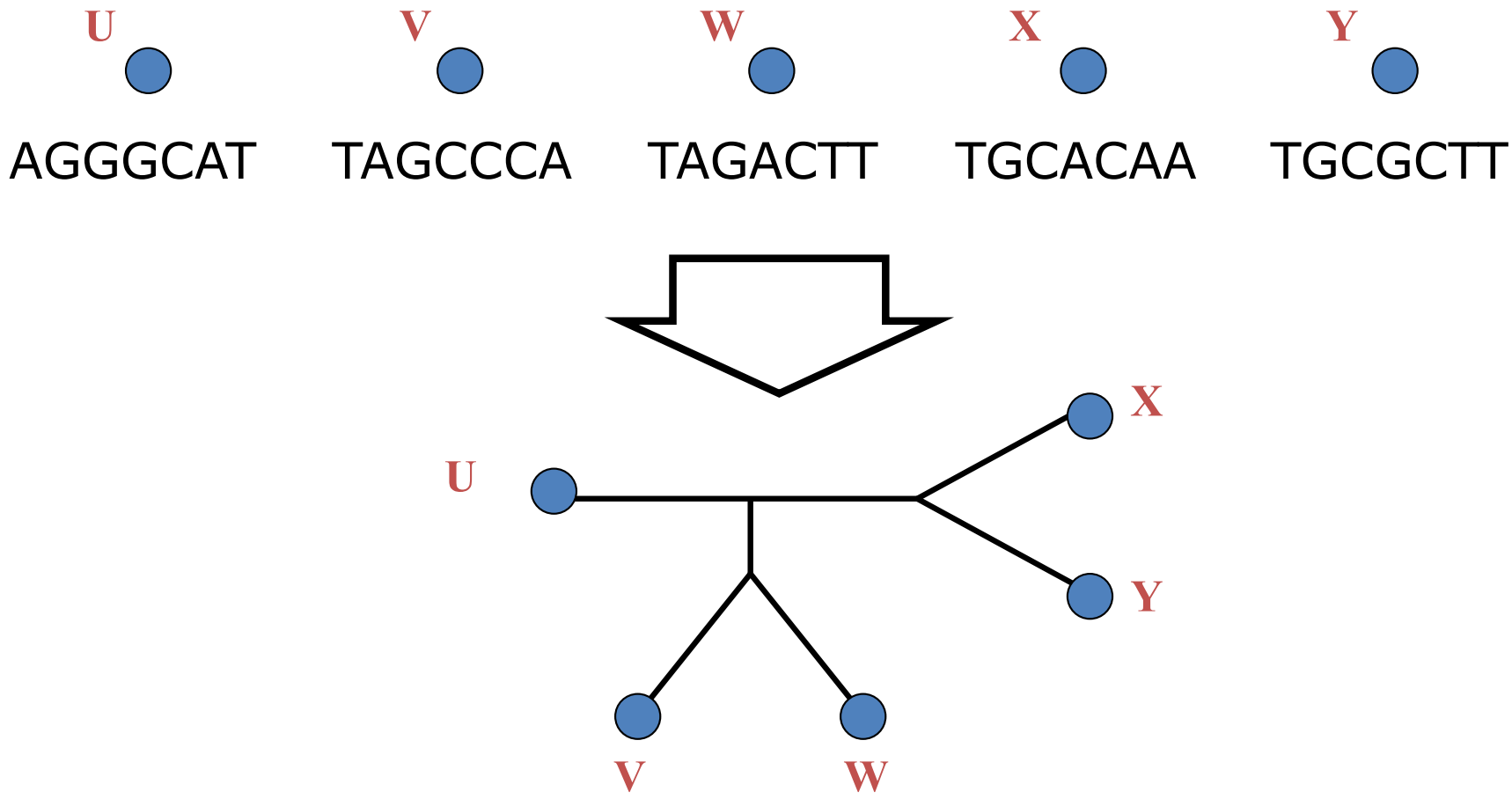
Tandy Warnow

Today

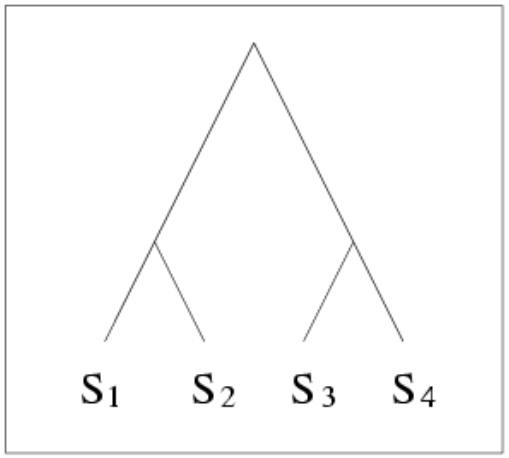
- Additive matrices
- The Four Point Condition
- The Four Point Method
- The Naïve Quartet Method
- The Cavender-Farris-Neyman model
- Estimating Cavender-Farris-Neyman model trees
- Estimating Jukes-Cantor model trees
- More complicated DNA sequence evolution models

See textbook Chapter 1, 8.1-8.2

Phylogeny Problem



Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

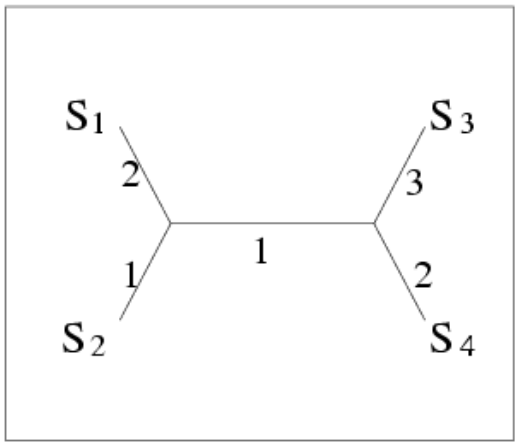
DNA SEQUENCES

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 Matrices

- A square matrix $D=[d_{ij}]$ is additive if and only if there is a tree T and edge-weighting w such that for all pairs i,j of leaves, d_{ij} is the path distance in T between i and j.

- We note this by saying D corresponds to (T,w).

Four Point Condition

- Theorem: Let D = $[d_{ij}]$ be an additive matrix. Then, for every four indices i,j,k,l, the median and maximum of the three pairwise sums are the same:

  $d_{ij} + d_{kl}$

  $d_{ik} + d_{jl}$

  $d_{il} + d_{jk}$

Proof of the Four Point Condition

Using the Four Point Condition

- Given a 4x4 additive matrix D, can you find the tree T (and edge-weighting w) that corresponds to D?

Using the Four Point Condition

- How would you construct a tree on a set of n>4 leaves, if you had an additive matrix?

Four Point Method

- Task: Given 4x4 dissimilarity matrix, compute a tree on four leaves
- Solution: Compute the three pairwise sums, and take the split ij|kl that gives the minimum!

Does this work? Why?

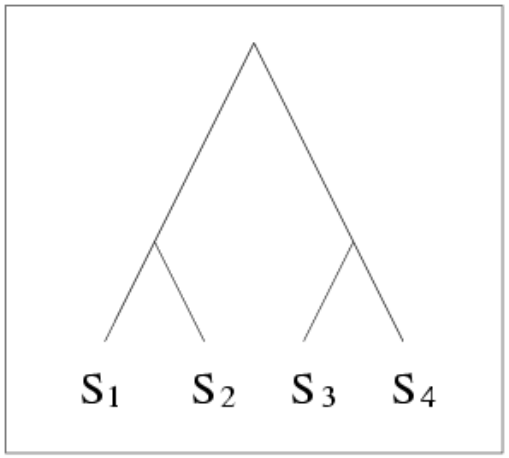
Naïve Quartet Method

- Compute the tree on each quartet using the four-point method
- Merge them into a tree on the entire set if they are compatible:
  - 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

Naïve Quartet Method, cont.

- Theorem: Let $D=[d_{ij}]$ be an additive matrix corresponding to an edge-weighted tree (T,w). Then the Naïve Quartet Method applied to D returns T.
- Proof: all estimated quartet trees are correct (by the Four Point Condition), and an induction proof shows the Naïve Quartet Method returns T.

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

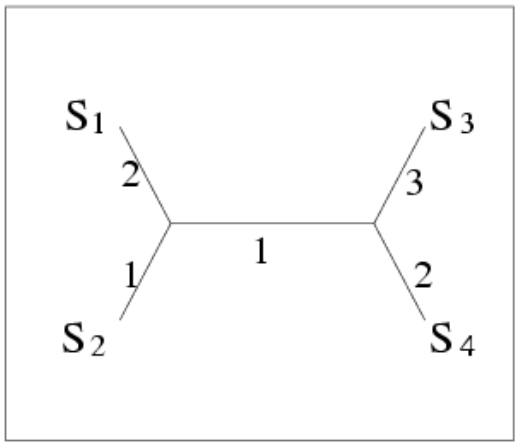
DNA SEQUENCES

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

Dissimilarity Matrices

- A square matrix that is symmetric and zero on the diagonal is called a **dissimilarity matrix.**
- A dissimilarity matrix may not satisfy the triangle inequality.

In phylogenetics, the distance matrices we calculate are dissimilarity matrices.

Can we construct a tree from a dissimilarity matrix?

Error tolerance for NQM

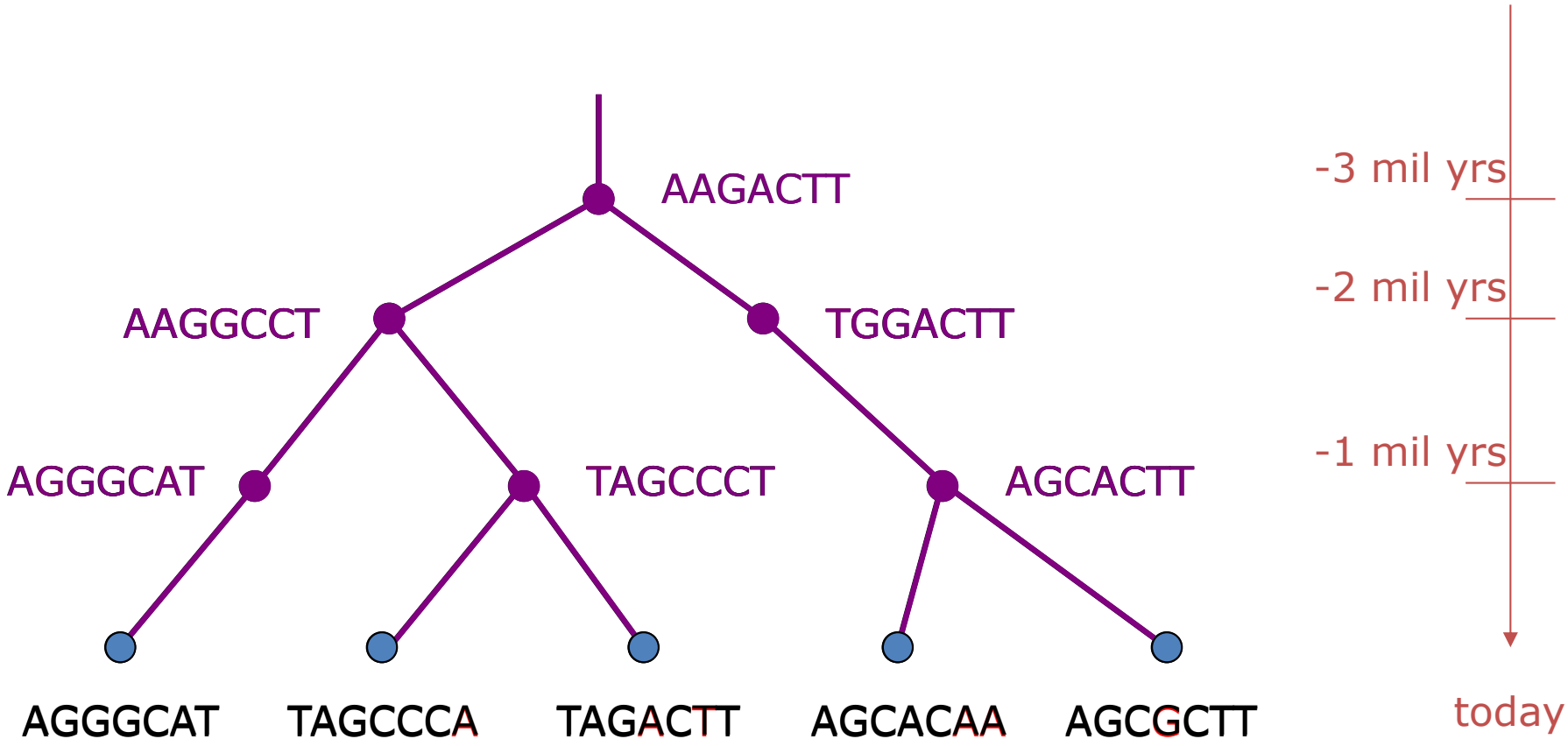
- Suppose every pairwise distance is estimated well enough (within f/2, for f the minimum length of any edge).
- Then the Four Point Method returns the correct tree on every quartet.
- And so all quartet trees are compatible, and NQM returns the true tree.

Phylogeny estimation as a statistical inverse problem

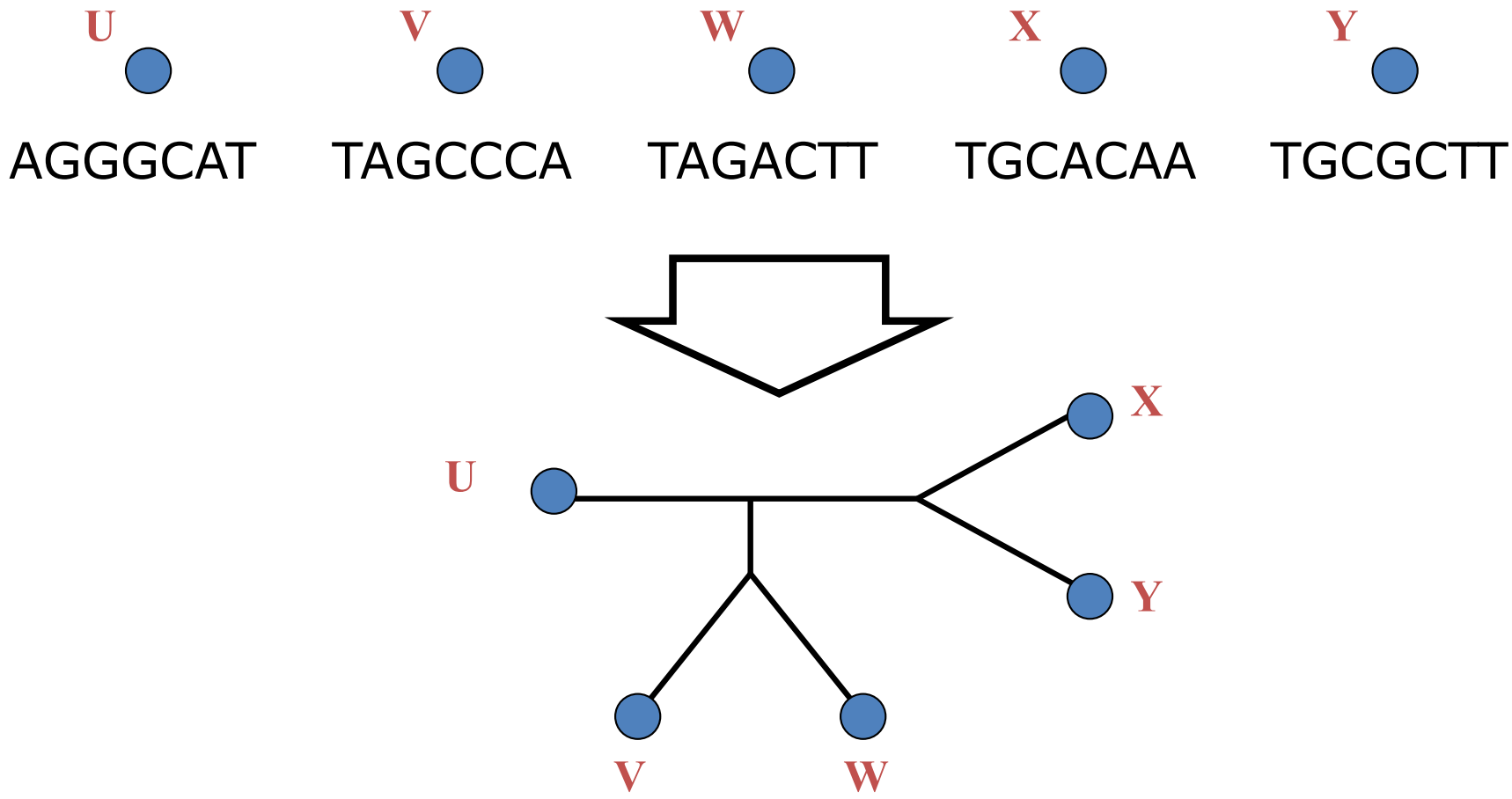
Estimation of evolutionary trees as a statistical inverse problem

- We can consider characters as properties that evolve down trees.
- We observe the character states at the leaves, but the internal nodes of the tree also have states.
- The challenge is to estimate the tree from the properties of the taxa at the leaves. This is enabled by characterizing the evolutionary process as accurately as we can.

DNA Sequence Evolution



Phylogeny Problem



Jukes-Cantor (1969) Model

- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Time Reversible model, or the General Markov model) are also considered, often with little change to the theory.

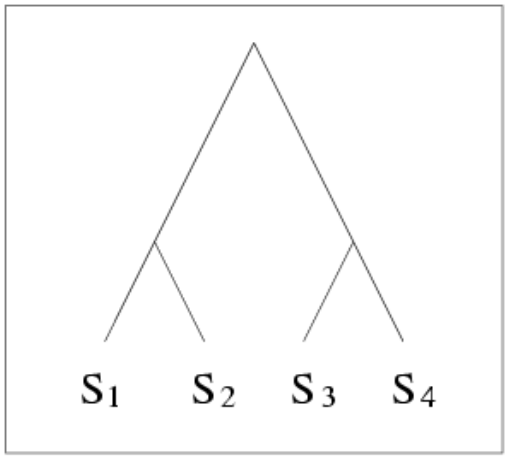
Questions about model trees

- Is the model tree topology identifiable?
- Are the branch lengths and other numeric parameters of the model tree identifiable?
- Is the root of the model tree identifiable?

Answers about model trees

- Is the model tree topology identifiable? – yes
- Are the branch lengths and other numeric parameters of the model tree identifiable? – yes
- Is the root of the model tree identifiable? – no

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

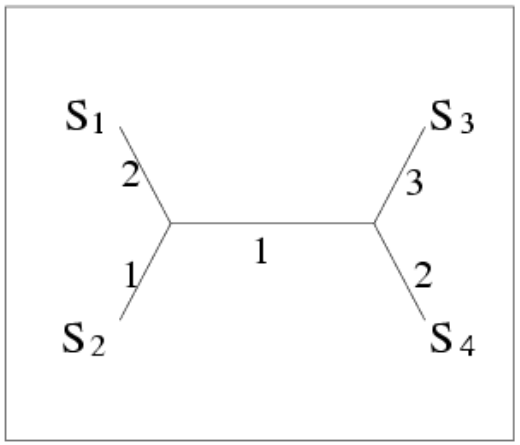
DNA SEQUENCES

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

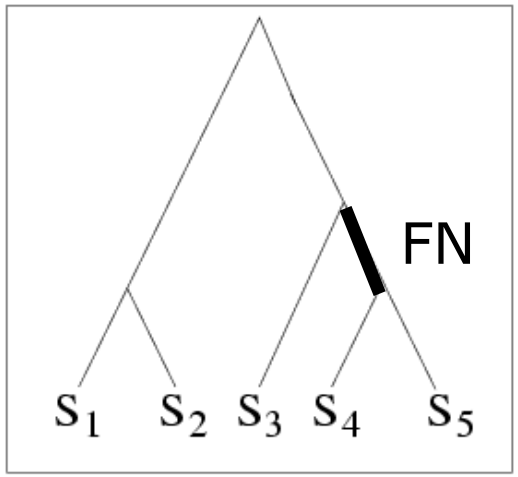
METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Performance criteria

- Running time
- Space
- Statistical performance issues (e.g., statistical consistency and sequence length requirements)
- “Topological accuracy” with respect to the underlying true tree, typically studied in simulation.
- Accuracy with respect to a mathematical score (e.g. tree length or likelihood score) on real data

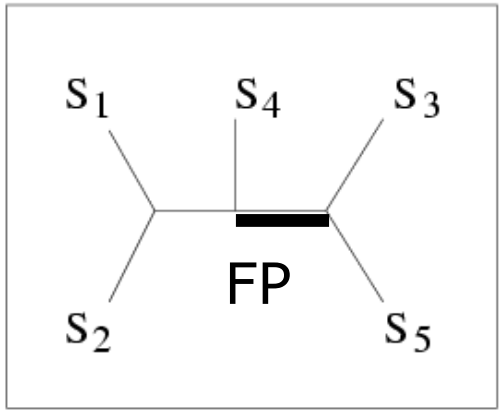


TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

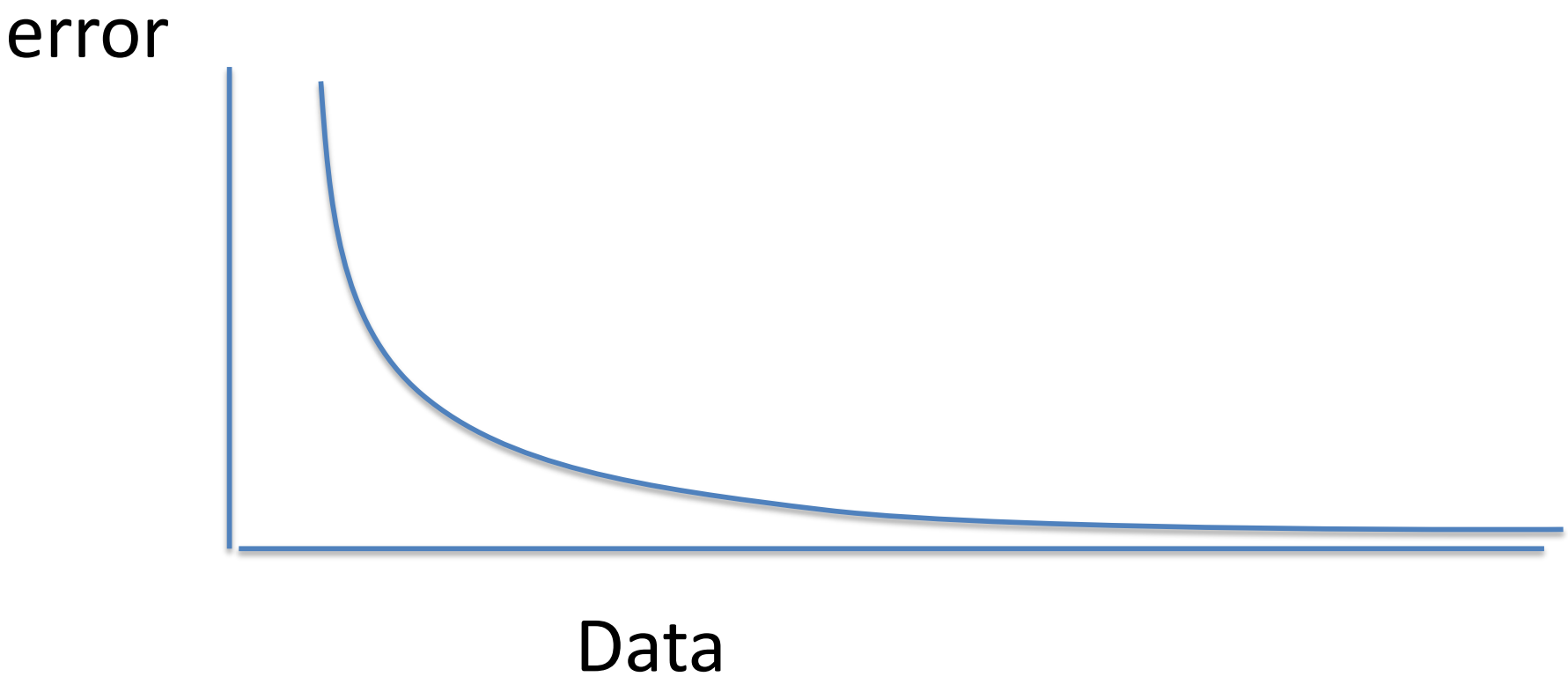
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)



INFERRED TREE

50% error rate

Statistical Consistency



Statistical models

- Simple example: coin tosses.
- Suppose your coin has probability p of turning up heads, and you want to estimate p. How do you do this?

Estimating p

- Toss coin repeatedly
- Let your estimate q be the fraction of the time you get a head

- Obvious observation: *q will approach p* as the number of coin tosses increases
- This algorithm is a *statistically consistent* estimator of p. That is, your error |q-p| goes to 0 (with high probability) as the number of coin tosses increases.

Another estimation problem

- Suppose your coin is biased either towards heads or tails (so that p is not 1/2).
- How do you determine which type of coin you have?

- Same algorithm, but say “heads” if q>1/2, and “tails” if q<1/2. For large enough number of coin tosses, *your answer will be correct with high probability.*

Phylogeny Estimation

- Simplest type of data: presence/absence of a property (e.g., has wings, has hair, has a particular amino acid)
- Treat this as binary character evolution, with 0 representing absence and 1 representing presence.
- How do we model the evolution of these binary characters?

Jukes-Cantor (1969) Model

- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Time Reversible model, or the General Markov model) are also considered, often with little change to the theory.

Cavender-Farris-Neyman (CFN)

- Models binary sequence evolution
- 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 unrooted CFN tree topologies are identifiable).
- Every position evolves under the same process, independently of the others.

Estimating trees under statistical models…

- 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.

CFN pattern probabilities

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

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

And then take logarithms

- The theorem gave us:
  $1-2p_{xy} = (1-2p_1)(1-2p_2)\ldots(1-2p_k)$

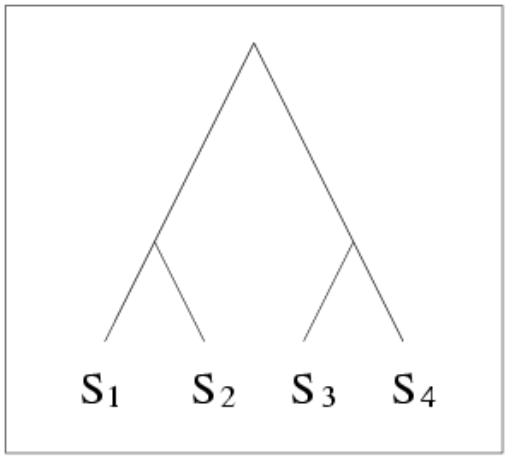
- If we take logarithms, we obtain
  $\ln(1-2p_{xy}) = \ln(1-2p_1) + \ln(1-2p_2)+\ldots+\ln(1-2p_k)$

- Since these probabilities lie between 0 and 0.5, these logarithms are all negative. So let's multiply by -1 to get positive numbers.

An additive matrix!

- Consider a matrix $D(x,y) = -\ln(1-2p_{xy})$
- This matrix is additive (i.e., fits a tree exactly)!
- Can we estimate this additive matrix from what we observe at the leaves of the tree?
- Key issue: how to estimate $p_{xy.}$
- (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

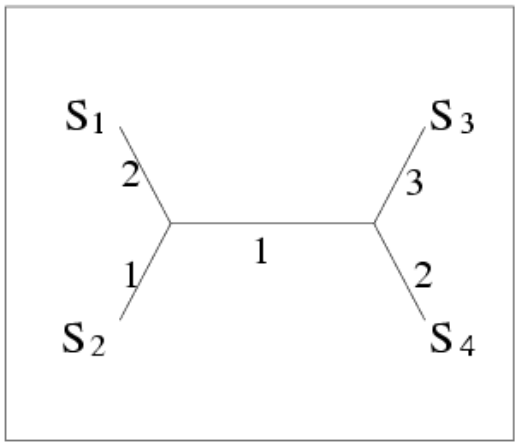
DNA SEQUENCES

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 CFN distances

- Consider

  d_{ij}= -1/2 ln(1-2H(i,j)/k),

  where k is the number of characters, and H(i,j) is the Hamming distance between s_i and s_j.

- Theorem: as k increases,

  d_{ij} converges to D_{ij} = -1/2 ln(1-2p_{ij}),

  which is an additive matrix.

Four Point Method (FPM)

- Task: Given 4x4 dissimilarity matrix, compute a tree on four leaves
- Solution: Compute the three pairwise sums, and take the split ij|kl that gives the minimum!
- When is this guaranteed accurate?

Error tolerance for FPM

- 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 (i.e., ij+kl remains the minimum)

Naïve Quartet Method (NQM)

- Compute the tree on each quartet using the four-point method
- Merge them into a tree on the entire set if they are compatible:
  - 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

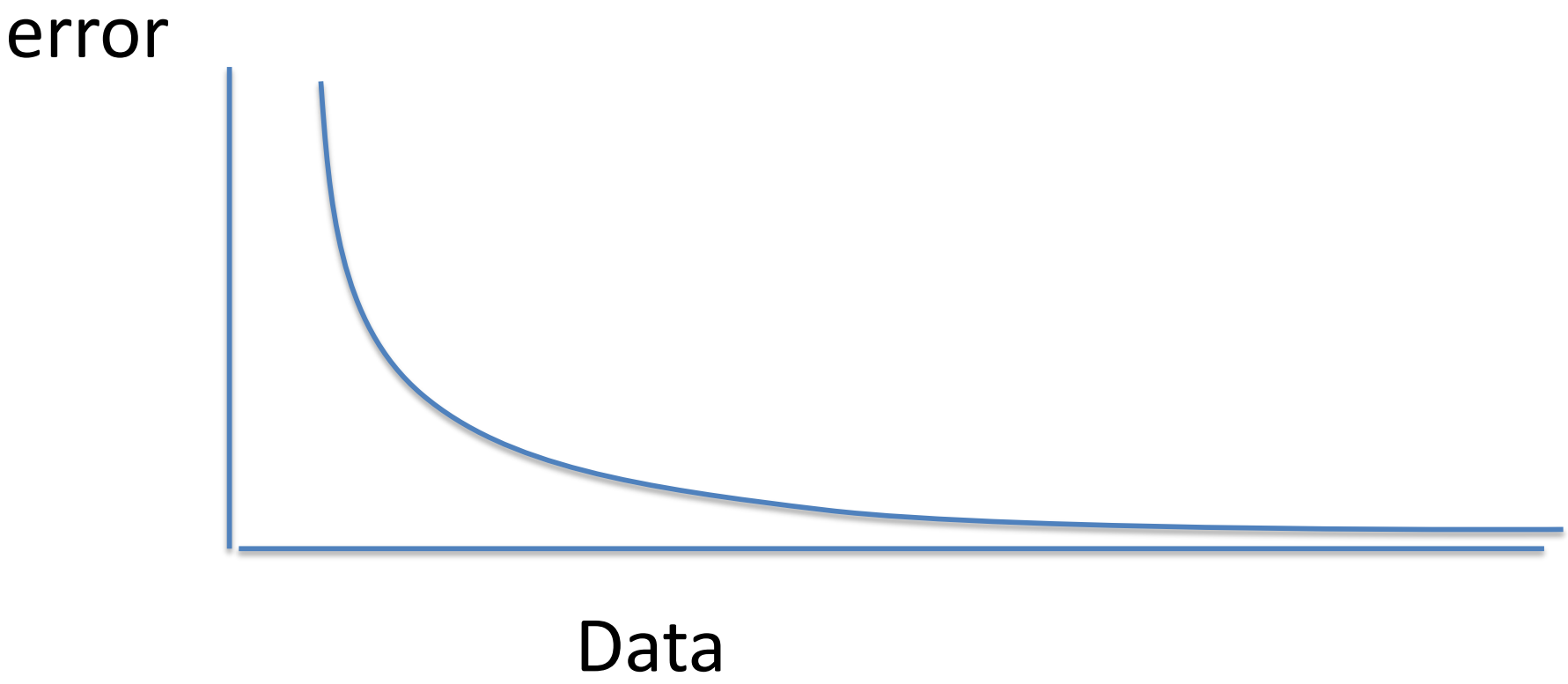
Error tolerance for NQM

- 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 CFN trees!
- Plus it is polynomial time!

Statistical Consistency



What about DNA sequence evolution?

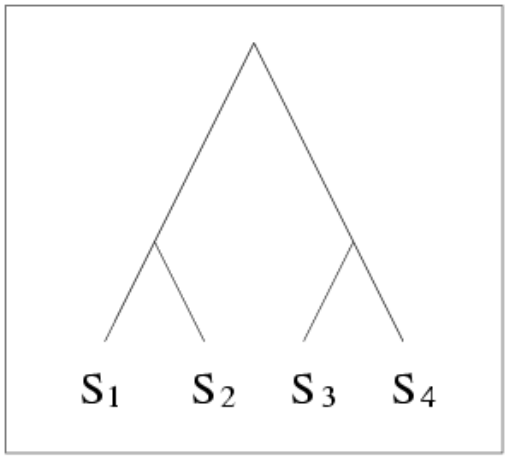
- The proof of statistical consistency for the NQM under the CFN model only really depended on the guarantee that CFN estimated distances converge, as the sequence length increase, to an additive matrix.
- What about DNA sequence evolution models?

Jukes-Cantor (1969) Model

- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Time Reversible model, or the General Markov model) are also considered, often with little change to the theory.

Distance-based Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

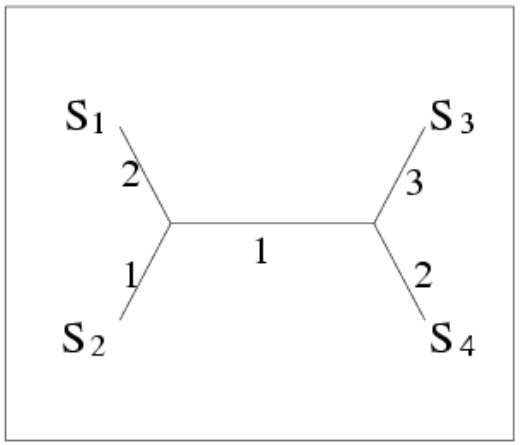
DNA SEQUENCES

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

Jukes-Cantor Tree Estimation

- Step 1: Compute Hamming distances
- Step 2: Correct the Hamming distances, using the JC distance calculation
- Step 3: Use NQM to construct the tree

In other words:

Theorem: The NQM method is statistically consistent methods for estimating JC trees, and uses polynomial time!

Notes:

- This is true for other models – all you need is a statistically consistent technique to estimate an additive matrix that corresponds to an edge-weighting of the model tree.
- This is also true for other distance-based methods (e.g., neighbor joining).

Standard DNA site evolution models

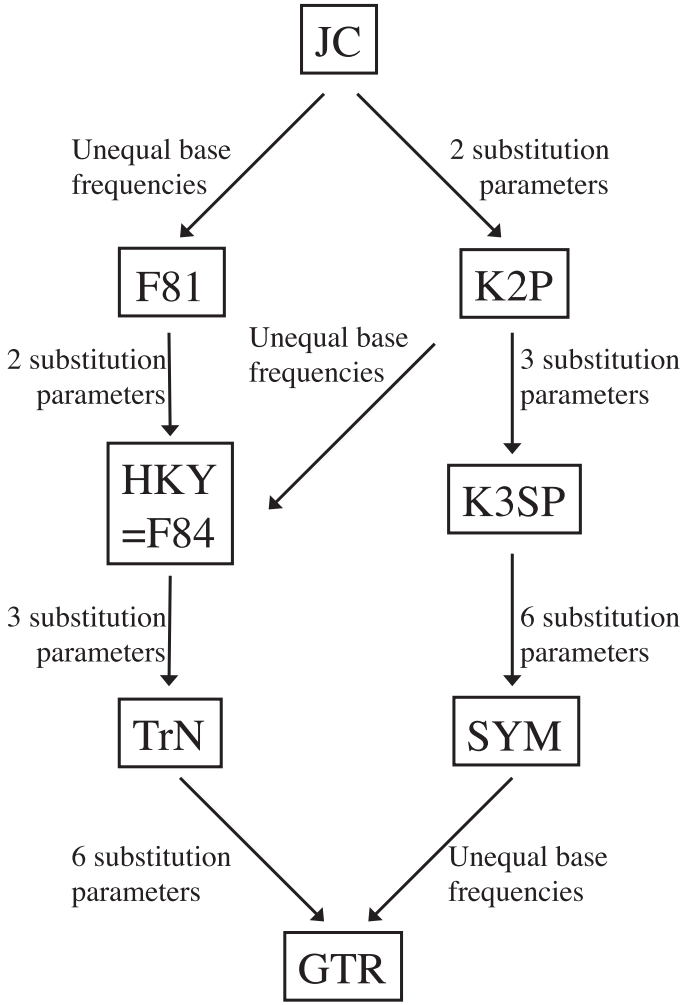


Figure 3.9 from Huson et al., 2010

Homework (due next week)

- Read Chapters 1, 2, and 5.1-5.5
- Homework problems:
  - Chapter 1, problems 1, 5, 10, and 11
  - Chapter 2, problems 1, 5, and 21
- The written homework is due in Moodle, and can be submitted ahead of time (and then revised and replaced as often as you like) before deadline.
- Note the penalty for late homework.
- You can collaborate on homework as long as you write it up yourself, and clearly indicate who you worked with.