CMP 464/788 Lecture Notes

Experimental Testing of Algorithms
Introduction to Computational Biology
30 October 2003

Testing Methods Empirically

- How accurate are the methods at reconstructing trees?
- In biological applications, the true, historical tree is almost never known, which makes assessing the quality of phylogenetic reconstruction methods problematic.
  (an exception: Hillis '92 created an evolutionary tree in the laboratory)
- Simulation is used instead to evaluate methods, given a model of evolution.

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

4. Evaluate the accuracy of the constructed tree.

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

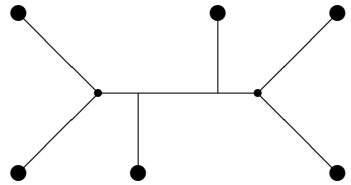
3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

4. Evaluate the accuracy of the constructed tree.

Simulating Data: Choosing Trees

- Usually chosen from a random distribution on trees: Uniform, or Yule-Harding (birth-death trees)



- Can view this as two different random processes:
  - generate the tree shape, and then
  - assign weights or branch lengths to the shape.

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

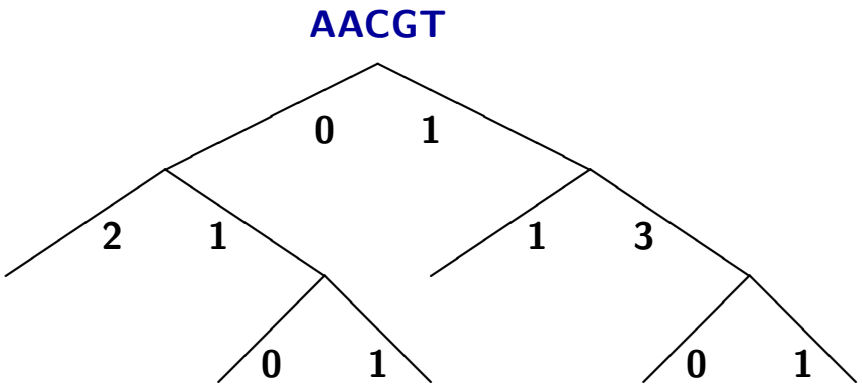
3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

4. Evaluate the accuracy of the constructed tree.

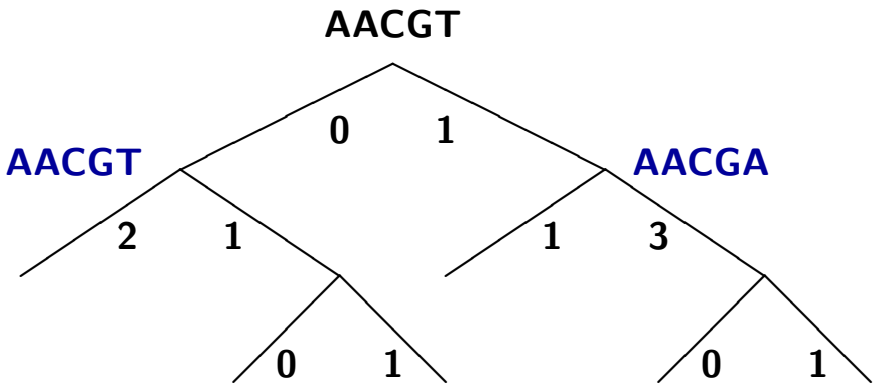
Simulating Data: Evolving Sequences

- The *Jukes-Cantor* (JC) model is the simplest Markov model of biomolecular sequence evolution.
- A DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a rooted binary tree T.



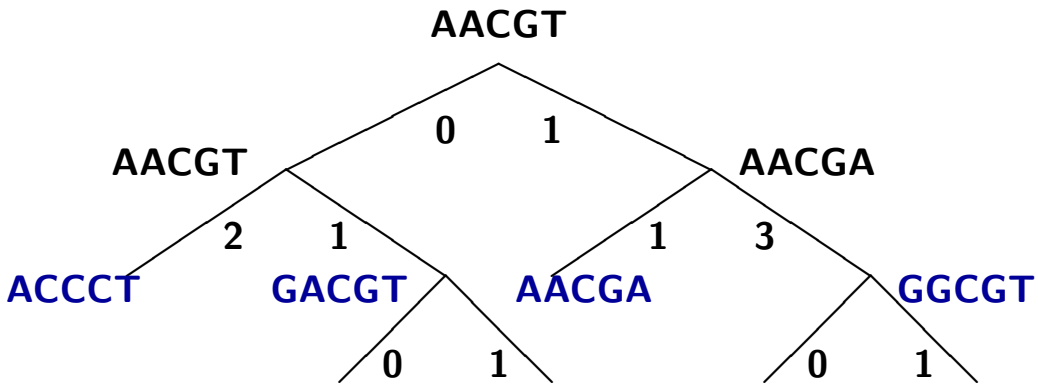
Simulating Data: Evolving Sequences

- The *Jukes-Cantor* (JC) model is the simplest Markov model of biomolecular sequence evolution.
- A DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a rooted binary tree T.



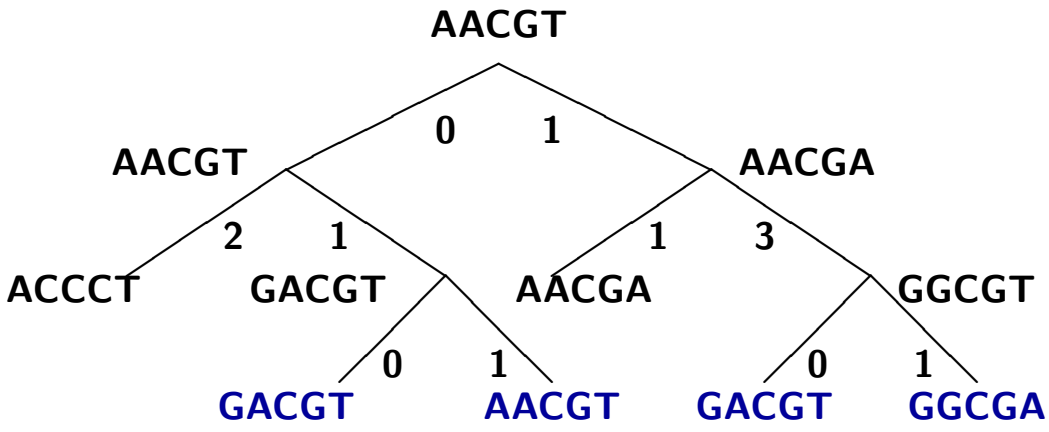
Simulating Data: Evolving Sequences

- The *Jukes-Cantor* (JC) model is the simplest Markov model of biomolecular sequence evolution.
- A DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a rooted binary tree T.



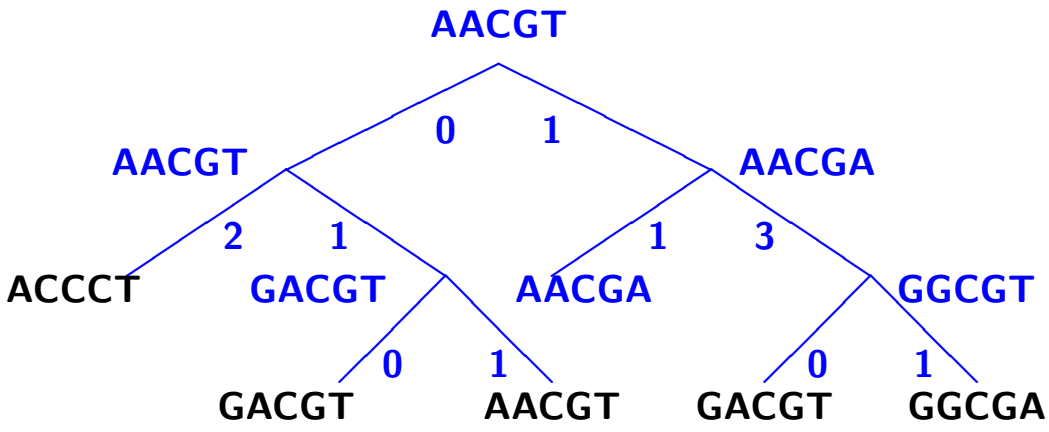
Simulating Data: Evolving Sequences

- The *Jukes-Cantor* (JC) model is the simplest Markov model of biomolecular sequence evolution.
- A DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a rooted binary tree T.



Simulating Data: Evolving Sequences

- The *Jukes-Cantor* (JC) model is the simplest Markov model of biomolecular sequence evolution.
- A DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a rooted binary tree T.



Simulating Data: Evolving Sequences

- The *Jukes-Cantor* (JC) model is the simplest Markov model of biomolecular sequence evolution.
- A DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a rooted binary tree T.
- The assumptions of the model are:
  1. the sites (i.e., the positions within the sequences) evolve independently and identically
  2. if a site changes state it changes with equal probability to each of the remaining states, and
  3. the number of changes of each site on an edge e is a Poisson random variable with expectation $\lambda(e)$ (this is also called the "length" of the edge e).

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

4. Evaluate the accuracy of the constructed tree.

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

3. Reconstruct the tree using method.

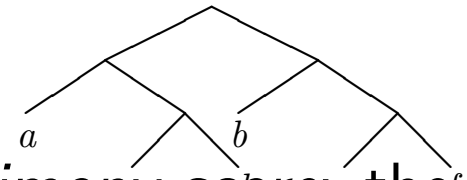
A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

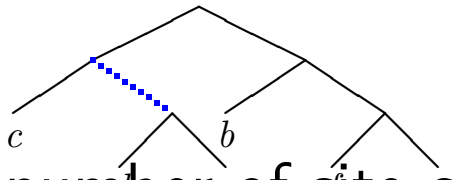
4. Evaluate the accuracy of the constructed tree.

Evaluating Accuracy

- To compare reconstructed tree to model tree, the *Robinson-Foulds Score* is often used:

$$\frac{\text{False Positives} + \text{False Negatives}}{\text{total edges}}$$

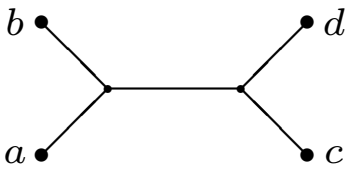




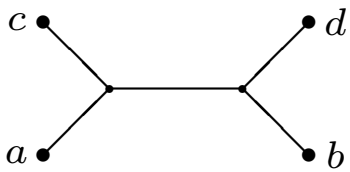
- best *parsimony score*: the sum of the number of site changes acrosss the edges in the tree.

Case Study: Quartet Methods

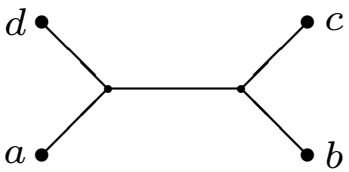
- A *quartet* is an unrooted binary tree on four taxa:



$\{ab|cd\}$



$\{ac|bd\}$



$\{ad|bc\}$

- Let $Q(T) =$ all quartets that agree with T.
  [Erdős *et al.* 1997]: T can be reconstructed from $Q(T)$ in polynomial time.

Case Study: Quartet Methods

- Quartet-based methods operate in two phases:
  - Construct quartets on all four taxa sets.
  - Combine these quartets into a tree.
- Running time:
  - For most optimizations, determining a quartet is fast.
  - There are $\Theta(n^4)$ quartets, giving $\Omega(n^4)$ running time.
  - In practice, the input quality is insufficient to ensure that all quartets are accurately inferred.
  - Quartet methods have to handle incorrect quartets.

Popular Quartet Methods

- Q^* or Buneman Method [Berry & Gascuel '97, Buneman '71]:
  Only add edges that agree with all input quartets.
  Doesn't tolerate errors– outputs conservative, but unresolved tree.
- Quartet Cleaning (QC) [Berry *et al.* 1999]: Add edges with a small number of errors proportional to q_e.
  Many variants: all handle a small number of errors.
- Quartet Puzzling [Strimmer & von Haeseler 1996]: "Order taxa randomly, greedily add edges, repeat 1000 times." Output majority tree.
  Most popular with biologists.

Standard Method: Neighbor Joining (NJ)

- [Saitou & Nei 1987]: very popular and fast: $O(n^3)$.
  - Based on the distance between nodes, join *neighboring leaves*, replace them by their parent, calculate distances to this node, and repeat.
  - This process eventually returns a binary (fully resolved) tree.
  - Joining the leaves with the minimal distance does not suffice, so subtract the averaged distances to compensate for long edges.
  - Experimental work shows that NJ trees are reasonably accurate, given a rate of evolution is neither too low nor too high.

Previous Experimental Studies

- Berry *et al.* [1999] studied various QC methods:
  - Showed that QC methods outperform the Q^*(didn't compare to any other methods)
  - By design, the QC methods recover all edges recovered by Q^*. Noteworthy that the QC methods *obtained* additional edges.
  - Varied evolutionary rates and sequence lengths, but studied only *10 taxa trees*.
  - The theoretical bounds we derive and experiments on larger n suggest that performance on very large n may be poor.

Our Study

- A detailed, large-scale experimental study of quartet methods and NJ under the Jukes-Cantor model of evolution:
  - Our results indicate that NJ always outperforms the quartet-based methods we examined, in terms of both accuracy and speed.
  - Give new theory about convergence rates of quartet-based methods which helps explain our observations.

Experimental Design

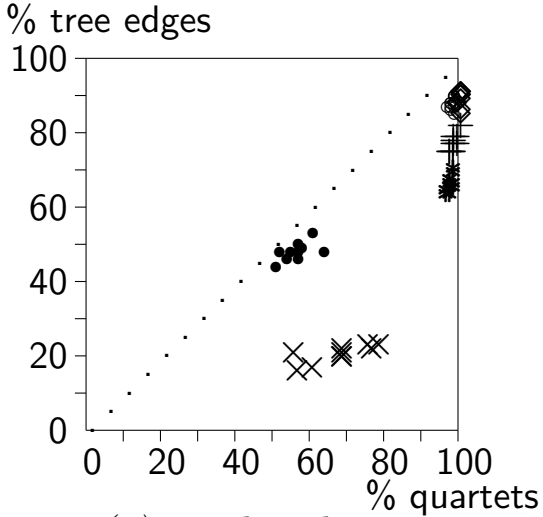
- Generated a large number of datasets, varying number of taxa, rates of evolution, and sequence lengths.
- For each dataset generated, we computed
  - the NJ and QP trees on the entire dataset, and
  - two sets of quartets, Q_{ML}, and Q_{NJ}.
- We applied cleaning methods to Q_{ML} and Q_{NJ} and compared quartets of Q_{ML}, of Q_{NJ}, and of the reconstructed trees against the model tree for accuracy.

Experimental Design: Parameter Space

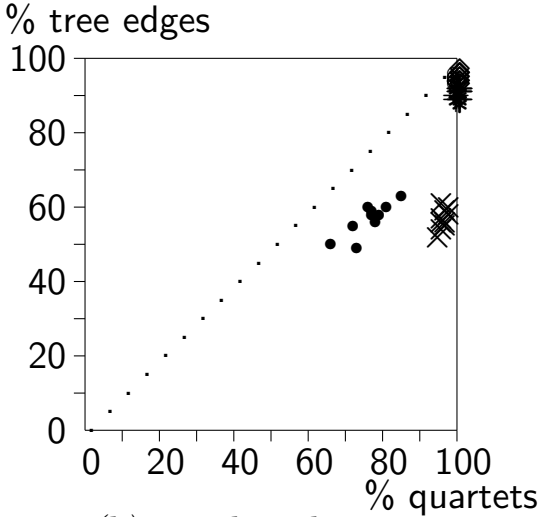
- In all, our study used 16,000 datasets and required many months of computation on the two clusters.
  - Taxa: 5, 10, 20, 40.
  - 8 expected evolutionary rates: from 5×10^{-5} to 5×10^{-1} per tree edge.
  - For each, we generated 100 tree shapes, grouped into 10 runs of 10 trials.
  - Sequence lengths: 500, 2,000, 8,000, and 32,000.

Measuring Accuracy: Quartets and Edges

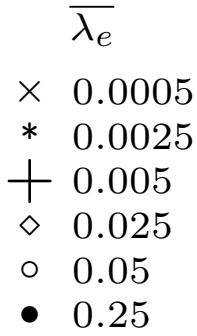
- Topological accuracy is a more demanding criterion than quartet accuracy.



(a) seq. length 500



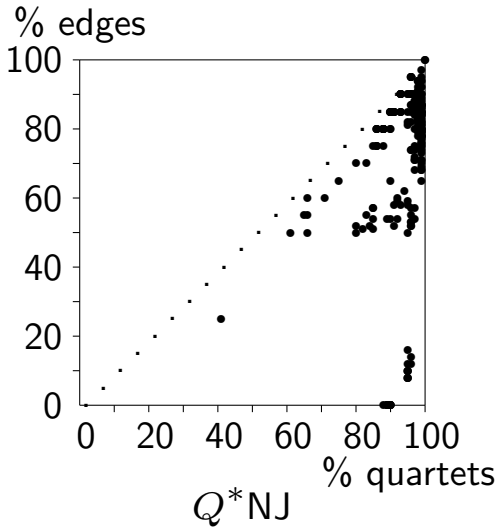
(b) seq. length 2000

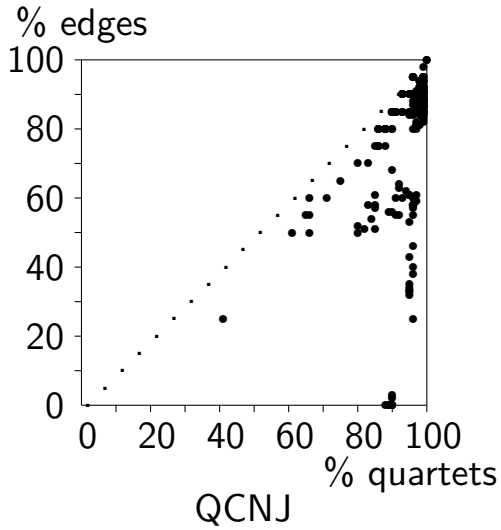


(Percent of true tree edges recovered by Quartet Puzzling for 40 taxa and two sequence lengths)

Sensitivity to Input Quality

- Methods that estimate quartets and then combine them into a single tree can be greatly affected by the quality of the input quartets.





Running Times

- NJ was clearly the fastest method tested.
- QCML and QP were by far the slowest of the methods tested, slow enough that running them on more than a fifty taxa appears infeasible at present.
- With default settings, QP takes more than 200 days of computation to analyze ten runs of ten trials each for a single set of parameters on 80 taxa with a sequence length of 500. (30 minutes for NJ).