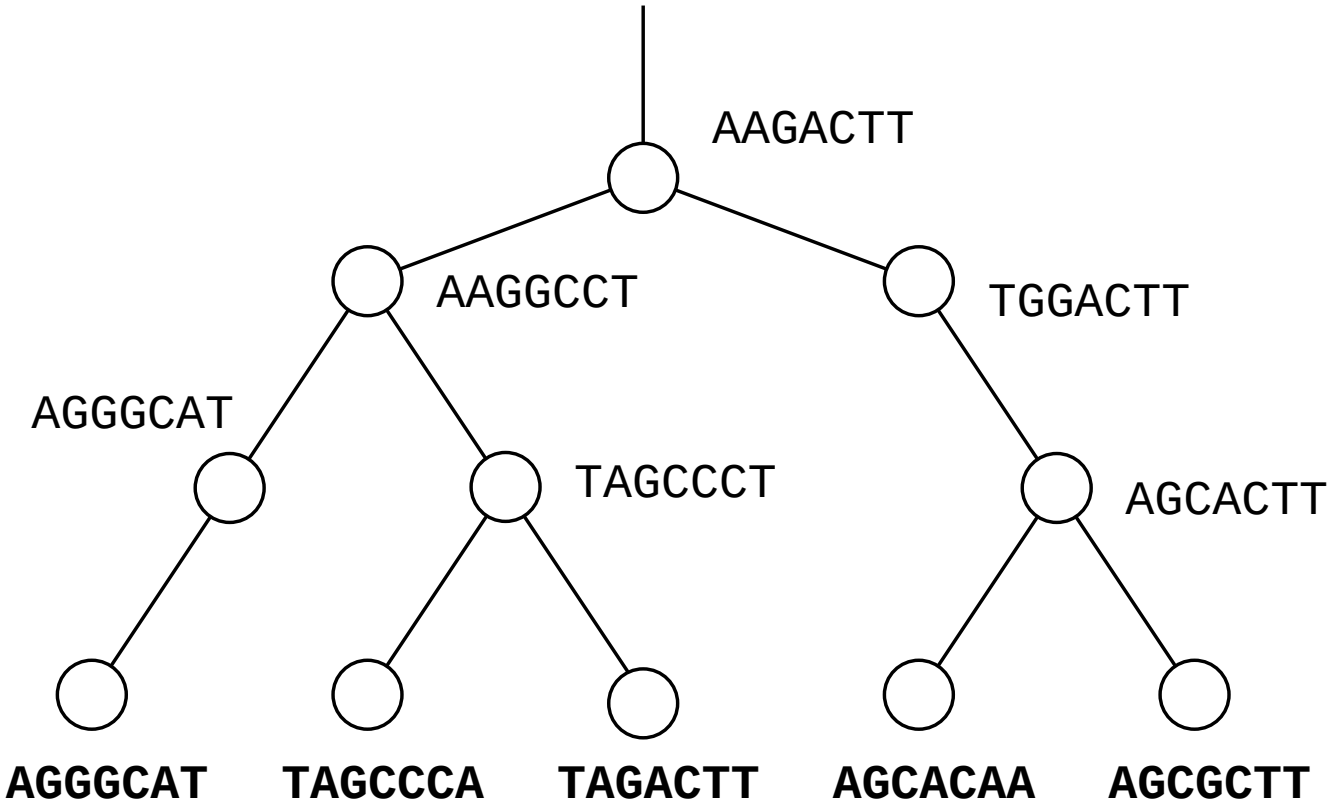
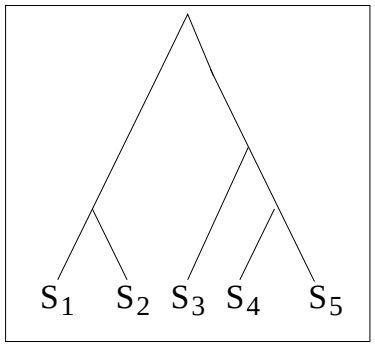
"Fast converging methods" in phylogeny reconstruction

Tandy Warnow
Dept. of Computer Sciences
University of Texas at Austin

Molecular Evolution



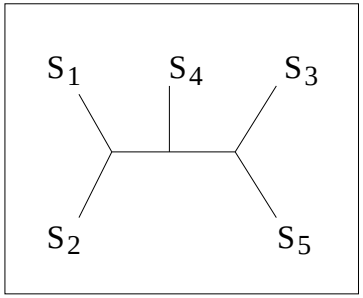
Phylogeny Reconstruction



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES



INFERRED TREE

Applications

- Big genome sequencing projects are producing a lot of data, but the data need to be analyzed - and a phylogeny helps with the analysis.
- Evolutionary history relates all organisms and genes, and helps us understand and predict:
  - interactions between genes (genetic networks)
  - drug design
  - predicting functions of genes
  - influenza vaccine development
  - origins and spread of disease
  - origins and and migrations of humans

Phylogeny reconstruction as a statistical estimation problem

Initially phylogeny reconstruction was based upon maximum parsimony analyses of morphology, or simple distance-based analyses of molecular sequences.

However, phylogeny reconstruction changed dramatically beginning in the 1960's with the introduction of stochastic models of evolution (Jukes-Cantor, Kimura 2-parameter, HKY, etc.).

Markov models of DNA sequence evolution

A Jukes-Cantor model tree is a pair (T, λ) modelling how a single site evolves:

- T is a rooted binary tree,
- λ is a function mapping edges to real numbers, so that $\lambda(e)$ is the expected number of mutations of the site on edge e

Assumptions:

1. The state at the root of T is drawn from the uniform distribution.
2. The number of times the site changes on each edge obeys a Poisson distribution
3. If the state at the site changes, it changes with equal probability to the other states.

Modelling site variation

Almost all proposed models, and all models in use, make the strong assumption of *i.i.d.* site evolution.

(The "rates-across-sites" assumption usually has the rates drawn from a distribution, and so it is still *i.i.d.*.)

Performance issues

- For a biologist: *how accurate* are the estimations of evolutionary history? (Mostly studied in simulation)
- For a statistician:
  - Is the model *identifiable?*
  - Is a given phylogeny reconstruction method *statistically consistent* under the model?
  - *How much data* does a given method need to reconstruct a given model tree correctly with high probability?

  The first two questions were fairly well understood, but the last question remained largely unanswered until the late 1990's.

This talk

- Mathematical techniques for bounding the sequence length requirements of phylogeny reconstruction methods.
- The first methods guaranteed to reconstruct the tree with high probability from "polynomial" sequence lengths.
- More recent methods with the same theory but better performance in simulation (lower topological error).

Warm-up

Questions:

1. Given a coin and $\epsilon > 0$, compute $Pr[head]$ exactly, with probability at least $1 - \epsilon$ of being correct.
2. Given a coin with $Pr[head] \neq \frac{1}{2}$ and $\epsilon > 0$, determine whether $Pr[head] \geq \frac{1}{2}$, with probability at least $1 - \epsilon$ of being correct.

Solution: both have simple solutions, but Question 1 needs an infinite number of coin tosses, while Question 2 can be done with a finite number of coin tosses (the number of coin tosses will depend upon both ϵ and $Pr[head]$).

Tree Estimation

Questions:

1. Given sequences generated by an unknown but fixed JC model tree (T, λ) and $\epsilon > 0$, determine (unrooted) T and λ exactly, with probability at least $1 - \epsilon$ of being correct.
2. Given sequences generated by an unknown but fixed JC model tree (T, λ) and $\epsilon > 0$, determine (unrooted) T exactly, with probability at least $1 - \epsilon$ of being correct.

Solution: both have solutions, but Question 1 needs infinite sequence length, while Question 2 can be done with finite sequence length.

We explore the performance of algorithms for Question 2 with respect to running time and the amount of data they need.

A brief history of mathematical phylogenetics

- 1960's and on: stochastic models of evolution, with *i.i.d.* evolution between sites
- 1978: Maximum Parsimony and Maximum Compatibility are not statistically consistent (Felsenstein)
- Mid-1990's and on:
  - Proofs of statistical consistency for basic methods (neighbor joining and maximum likelihood)
  - First mathematical analyses bounding the sequence length requirements of different methods
  - The Short Quartet Methods (the first "fast converging" methods) (Co-authors Peter Erdos, Laszlo Szekely, and Mike Steel)
  - The Disk-Covering Methods: turning exponentially converging methods into fast converging methods

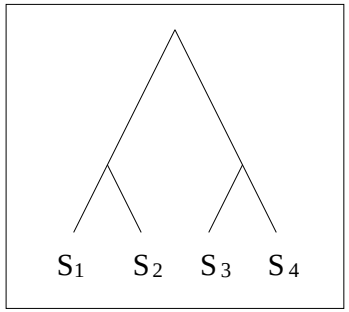
2000 and since

- Work at Berkeley (Mossel et al.) has looked at model parameters for which logarithmic sequence length suffices, and has also looked at the problem of constructing forests rather than trees
- Mike Steel and Laszlo Szekely: "teasing apart" two trees

Outline

- Distance-based phylogeny reconstruction
- Sketched proof of statistical consistency and exponential convergence rate for a simple method
- The Dyadic Closure (Short Quartet) method, and a sketch of the proof of its polynomial convergence
- The Disk-Covering method, and its properties

Distance-based Phylogenetic Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

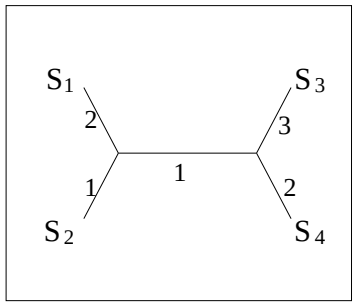
DNA SEQUENCES

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



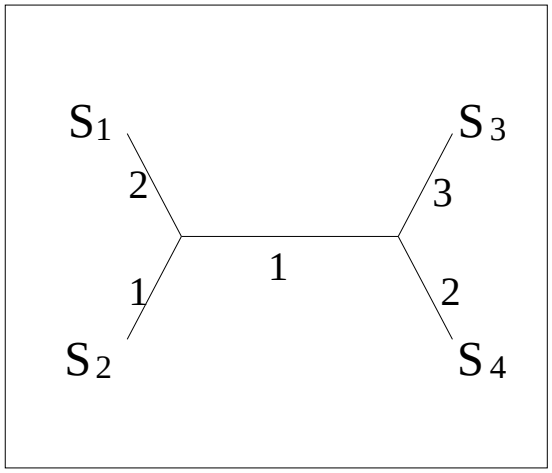
INFERRED TREE

Additive matrices define trees

Given any additive matrix $[D_{ij}]$ we can construct the *unrooted* version of T in polynomial time, along with the edge weights $w(e)$ realizing $[D_{ij}]$. Furthermore, (T, w) is unique up to nodes of degree two.

	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

POLYTIME
INVERTIBLE

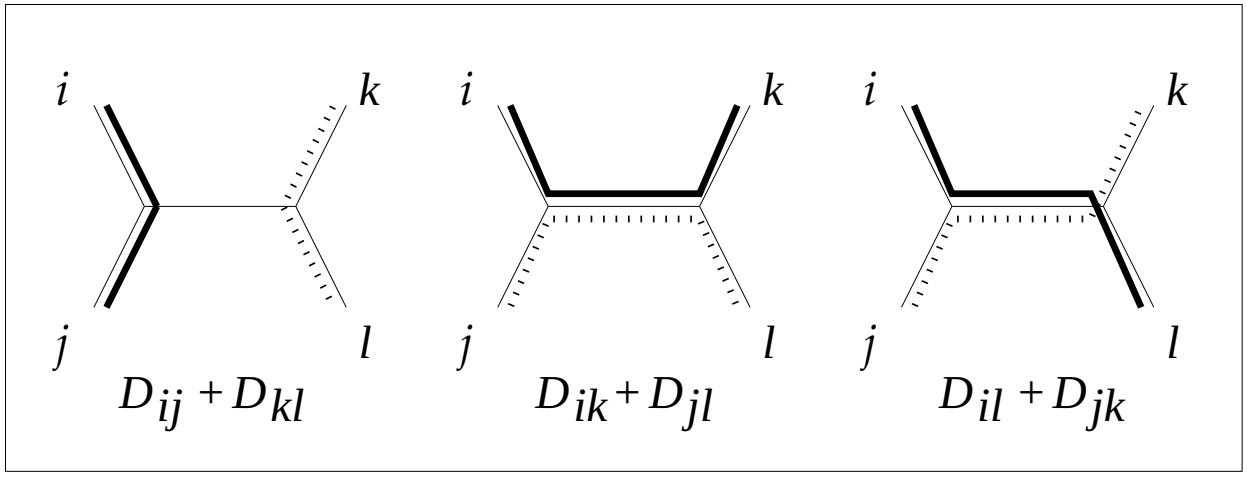


Four Point Condition: $[D_{ij}]$ is an additive matrix *if and only if* for all i, j, k, l, the median and maximum of the three pairwise sums are identical:

$$D_{ij} + D_{kl}$$

$$D_{ik} + D_{jl}$$

$$D_{il} + D_{jk}$$



The Four Point Method (FPM)

The Four Point Method can be used to infer trees on quartets of leaves from a dissimilarity matrix, $[D_{ij}]$ (a matrix satisfying $D_{ii} = 0$ and $D_{ij} = D_{ji}$, but not necessarily satisfying the triangle inequality).

Given the dissimilarity matrix $[D_{ij}]$ and four indices i, j, k, l, the FPM return the tree $ij|kl$ such that

$$D_{ij} + D_{kl} = \min\{D_{ij} + D_{kl}, D_{ik} + D_{jl}, D_{il} + D_{jk}\}.$$

Naive Quartet Method (NQM)

Let $[D_{ij}]$ be a dissimilarity matrix.

- For each quartet i, j, k, l, compute the subtree on i, j, k, l using the Four Point Method.
- If all the quartet trees are compatible, merge them into a single tree. Else return *Fail*.

Error Tolerance of NQM

Theorem: Let $[A_{ij}]$ be an $n \times n$ additive matrix for a tree (T, w) and let $f = min\{w(e)\}$.

Let $[D_{ij}]$ be an $n \times n$ dissimilarity matrix such that $L_\infty(D, A) < f/2$.

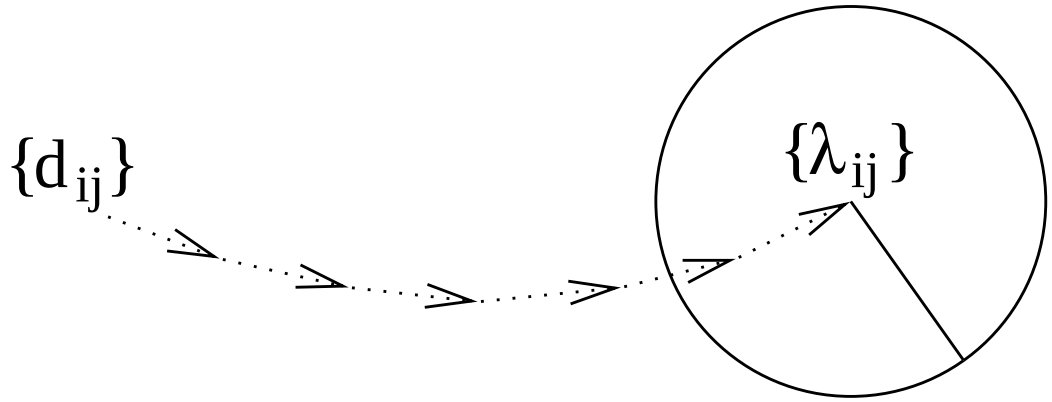
Then $NQM(D) = T$.

Proof: The smallest pairwise sum stays the smallest, and so the Four Point Method makes no mistakes on any quartet tree.

Statistical consistency

A phylogeny reconstruction method Φ is said to be *statistically consistent* under the JC model if for all JC model trees (T, λ), $Pr[\Phi(S) = T] \rightarrow 1$ as the sequence length $k \rightarrow \infty$.

Is NQM statistically consistent under the JC model?



Statistical consistency of distance-based methods

There are statistically consistent techniques for estimating Jukes-Cantor model distances, as well as for estimating distances under other models.

Sequence length requirements

Question: Let Φ be a phylogeny reconstruction method, (T, λ) be a Jukes-Cantor model tree, and $\epsilon > 0$. For what sequence length k will $Pr[\Phi(S) = T] > 1 - \epsilon$, for S a set of sequences of length k generated on (T, λ)?

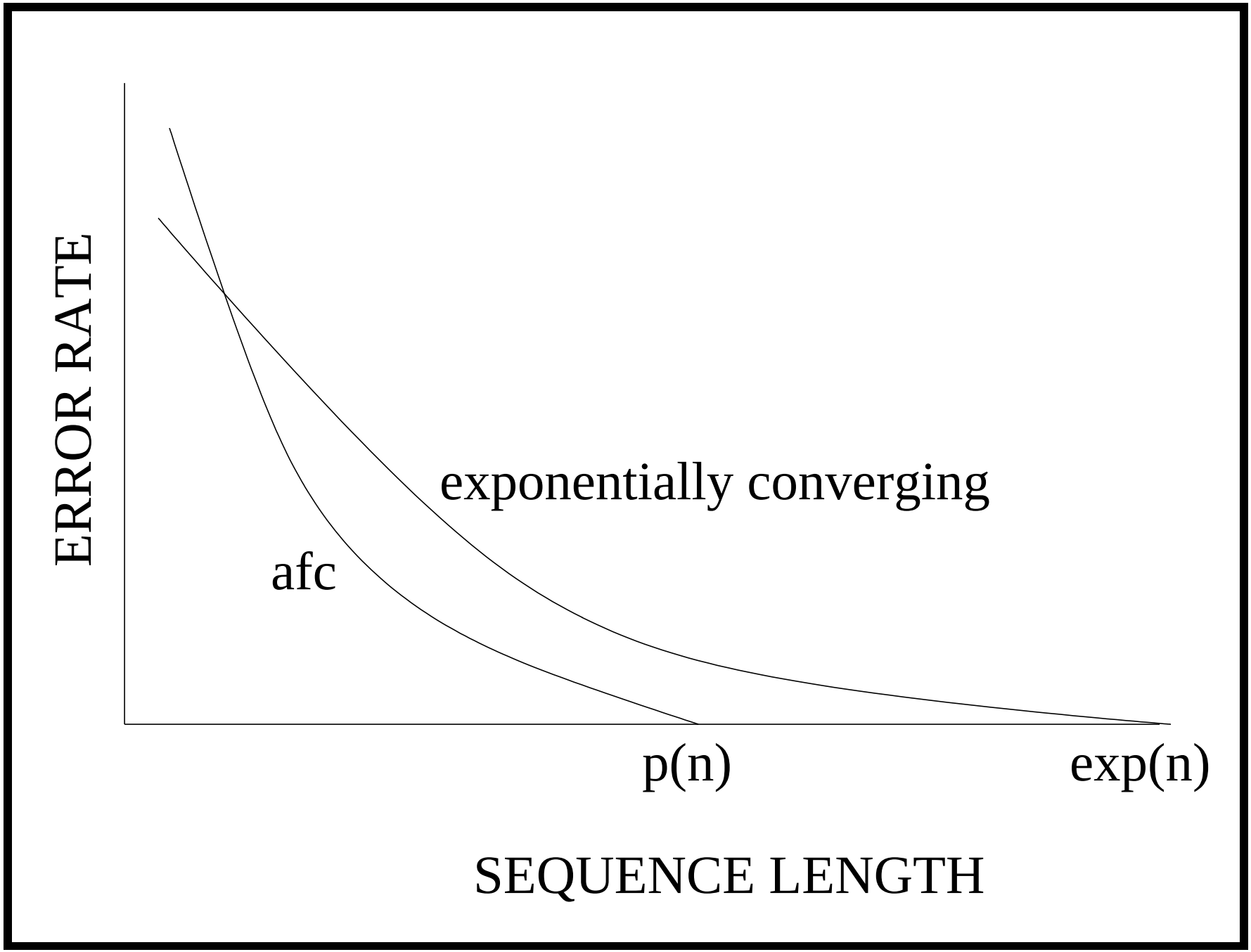
Factors affecting this:

- ϵ,
- $f = \min \lambda(e)$,
- $g = \max \lambda(e)$,
- n, the number of leaves in the tree, and
- Φ.

$\mathrm{JC}_{f,g}$ contains those JC model trees (T, λ) s.t. $f \leq \lambda(e) \leq g$ holds for all edges $e \in E(T)$.

Φ is *absolute fast-converging (afc)* for the JC model if, for all positive f, g, ε, there is a polynomial p such that, for all (T, λ) in the $JC_{f,g}$ model, on set S of n sequences of length at least $p(n)$ generated on T, we have

$$Pr[\Phi(S) = T] > 1 - \varepsilon.$$


ERROR RATE
exponentially converging
afc
p(n)
exp(n)
SEQUENCE LENGTH

Question: Is NQM afc? Are there any afc methods?

Theorem 1: Let $(T, \lambda) \in JC_{f,g}$ and $\epsilon > 0$ be given. Then there is a constant $C > 0$ (that depends on ϵ and f), such that if the sequence length is at least

$$C \log n e^{O(\max \lambda_{ij})}$$

then, $Pr[NQM(d) = T] \geq 1 - \epsilon$, where d is the JC distance matrix computed on the sequences.

Proof (sketch): The condition that is needed is
$Pr[L_\infty(d, \lambda) < f/2] \geq 1 - \epsilon$.

Comments:

- Since $\max \lambda_{ij} = O(g \cdot diam(T))$, and $diam(T) \leq n - 1$, we say that NQM is *exponentially converging.*
- The same condition holds for neighbor joining (NJ).

Are there afc methods?

- To date, none of the standard methods has been shown to be afc for any model.
- NJ's performance in simulation is greatly superior to NQM (and almost all other distance methods), but it is not afc – a matching lower bound for a special case of JC model trees was proven a few years ago.
- The only known upper bound on the sequence length requirement of Maximum Likelihood is larger (Szekely and Steel).

Problem with NQM

The problem with NQM (and other methods) is that *every* entry D_{ij} must be sufficiently well estimated (i.e. all D_{ij} must satisfy $|D_{ij} - \lambda_{ij}| < f/2$).

What if we could identify the entries in the input matrix $[D_{ij}]$ which have sufficiently small error? Could we construct the tree from that subset of the matrix?

Example: The "caterpillar tree" on leaves labelled $1, 2, \ldots, 7$ can be constructed from 12|34, 23|45, 34|56, 45|67.

Conjectures: Perhaps (1) all trees can be constructed from a proper subset of their quartet trees, and (2) that proper subset might be more likely to be accurately constructed from short sequences, and (3) that proper subset might be relatively easy to identify?

Theorem 2 (Warnow, Moret, and St John 1999): Let $(T, \lambda) \in JC_{f,g}$. Define

$$L_{\infty}^{(q)}(D, \lambda) = \max\{|D_{i,j} - \lambda_{i,j}| : \min\{\lambda_{i,j}, D_{i,j}\} \leq q\}.$$

For every $\epsilon, \delta > 0$, there exists a constant C such that

$$\text{if } k \geq C \log n e^{O(q)}$$

$$\text{then } Pr[L_{\infty}^{(q)}(D, \lambda) < \delta] > 1 - \epsilon.$$

where $[D_{ij}]$ is the JC distance matrix obtained for a set of sequences of length k generated on (T, λ).

This suggest that *estimates of small distances are more accurate than estimates of large distances.*

Some afc methods:

- The *Short-Quartet* methods [Erdös, Steel, Szekely, and Warnow, ICALP 1997]
- An unnamed method [Cryan, Goldberg and Goldberg, FOCS 1998]
- Harmonic Greedy Triplets plus the Four Point Method [Csuros, 2002]
- $DCM_{NJ} + SQS$, and other such "DCM-boosted" methods [Warnow, St. John, and Moret, SODA 2001]

Comment: The Short Quartet Methods have the simplest theory and best convergence rate, but $DCM + SQS$ has the best empirical performance.

Simple idea for constructing a tree in $JC_{f,g}$ given $[D_{ij}]$

- Guess q so that if $D_{ij} \leq q$ then $|D_{ij} - \lambda_{ij}| < f/2$
- For all quartets i, j, k, l such that all pairwise D-distances are at most q, construct a quartet tree using the Four Point Method.
- Compute a tree (if it exists) which is consistent with all the quartet trees.

Issues:

- If q is too large, then some entries might have too much error.
- If q is too small, then there may be insufficient coverage to reconstruct the tree.
- The subtree compatibility problem is NP-Complete.

Question: How small can q be, and still identify the tree?

Short Quartets

Let $[\lambda_{ij}]$ be the additive matrix for the binary tree (T, λ). Let e be an edge in T with subtrees U, V, W, and X off e.

- The **short quartets** around e are obtained by picking a nearest leaf in each of the four subtrees U, V, W and X. (There can be more than one around an edge.)
- $Q_{short}(T, \lambda) = \{\text{short quartet trees around any edge of } T\}$.

Theorem: (Erdos et al, 1997) Let (T, λ) and (T', w) be two trees on the same leaf set. Suppose $Q_{short}(T, \lambda) \subseteq Q(T')$, where $Q(T')$ denotes the set of induced quartet trees of T'. Then $T = T'$.

How "big" are the short quartets?

Let $(T, \lambda) \in JC_{f,g}$.
Define λ-$width(T)$ to be $\max\{\lambda_{ij} : \text{i and j in a short quartet of } T\}$.

Theorem (Erdos et al., 1999):
For all trees $(T, \lambda) \in JC_{f,g}$, λ-width(T) $= O(g \log n)$, where T has n leaves.

Dyadic Closure

Dyadic Closure rules:

- Rule 1: $ij|kl$ and $jk|lm$ imply $ij|km$, $ij|lm$ and $ik|lm$.
- Rule 2: $ij|kl$ and $ij|lm$ imply $ij|km$.

Given set X of trees on four-leaves, repeatedly apply Dyadic Closure rules until no additional trees are obtained. The result is $cl(X)$, the *dyadic closure of* X.

Computing a tree from its short quartet trees

Theorem 3 (Erdos et al, 1997): Let T be a fixed edge-weighted tree, and let $Q_{short}(T)$ denote the set of trees induced by the short quartets of T. Let $Q(T)$ denote the set of four-leaf trees in T.

$$\text{If } Q_{short}(T) \subseteq X \subseteq Q(T) \text{ then } cl(X) = Q(T)$$

Corollary 1: T can be reconstructed from $Q_{short}(T)$ in polynomial time.

The Dyadic Closure Method

Let Q_w denote all the trees computed (using the Four Point Method) on quartets with maximum D-distance w. Construct trees on all quartets using the Four Point Method.

Binary search on $w \in \{D_{ij}\}$ (hoping to find a w such that $Q_{short}(T) \subseteq Q_w \subseteq Q(T)$, so that $cl(Q_w) = Q(T)$) as follows:

- Compute $cl(Q_w)$.
  - If $cl(Q_w)$ contains two trees on some quartet, mark w as too big, and decrease w
  - If $cl(Q_w)$ doesn't contain a tree on some quartet, mark w as too small, and increase w
  - If $cl(Q_w)$ is neither too big nor too small, then $cl(Q_w) = Q(T')$ for some tree T', and we can construct T' in polynomial time.

Theorem 4 (Erdos et al.): Let $(T, \lambda) \in JC_{f,g}$, and let d be a dissimilarity matrix given as input to the Dyadic Closure Method. Then the Dyadic Closure Method returns T if

$$L_{\infty}^{(d-width(T))}(d, \lambda) < \frac{f}{2}$$

where the d-width is the maximum d-distance in a short quartet.

Theorem 4 (Erdos et al, 1997): The Dyadic Closure Method is $O(n^5 \log n)$ time and fast-converging for JC tree reconstruction. Furthermore, polylogarithmic length sequences suffice for accuracy with high probability for random JC trees.

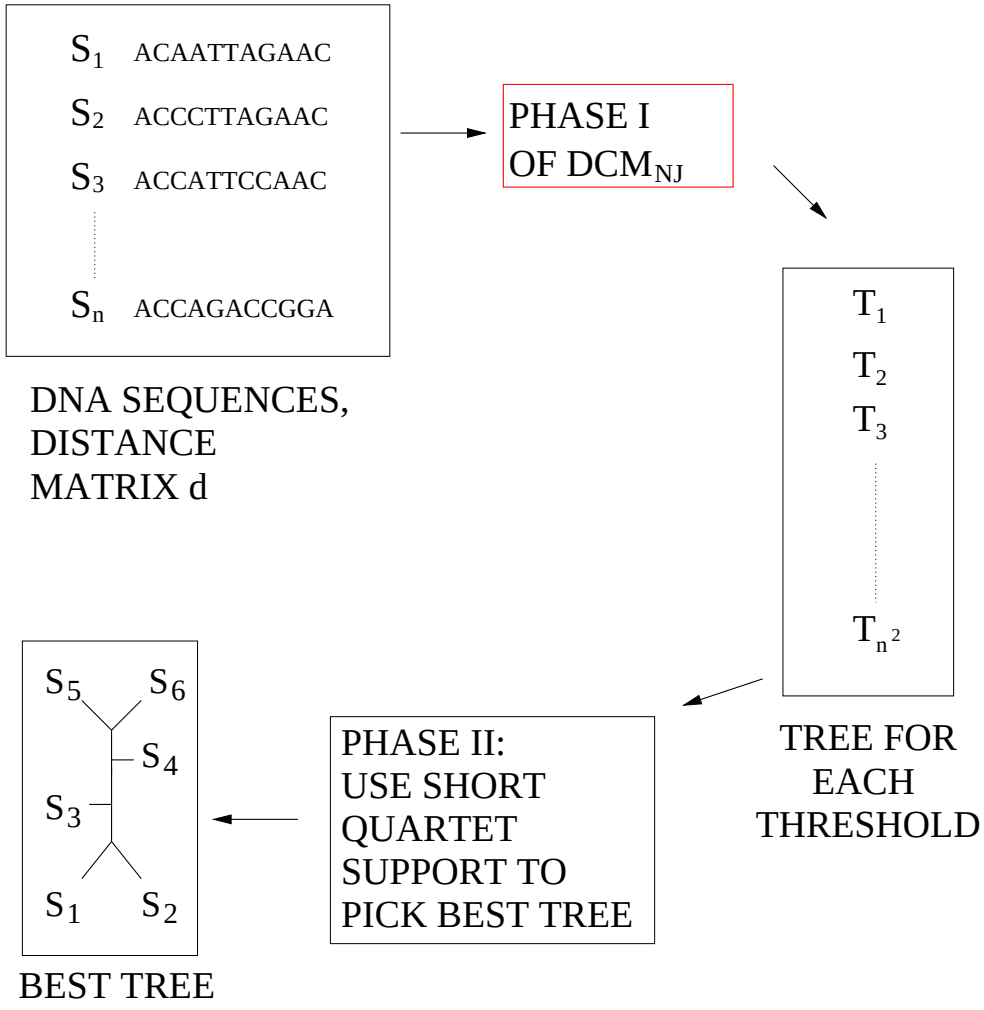
Sketch of proof: The running time is easy. We show that $\lambda\text{-}width(T) = O(g \cdot \log n)$ for all trees, so that by Theorem 2, the Dyadic Closure Method is fast converging. Also, random trees have $\lambda\text{-}width(T) = O(g \cdot \log \log n)$, so that by Theorem 2, the Dyadic Closure Method converges from *polylogarithmic* length sequences on random trees.

The performance of the Dyadic Closure Method

- The Dyadic Closure Method has excellent theory (with respect to its sequence length requirement) but does not perform well in practice: *it only succeeds in returning a tree if all the short quartets can be accurately reconstructed.*
- By comparison, NJ is better in simulation on model trees that look biological (unless they are extremely large trees, and we simulated evolution of short sequences).
- Even so, NJ is not afc.

These observations led us to develop a different kind of afc method, with the objective of obtaining an empirical improvement while maintaining theory.

afc method: DCM_{NJ}+SQS (Warnow *et al*, SODA 2001)



Theorem (Warnow et al.) If Φ is exponentially converging under JC, then $DCM_\Phi + SQS$ is absolute fast converging under JC.

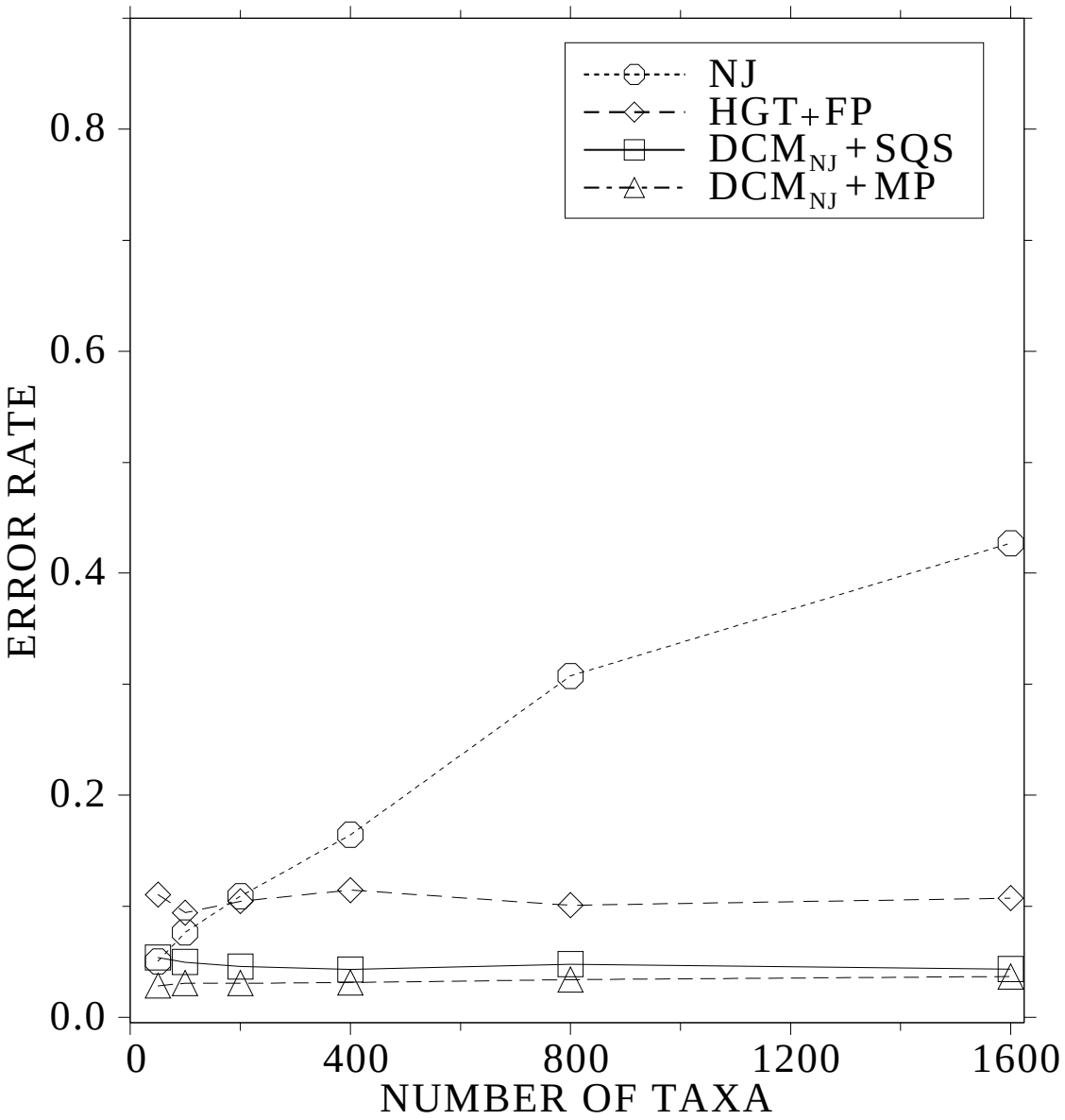
Outline of proof:

We need to show that for all f, g, and $\epsilon > 0$, there is a polynomial $p(n)$ such that for all model trees $(T, \lambda) \in JC_{f,g}$ on n leaves, if we are given a dataset of sequences of sequence length $k \geq p(n)$ then

- $Pr[T \in \{T_w : w \in \{D_{ij}\}\}] > 1 - \epsilon$
- If $T \in \{T_w : w \in \{D_{ij}\}\}$, then $Pr[SQS\ selects\ T] > 1 - \epsilon$.

Comments:

- The same result holds under the General Markov model.
- The empirical performance is dramatic
- SQS can be replaced by other methods for selecting a "best tree" given a set of trees, with evidently better performance – though without proof of theoretical performance. For example - maximum likelihood can be used. Surprisingly, maximum parsimony has comparable performance to ML.


NJ
HGT+FP
DCM$_{NJ}$+SQS
DCM$_{NJ}$+MP
ERROR RATE
0.0
0.2
0.4
0.6
0.8
0
400
800
1200
1600
NUMBER OF TAXA

Open problems:

- New techniques need to be developed to establish convergence rates, as clearly the mathematical bounds are loose for some methods (at least on "random" trees). In particular, what is the sequence length requirement for Maximum Likelihood?
- Why do Maximum Parsimony heuristics do so well?

Related work:

- Disk-Covering methods have also been developed to speed-up heuristics for hard optimization problems in phylogenetics (maximum likelihood and maximum parsimony, as well as problems in gene order phylogeny), obtaining speed-ups of up to several orders of magnitude.

Thanks to:

Collaborators:

- Theory: Mike Steel, Peter Erdos, and Laszlo Szekely (for the short quartet methods), Katherine St. John and Bernard Moret (for DCM+SQS), and Daniel Huson (for an earlier version of DCM).
- Implementation: Usman Roshan, Luay Nakhleh, and Jerry Sun.

Sponsors:

- NSF
- The David and Lucile Packard Foundation
- The Institute for Cellular and Molecular Biology at UT-Austin
- The Program for Evolutionary Dynamics at Harvard
- The Radcliffe Institute for Advanced Study