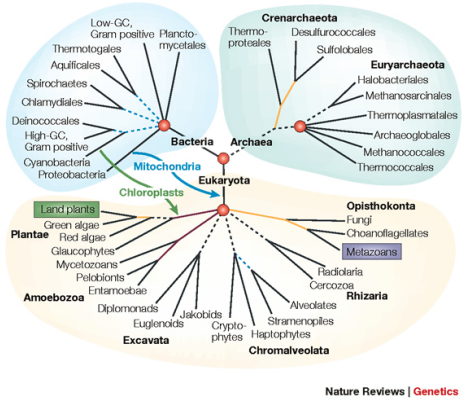
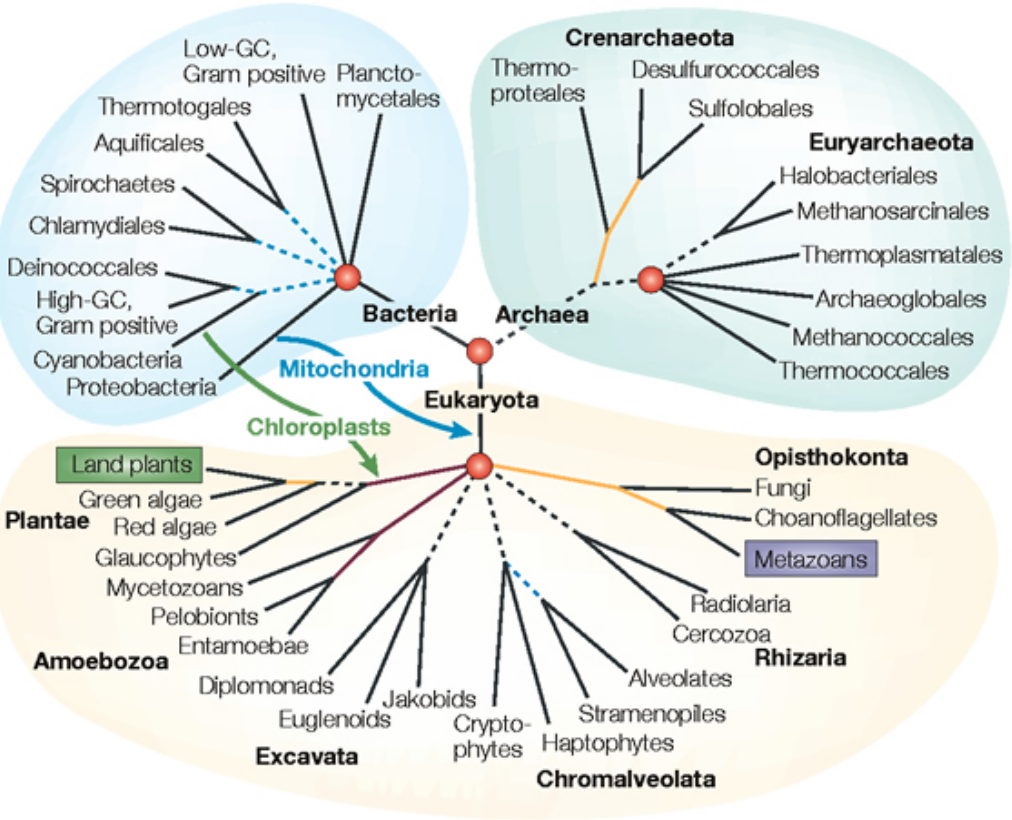


From Gene Trees to Species Trees

Tandy Warnow

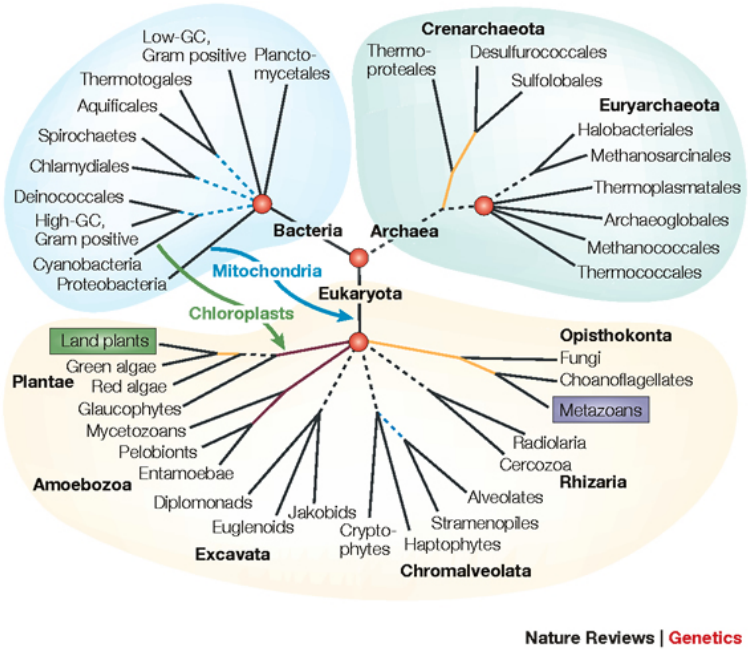
The University of Texas at Austin

The “Tree of Life”



Nature Reviews | Genetics

The Tree of Life: Importance to Biology



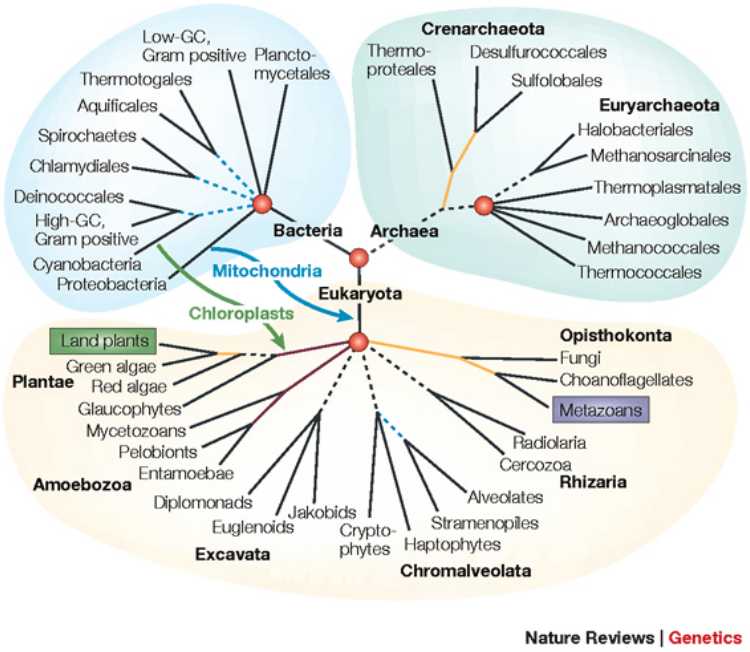
Nature Reviews | Genetics

Biomedical applications

- Mechanisms of evolution
- Environmental influences
- Drug Design
- Protein structure and function
- Human migrations

"Nothing in Biology makes sense except in the light of evolution" - Dobhzhansky

Estimating The Tree of Life: a *Grand Challenge*



Nature Reviews | Genetics

Biomedical applications

- Mechanisms of evolution
- Environmental influences
- Drug Design
- Protein structure and function
- Human migrations

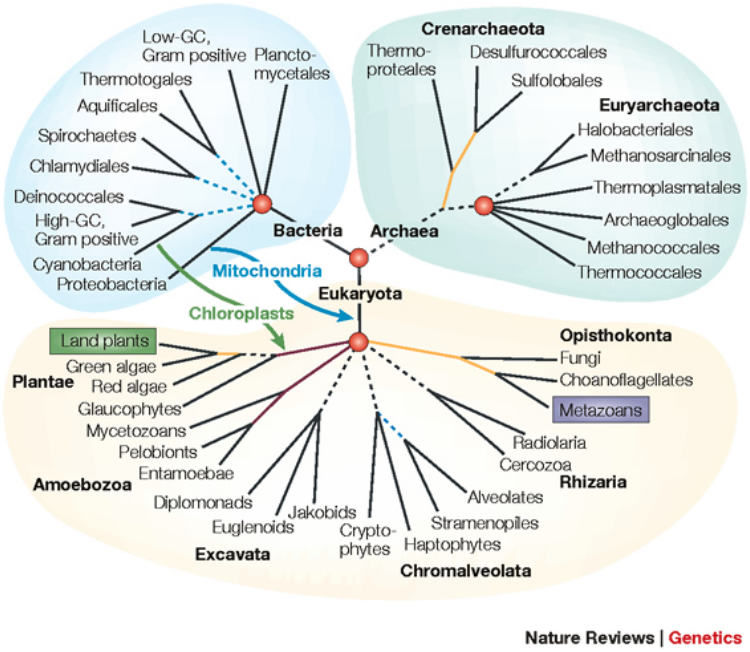
NP-hard problems and large datasets

Current methods do not provide good accuracy

HPC is insufficient

Novel techniques needed for scalability and accuracy

Estimating The Tree of Life: a *Grand Challenge*



Biomedical applications

- Mechanisms of evolution
- Environmental influences
- Drug Design
- Protein structure and function
- Human migrations

Canonical problem:

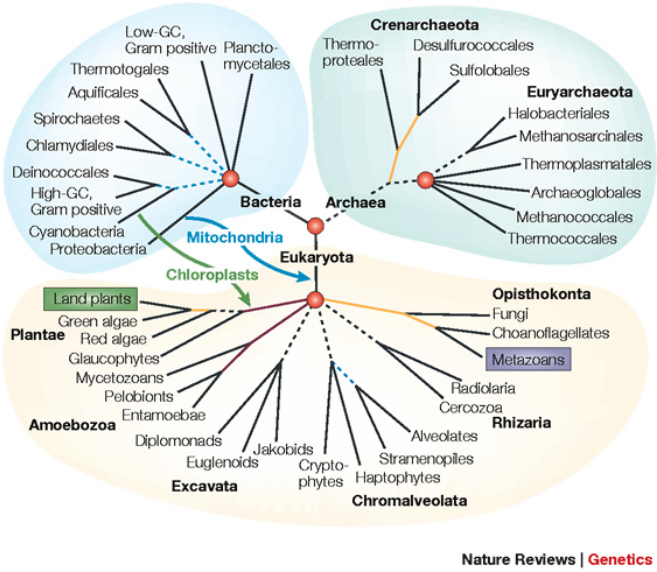
Given multiple sequence alignment, find the Maximum Likelihood Tree

NP-hard, and not solved exactly in practice

Good heuristics, but even these are computationally intensive

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

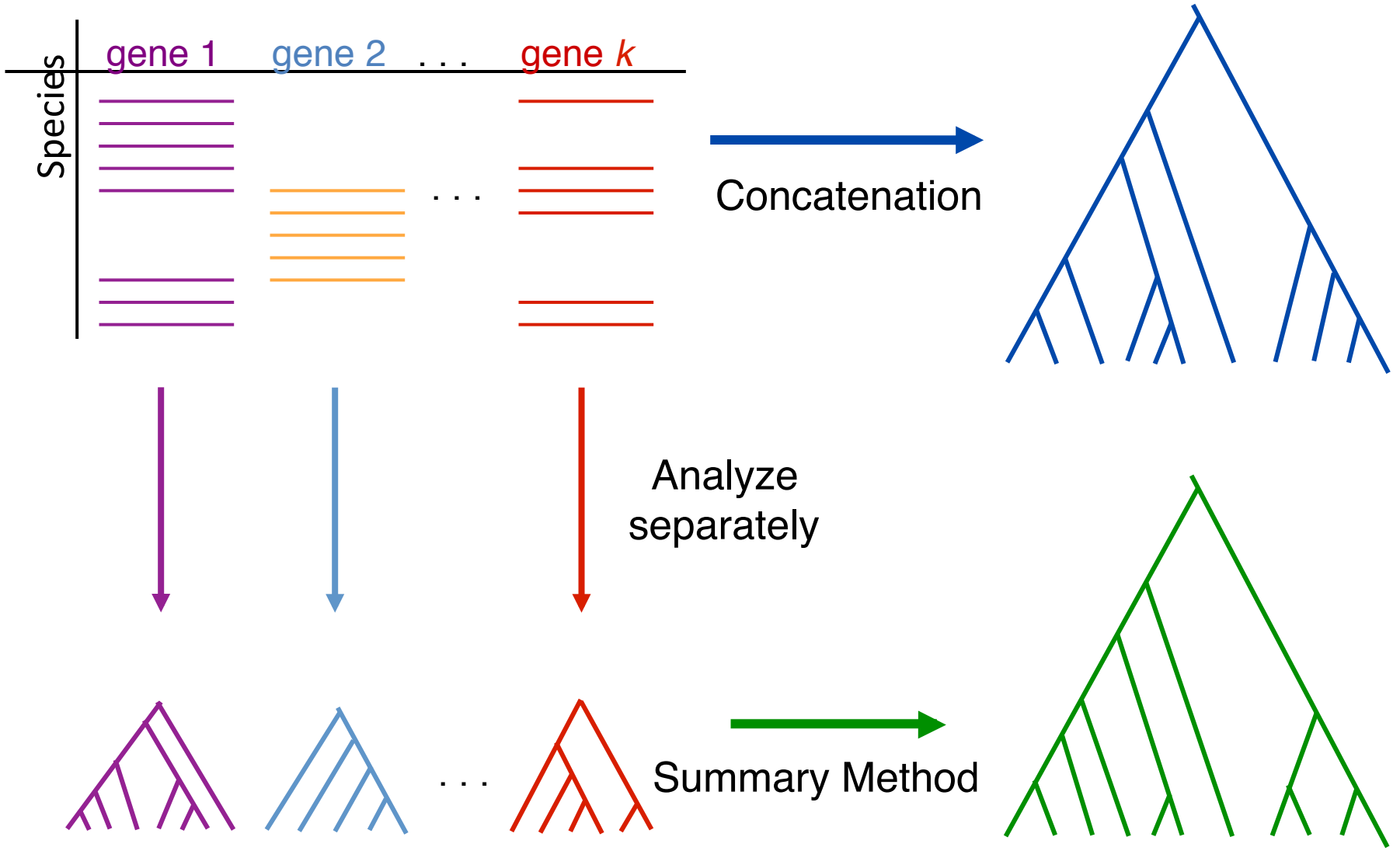
Using multiple genes

	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

Two competing approaches



1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

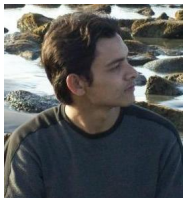
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin



Plus many many other people…

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenges:

Large-scale alignments of > 100,000 sequences

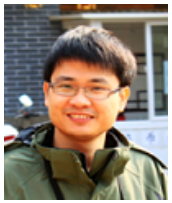
Gene tree incongruence

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

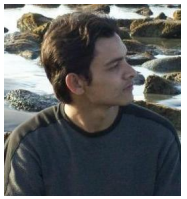
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

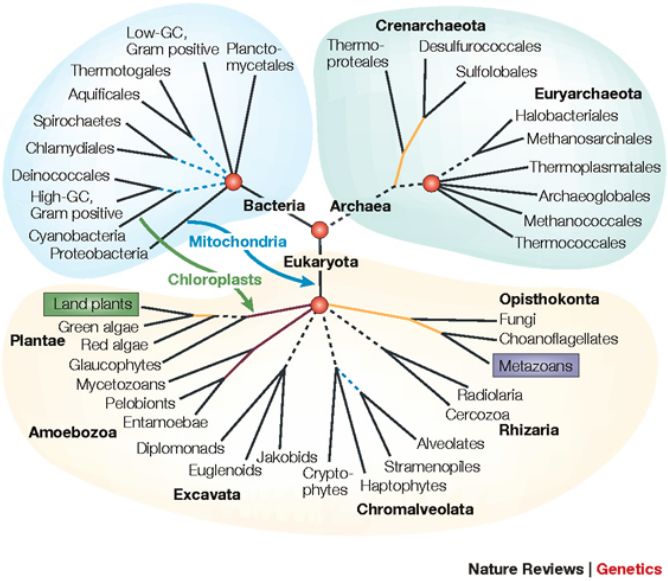
- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Challenges:

Successive radiations producing very short branches

Massive gene tree incongruence

The Tree of Life: *Multiple Grand Challenges*



Nature Reviews | Genetics

Large datasets:

100,000+ sequences

10,000+ genes

"BigData" complexity

Also:

Ultra-large multiple-sequence alignment

Estimating species trees from incongruent gene trees

Supertree estimation

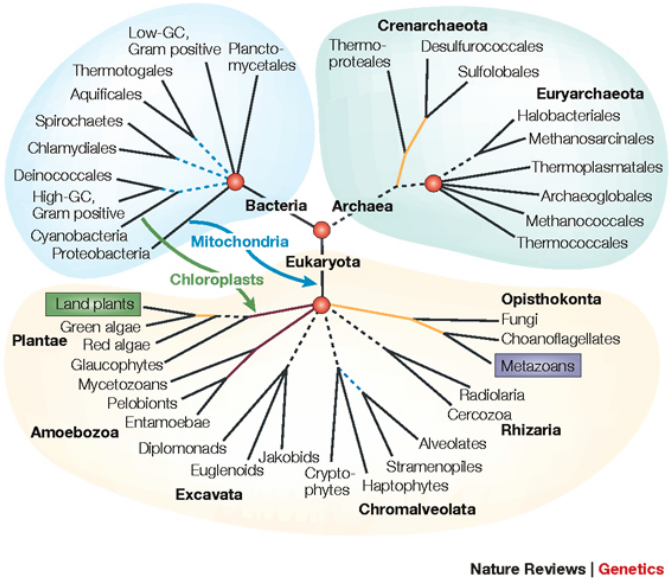
Genome rearrangement phylogeny

Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

The Tree of Life: *Multiple Grand Challenges*



Nature Reviews | Genetics

Large datasets:

100,000+ sequences

10,000+ genes

"BigData" complexity

SATé, Liu et al. Science 2009,
UPP and PASTA, in preparation

Also:

Ultra-large multiple-sequence alignment

Estimating species trees from incongruent gene trees

Supertree estimation

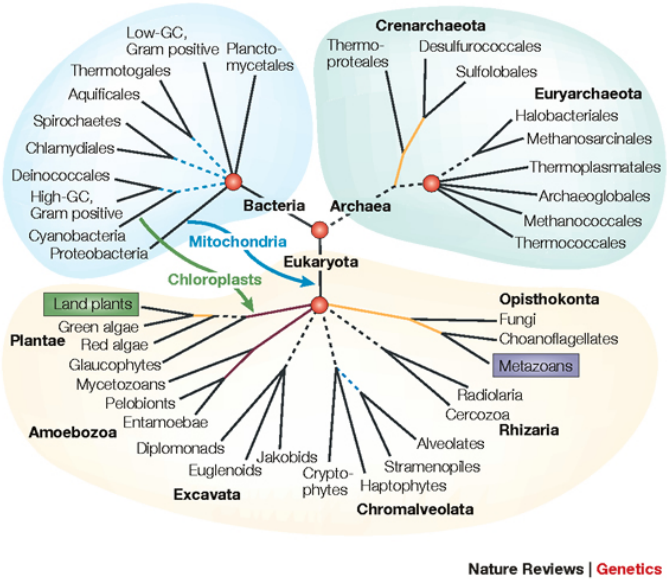
Genome rearrangement phylogeny

Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

The Tree of Life: *Multiple Grand Challenges*



Nature Reviews | Genetics

Large datasets:

100,000+ sequences

10,000+ genes

"BigData" complexity

Also

Ultra-large multiple-sequence alignment

Estimating species trees from incongruent gene trees

Supertree estimation

Genome rearrangement phylogeny

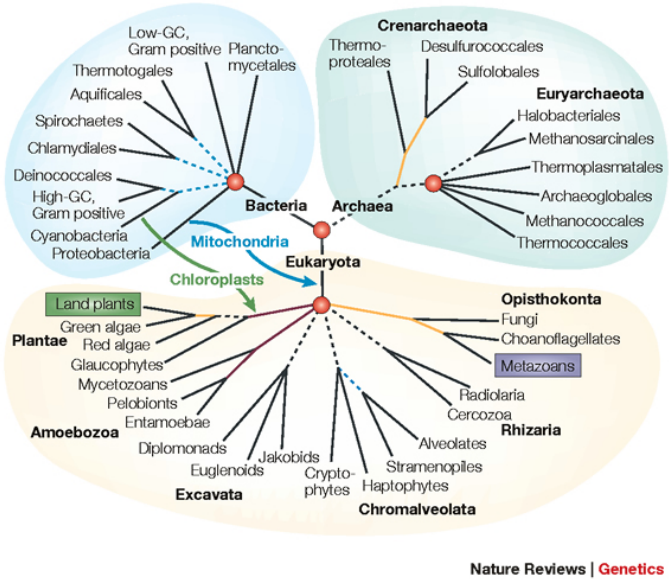
Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

Superfine, Swenson et al., 2012

The Tree of Life: *Multiple Grand Challenges*



Nature Reviews | Genetics

Large datasets:

100,000+ sequences

10,000+ genes

"BigData" complexity

This talk

Also:

Ultra-large multiple-sequence alignment

Estimating species trees from incongruent gene trees

Supertree estimation

Genome rearrangement phylogeny

Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

This talk

Species tree estimation from multiple genes

- Mathematical foundations
- Algorithms
- Data challenges
- New statistical questions
- Avian Phylogenomics

Computational Phylogenetics

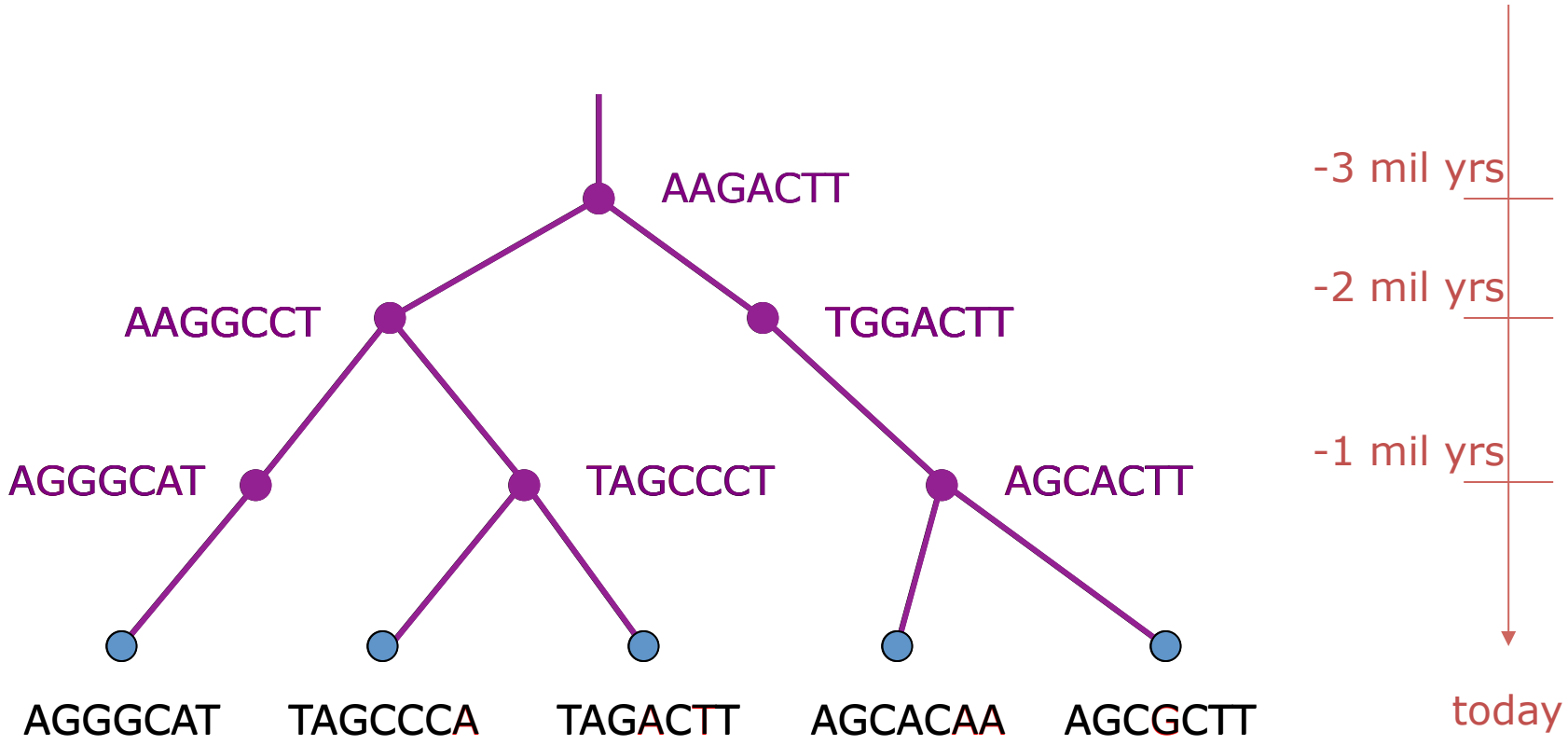
Interesting combination of different mathematics:

- statistical estimation under Markov models of evolution
- mathematical modelling
- graph theory and combinatorics
- machine learning and data mining
- heuristics for NP-hard optimization problems
- high performance computing

Testing involves massive simulations

Part I: Gene Tree Estimation

DNA Sequence Evolution (Idealized)

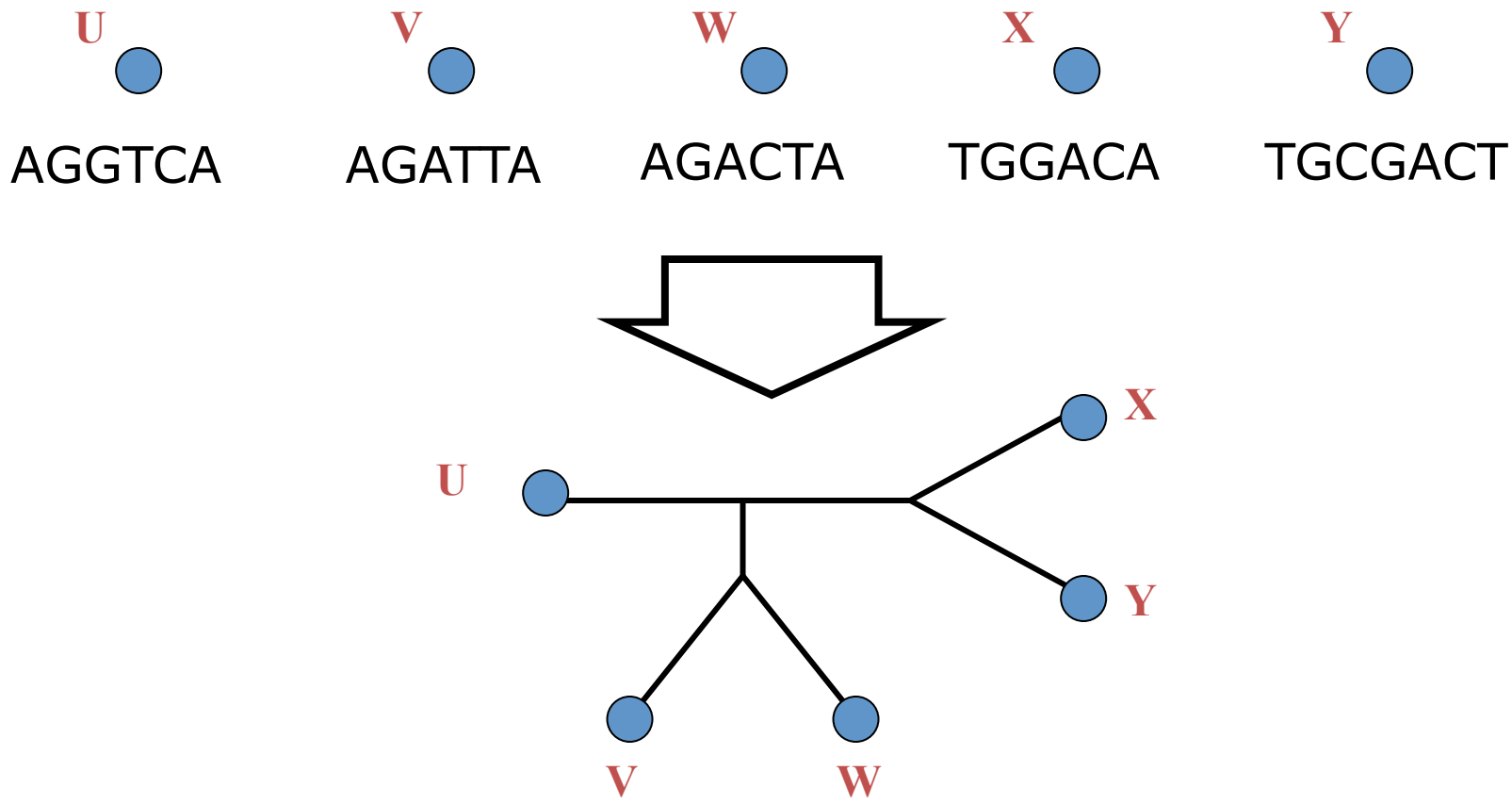


Markov Model of Site Evolution

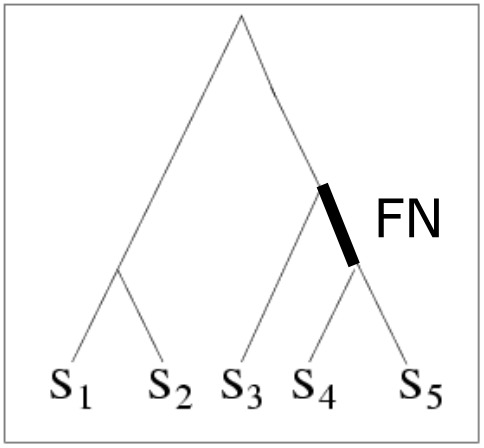
Simplest (Jukes-Cantor, 1969):

- 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.

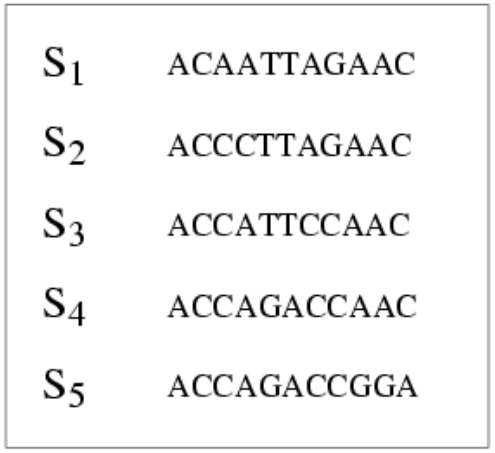
More complex models (such as the General Markov model) are also considered, often with little change to the theory.


U
V
W
X
Y
AGGTCA
AGATTA
AGACTA
TGGACA
TGCGACT
X
U
Y
V
W

Quantifying Error



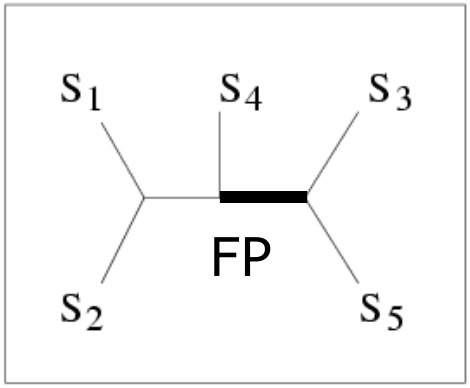
TRUE TREE



DNA SEQUENCES

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

50% error rate

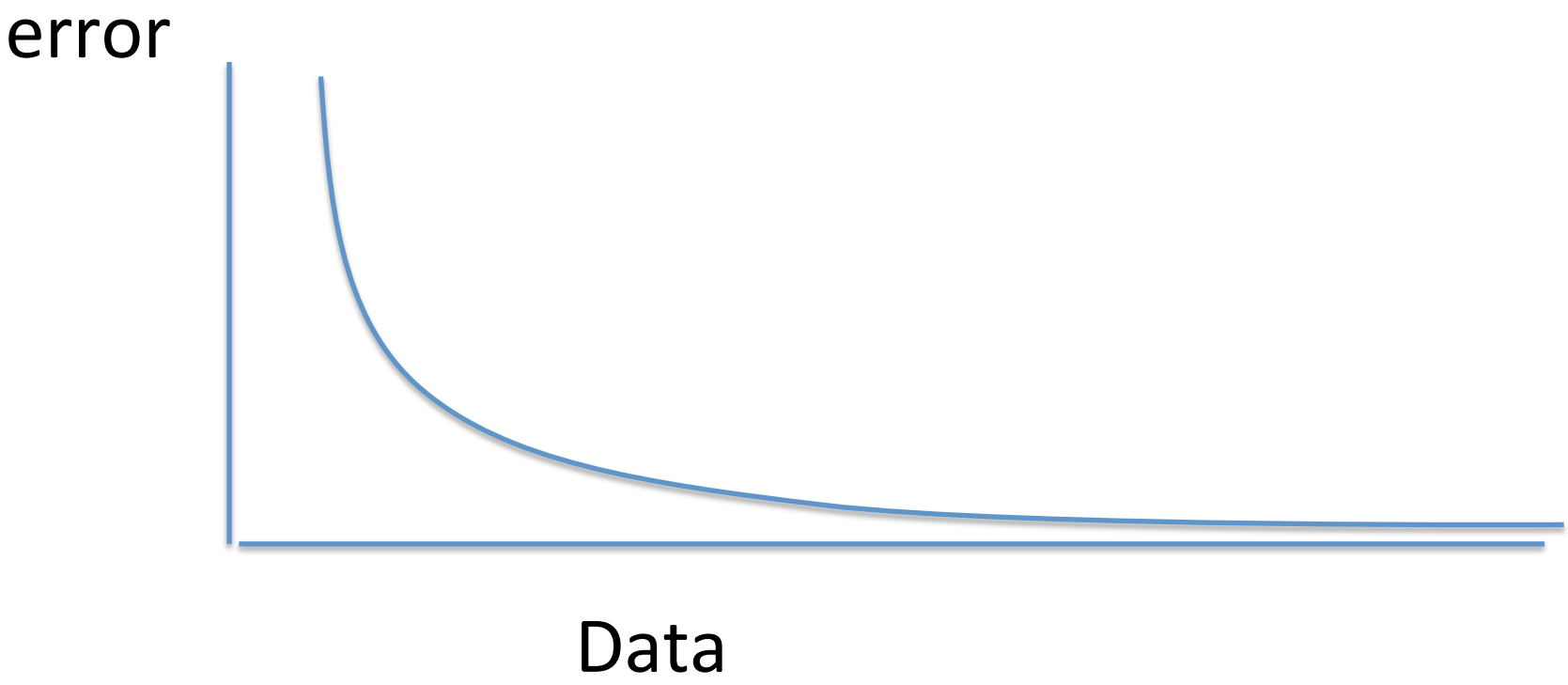


INFERRED TREE

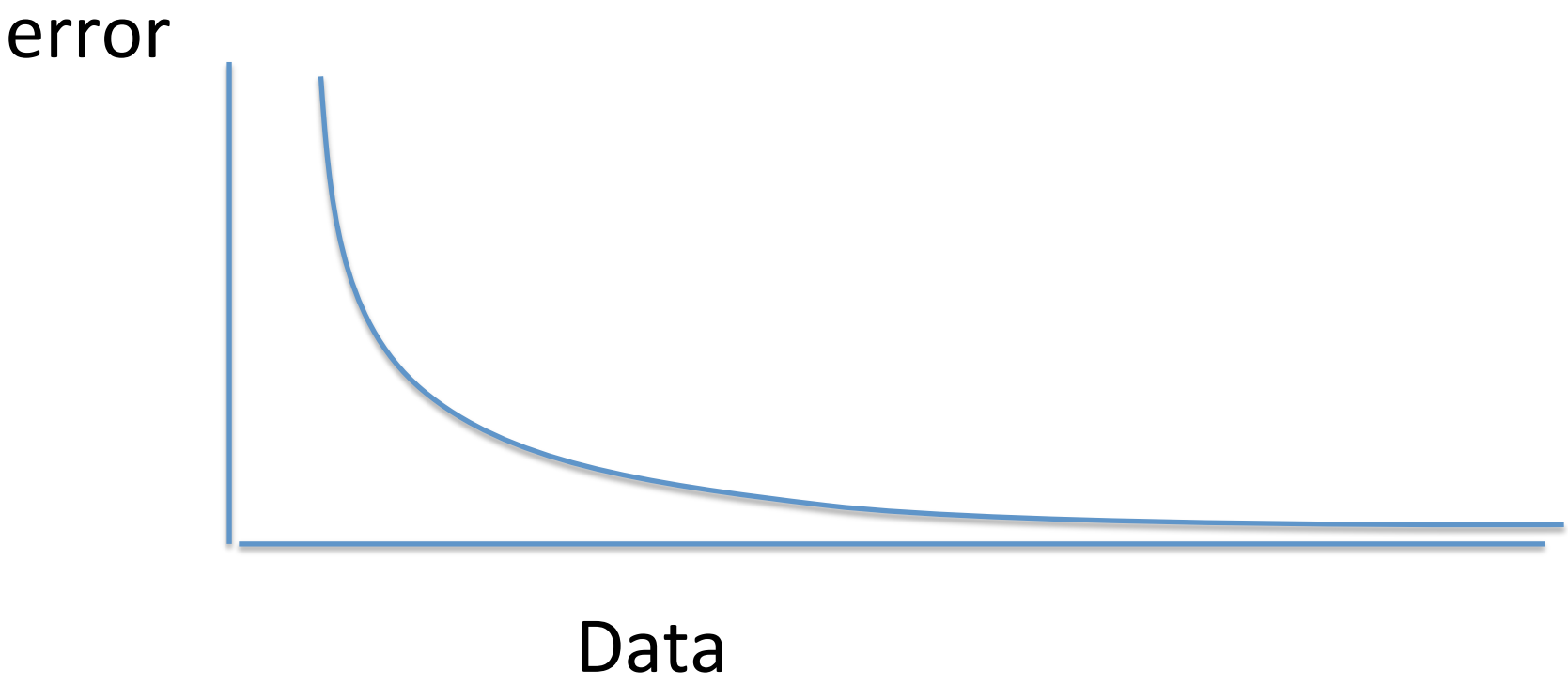
Questions

- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model tree correctly (with high probability)?
- What is the computational complexity of an estimation problem?

Statistical Consistency

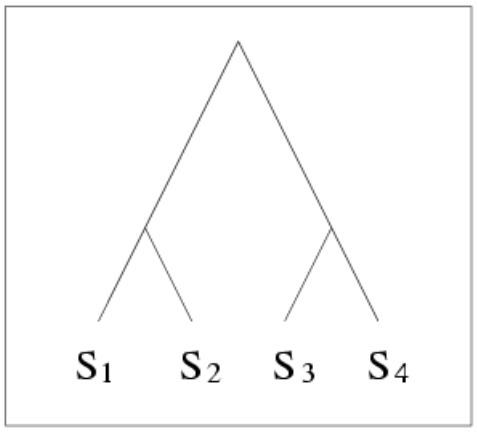


Statistical Consistency



Data are sites in an alignment

Distance-based estimation



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

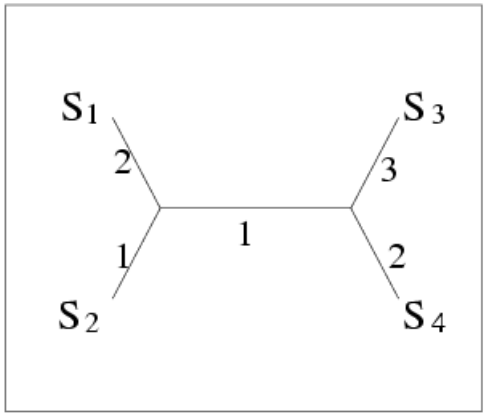
DNA SEQUENCES

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

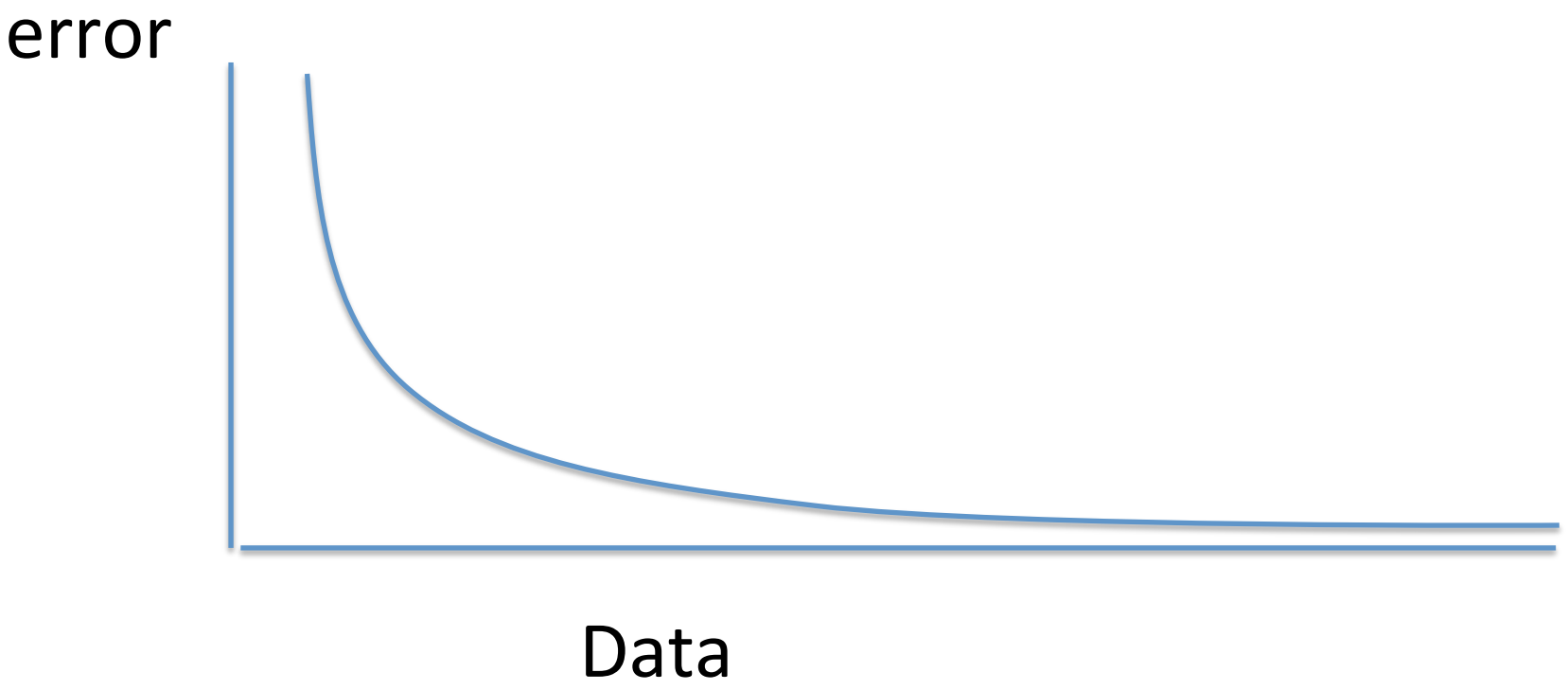
DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING

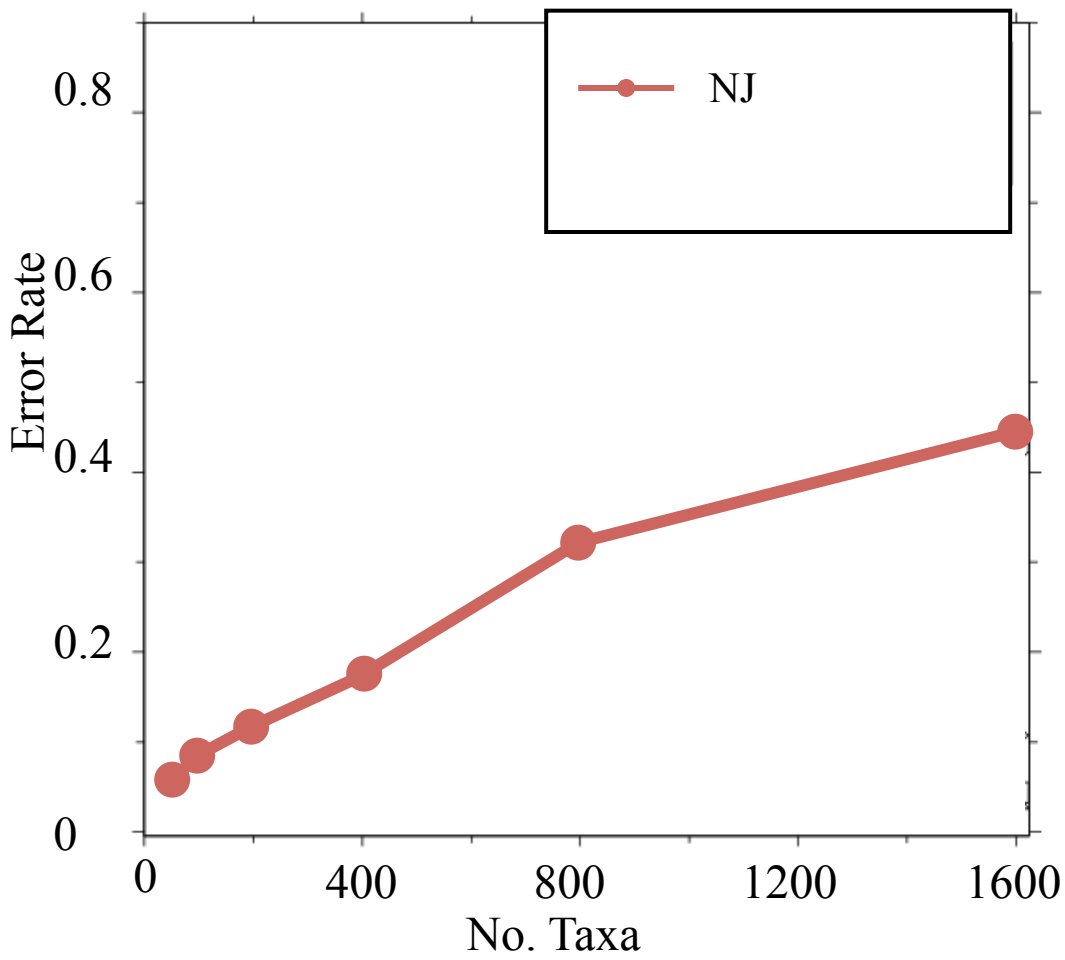


INFERRED TREE

Distance-based methods are statistically consistent under JC



Neighbor Joining on large diameter trees



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.

[Nakhleh et al. ISMB 2001]

"Convergence rate" or sequence length requirement

The sequence length (number of sites) that a phylogeny reconstruction method M needs to reconstruct the true tree with probability at least $1-\varepsilon$ depends on

- M (the method)
- ε
- f = min p(e),
- g = max p(e), and
- n = the number of leaves

We fix everything but n.

Theorem (Erdos et al. 1999, Atteson 1999):

Various distance-based methods (including Neighbor joining) will return the true tree with high probability given sequence lengths that are exponential in the evolutionary diameter of the tree.

Proof: show that

- the method returns the true tree if the estimated distance matrix is close to the model tree distance matrix
- the sequence lengths that suffice to achieve bounded error are exponential in the evolutionary diameter.

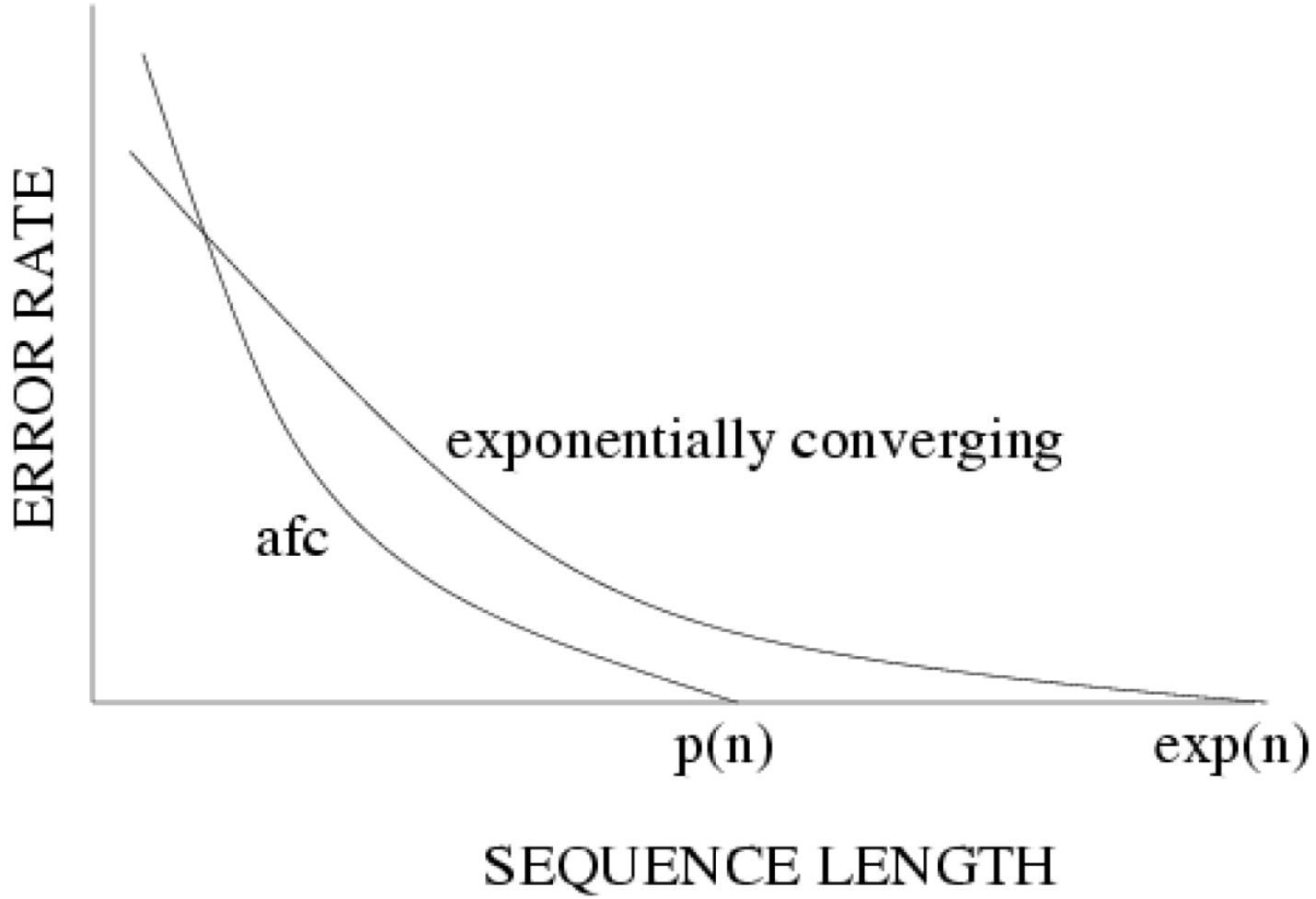
Afc methods (Warnow et al., 1999)

A method M is "absolute fast converging", or afc, if for all positive f, g, and ε, there is a polynomial p(n) s.t. $Pr(M(S)=T) > 1-\varepsilon$, when S is a set of sequences generated on T of length at least p(n).

Notes:

1. The polynomial p(n) will depend upon M, f, g, and ε.
2. The method M is not "told" the values of f and g.

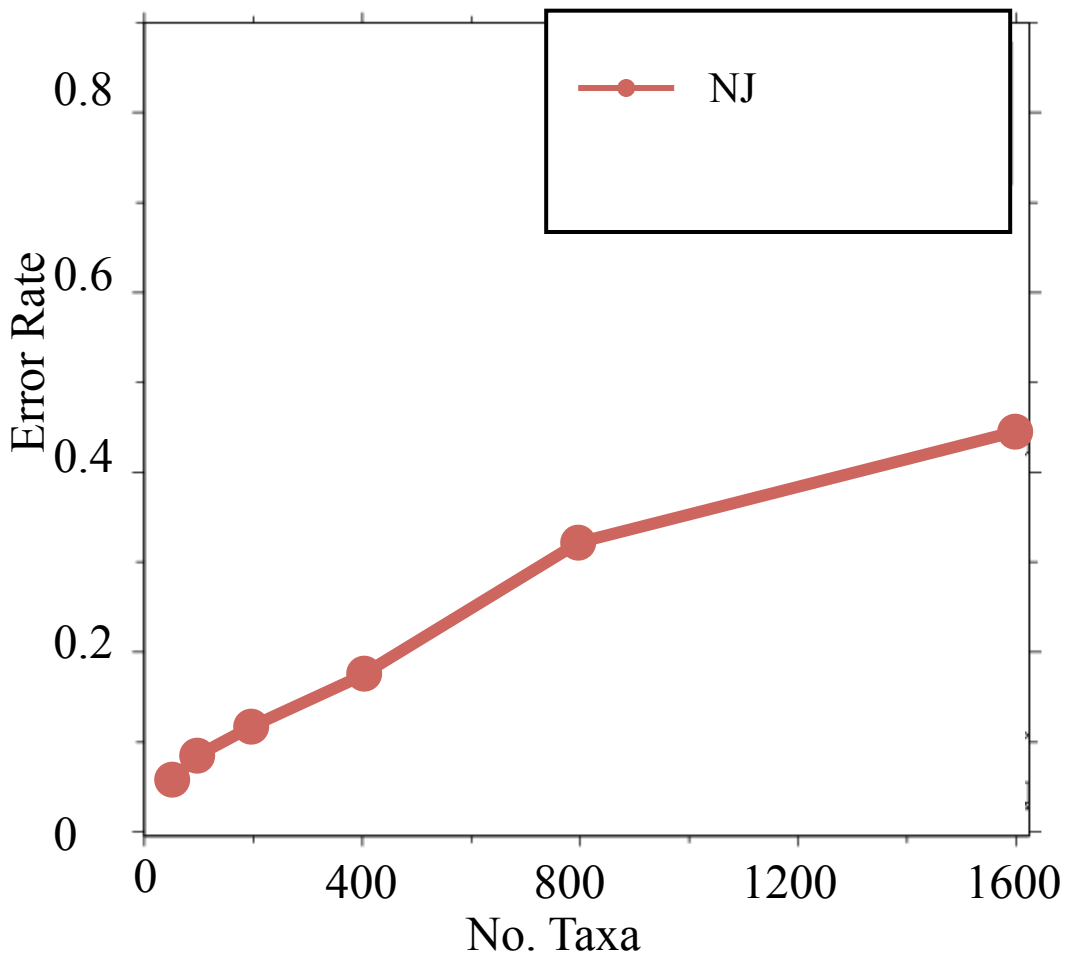
Statistical consistency, exponential convergence, and absolute fast convergence (afc)



Fast-converging methods (and related work)

- 1997: Erdos, Steel, Szekely, and Warnow (ICALP).
- 1999: Erdos, Steel, Szekely, and Warnow (RSA, TCS);
  Huson, Nettles and Warnow (J. Comp Bio.)
- 2001: Warnow, St. John, and Moret (SODA);
  Nakhleh, St. John, Roshan, Sun, and Warnow (ISMB)
  Cryan, Goldberg, and Goldberg (SICOMP);
  Csuros and Kao (SODA);
- 2002: Csuros (J. Comp. Bio.)
- 2006: Daskalakis, Mossel, Roch (STOC),
  Daskalakis, Hill, Jaffe, Mihaescu, Mossel, and Rao (RECOMB)
- 2007: Mossel (IEEE TCBB)
- 2008: Gronau, Moran and Snir (SODA)
- 2010: Roch (Science)
- 2013: Roch (in preparation)

Neighbor Joining on large diameter trees



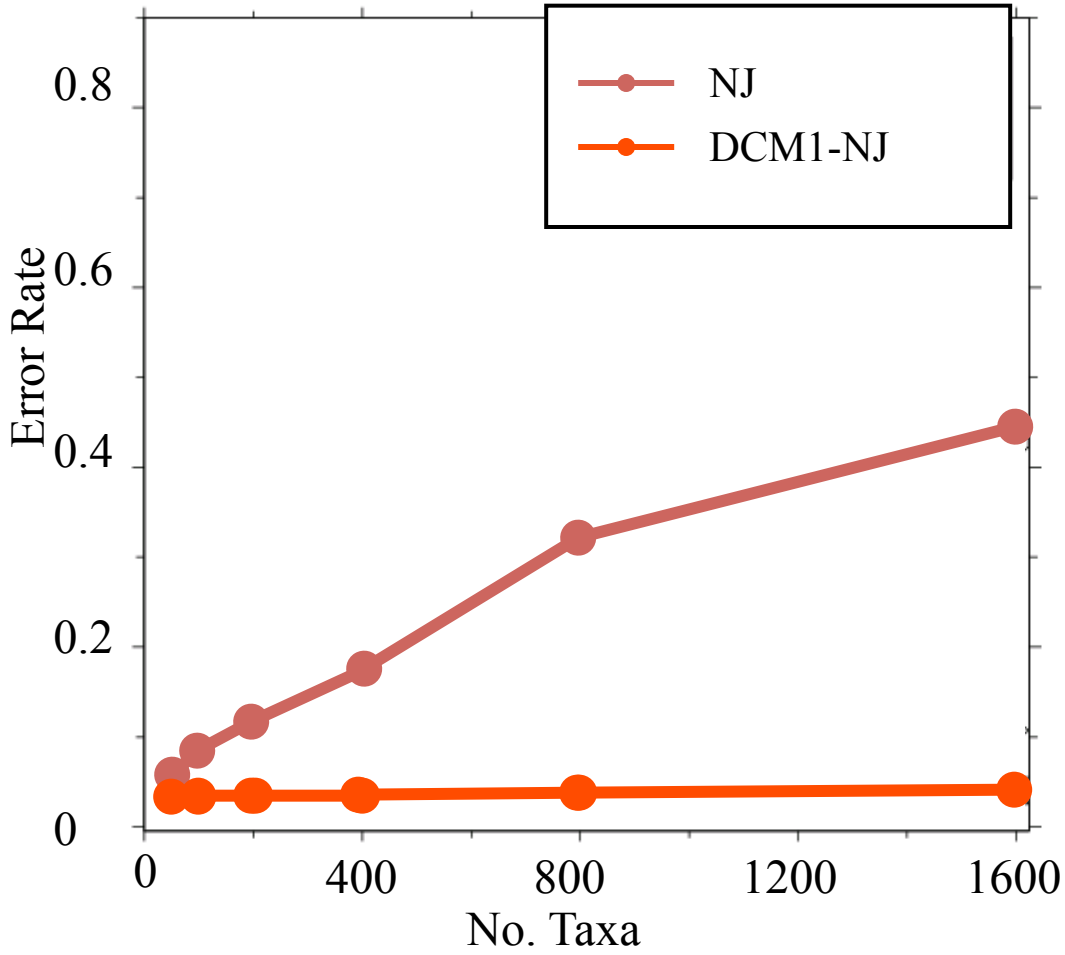
Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

[Nakhleh et al. ISMB 2001]

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. Hence DCM1-NJ is afc.

Key technique: chordal graph theory

Questions

- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model tree correctly (with high probability)?
- What is the computational complexity of an estimation problem?

Answers?

- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.

Answers?

- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.

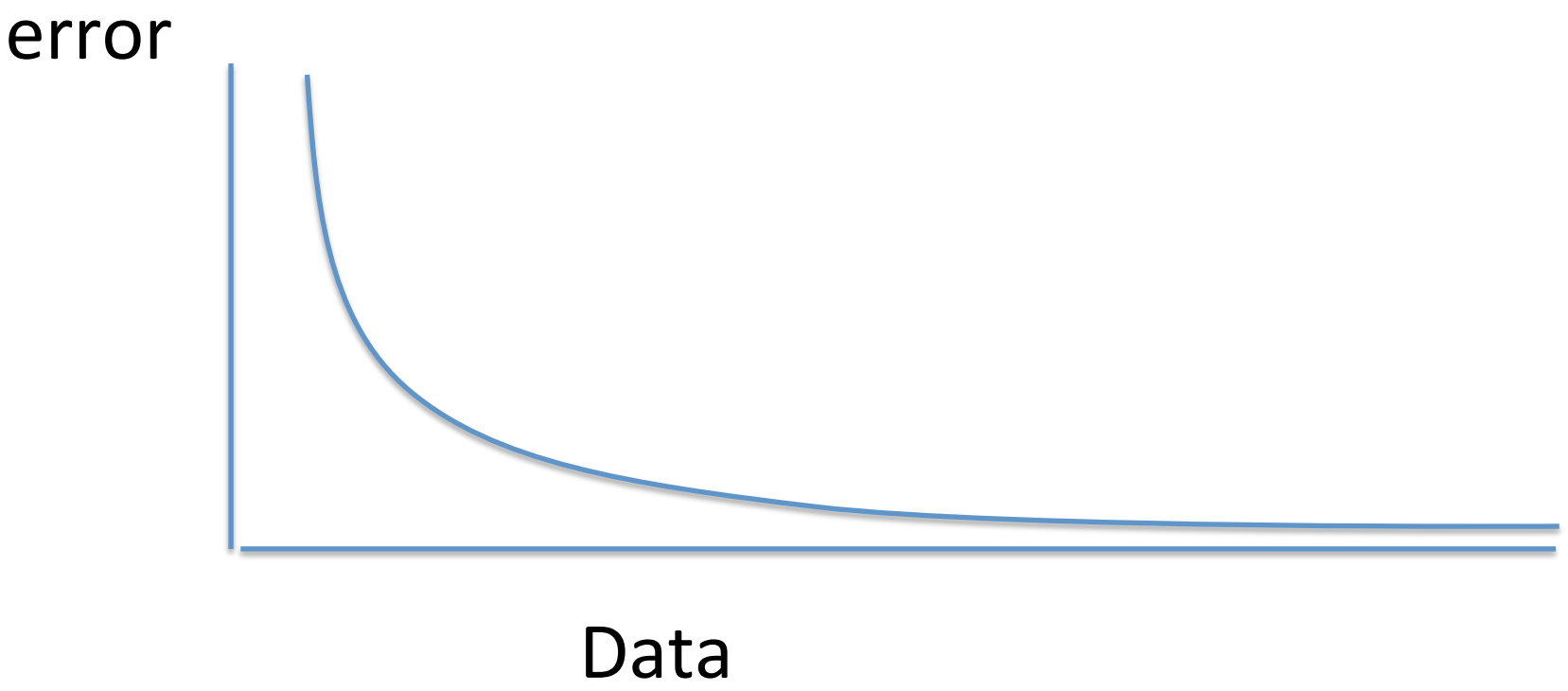
Answers?

- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.
- Just about everything is NP-hard, and the datasets are big.

Answers?

- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.
- Just about everything is NP-hard, and the datasets are big.
- Extensive studies show that even the best methods produce gene trees with some error.

In other words…

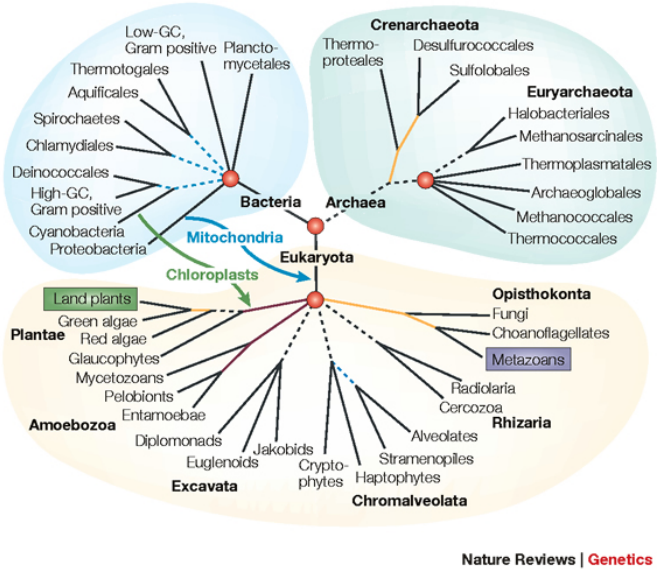


Statistical consistency doesn't guarantee accuracy w.h.p. unless the sequences ***are long enough.***

Part II: Species Tree Estimation from multiple genes

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Using multiple genes

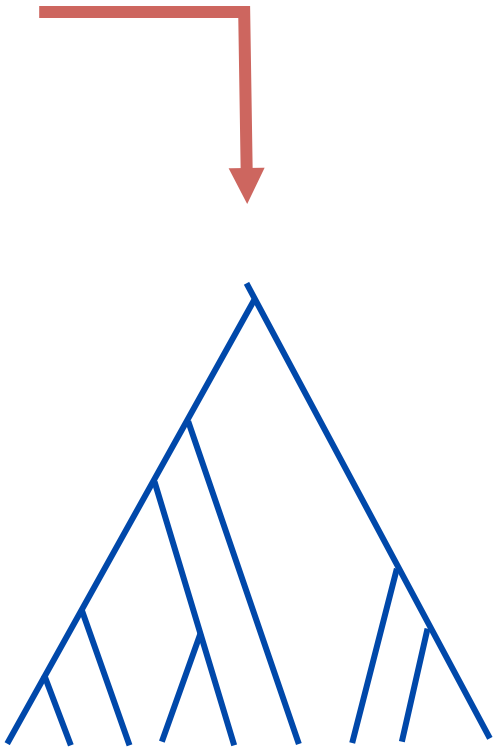
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

Concatenation

	gene 1	gene 2	gene 3
S_1	TCTAATGGAA	? ? ? ? ? ? ? ? ? ?	TATTGATACA
S_2	GCTAAGGGAA	? ? ? ? ? ? ? ? ? ?	? ? ? ? ? ? ? ? ? ?
S_3	TCTAAGGGAA	? ? ? ? ? ? ? ? ? ?	TCTTGATACC
S_4	TCTAACGGAA	GGTAACCCTC	TAGTGATGCA
S_5	? ? ? ? ? ? ? ? ? ?	GCTAAACCTC	? ? ? ? ? ? ? ? ? ?
S_6	? ? ? ? ? ? ? ? ? ?	GGTGACCATC	? ? ? ? ? ? ? ? ? ?
S_7	TCTAATGGAC	GCTAAACCTC	TAGTGATGCA
S_8	TATAACGGAA	? ? ? ? ? ? ? ? ? ?	CATTCATACC



Concatenation

Challenges:

- The most accurate tree estimation methods are heuristics for NP-hard problems (e.g., maximum likelihood).

Concatenation

Challenges:

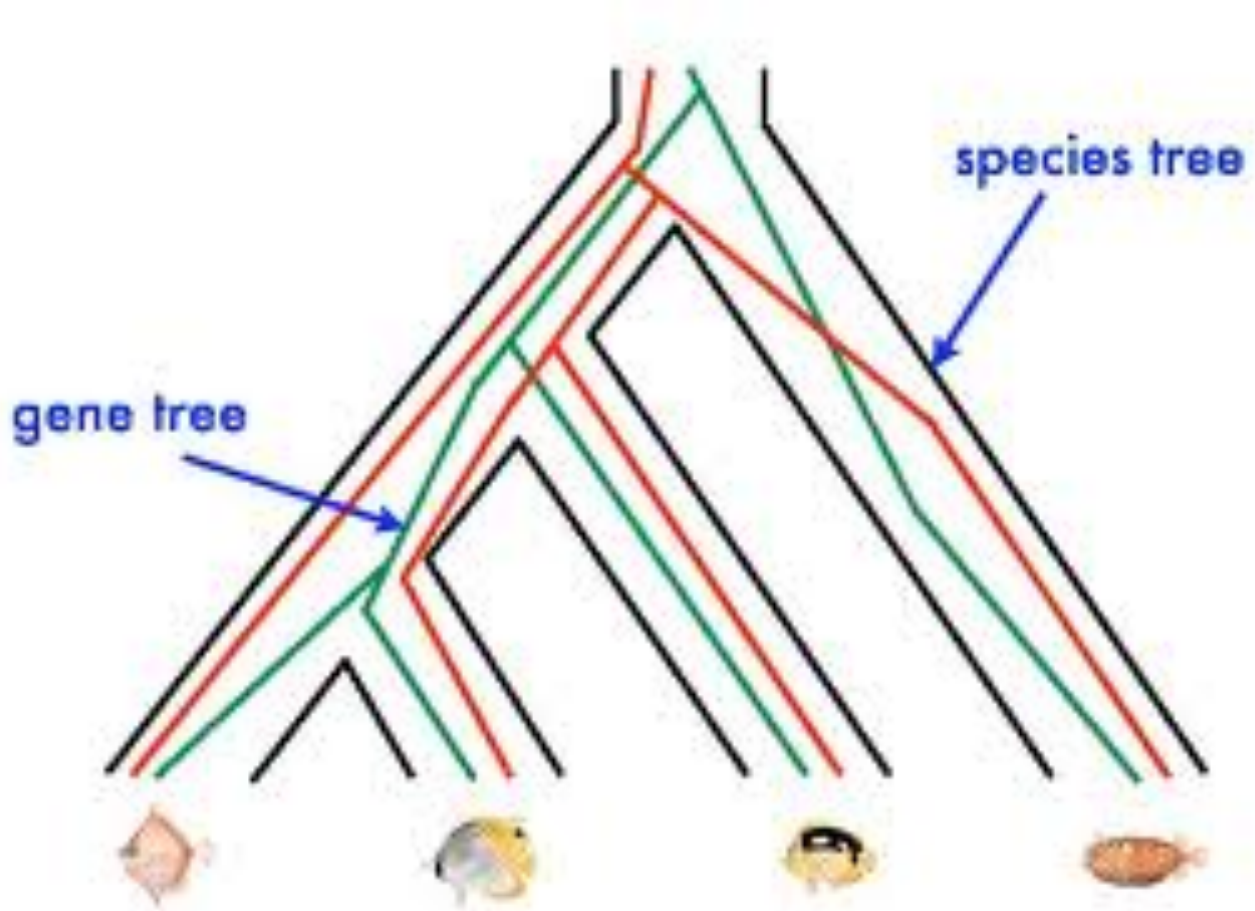
- The most accurate tree estimation methods are heuristics for NP-hard problems (e.g., maximum likelihood).
- The tree estimation can be computationally intensive if the concatenated alignment is very large (i.e., has many sequences or many sites).

Concatenation

Challenges:

- The most accurate tree estimation methods are heuristics for NP-hard problems (e.g., maximum likelihood).
- The tree estimation can be computationally intensive if the concatenated alignment is very large (i.e., has many sequences or many sites).
- Concatenation may not be statistically consistent if the genes evolve differently from the species!

Red gene tree ≠ species tree (green gene tree okay)

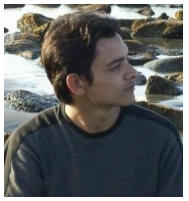


1KP: Thousand Transcriptome Project







G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin

Gene Tree Incongruence

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)
- iPLANT (NSF-funded cooperative)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Avian Phylogenomics Project

E Jarvis, HHMI

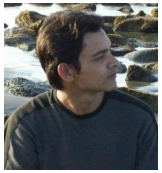
MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Gene Tree Incongruence

Plus many many other people…

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Part III: Species Tree Estimation

- Objective: species tree estimation from multiple incongruent gene trees

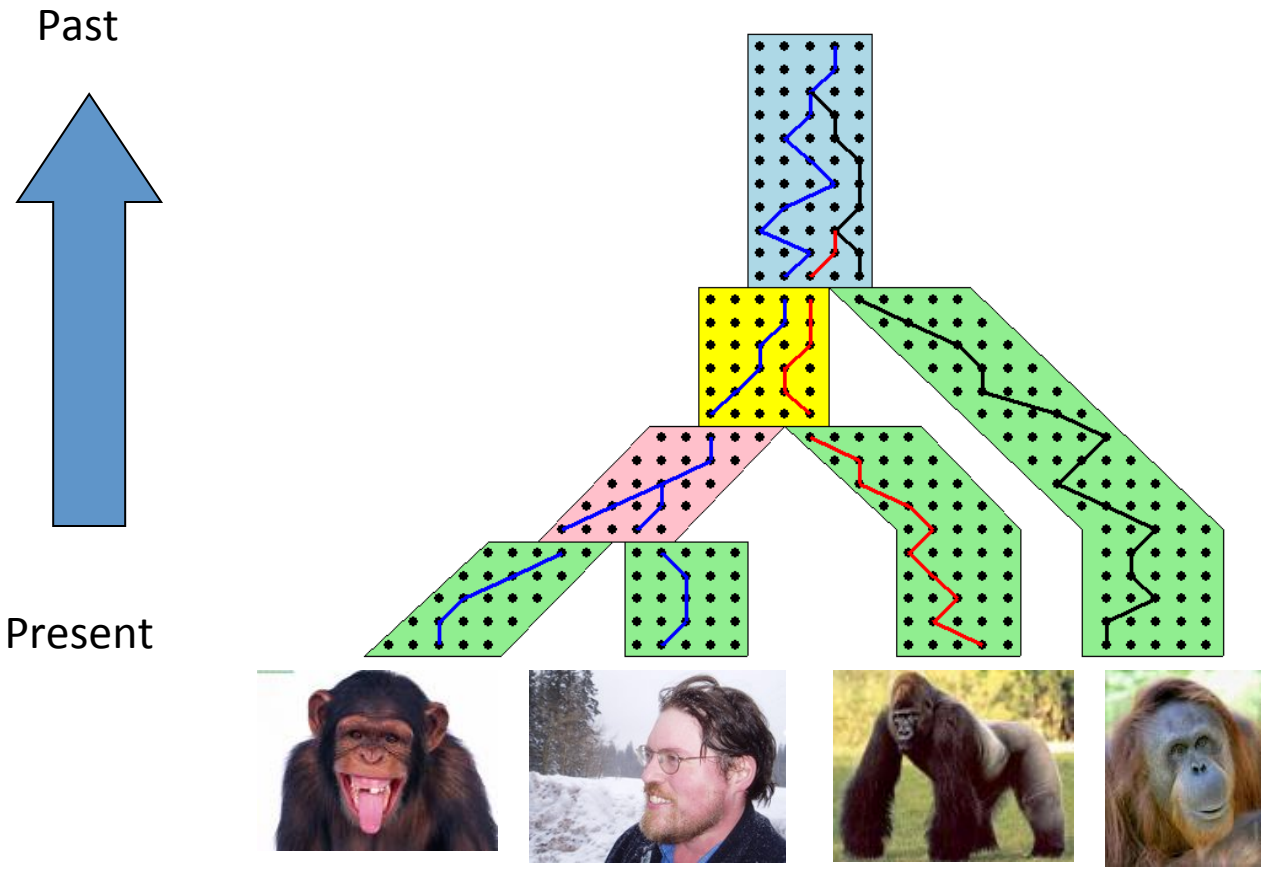
- Evaluation:
  - Theoretical guarantees
  - Performance on simulated data
  - Impact on biological data analysis

Gene Tree Incongruence

- Gene trees can differ from the species tree due to:
  - Duplication and loss
  - Horizontal gene transfer
  - Incomplete lineage sorting

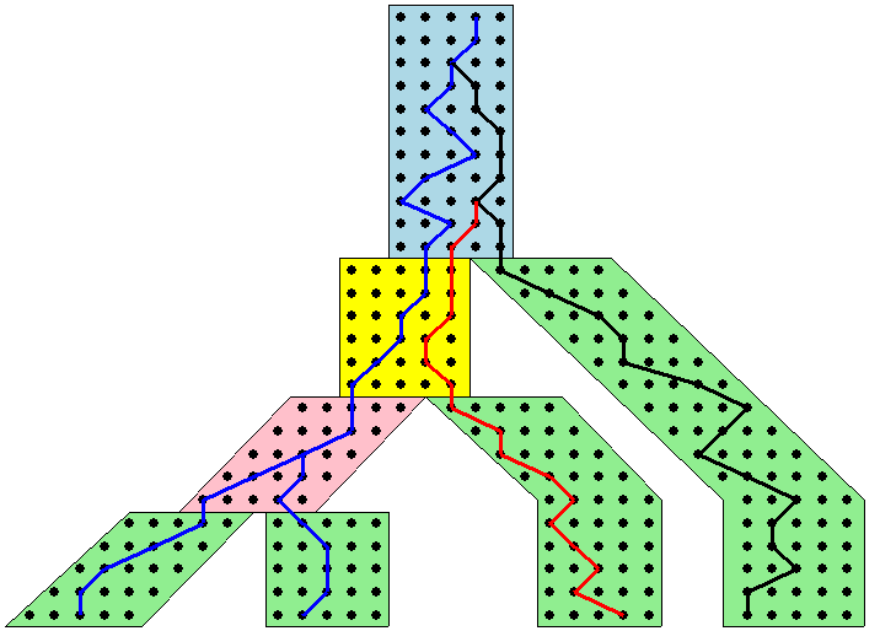
Multiple populations/species

Courtesy James Degnan



Gene tree in a species tree

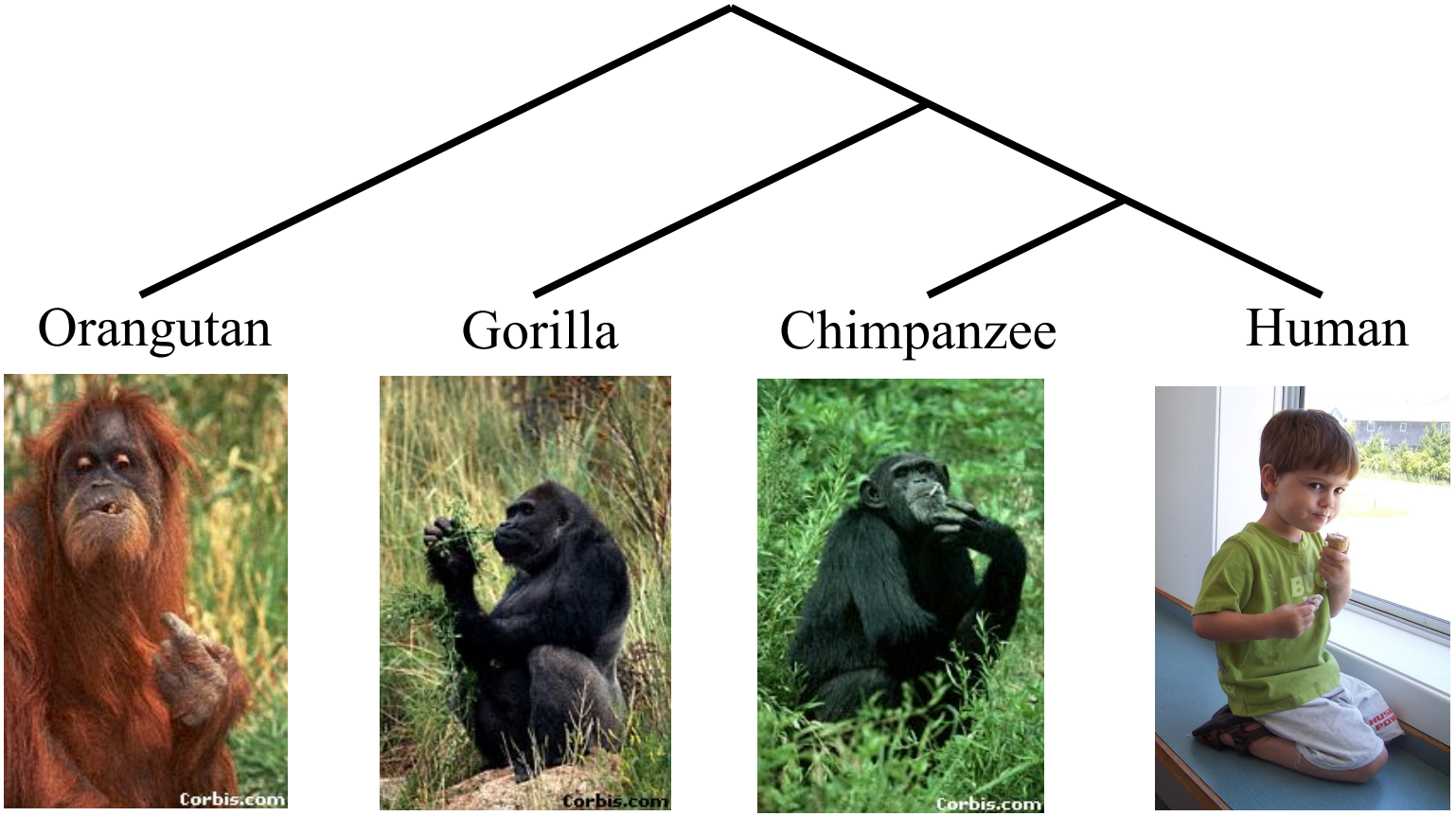
Courtesy James Degnan



Lineage Sorting

- Population-level process, also called the "Multi-species coalescent"
- Gene trees can differ from species trees due to short times between speciation events (population size also impacts this probability); this is called "Incomplete Lineage Sorting" or "Deep Coalescence".
- Causes difficulty in estimating some species trees (such as human-chimp-gorilla)

Phylogeny
(evolutionary tree)

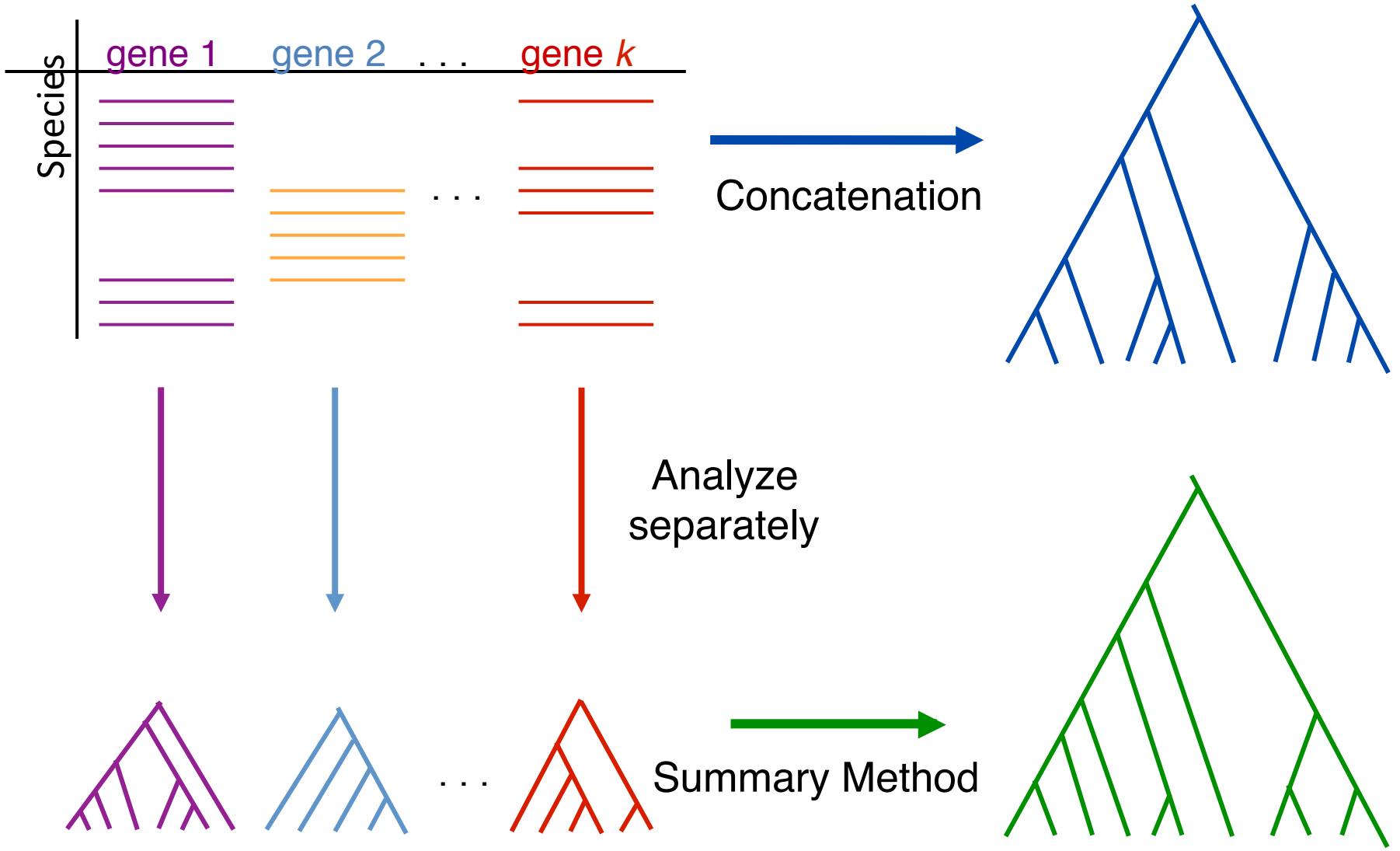


From the Tree of the Life Website,
University of Arizona

Incomplete Lineage Sorting (ILS)

- 2000+ papers in 2013 alone
- Confounds phylogenetic analysis for many groups:
  - Hominids
  - Birds
  - Yeast
  - Animals
  - Toads
  - Fish
  - Fungi
- There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS.

Two competing approaches



How to compute a species tree?

How to compute a species tree?

How to compute a species tree?

Techniques:

- Most frequent gene tree?
- Consensus of gene trees?
- Other?

How to compute a species tree?

Theorem (Degnan et al., 2006, 2009):
Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on {A,B,C} is identical to the species tree induced on {A,B,C}.

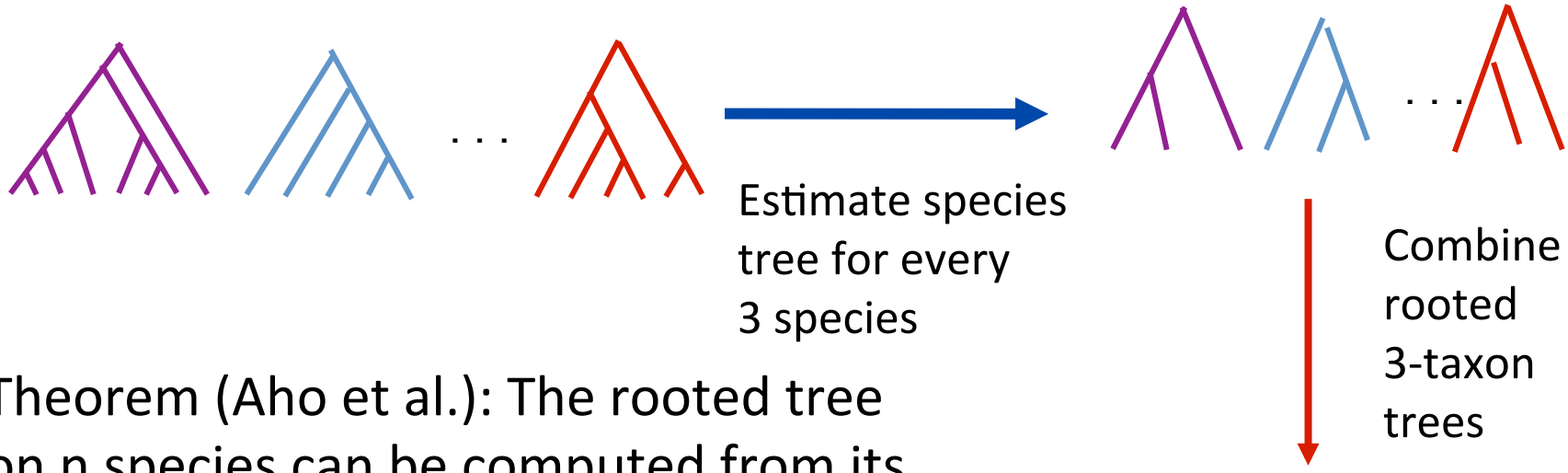
How to compute a species tree?

Theorem (Degnan et al., 2006, 2009):
Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on {A,B,C} is identical to the species tree induced on {A,B,C}.

How to compute a species tree?

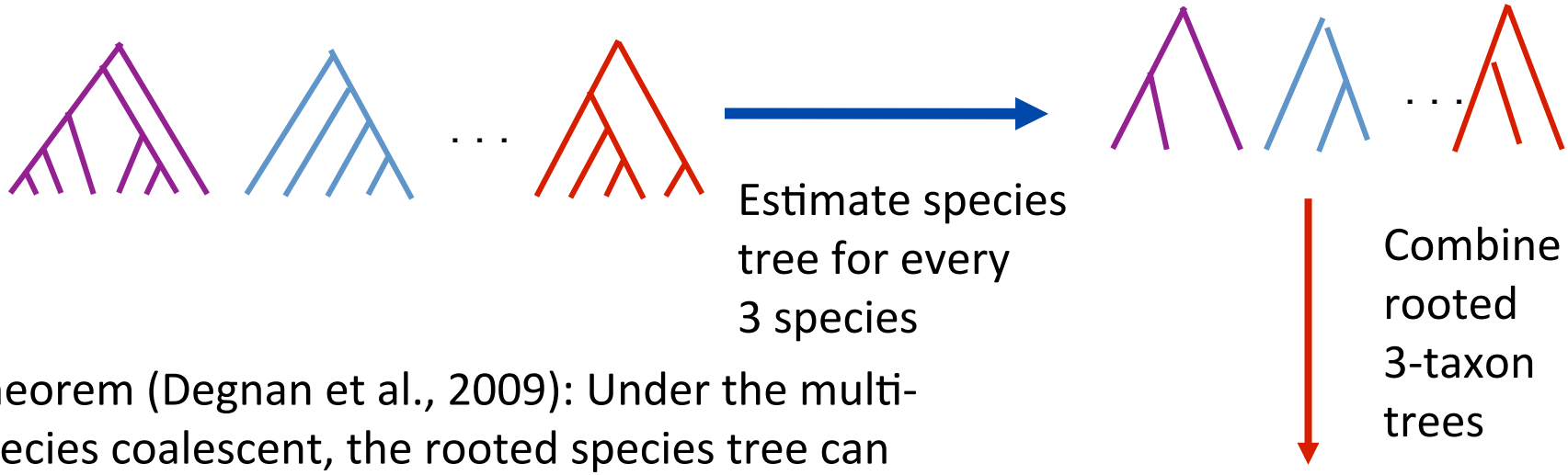
Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.

How to compute a species tree?



Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.

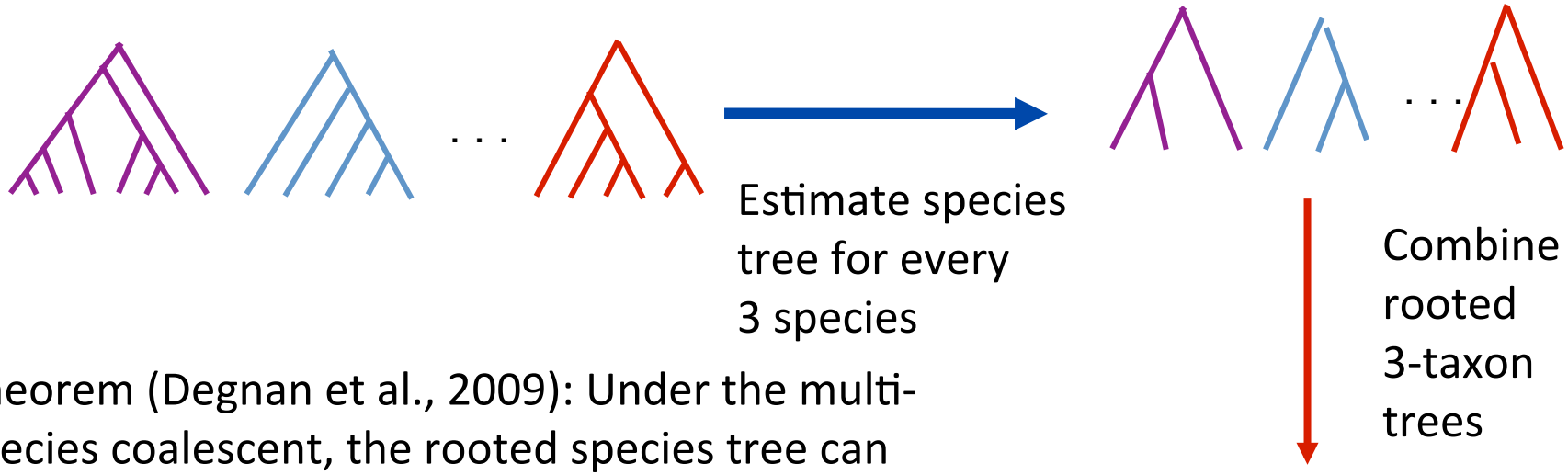
How to compute a species tree?



Theorem (Degnan et al., 2009): Under the multi-species coalescent, the rooted species tree can be estimated correctly (with high probability) given a large enough number of true rooted gene trees.

Theorem (Allman et al., 2011): the unrooted species tree can be estimated from a large enough number of true unrooted gene trees.

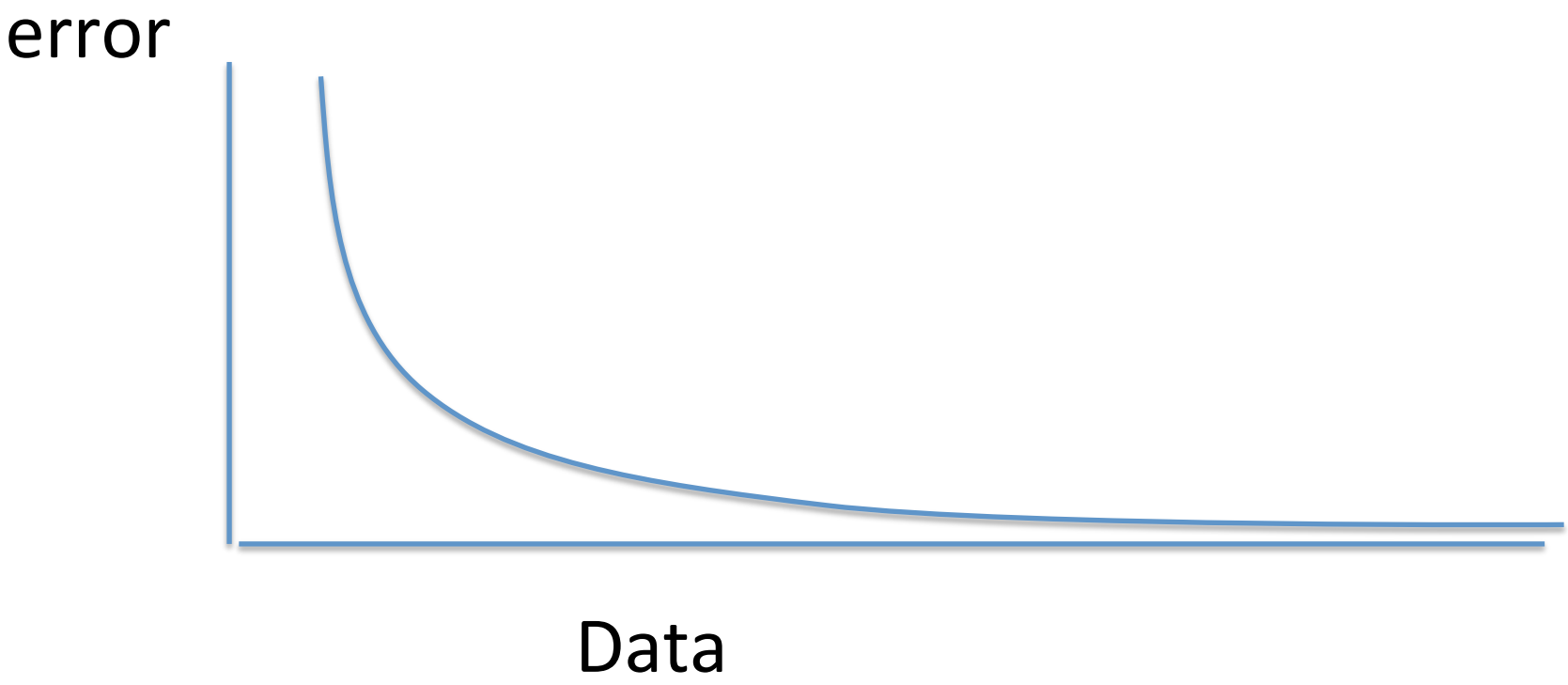
How to compute a species tree?



Theorem (Degnan et al., 2009): Under the multi-species coalescent, the rooted species tree can be estimated correctly (with high probability) given a large enough number of true rooted gene trees.

Theorem (Allman et al., 2011): the unrooted species tree can be estimated from a large enough number of true unrooted gene trees.

Statistical Consistency



Data are gene trees, presumed to be randomly sampled <u>true gene trees.</u>

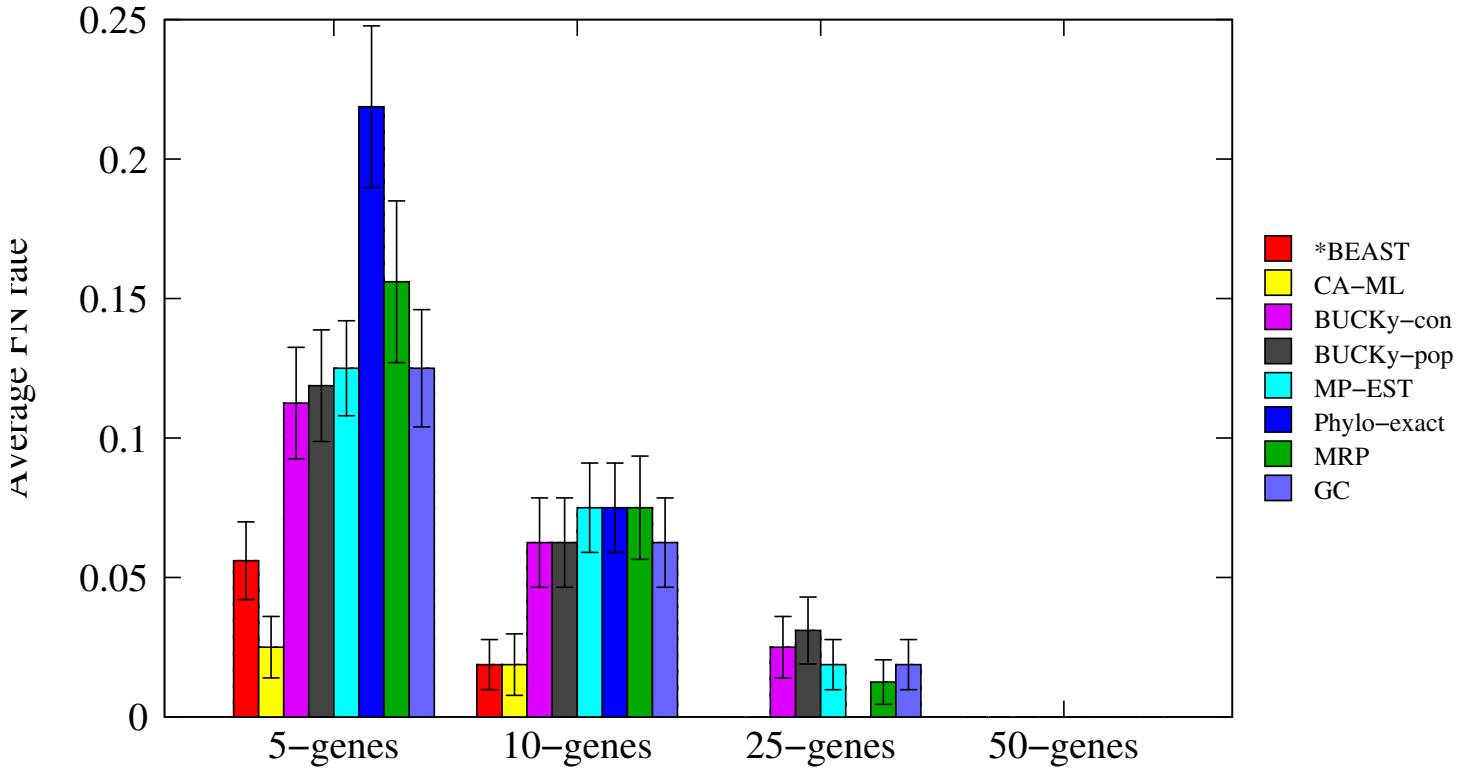
Statistically consistent methods

Input: Set of estimated gene trees or alignments, one (or more) for each gene

Output: estimated species tree

- ***BEAST** (Heled and Drummond 2010): Bayesian co-estimation of gene trees and species trees given sequence alignments
- **MP-EST** (Liu et al. 2010): maximum likelihood estimation of rooted species tree
- **BUCKy-pop** (Ané and Larget 2010): quartet-based Bayesian species tree estimation

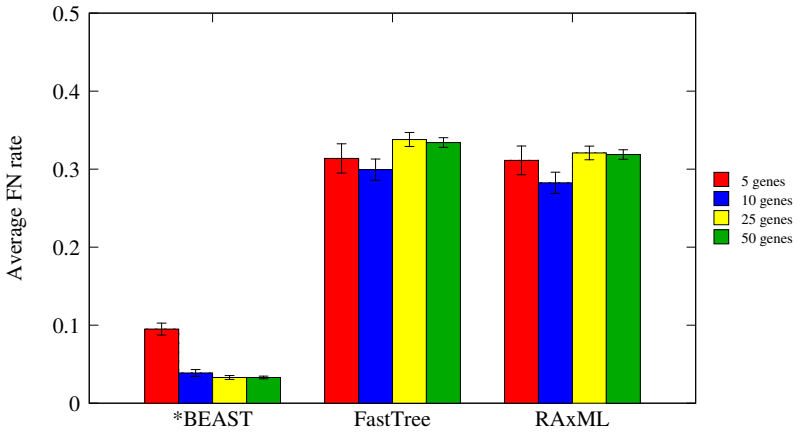
Results on 11-taxon datasets with weak ILS



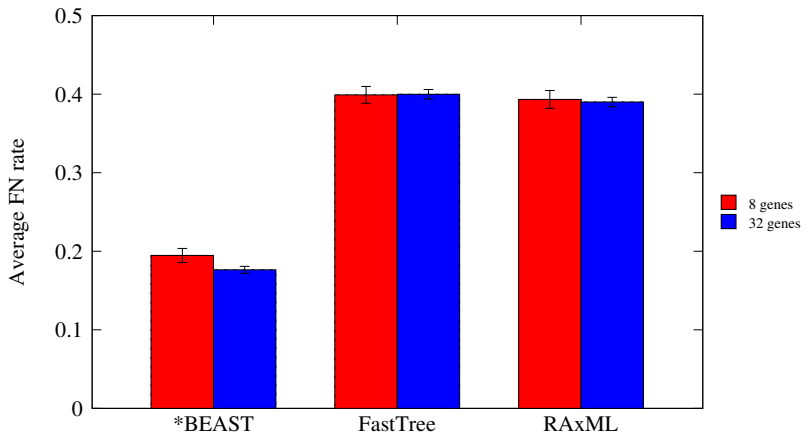
***BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

*BEAST better than Maximum Likelihood



11-taxon weakILS datasets

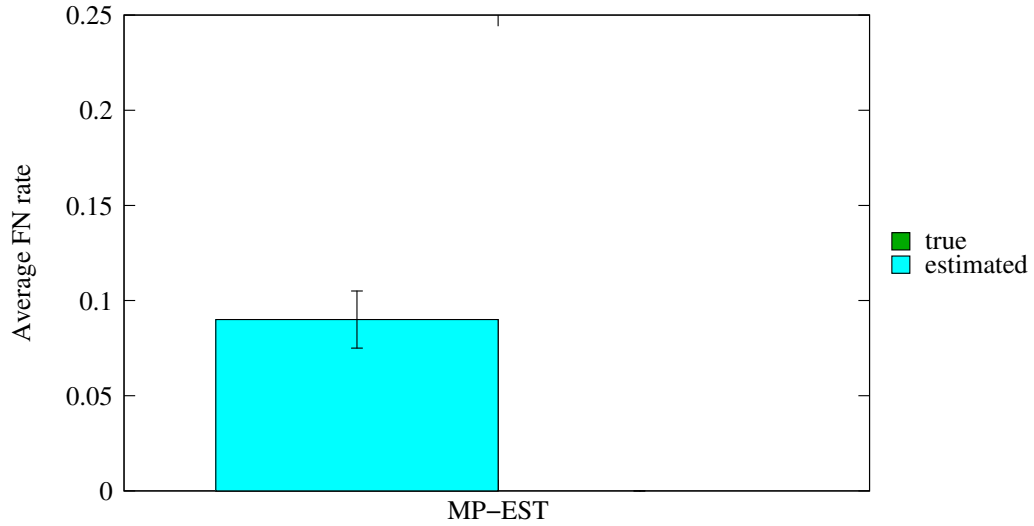


17-taxon (very high ILS) datasets

*BEAST produces more accurate gene trees than ML on gene sequence alignments

11-taxon datasets from Chung and Ané, Syst Biol 2012
17-taxon datasets from Yu, Warnow, and Nakhleh, JCB 2011

Impact of Gene Tree Estimation Error on MP-EST



MP-EST has no error on true gene trees, but
MP-EST has 9% error on estimated gene trees

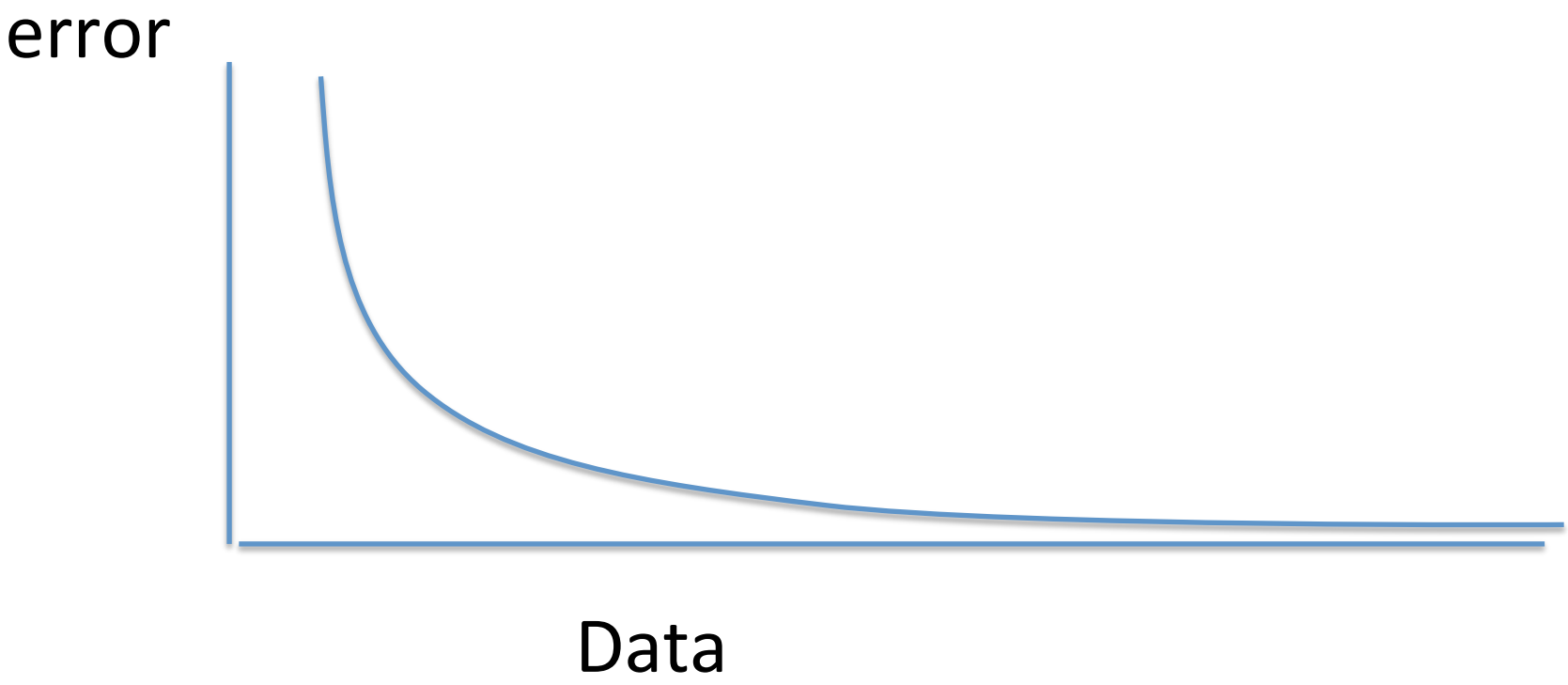
Datasets: 11-taxon strongILS conditions with 50 genes

Similar results for other summary methods (MDC, Greedy, etc.).

Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
- Species trees obtained by combining poorly estimated gene trees have poor accuracy.

Statistical Consistency



Data are gene trees, presumed to be randomly sampled <u>true gene trees.</u>

Questions

- Is the model species tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model species tree correctly (with high probability)?
- What is the computational complexity of an estimation problem?

Questions

- Is the model species tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model species tree correctly (with high probability)?
- What is the computational complexity of an estimation problem?
- What is the impact of error in the input data on the estimation of the model species tree?

Addressing gene tree estimation error

- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- “Bin-and-conquer”

Addressing gene tree estimation error

- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- “Bin-and-conquer”

Technique #1: Modify gene trees

- Idea: Use statistical technique to identify unreliable aspects of the tree, and modify tree, to produce “constraint tree”.
- Example: ignore root, collapse edges in the tree with poor statistical support, remove “rogue taxa”.
- Note: need to modify the optimization problems and algorithms.

MDC Problem

- MDC (minimize deep coalescence) problem:
  - given set of true gene trees, find the species tree that implies the *fewest deep coalescence events*

- Posed by Wayne Maddison, Syst Biol 1997
- NP-hard
- Than and Nakhleh (PLoS Comp Biol 2009) gave Dynamic Programming algorithm for MDC on binary, rooted gene trees on same set of taxa

MDC*: Extending MDC

Input: Set X of k *unrooted, not necessarily binary, not necessarily complete* gene trees on taxon set S (and optionally a set C of bipartitions on the taxon set).

Output: Species tree T with Bipartitions(T) drawn from C, and set X* = {t*: t in X}, where each t* is a rooted refinement of t, so as to optimize MDC(X*,T).

MDC*: Extending MDC

Input: Set X of k *unrooted, not necessarily binary, not necessarily complete* gene trees on taxon set S (and optionally a set C of bipartitions on the taxon set).

Output: Species tree T with Bipartitions(T) drawn from C, and set X* = {t*: t in X}, where each t* is a rooted refinement of t, so as to optimize MDC(X*,T).

Thus, each tree t in X is a constraint on the gene tree t*, and C is a constraint on the species tree T. If C is not provided, then C = {all bipartitions on S}.

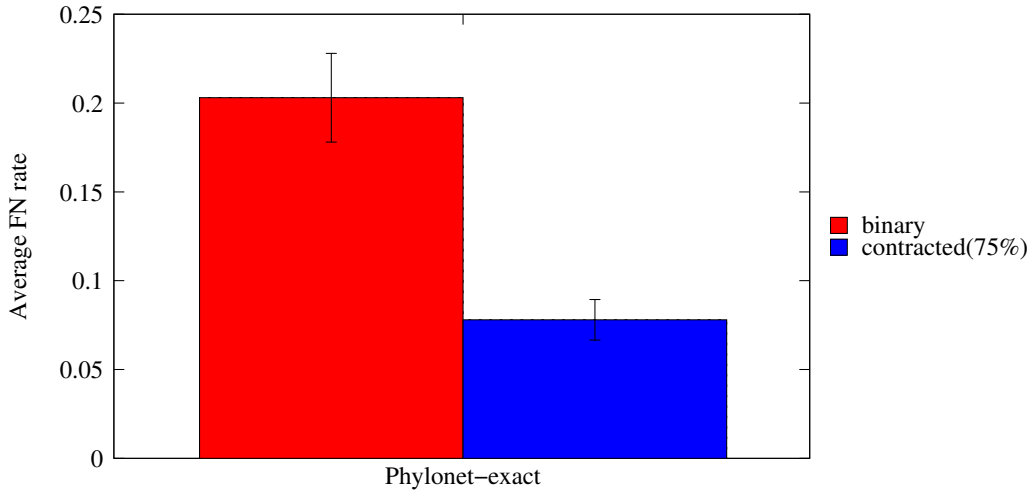
Solving MDC*

Theorem (Yu, Warnow, and Nakhleh, 2011): The optimal solution to constrained MDC* can be found in $O(|C|^2nk)$ time, where $|S|=n$ and k is the number of gene trees. Hence the optimal solution to MDC* can be found in $O(2^{2n}nk)$.

Bayzid and Warnow J Comp Biol 2012 proves that the YWN 2011 DP algorithm correctly handles incomplete gene trees.

MDC* implemented in Phylonet package (Nakhleh et al.).
Default heuristic uses C = {bipartitions in input gene trees}, and is polynomial in n and k.

MDC vs. MDC*: Impact of collapsing low support branches



- 11-taxon datasets with strongILS, 50 gene trees estimated using maximum likelihood.
- Phylo-exact solves MDC* optimally, with contracted gene trees are based on 75% bootstrap support threshold.
- **Similar improvements shown for some other methods.**

Technique #2: Bin-and-Conquer?

1. Assign genes to "bins"

2. Estimate trees on each supergene alignment using ML

3. Combine the supergene trees together using a summary method

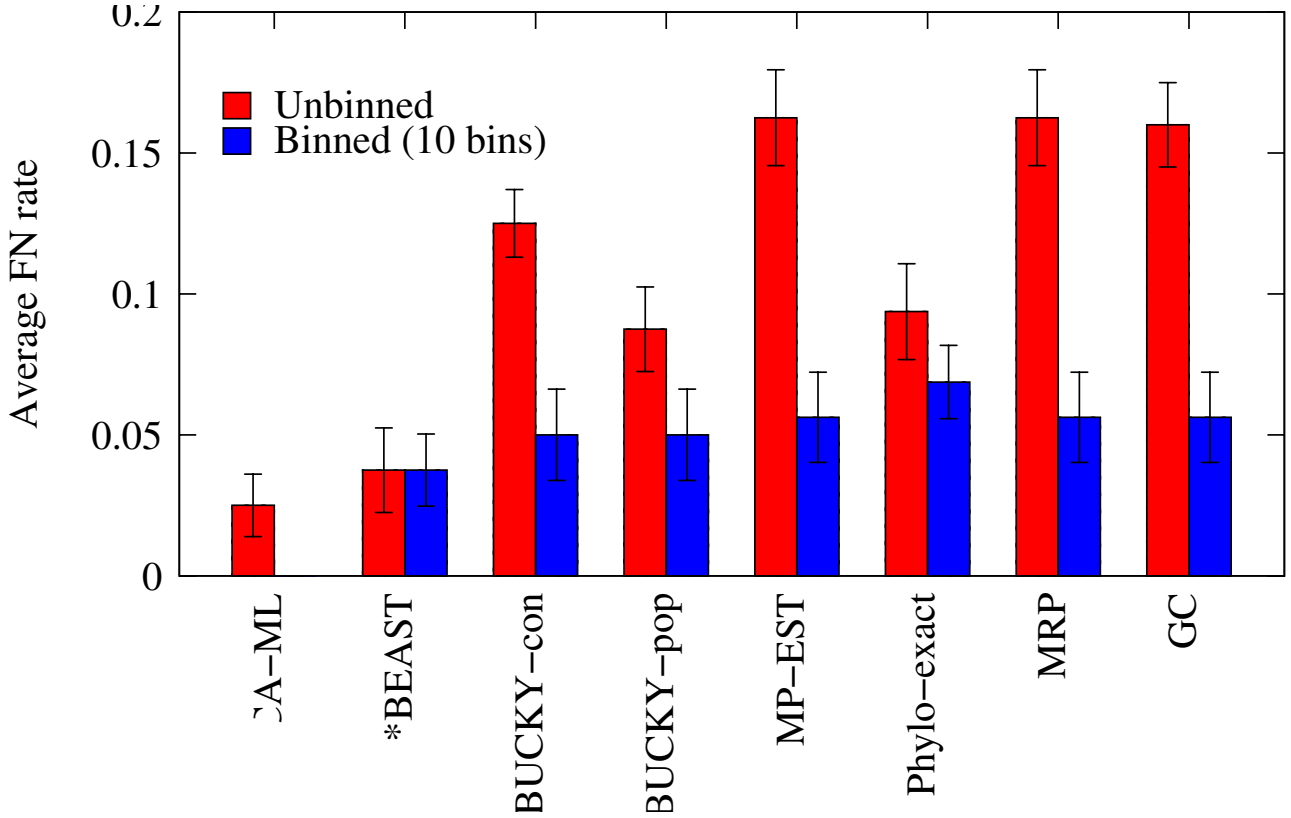
   or

   Run *BEAST on the new supergene alignments.

Variants:

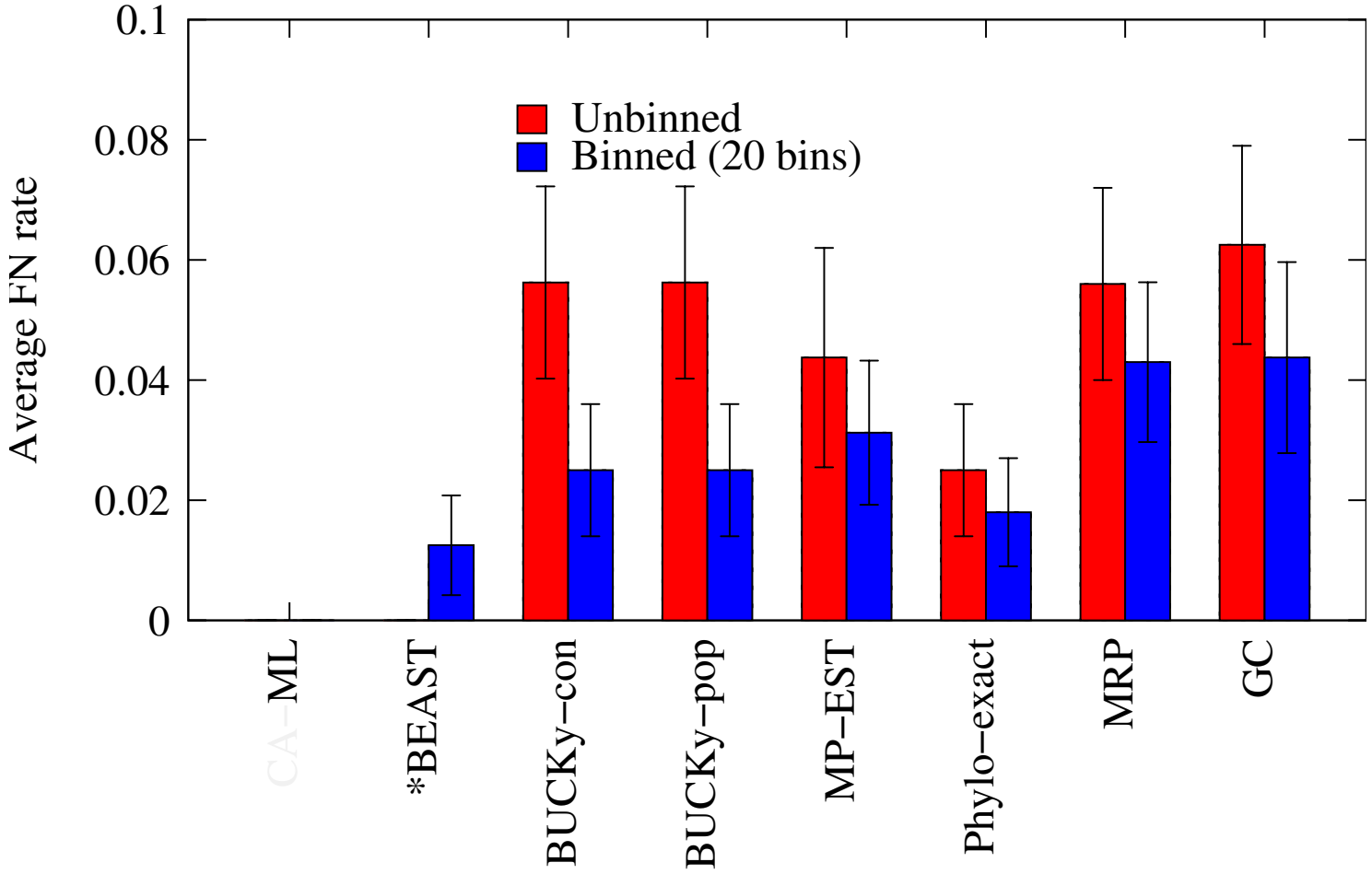
- Naïve binning (Bayzid and Warnow, Bioinformatics 2013)
- Statistical binning (Mirarab, Bayzid, and Warnow, in preparation)

Naïve binning vs. unbinned: 50 genes



Bayzid and Warnow, Bioinformatics 2013
11-taxon strongILS datasets with 50 genes, 5 genes per bin

Naïve binning vs. unbinned, 100 genes



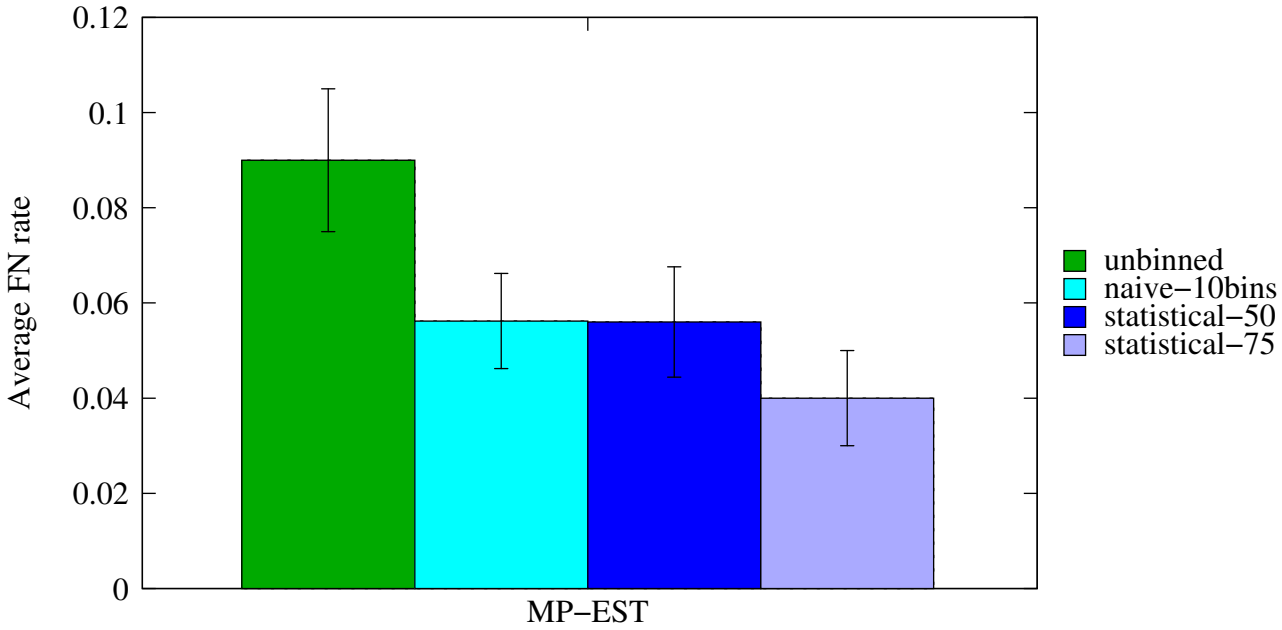
*BEAST did not converge on these datasets, even with 150 hours.
With binning, it converged in 10 hours.

Statistical binning

- Naïve binning does not consider "combinability" – bins are randomly defined, but have the same size.
- We are testing statistically-informed binning techniques, that check for combinability before binning. This creates a graph, in which vertices correspond to genes and edges correspond to "not combinable".
- We use a heuristic for "minimum balanced vertex coloring" to produce the bins.

Mirarab, Bayzid, and Warnow (in preparation)

Impact of binning on MP-EST



Binning improves MP-EST, and statistical binning can be better than naive

11-taxon strong ILS datasets, 50 genes
Statistical binning based on two different thresholds for “combinability”

Avian Phylogenomics Project

E Jarvis, HHMI

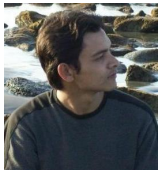
MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Gene Tree Incongruence

Plus many many other people…

- Strong evidence for substantial ILS, suggesting need for coalescent-based species tree estimation.
- But MP-EST on full set of 14,000 gene trees was considered unreliable, due to poorly estimated exon trees (very low phylogenetic signal in exon sequence alignments).

Avian Phylogeny

- RAxML analysis of 37 million bp alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.
- Extremely computationally intensive! More than 6000 days (150,000 hours) of compute time, and 256 GB. Run at HPC centers.
- Statistical binning version of MP-EST on 14000+ gene trees – highly resolved tree, slightly lower bootstrap support.
- Largely congruent with the concatenated analysis.
- Very fast to compute (after the gene trees are computed).

To consider

- Binning *reduces the amount* of data (number of gene trees) but can improve the accuracy of individual "supergene trees". The response to binning differs between methods. Thus, there is a trade-off between data *quantity* and *quality, and not all methods respond the same to the trade-off*.

- We know very little about the impact of data *error* on methods. We do not even have proofs of statistical consistency in the presence of data error.

Basic Questions

- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model tree correctly (with high probability)?
- What is the computational complexity of an estimation problem?

Additional Statistical Questions

- Trade-off between data quality and quantity
- Impact of data selection
- Impact of data error
- Performance guarantees on finite data (e.g., prediction of error rates as a function of the input data and method)

We need a solid mathematical framework for these problems.

Summary

- DCM1-NJ: an absolute fast converging (afc) method, uses chordal graph theory and probabilistic analysis of algorithms to prove performance guarantees

- MDC*: species tree estimation from multiple gene trees, uses graph theory to prove performance guarantee

- Binning: species tree estimation from multiple genes, suggests new questions in statistical estimation

Other Research in my lab

Method development for

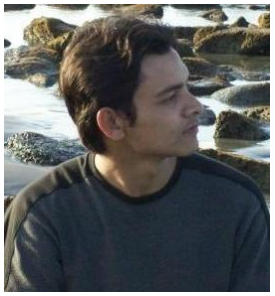
- Supertree estimation
- Multiple sequence alignment
- Metagenomic taxon identification
- Genome rearrangement phylogeny
- Historical Linguistics

Techniques:

- Statistical estimation under Markov models of evolution
- Graph theory and combinatorics
- Machine learning and data mining
- Heuristics for NP-hard optimization problems
- High performance computing
- Massive simulations

Warnow Laboratory





PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: http://www.cs.utexas.edu/users/phylo

Funding: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, and TACC (Texas Advanced Computing Center)

TACC and UTCS computational resources

* Supported by HHMI Predoctoral Fellowship

** Supported by Fulbright Foundation Predoctoral Fellowship