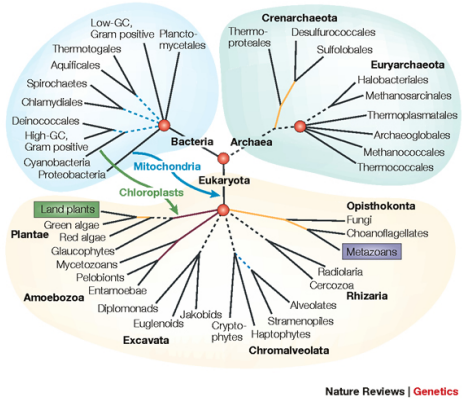
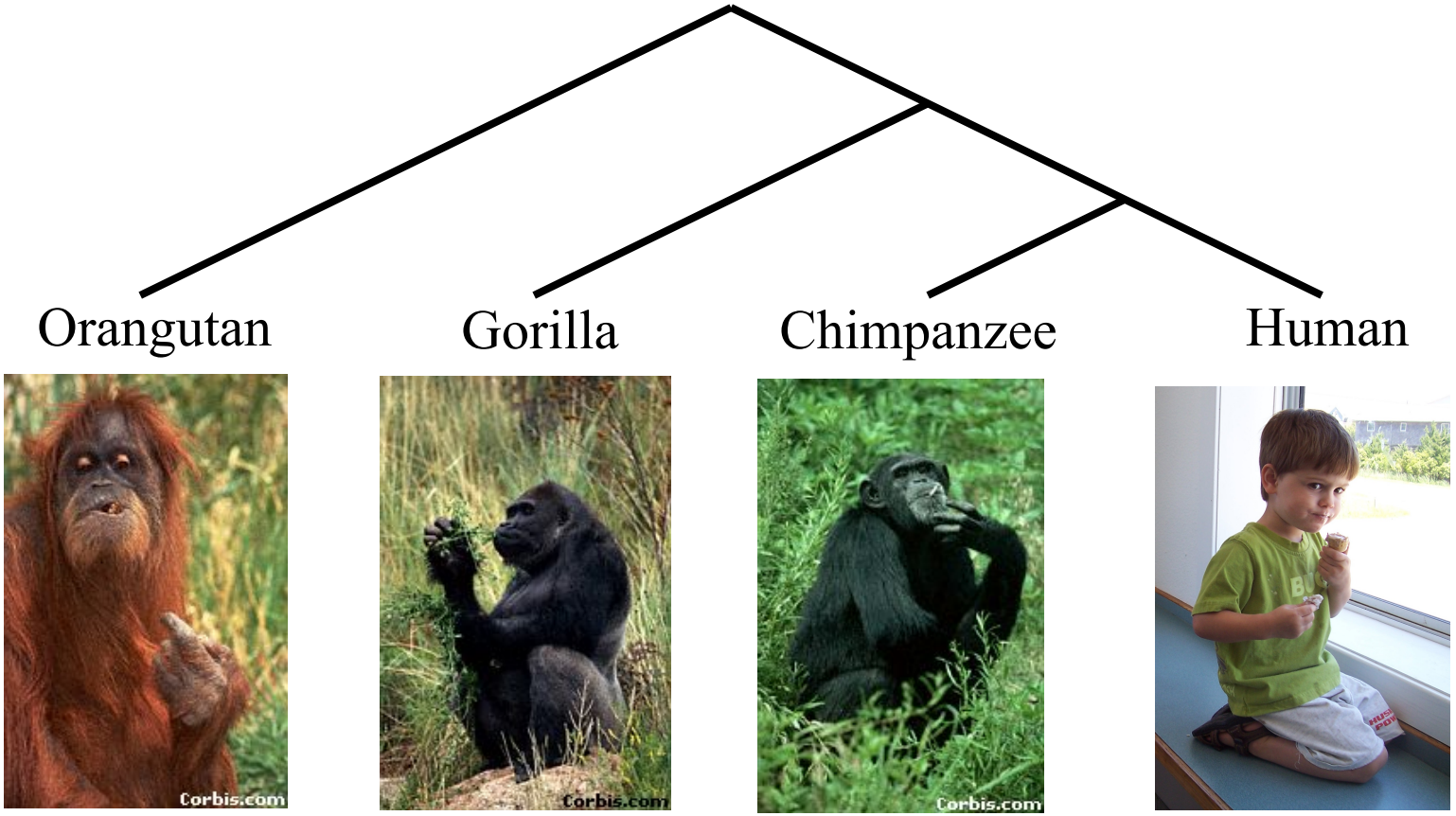


From Gene Trees to Species Trees

Tandy Warnow

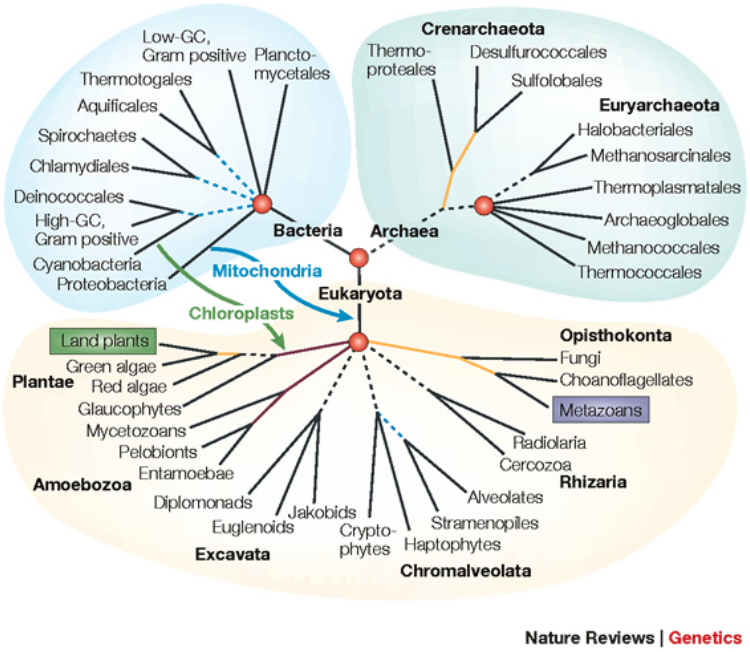
The University of Texas at Austin

Phylogeny
(evolutionary tree)



From the Tree of the Life Website,
University of Arizona

The Tree of Life: Importance to Biology



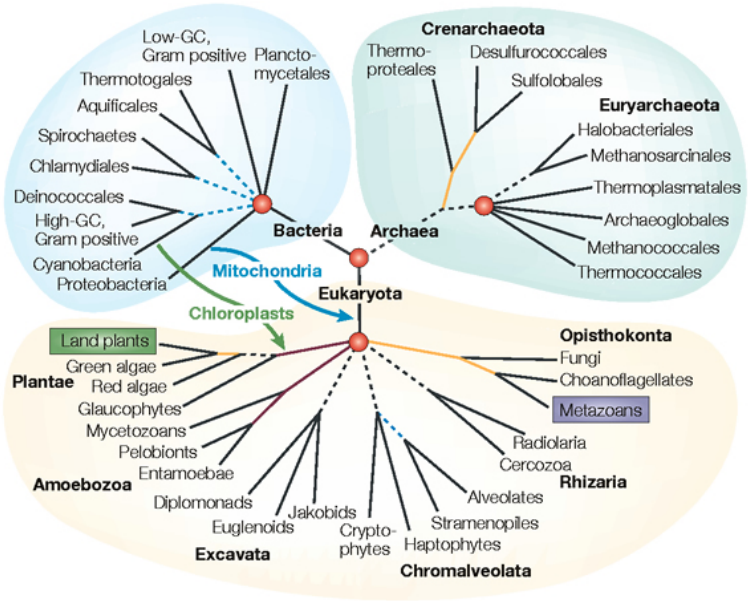
Nature Reviews | Genetics

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*



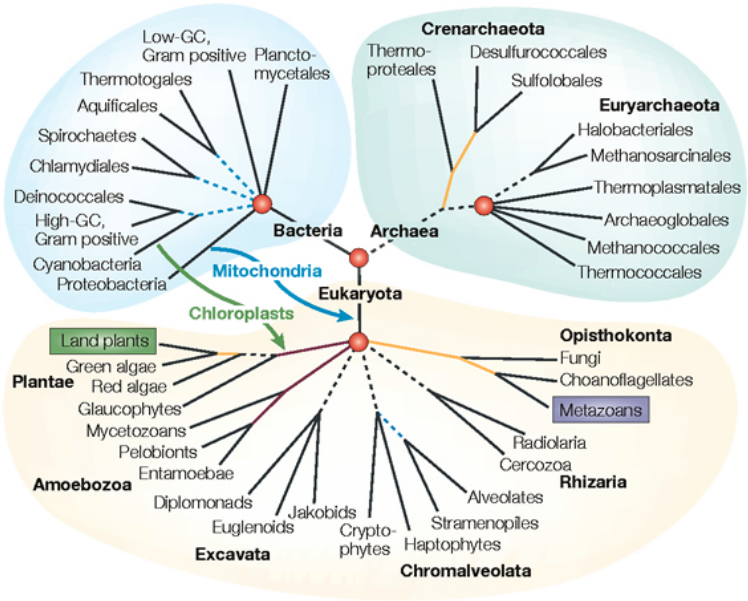
Nature Reviews | Genetics

Most well studied problem:

Given DNA sequences, find the Maximum Likelihood Tree

NP-hard, lots of software (RAxML, FastTree-2, GARLI, etc.)

Estimating The Tree of Life: a *Grand Challenge*



Nature Reviews | Genetics

Novel techniques needed for scalability and accuracy:

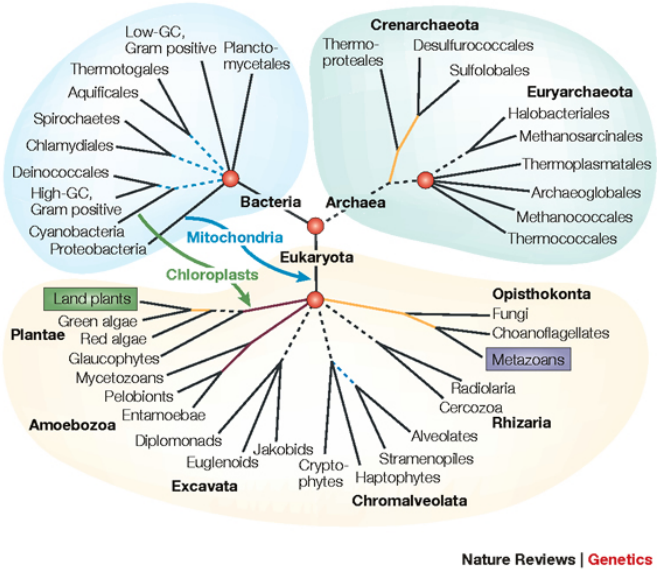
NP-hard problems and large datasets

Current methods not good enough on large datasets

HPC is necessary but not sufficient

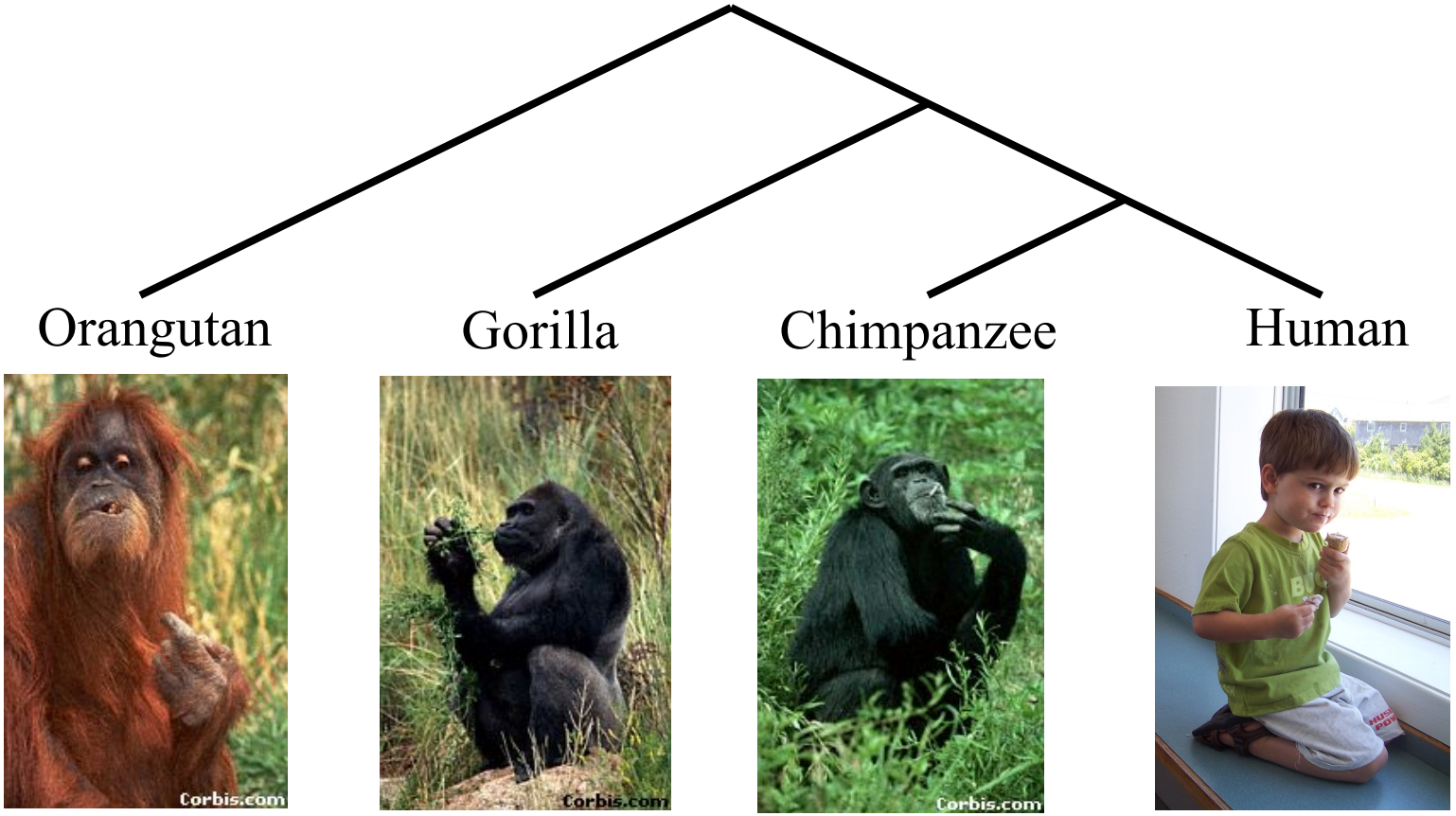
Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Sampling multiple genes from multiple species



From the Tree of the Life Website,
University of Arizona

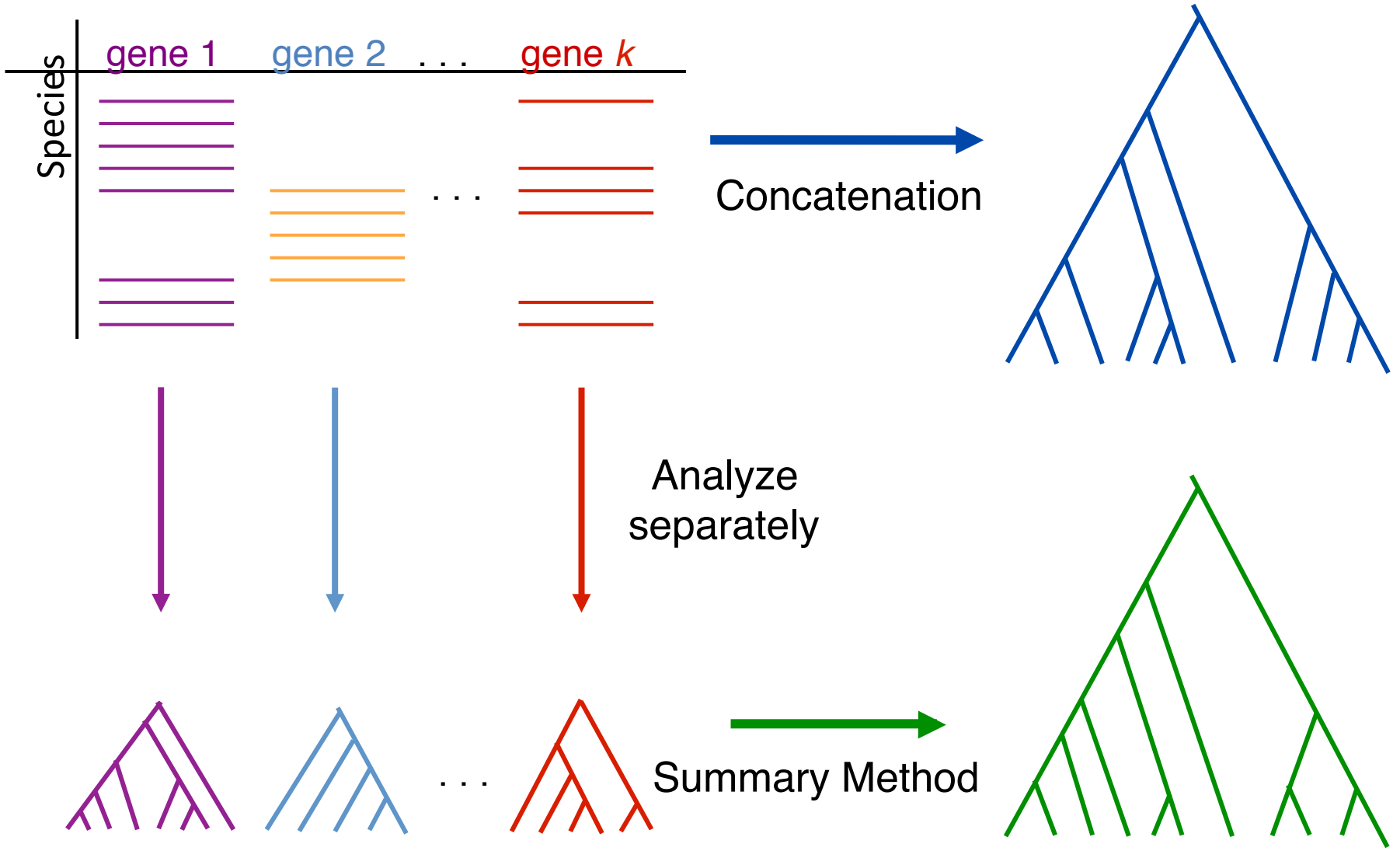
Using multiple genes

	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

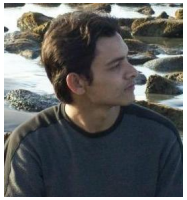
N. Matasci
iPlant

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)
- Gene sequence alignments and trees computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Challenges:

Multiple sequence alignments of > 100,000 sequences

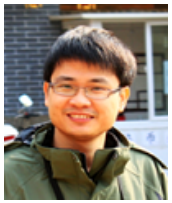
Gene tree incongruence

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

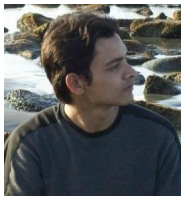
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

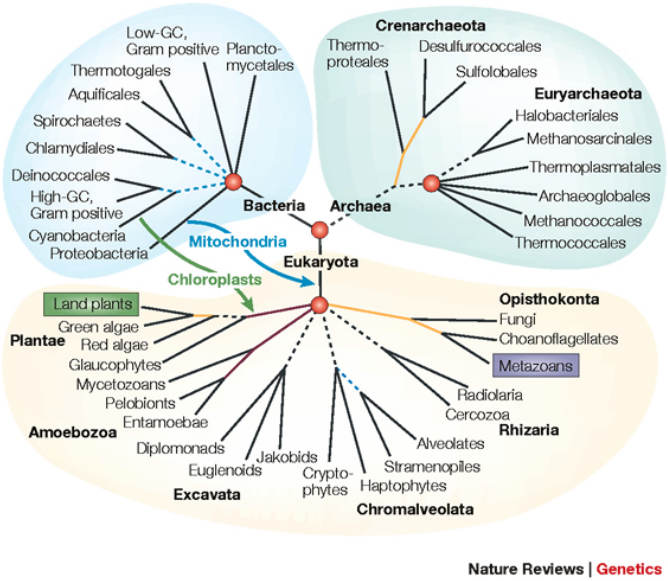
- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Challenges:

Maximum likelihood on multi-million-site sequence alignments

Massive gene tree incongruence

The Tree of Life: *Multiple 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

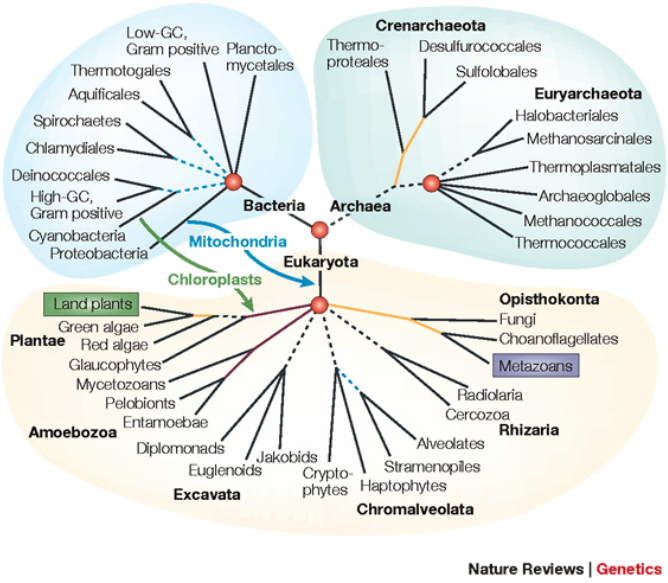
Genome rearrangement phylogeny

Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

The Tree of Life: *Multiple 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

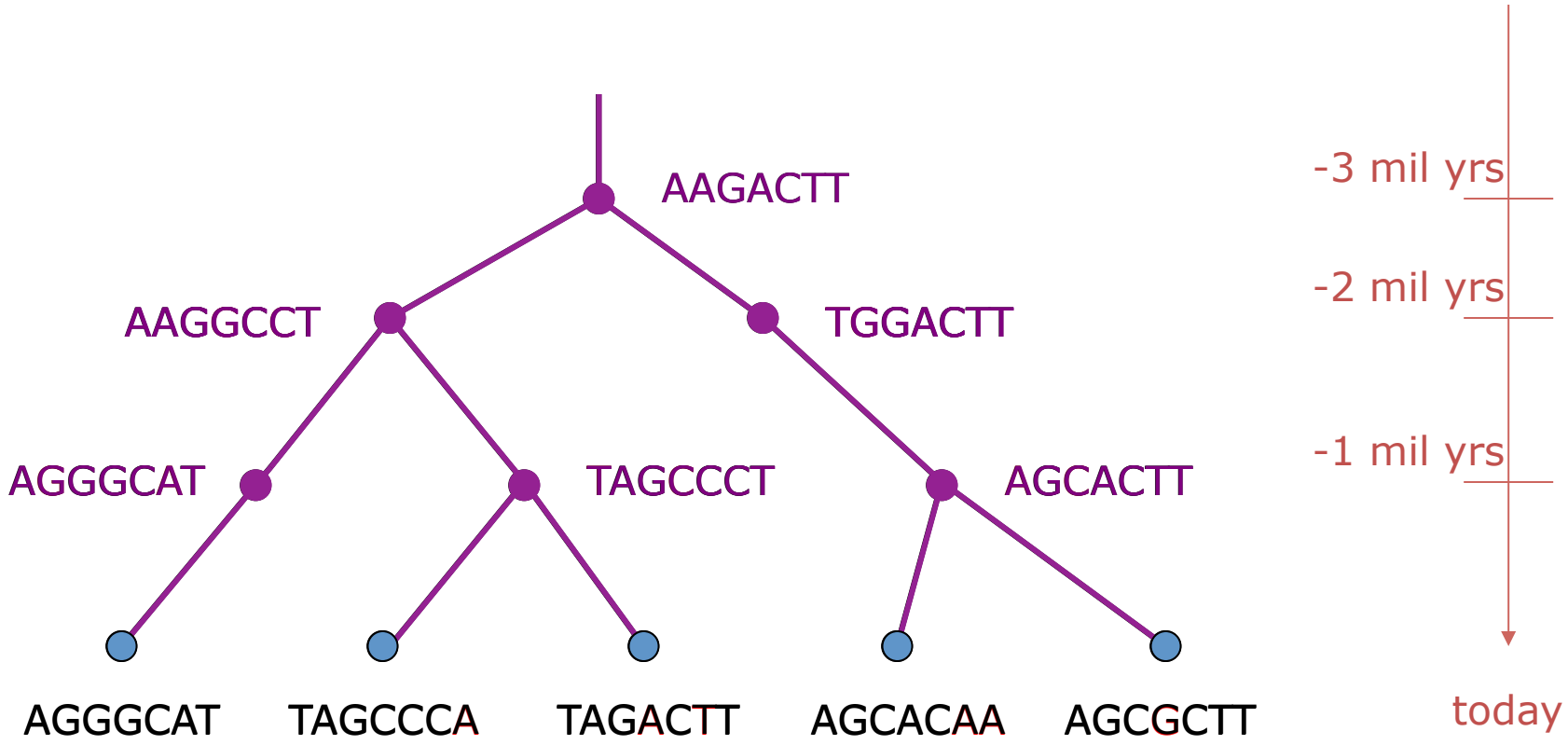
Interesting combination of different mathematics:

- 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)

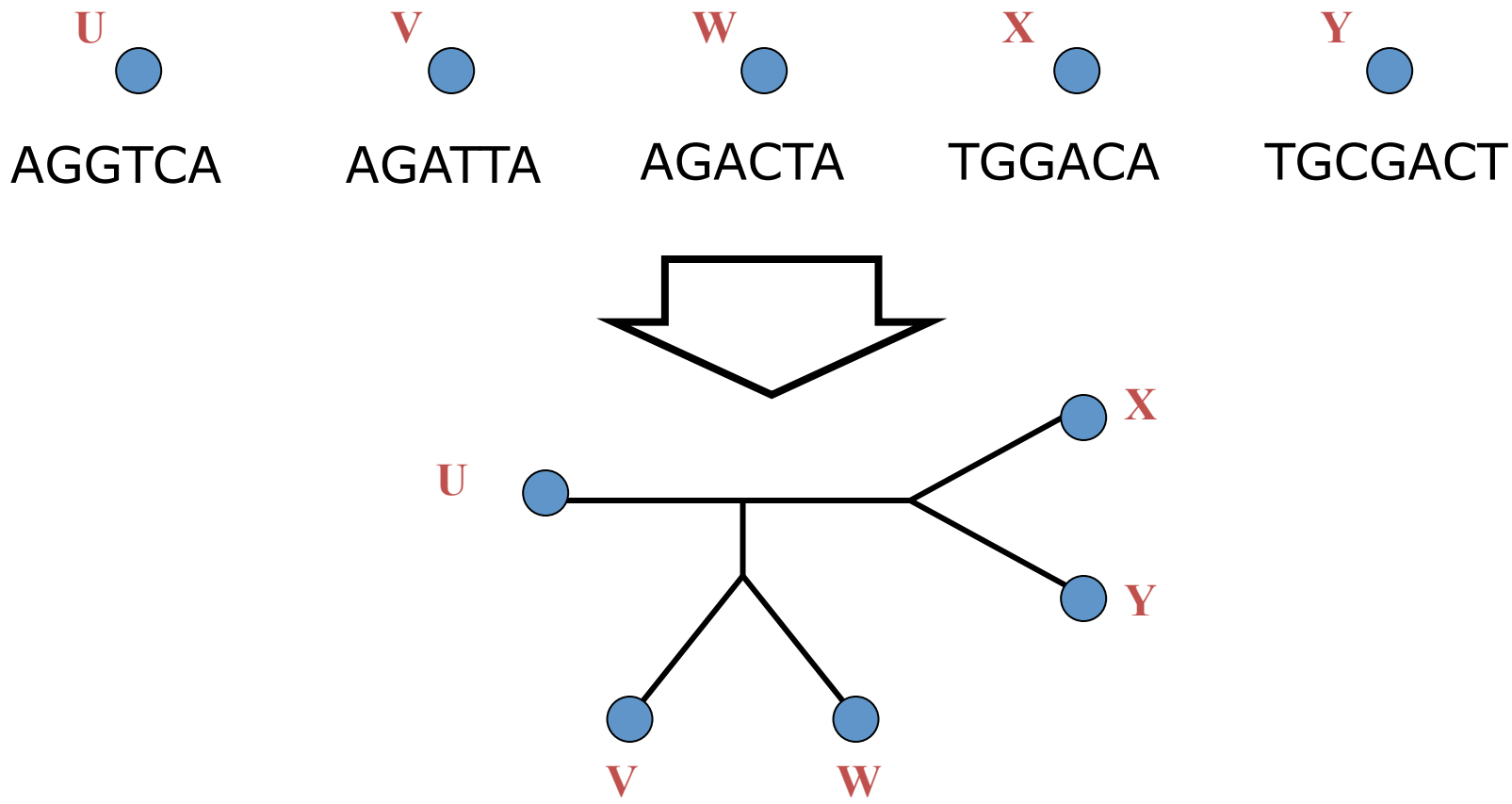


Markov Model of Site Evolution

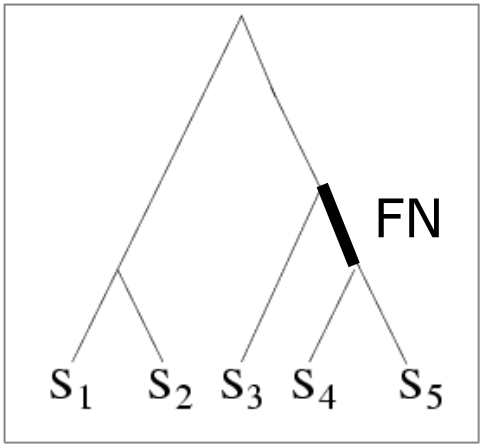
Simplest (Jukes-Cantor, 1969):

- 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.

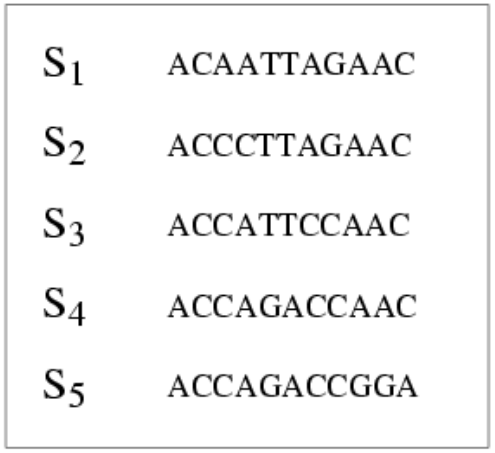
More complex models (such as the General Markov model) are also considered, often with little change to the theory.


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

Quantifying Error



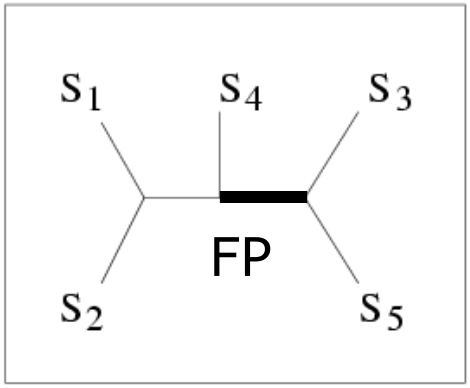
TRUE TREE



DNA SEQUENCES

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

50% error rate

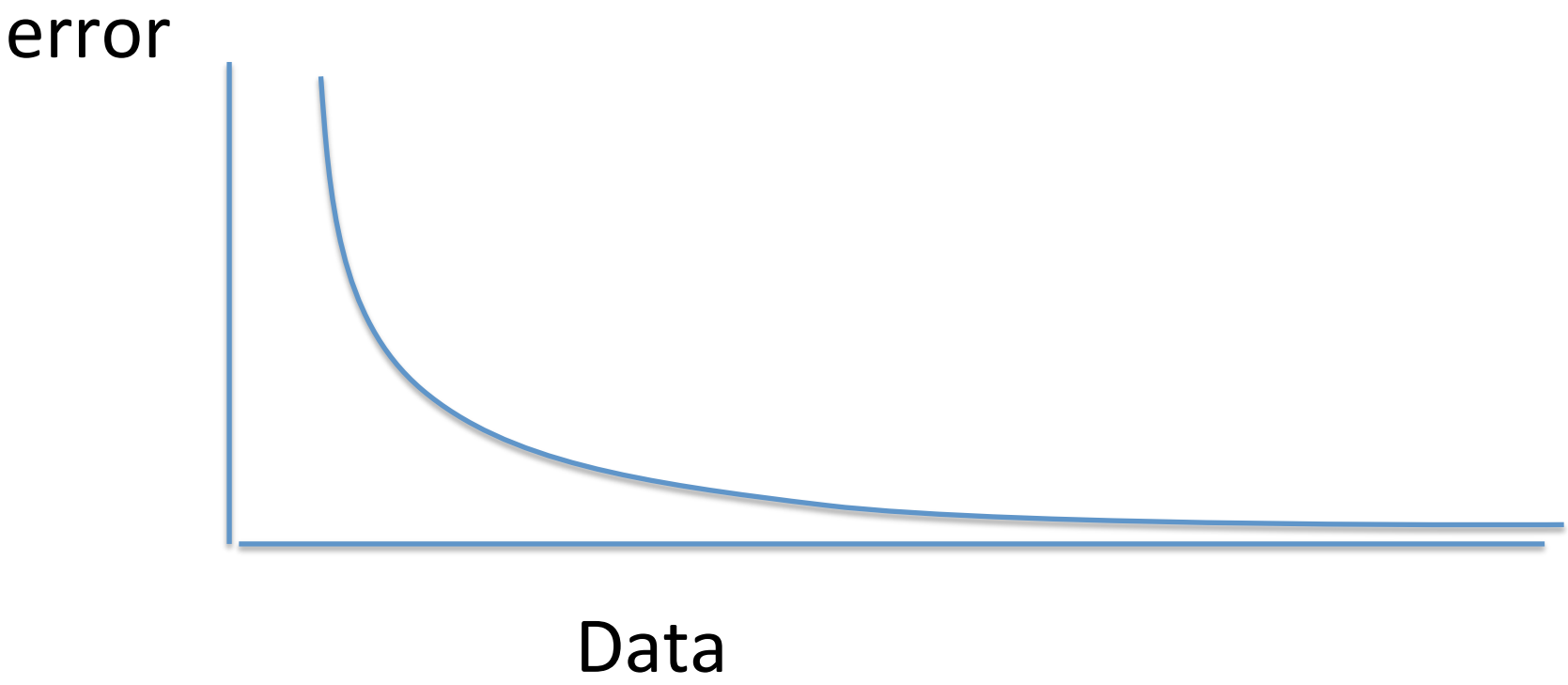


INFERRED TREE

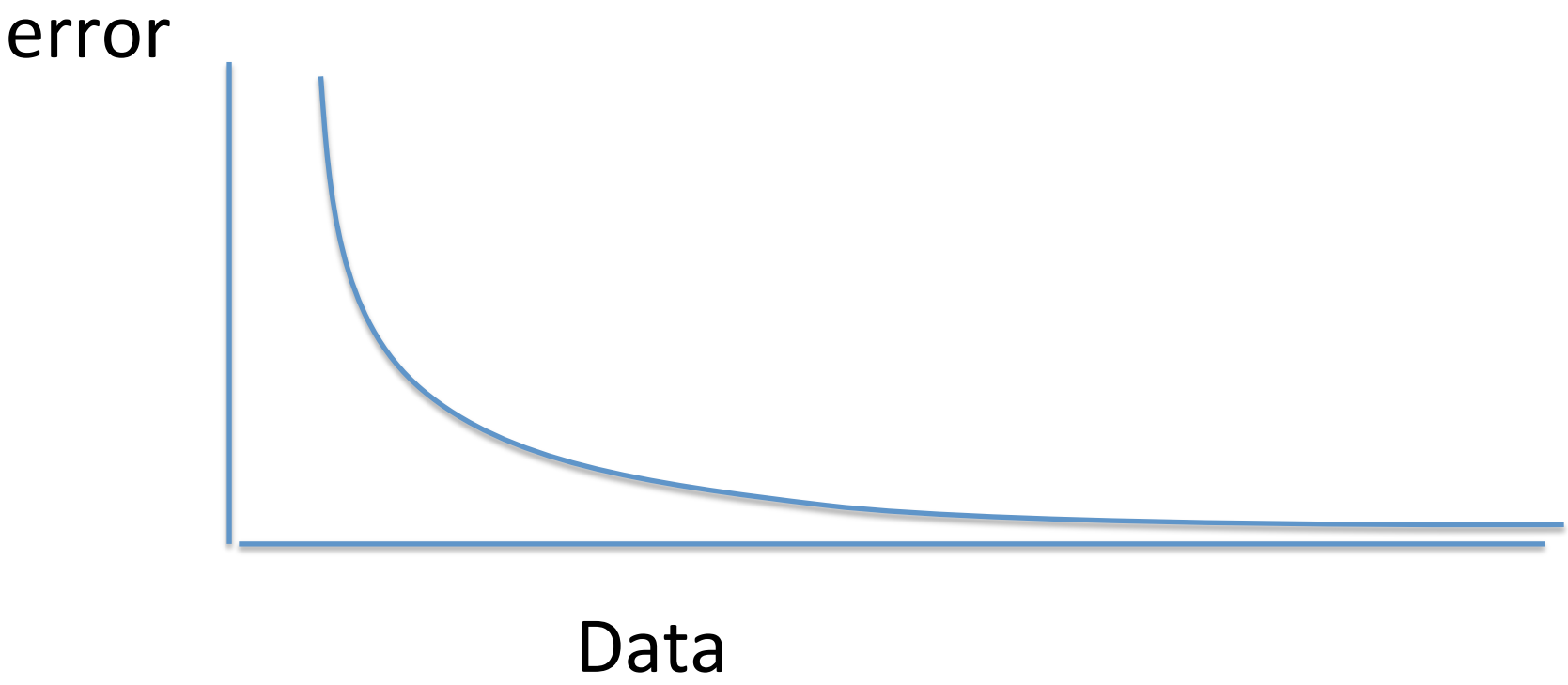
Questions

- 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?

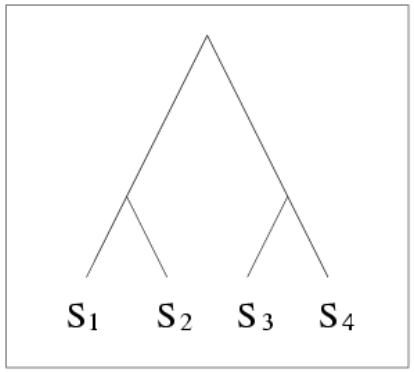
Statistical Consistency



Statistical Consistency



Data are sites in an alignment



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

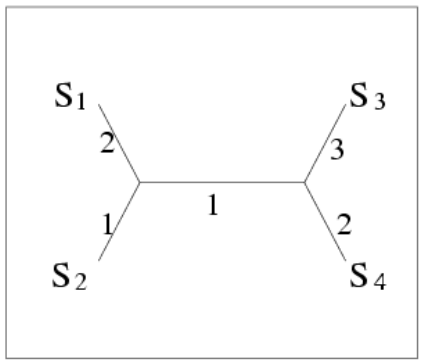
DNA SEQUENCES

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

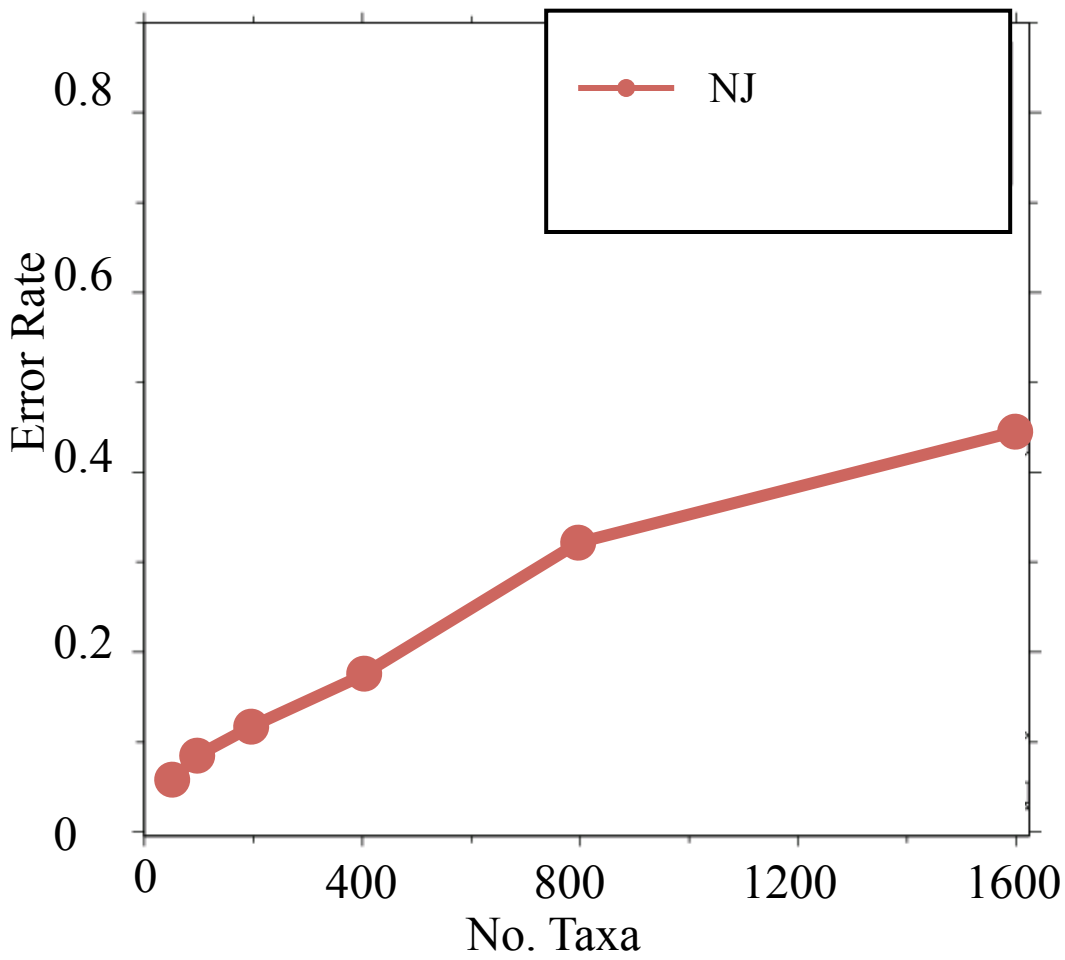
METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Neighbor Joining (and many other distance-based methods) are statistically consistent under Jukes-Cantor

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 (hence, exponential in n).

Proof:

- 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.

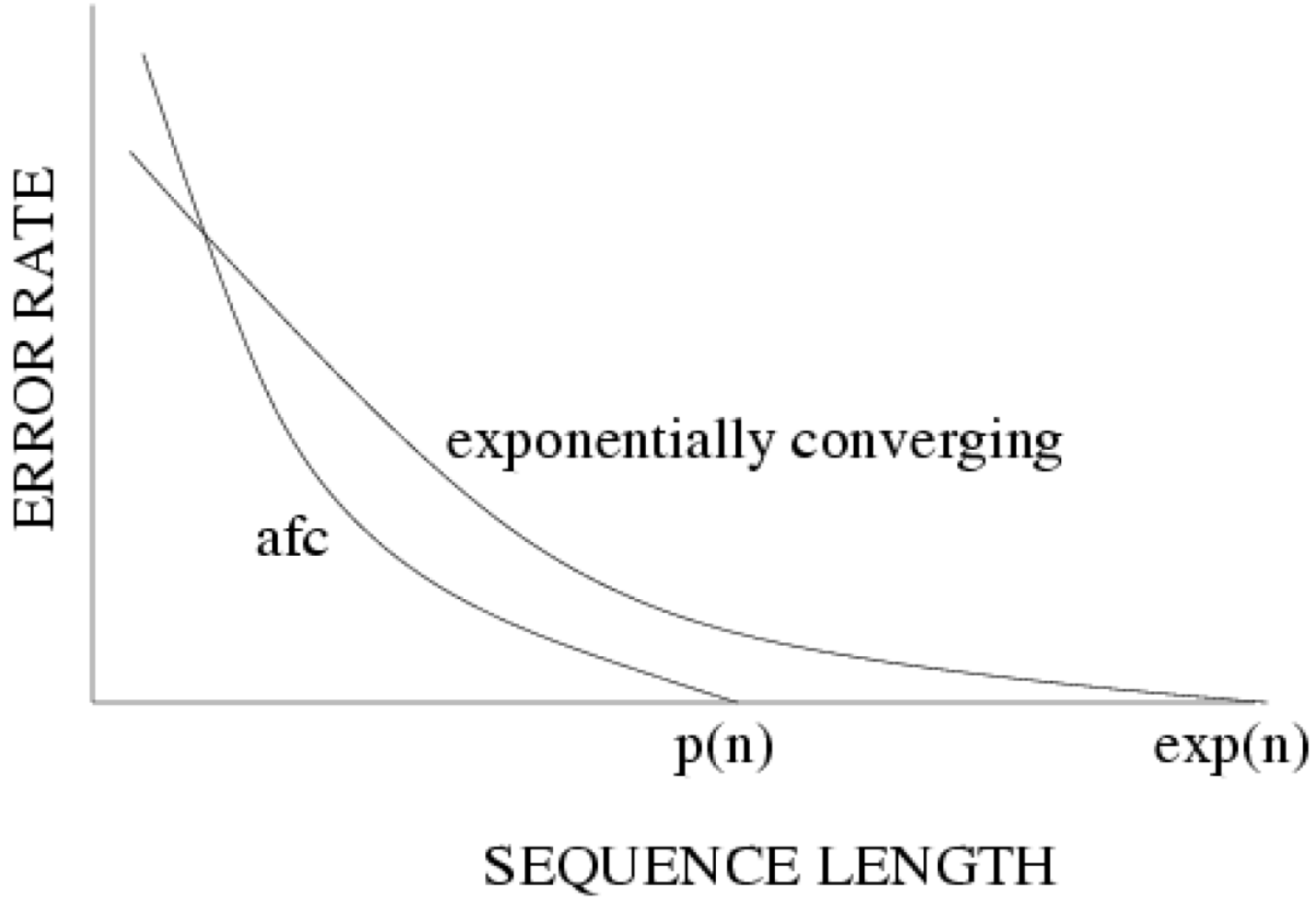
Afc methods (Warnow et al., 1999)

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.

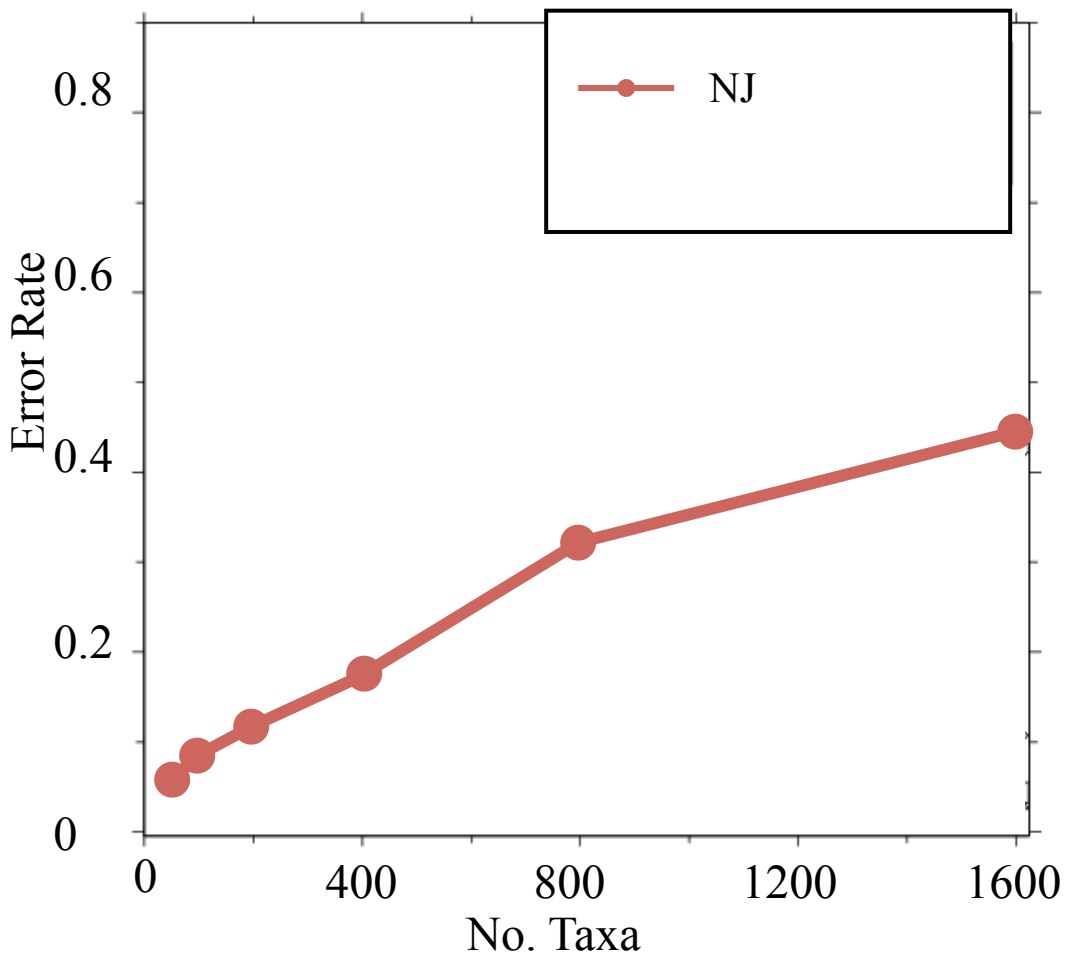
Statistical consistency, exponential convergence, and absolute fast convergence (afc)



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



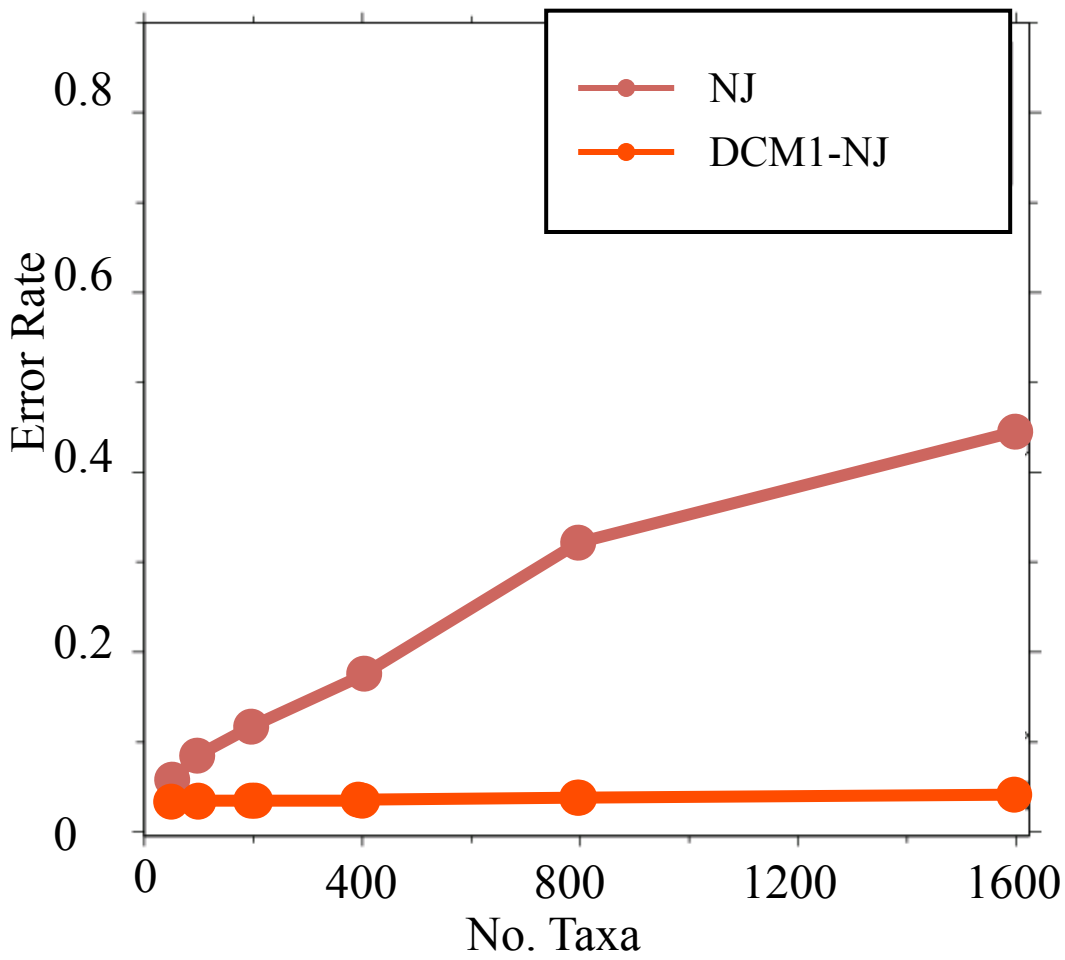
Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

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.

Proof: uses chordal graph theory and probabilistic analysis of algorithms

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.

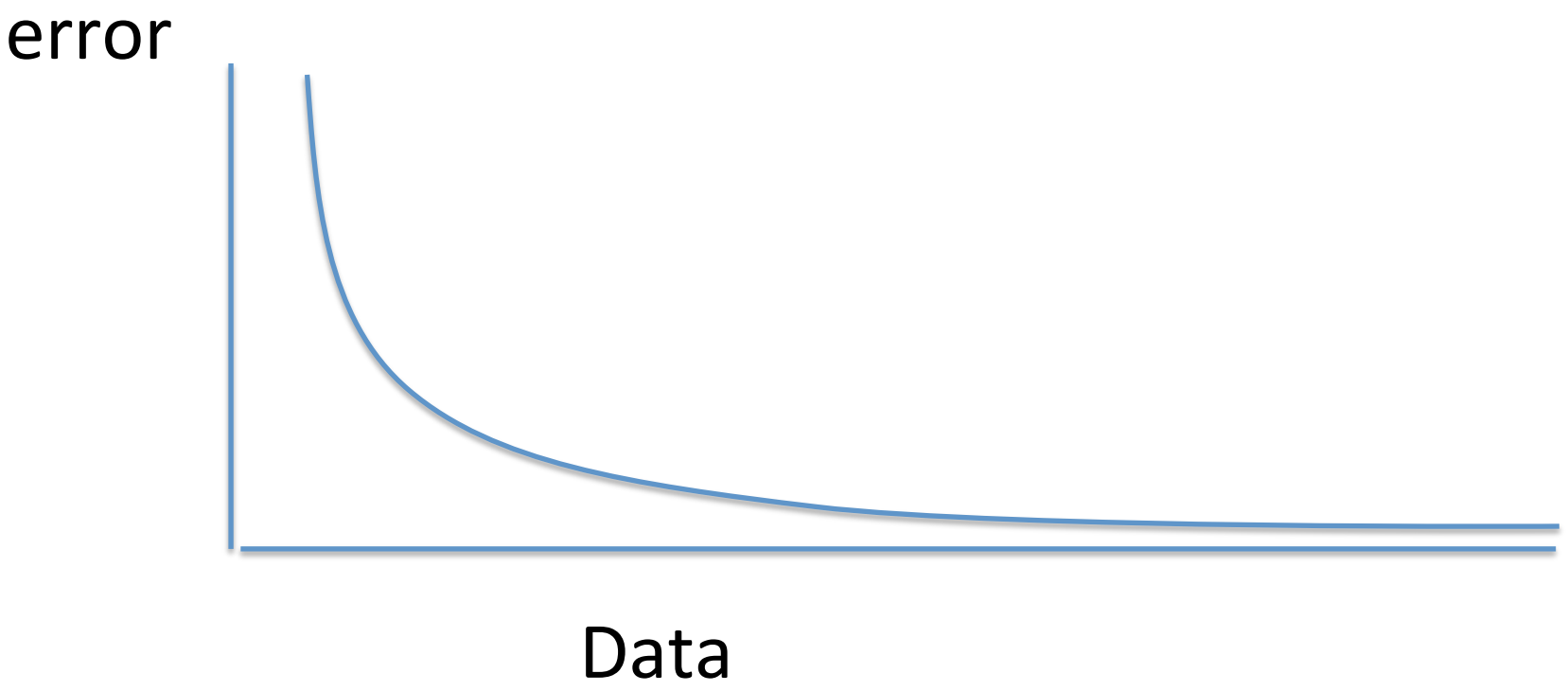
Answers?

- 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.

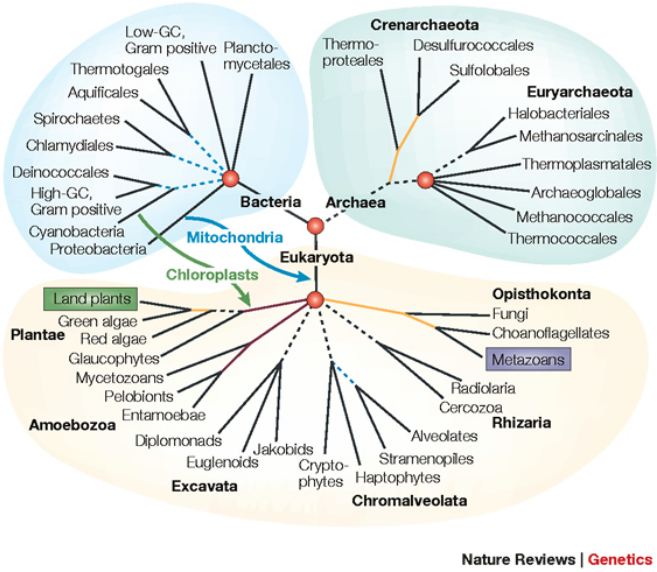
In other words…



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

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Using multiple genes

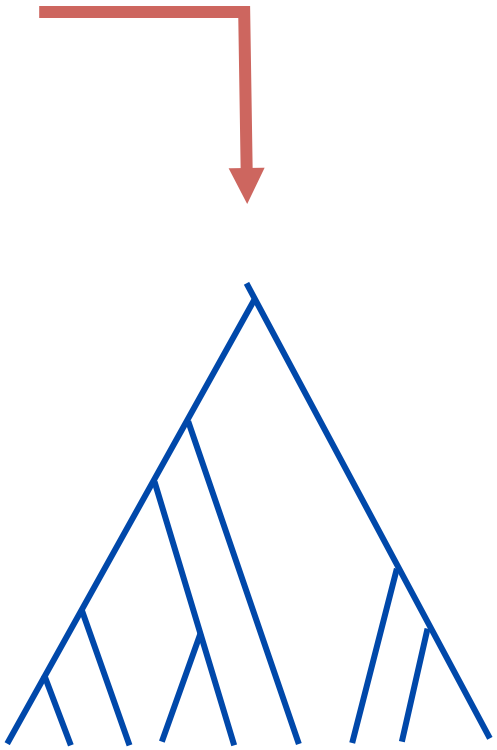
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

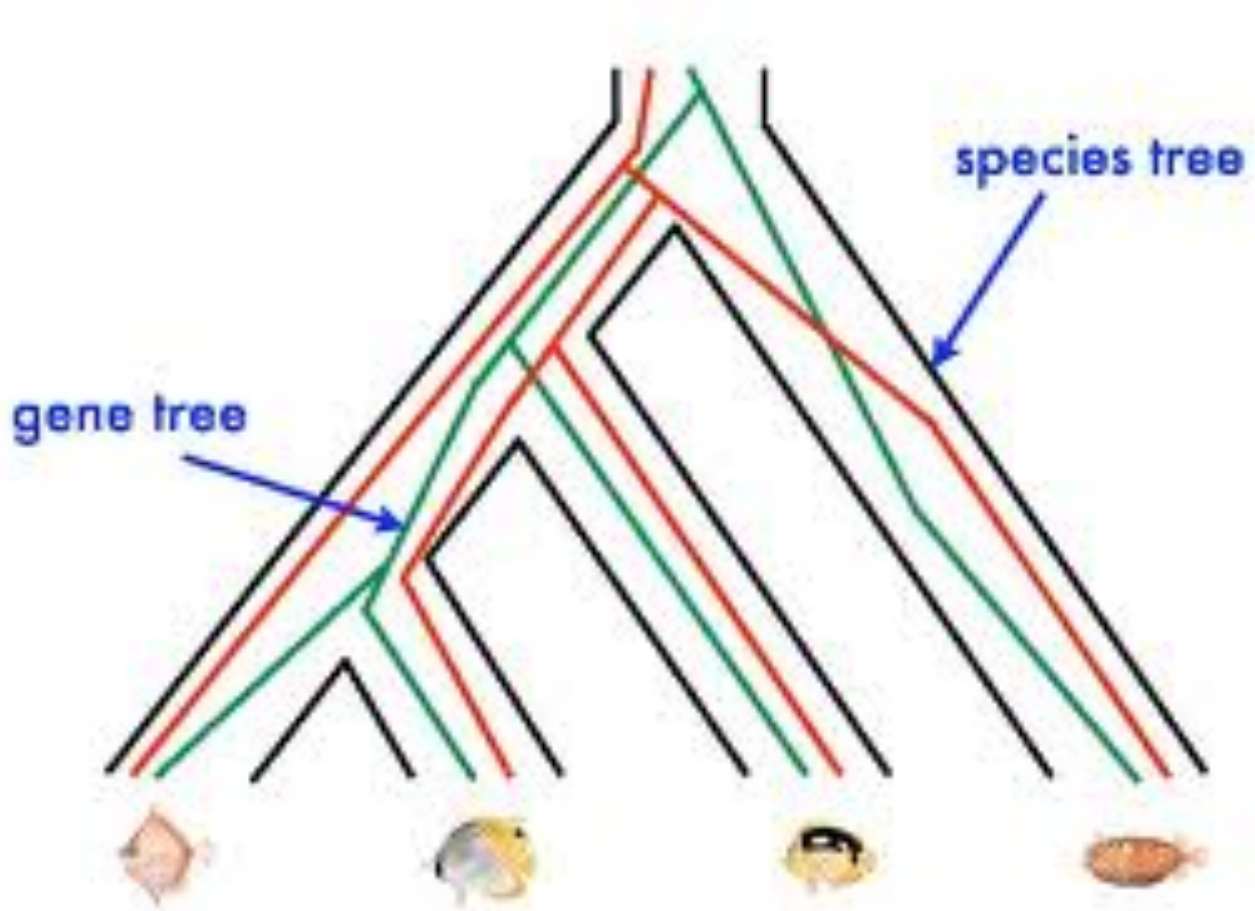
	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

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



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

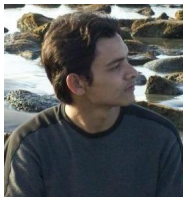


1KP: Thousand Transcriptome Project







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

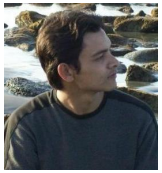
MTP Gilbert, Copenhagen

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)

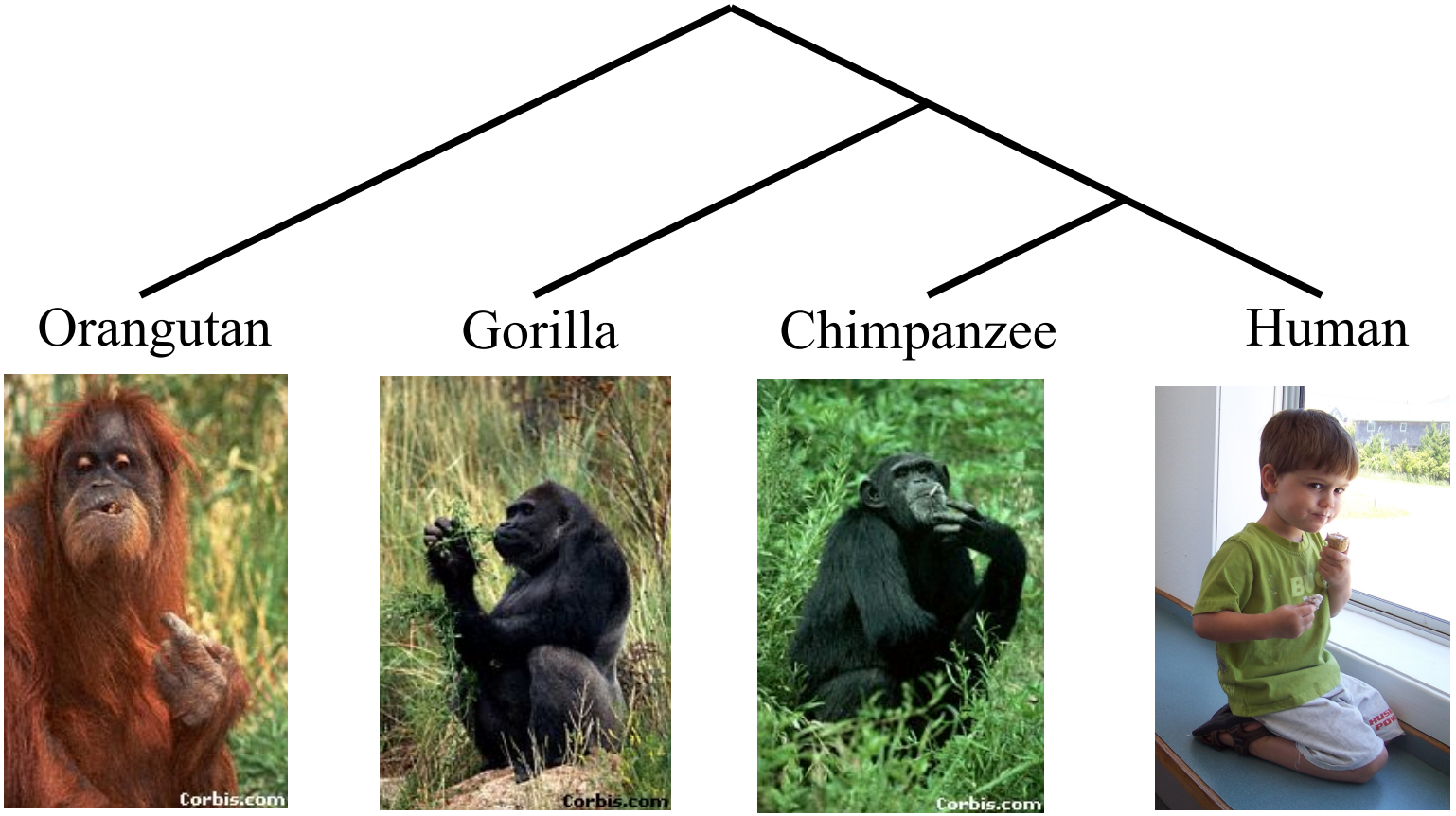
Gene Tree Incongruence

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

Part II: Species Tree Estimation in the presence of ILS

- Mathematical model: Kingman's coalescent
- "Coalescent-based" species tree estimation methods
- Simulation studies evaluating methods
- New techniques to improve methods
- Application to the Avian Tree of Life

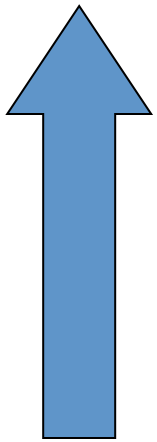
Species tree estimation: difficult, even for small datasets!



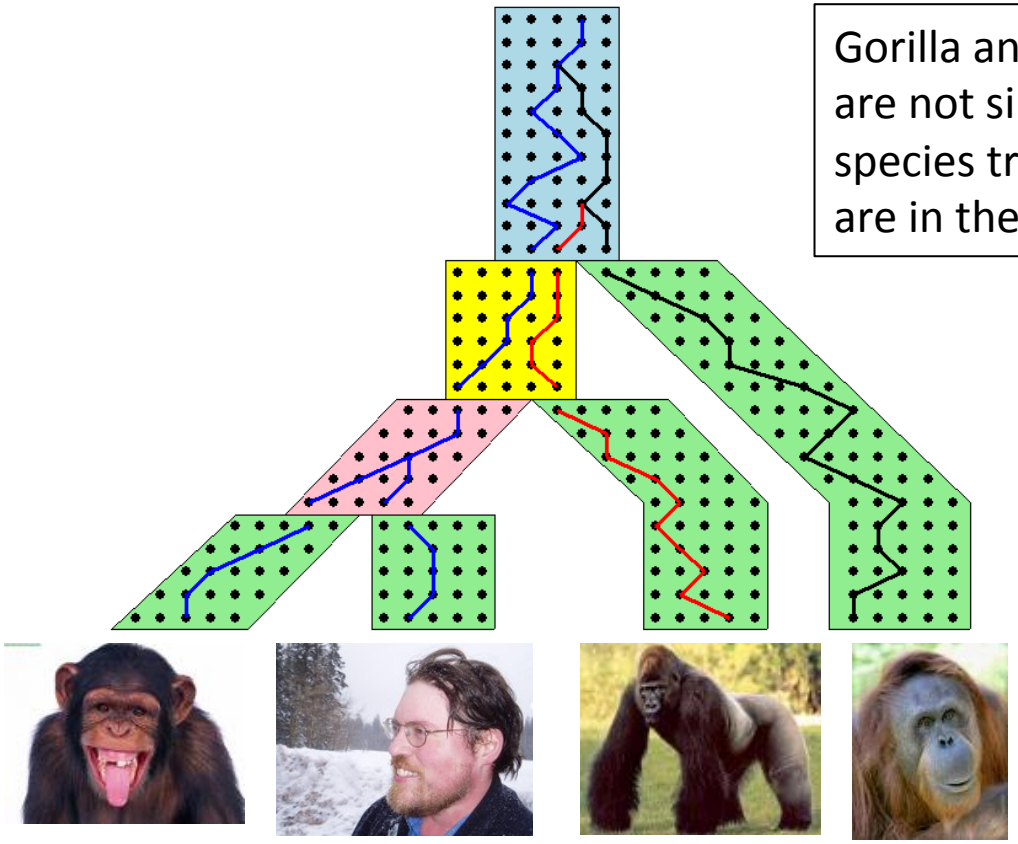
From the Tree of the Life Website, University of Arizona

The Coalescent

Past



Present

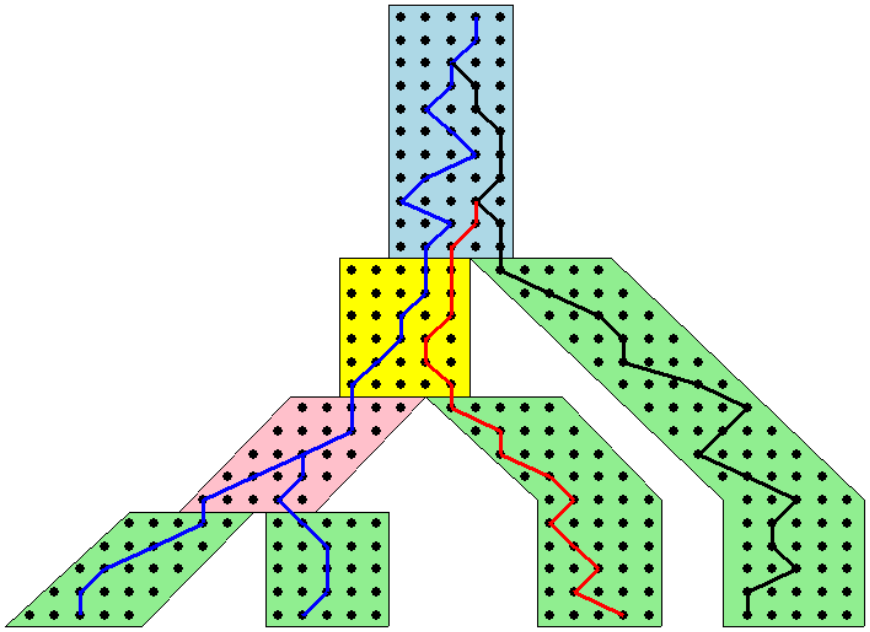


Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Courtesy James Degnan

Gene tree in a species tree

Courtesy James Degnan

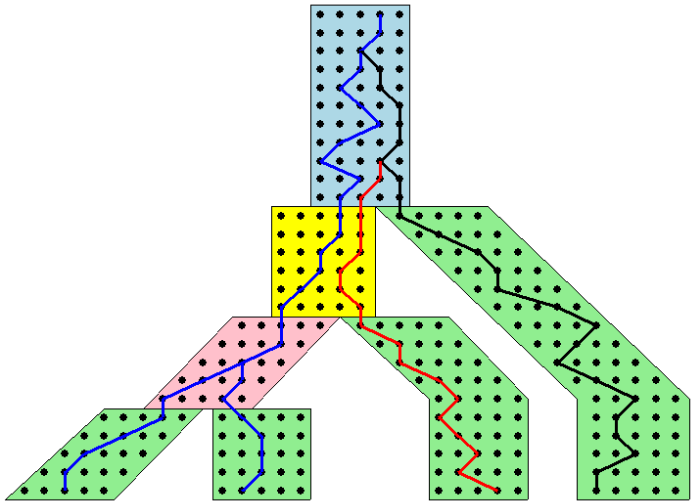


Lineage Sorting

- Lineage sorting is a Population-level process, also called the "Multi-species coalescent" (Kingman, 1982).
- The probability that a gene tree will differ from species trees increases for short times between speciation events or large population size.
- When a gene tree differs from the species tree, this is called "Incomplete Lineage Sorting" or "Deep Coalescence".

Key observation:

Under the multi-species coalescent model, the species tree defines a *probability distribution on the gene trees*

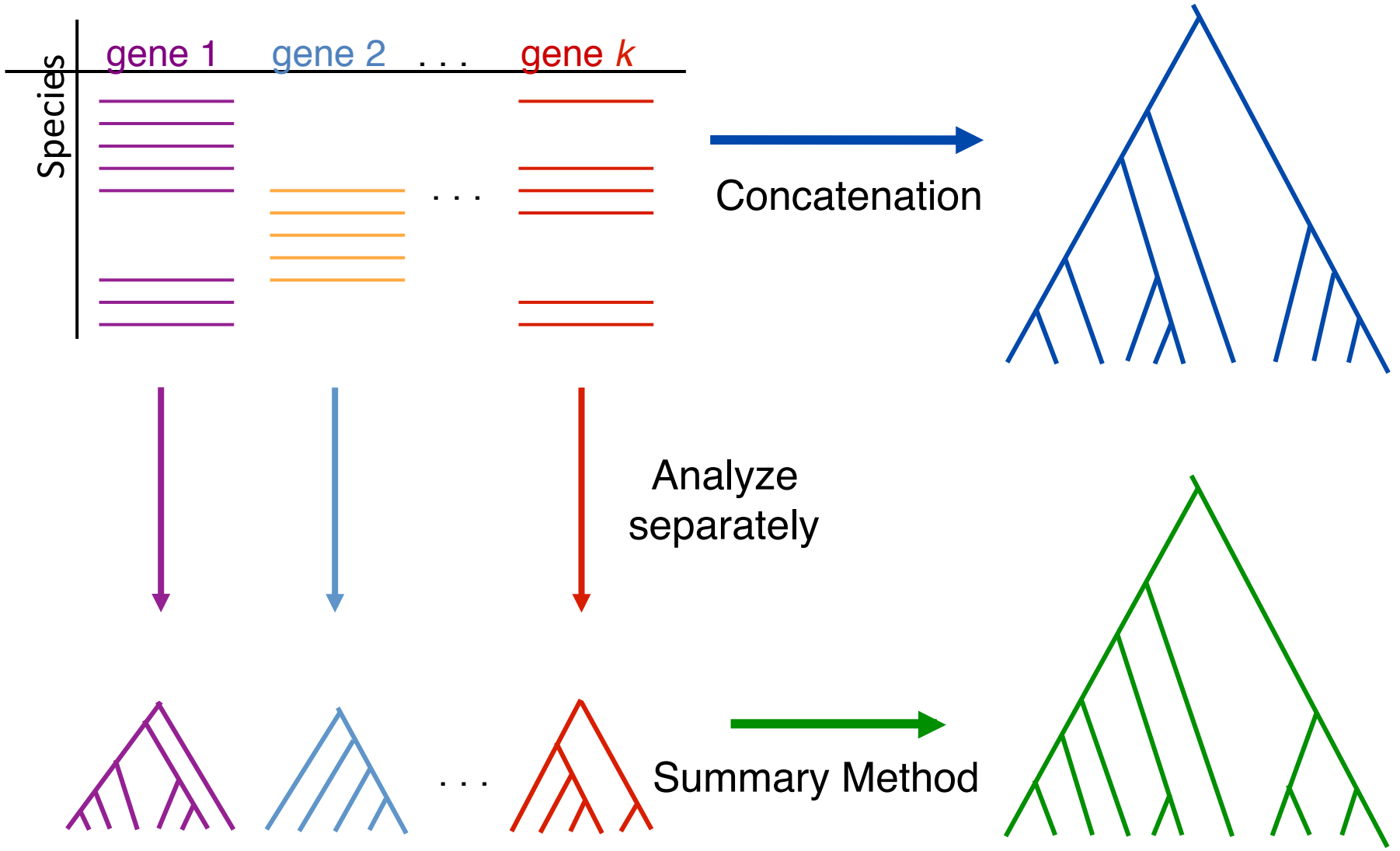


Courtesy James Degnan

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

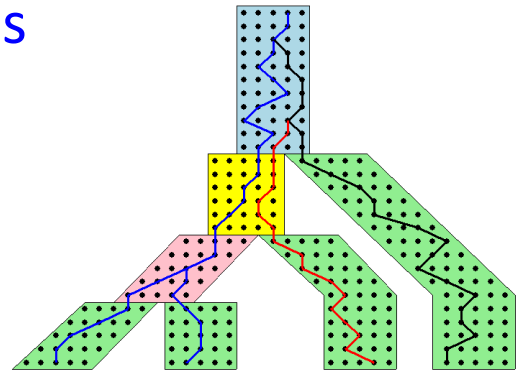


How to compute a species tree?

MDC Problem (Maddison 1997)

XL(T,t) = the number of extra lineages on the species tree T with respect to the gene tree t. In this example, XL(T,t) = 1.

Courtesy James Degnan



MDC (minimize deep coalescence) problem:

Given set X = $\{t_1, t_2, \ldots, t_k\}$ of gene trees find the species tree T that implies the *fewest extra lineages (deep coalescences)* with respect to X, i.e.,

minimize MDC(T, X) = Σ_i XL(T,t_i)

MDC Problem

- MDC is NP-hard
- Exact solution to MDC that runs in exponential time (Than and Nakhleh, PLoS Comp Biol 2009).
- Popular technique, often gives good accuracy.
- However, not statistically consistent under ILS, even if solved exactly!

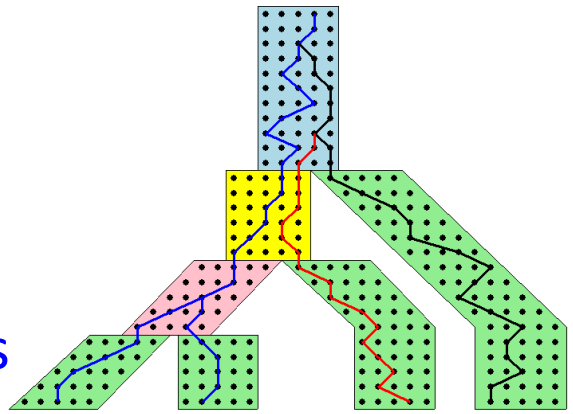
Statistically consistent under ILS?

- MDC – NO
- Greedy – NO
- Most frequent gene tree - NO
- Concatenation under maximum likelihood – open
- MRP (supertree method) – open

Under the multi-species coalescent model, the species tree defines a probability distribution on the gene trees

Courtesy James Degnan

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 rooted species tree induced on {A,B,C}.



How to compute a species tree?

Techniques:

- MDC?
- 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 rooted species tree induced on {A,B,C}.

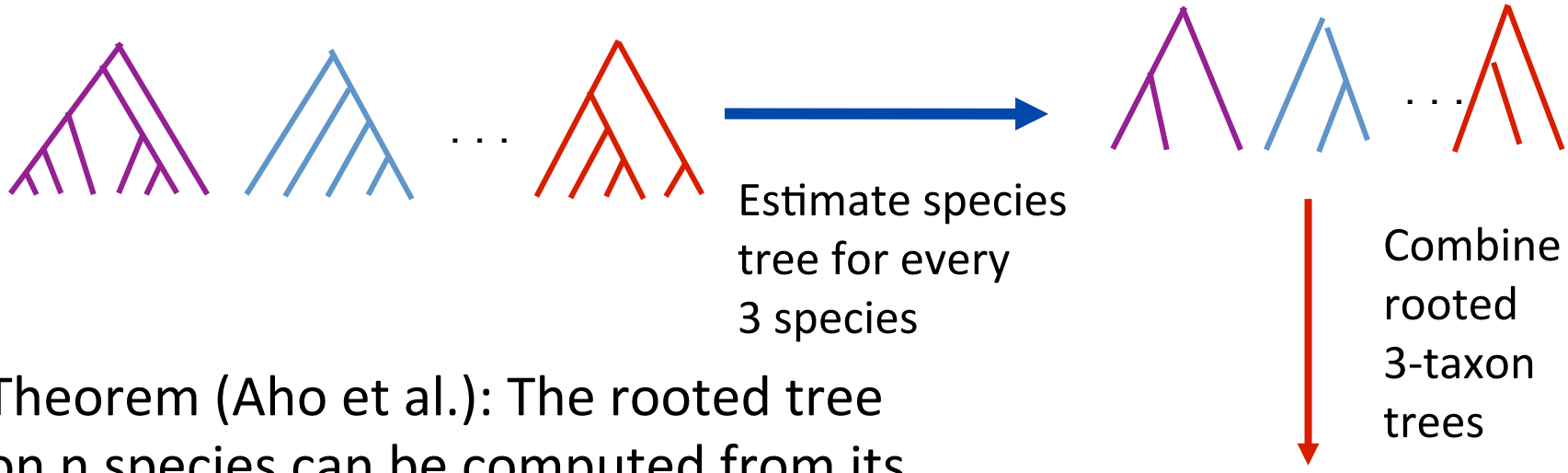
How to compute a species tree?

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 rooted species tree induced on {A,B,C}.

How to compute a species tree?

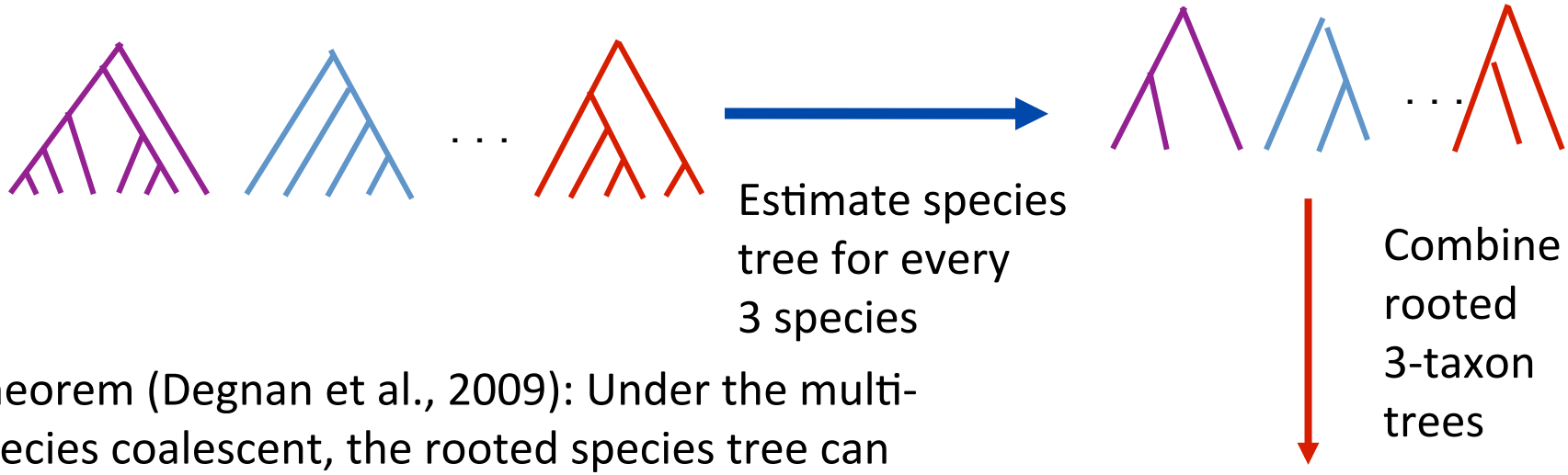
Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.

How to compute a species tree?



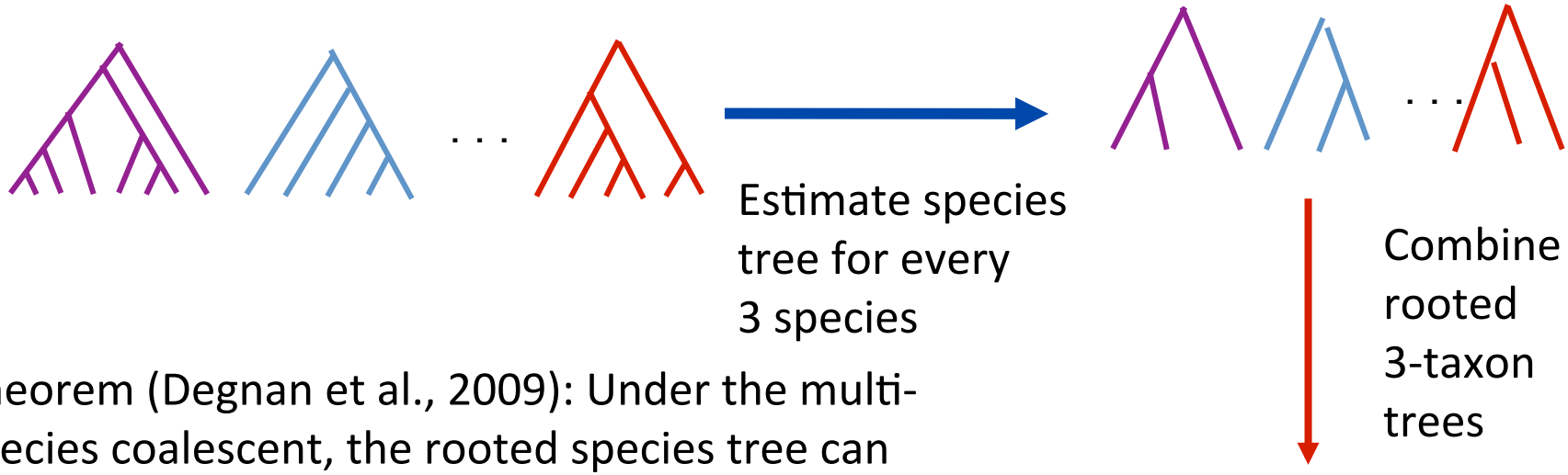
Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.

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.

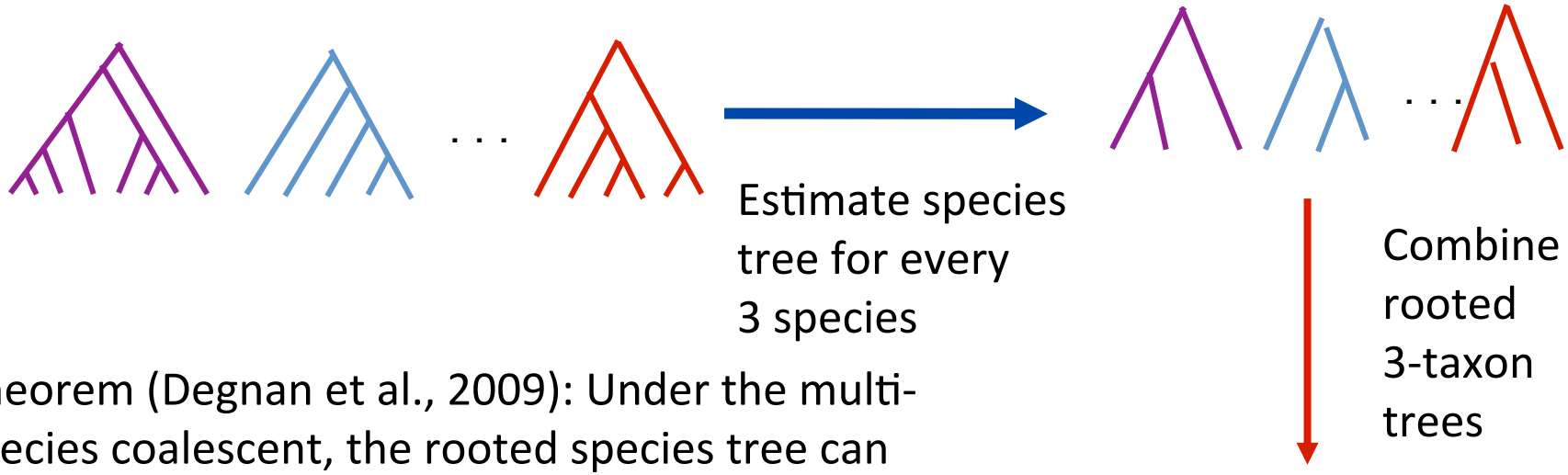
How to compute a species tree?



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.

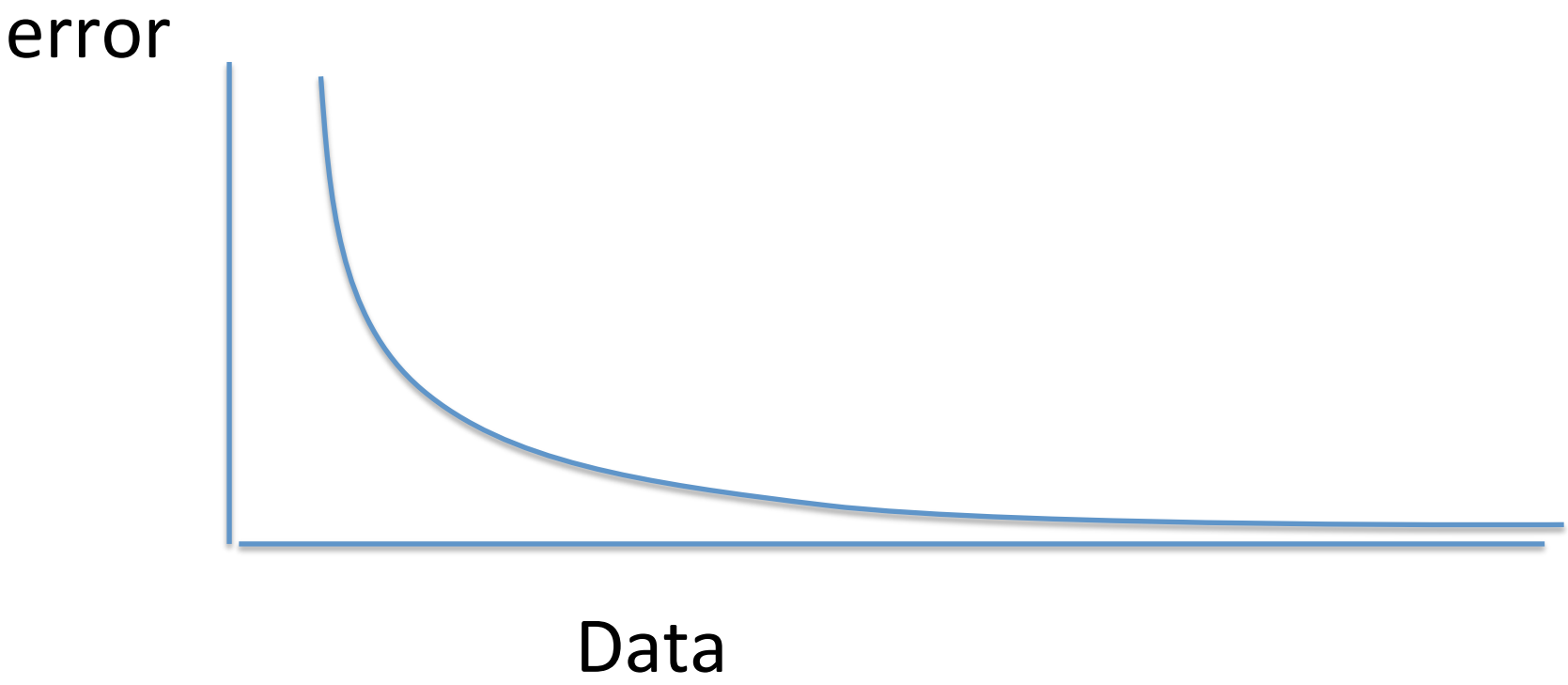
How to compute a species tree?



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 under ILS

Quartet-based methods (e.g., BUCKy-pop (Ané and Larget 2010)) for unrooted species trees

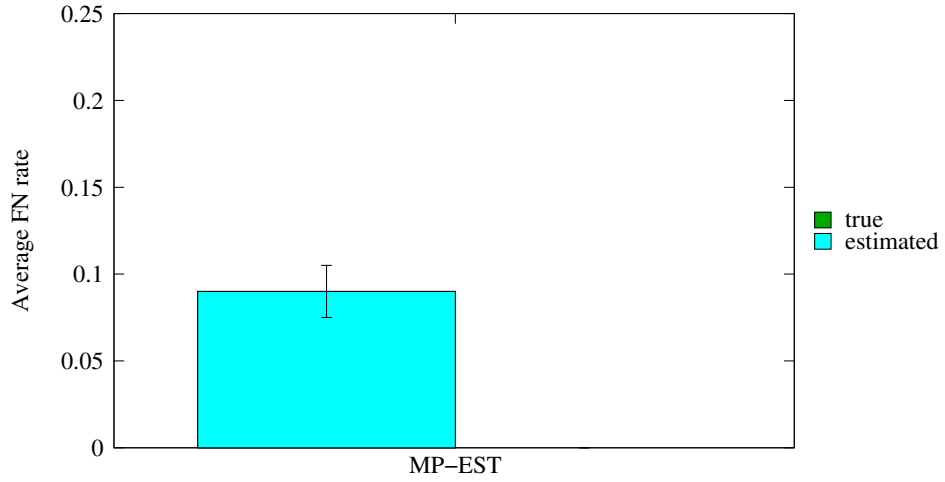
MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree for rooted species trees

(and some others)

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?

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
Similar results for other summary methods (e.g., MDC)

Datasets: 11-taxon 50-gene datasets with high ILS (Chung and Ané 2010).

Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.

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.

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.

TYPICAL PHYLOGENOMICS PROBLEM: many 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.

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".

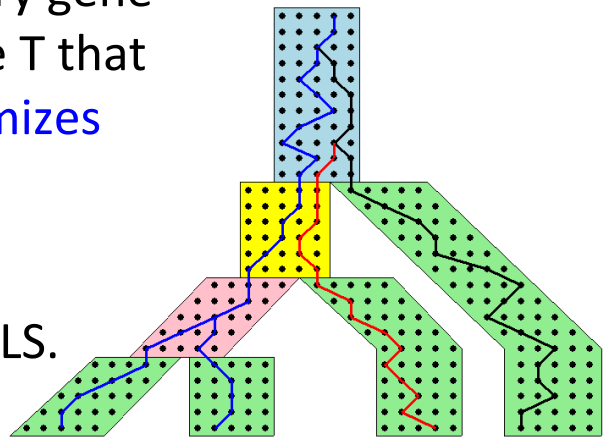
Example:

- Use bootstrapping to identify "low support edges", and contract them.
- The result is a unresolved tree, and you expect the true gene tree to be a refinement of the constraint tree.

MDC Problem (Maddison 1997)

MDC problem: Given set X = $\{t_1,t_2,\ldots,t_k\}$ of rooted binary gene trees on the same set S of leaves, find the species tree T that implies the *fewest extra lineages*, i.e., find T that minimizes $MDC(T,X) = \Sigma_i XL(T,t_i)$.

MDC produces very accurate species trees if the gene trees are highly accurate and there is not "too much" ILS.



Courtesy James Degnan

But MDC produces poor estimates of species trees if the gene trees have high error.
Sources of error:

- root location
- incorrect edges (due to insufficient sequence length)
- "rogue taxa"

Technique #1: Modify gene trees

- Identify and collapse edges with low support.
- Unroot the tree.
- Remove “rogue taxa”.

- The result is a “constraint tree”: we expect the true gene tree to be obtained by
  - Rooting
  - Refining
  - Adding in missing taxa

MDC*: Extending MDC

Input:

- Set X of k gene trees (*unrooted, not necessarily binary, not necessarily complete)* on set S
- Optional: 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 binary tree that completes and refines t,

so as to minimize MDC(T,X*).

We use the set C to constrain the search space, in order to achieve faster running times.

If C is not provided, then there is no constraint on T.

Solving MDC*

Theorem (Yu, Warnow, and Nakhleh, 2011): When all gene trees have the same set of leaves, 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, using dynamic programming. Thus, the optimal solution to MDC* can be found in $O(2^{2n}nk)$.

Solving MDC*

Theorem (Yu, Warnow, and Nakhleh, 2011): When all gene trees have the same set of leaves, 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, using dynamic programming. Thus, the optimal solution to MDC* can be found in $O(2^{2n}nk)$.

Proof (sketch): *The optimal solution can be computed by finding a maximum weight clique in a graph with vertex set C, and edges between compatible bipartitions. This can be found in $O(|C|^2nk)$ time, using the structure of the graph.*

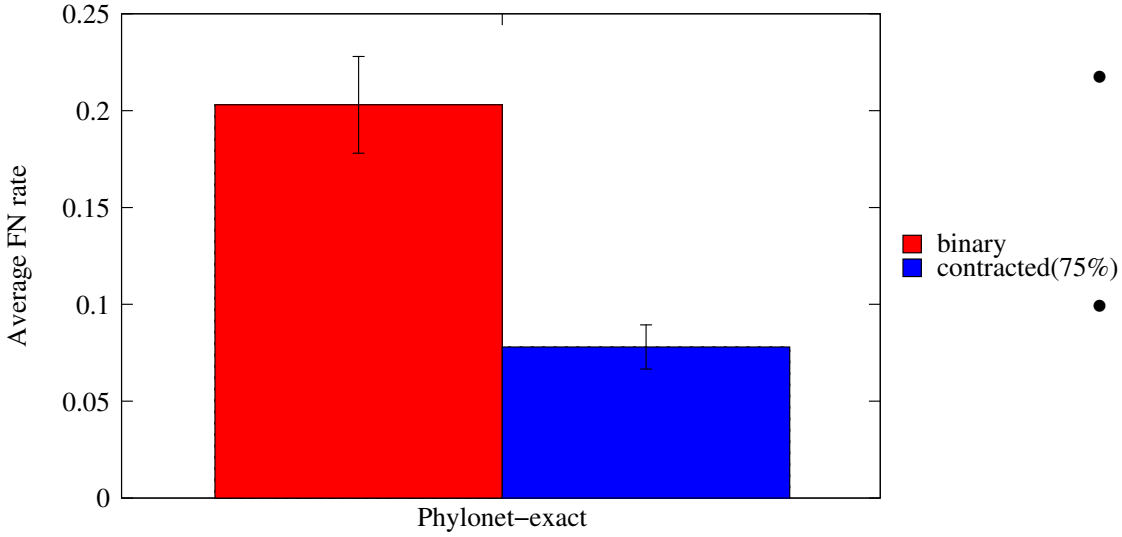
Solving MDC*

Theorem (Yu, Warnow, and Nakhleh, 2011): When all gene trees have the same set of leaves, 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, using dynamic programming. Thus, the optimal solution to MDC* can be found in $O(2^{2n}nk)$.

Proof (sketch): *The optimal solution can be computed by finding a maximum weight clique in a graph with vertex set C, and edges between compatible bipartitions. This can be found in $O(|C|^2nk)$ time, using the structure of the graph.*

Theorem (Bayzid and Warnow, 2012): YWN 2011 correctly handles case where some gene trees can miss some species.

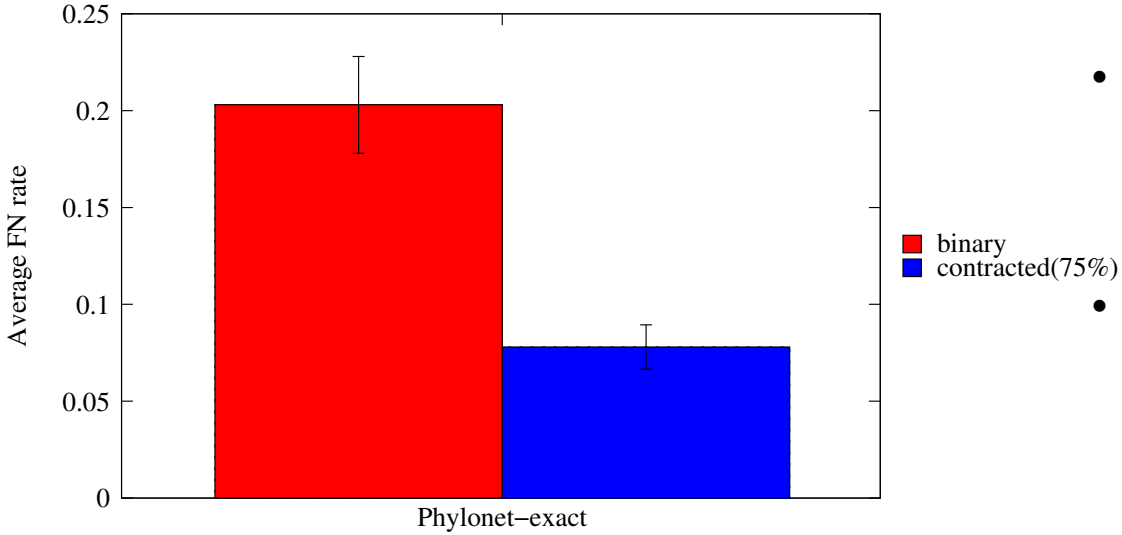
MDC vs. MDC*



- Data: 11-taxon 50-gene datasets with high levels of ILS, 50 ML gene trees, Chung and Ané, Syst Biol
- Phylo-exact solves MDC* optimally, with contracted gene trees based on 75% bootstrap support threshold (Phylonet software)

Contracting low support edges improves accuracy.

MDC vs. MDC*



- Data: 11-taxon 50-gene datasets with high levels of ILS, 50 ML gene trees, Chung and Ané, Syst Biol
- Phylo-exact solves MDC* optimally, with contracted gene trees based on 75% bootstrap support threshold (Phylonet software)

But not all methods are improved: In particular, this technique *does not help MP-EST*.

Technique #2: Bin-and-Conquer?

1. Assign genes to "bins", creating "supergene alignments"
2. Estimate trees on each supergene alignment using maximum likelihood
3. Combine the supergene trees together using a summary method

Technique #2: Bin-and-Conquer?

1. Assign genes to "bins", creating "supergene alignments"
2. Estimate trees on each supergene alignment using maximum likelihood
3. Combine the supergene trees together using a summary method

Variants:

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

Statistical binning

Input: estimated gene trees with bootstrap support, and minimum support threshold t

Output: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible.

Statistical binning

Input: estimated gene trees with bootstrap support, and minimum support threshold t

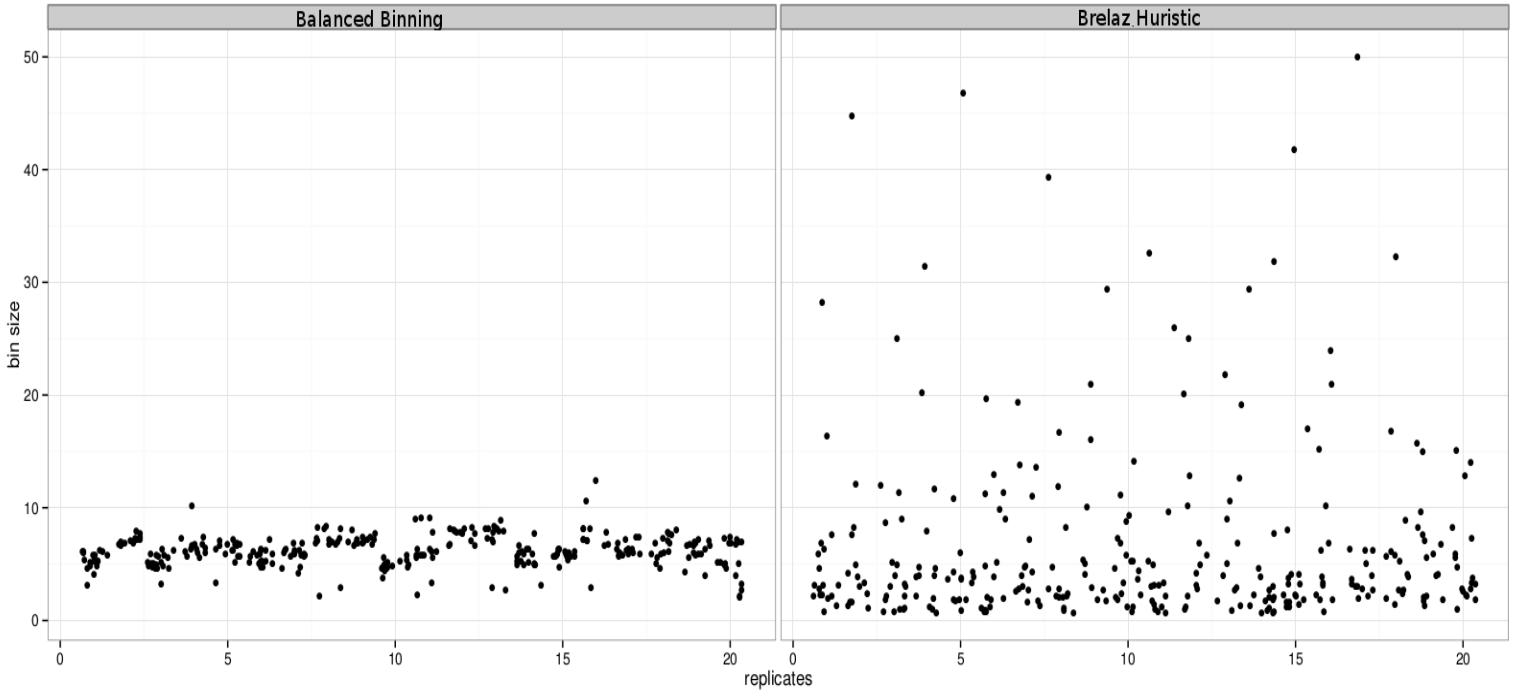
Output: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible.

Vertex coloring problem (NP-hard),

but good heuristics are available (e.g., Brélaz 1979)

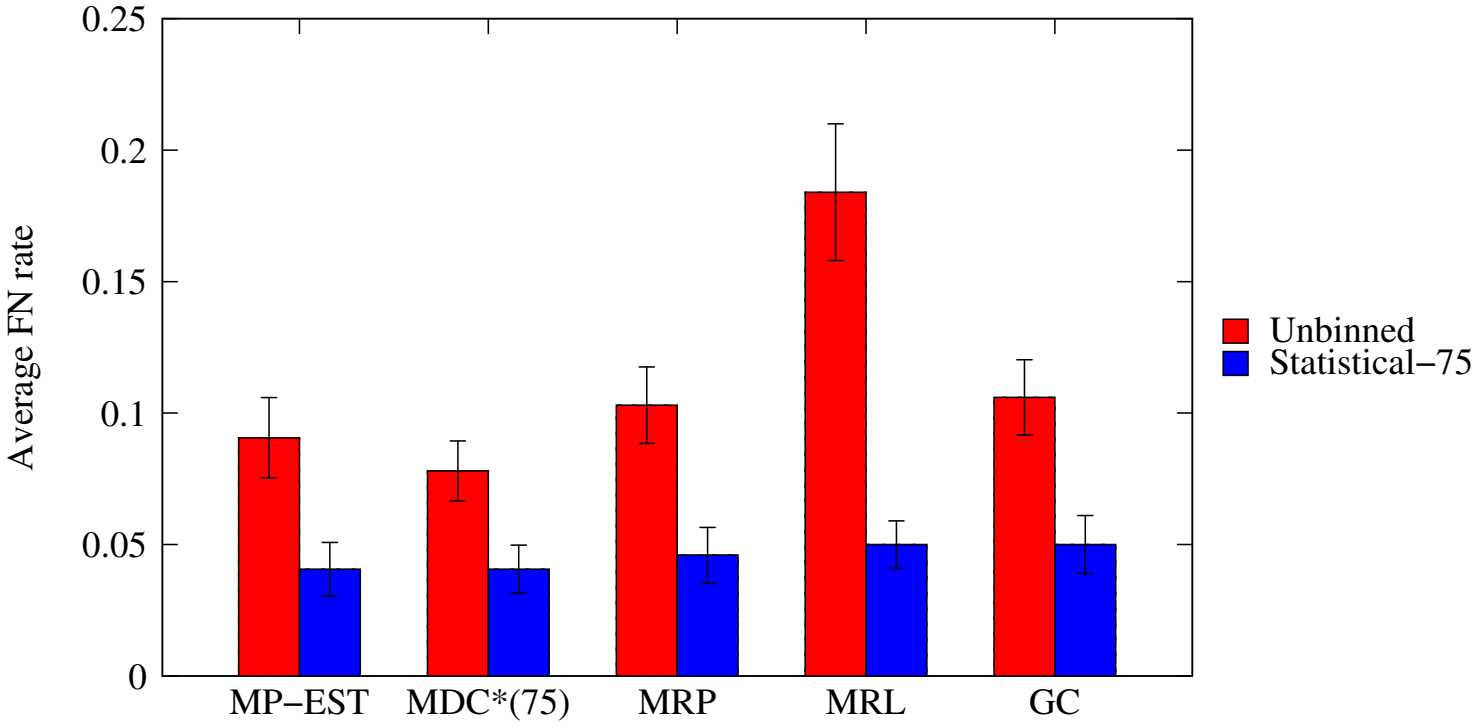
However, for statistical inference reasons, we need balanced vertex color classes

Balanced Statistical Binning



Mirarab, Bayzid, and Warnow, in preparation
Modification of Brélaz Heuristic for minimum vertex coloring.

Statistical binning vs. unbinned



Mirarab, et al. in preparation

Datasets: 11-taxon strongILS datasets with 50 genes, Chung and Ané, Systematic Biology

Avian Phylogenomics Project

E Jarvis, HHMI

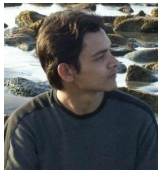
MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Gene Tree Incongruence

Plus many many other people…

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

Avian Phylogenomics Project, in preparation

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

Avian Phylogenomics Project, in preparation

Avian Simulation – 14,000 genes

- **MP-EST:**
  - Unbinned ~ 11.1% error
  -
- **Greedy:**
  - Unbinned ~ 26.6% error
  -

- 8250 exon-like genes (27% avg. bootstrap support)
- 3600 UCE-like genes (37% avg. bootstrap support)
- 2500 intron-like genes (51% avg. bootstrap support)

Avian Simulation – 14,000 genes

- **MP-EST:**
  - Unbinned ~ 11.1% error
  - Binned ~ 6.6% error
- **Greedy:**
  - Unbinned ~ 26.6% error
  - Binned ~ 13.3% error

- 8250 exon-like genes (27% avg. bootstrap support)
- 3600 UCE-like genes (37% avg. bootstrap support)
- 2500 intron-like genes (51% avg. bootstrap support)

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

Avian Phylogenomics Project, in preparation

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.
- More than 17 years of compute time, and used 256 GB. Run at HPC centers.
- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.
- Statistical binning version of MP-EST on 14000+ gene trees – highly resolved tree, largely congruent with the concatenated analysis, good bootstrap support

Avian Phylogenomics Project, in preparation

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 guarantees.

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

All methods provide improved accuracy compared to existing methods, as shown on simulated and biological datasets.

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

Research Agenda

Major scientific goals:

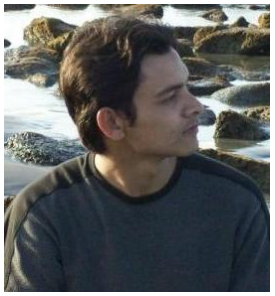
- Develop methods that produce more accurate alignments and phylogenetic estimations for *difficult-to-analyze datasets*
- Produce mathematical theory for statistical inference under complex models of evolution
- Develop novel machine learning techniques to boost the performance of classification methods

Software that:

- Can run efficiently on *desktop* computers on large datasets
- Can analyze ultra-large datasets (100,000+) using multiple processors
- Is freely available in *open source* form, with biologist-friendly GUIs

Warnow Laboratory





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

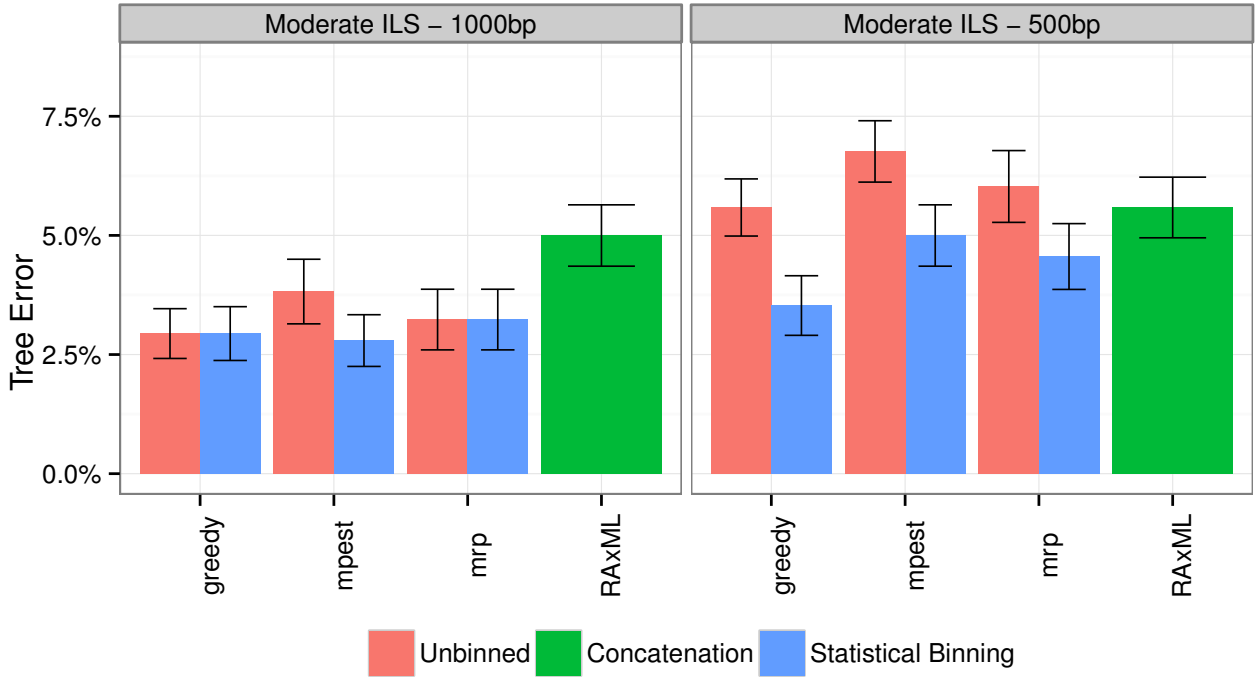
Funding: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, and TACC (Texas Advanced Computing Center)

TACC and UTCS computational resources

* Supported by HHMI Predoctoral Fellowship

** Supported by Fulbright Foundation Predoctoral Fellowship

Mammalian Simulation Study



Observations:

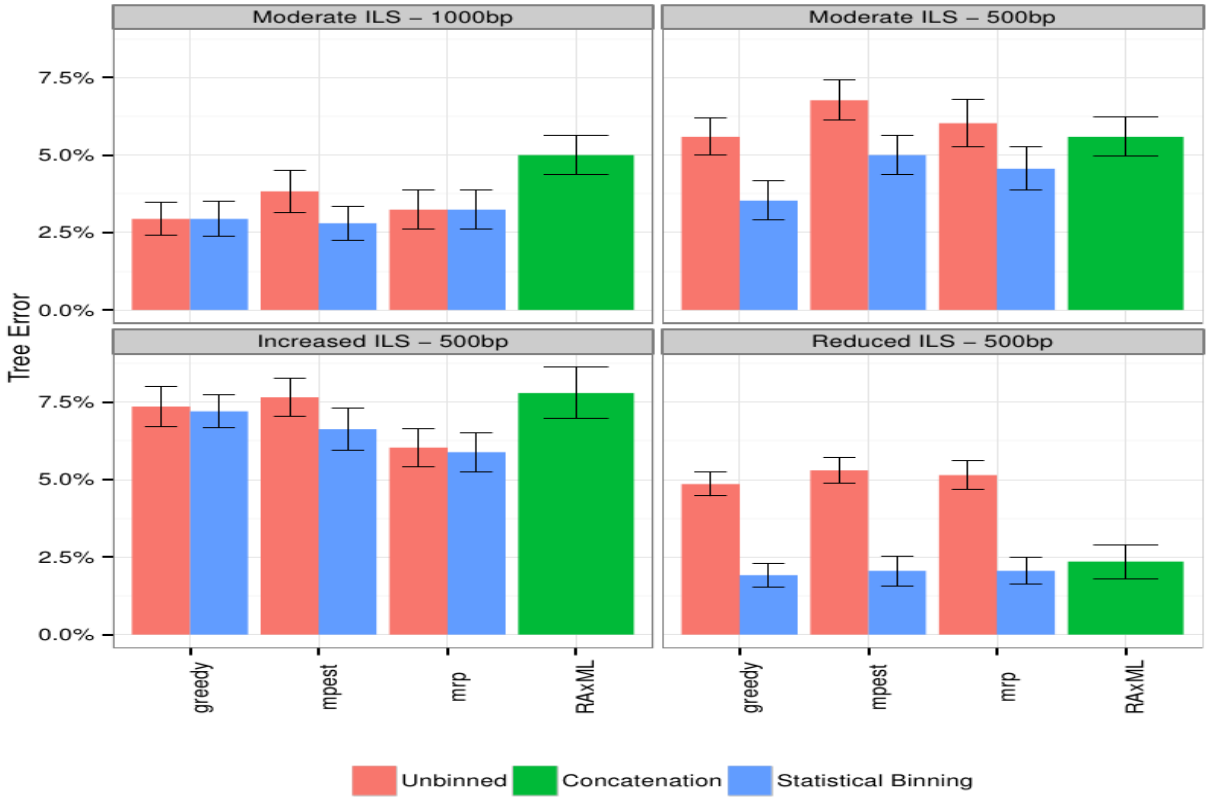
Binning can improve accuracy, but impact depends on accuracy of estimated gene trees and phylogenetic estimation method.

Binned methods can be more accurate than RAxML (maximum likelihood), even when unbinned methods are less accurate.

Data: 200 genes, 20 replicate datasets, based on Song et al. PNAS 2012

Mirarab et al., in preparation

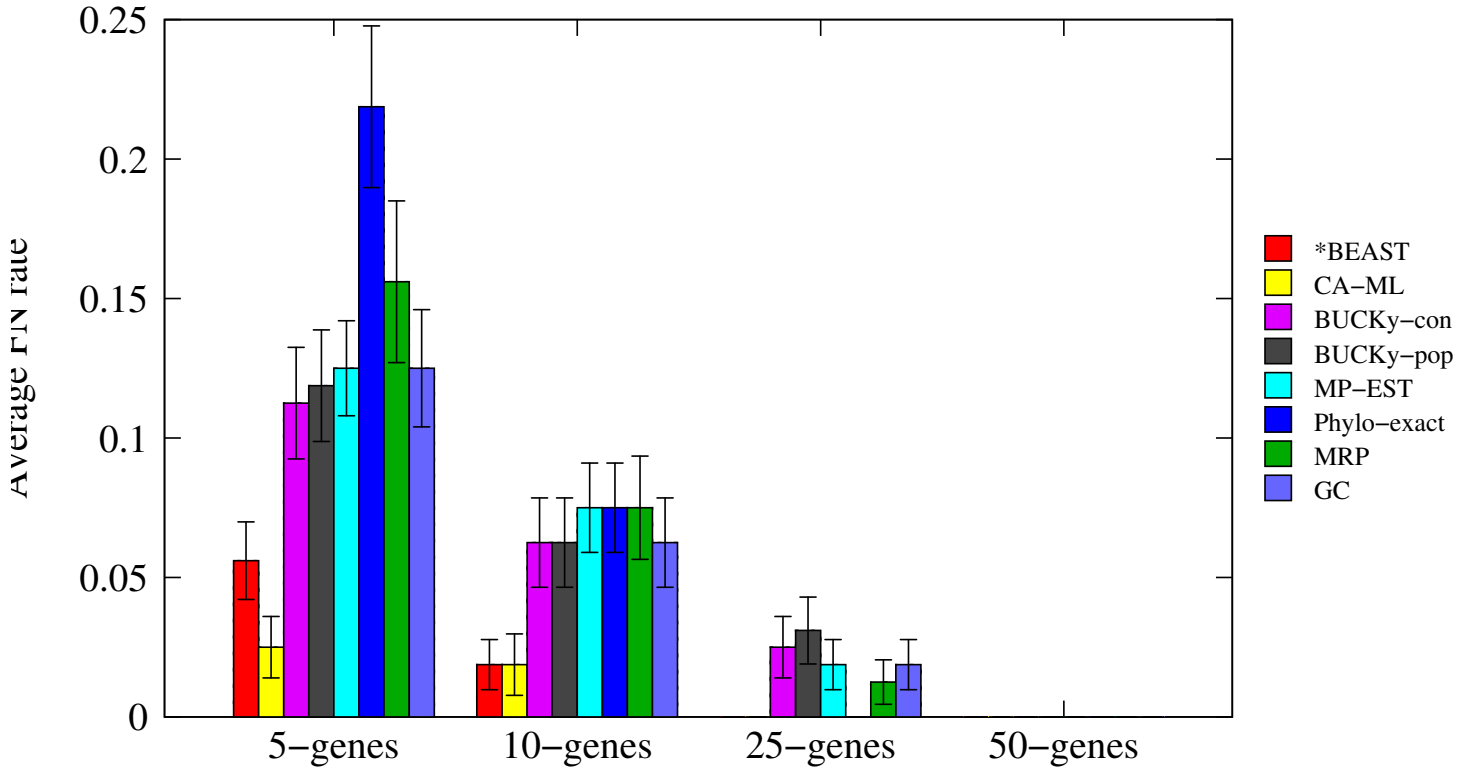
Mammalian simulation



Observation:
Binning can improve summary methods, but amount of improvement depends on: method, amount of ILS, and accuracy of gene trees.

MP-EST is statistically consistent in the presence of ILS; Greedy is not, unknown for MRP And RAxML.
Data (200 genes, 20 replicate datasets) based on Song et al. PNAS 2012

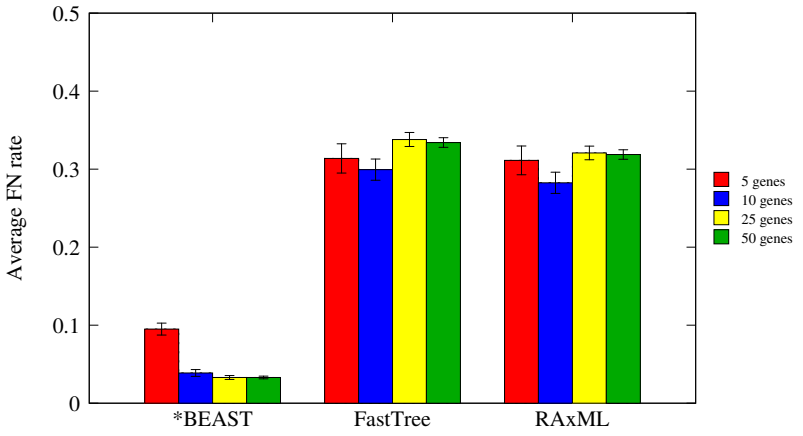
Results on 11-taxon datasets with weak ILS



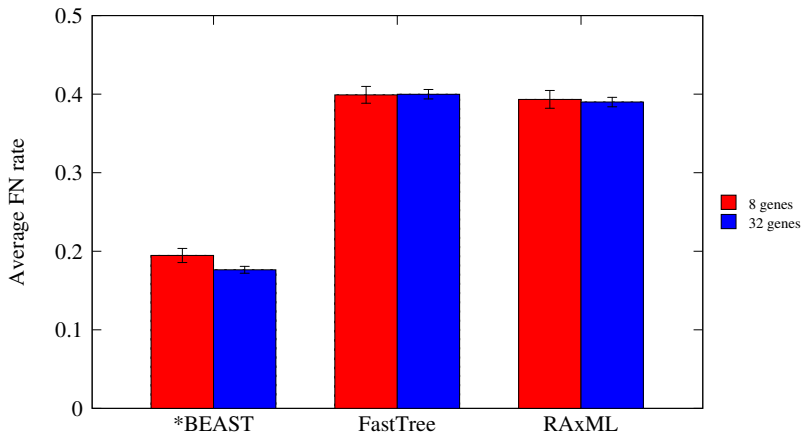
***BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: concatenated analysis) most accurate

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

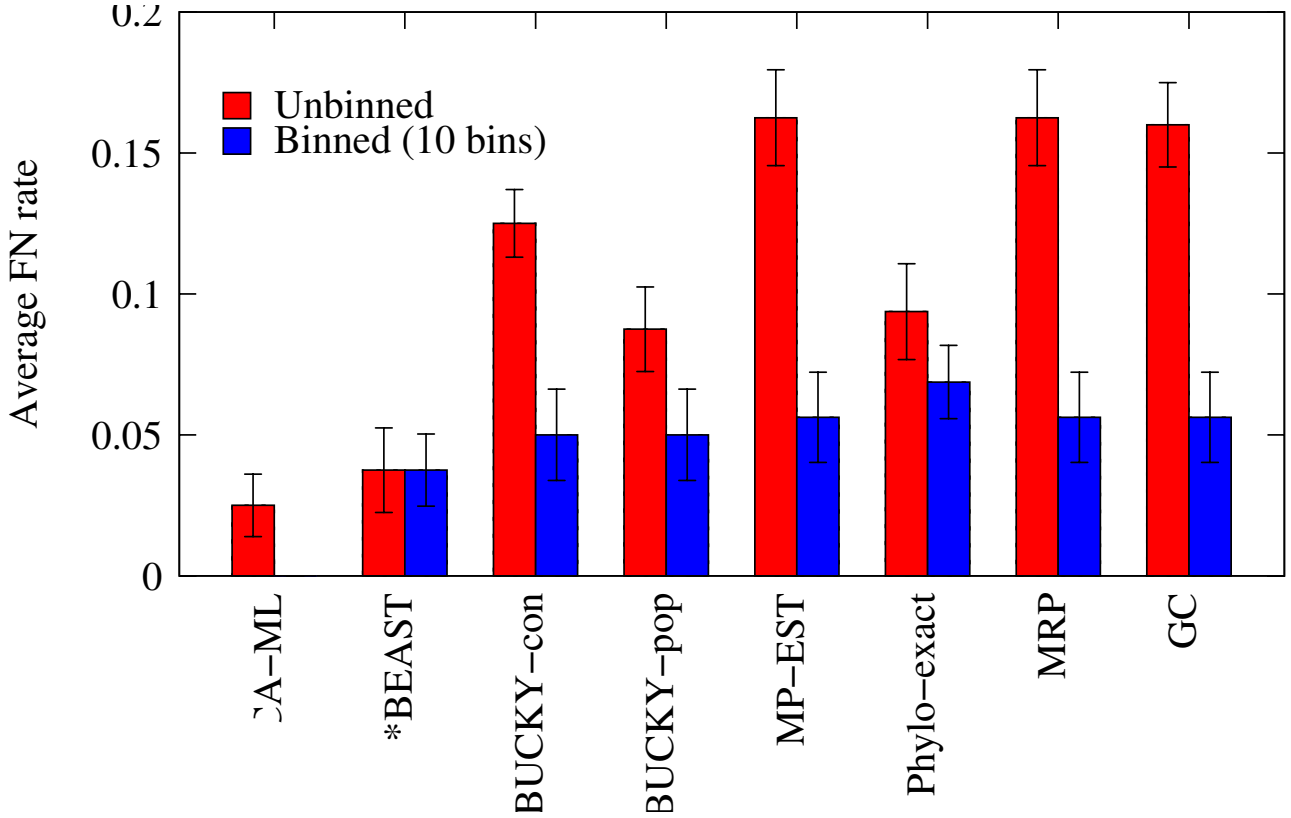
Statistically consistent methods

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

Output: estimated species tree

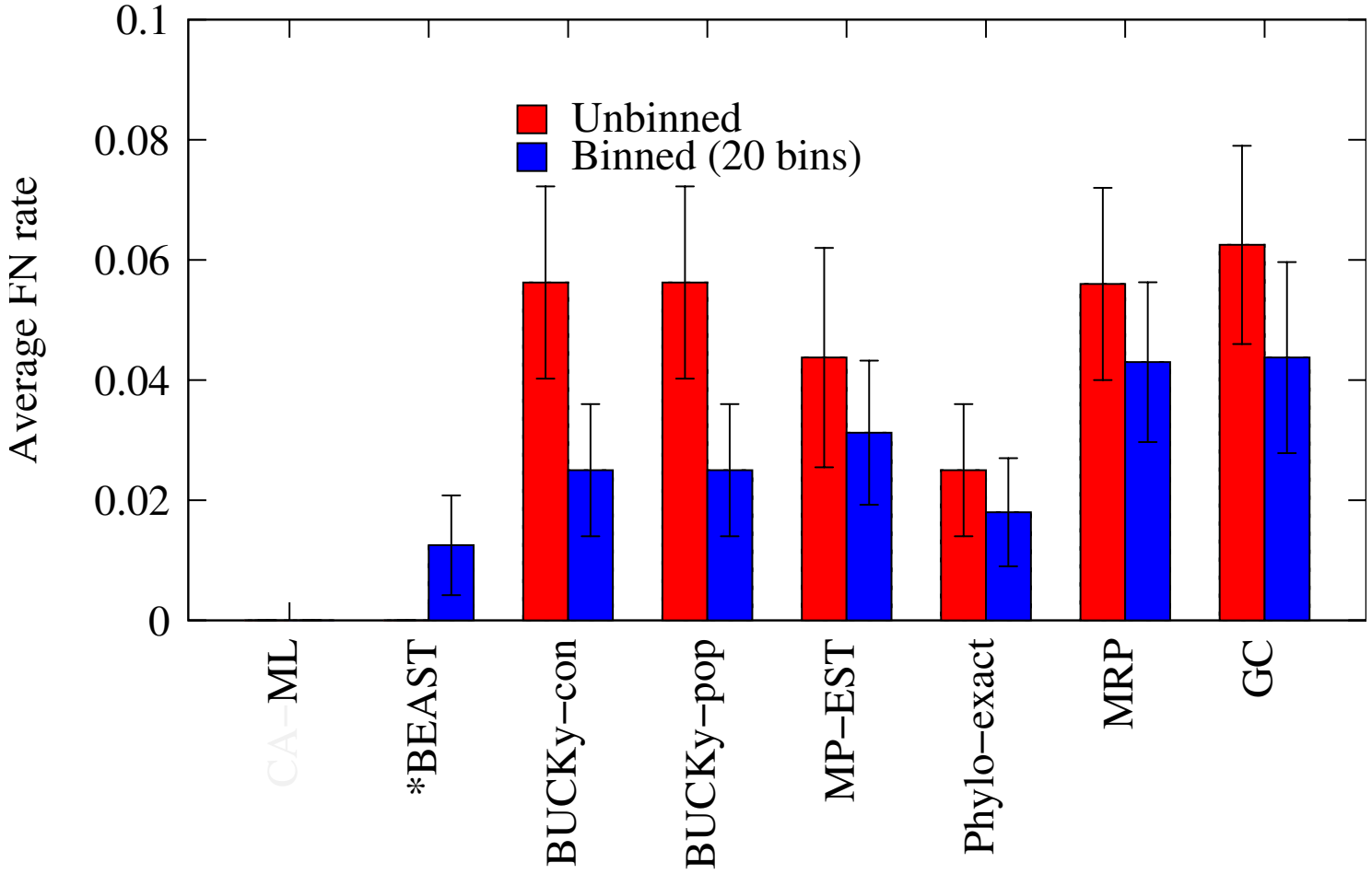
- ***BEAST** (Heled and Drummond 2010): Bayesian co-estimation of gene trees and species trees given sequence alignments
- **MP-EST** (Liu et al. 2010): maximum likelihood estimation of rooted species tree
- **BUCKy-pop** (Ané and Larget 2010): quartet-based Bayesian species tree estimation

Naïve binning vs. unbinned: 50 genes



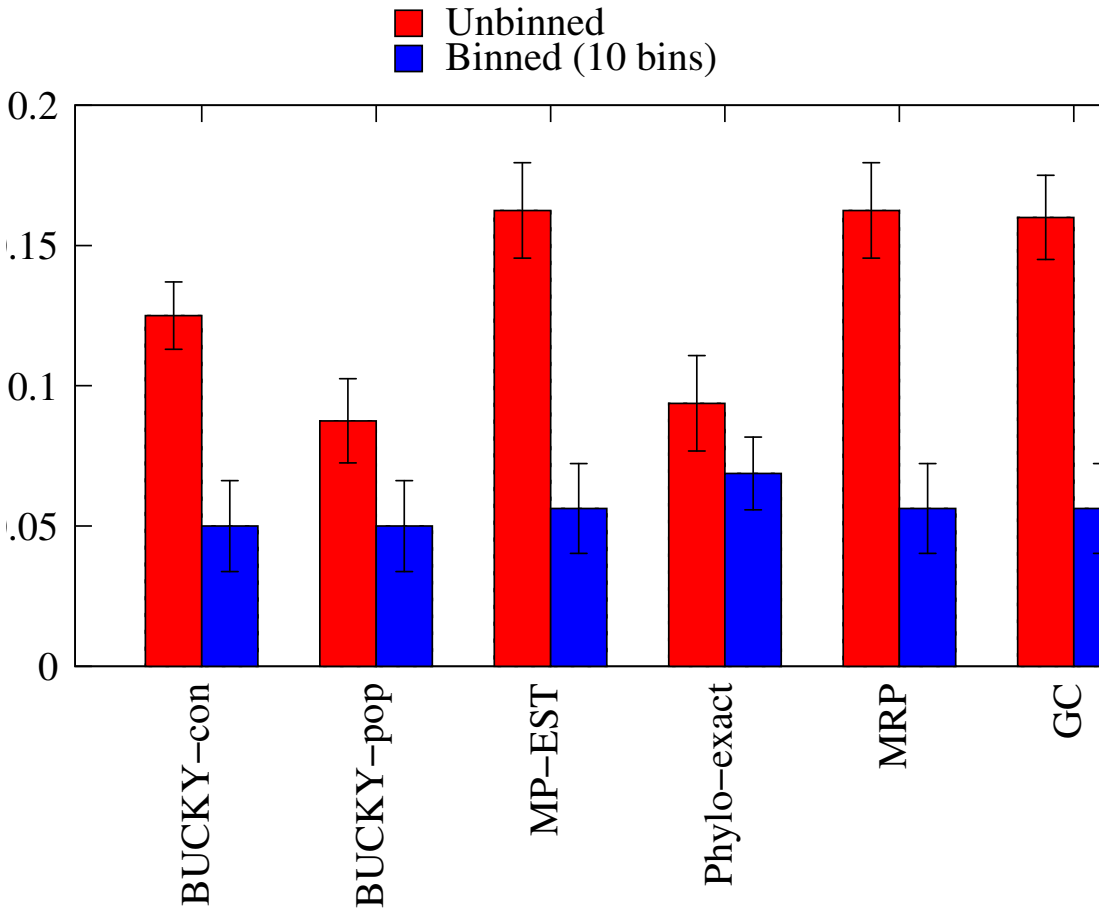
Bayzid and Warnow, Bioinformatics 2013
11-taxon strongILS datasets with 50 genes, 5 genes per bin

Naïve binning vs. unbinned, 100 genes



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

Naïve binning vs. unbinned: 50 genes



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