Hoatzin Chick
Opisthocomus hoazin
This young bird closely resembles its parents which are about a third larger. The white flecks on the face are mallophaga eggs. Most bird lice taxa found on Hoatzins are unique to this host.
© 2009 Photo and Comment by Petroglyph
http://www.flickr.com/photos/28113115@N00/
Licensed under Creative Commons
Attribution 2.0 or later version

New methods for inferring species trees in the presence of incomplete lineage sorting

Tandy Warnow

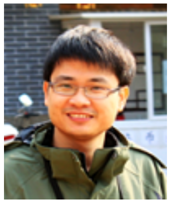
The University of Illinois

Avian Phylogenomics Project

Erich Jarvis,
HHMI

MTP Gilbert,
Copenhagen

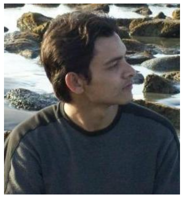
G Zhang,
BGI



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

Challenges:

- Massive gene tree conflict consistent with incomplete lineage sorting
- Maximum likelihood estimation on a million-site genome-scale alignment

In press

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

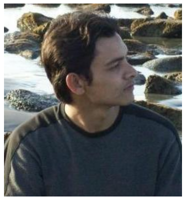
N. Matasci
iPlant

T. Warnow,
UIUC

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin



Plus many many other people…

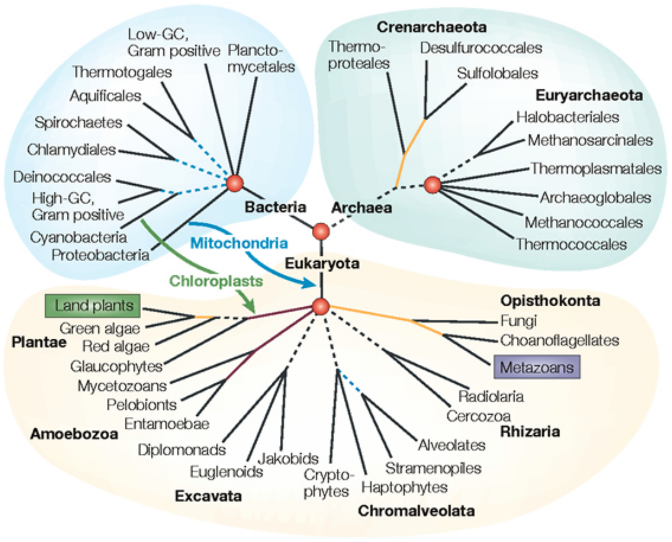
- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenges:

- Massive gene tree conflict consistent with incomplete lineage sorting
- Multiple sequence alignment of >100,000 sequences (with lots of fragments!)

In press

The Tree of Life: *Multiple Challenges*



Nature Reviews | Genetics

Large datasets:

100,000+ sequences

10,000+ genes

“BigData” complexity

Large-scale statistical phylogeny estimation

Ultra-large multiple-sequence alignment

Estimating species trees from incongruent gene trees

Supertree estimation

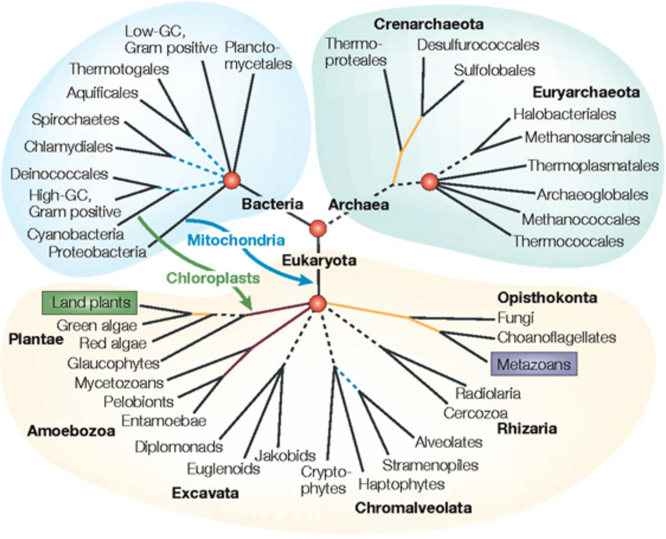
Genome rearrangement phylogeny

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

Large-scale statistical phylogeny estimation

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

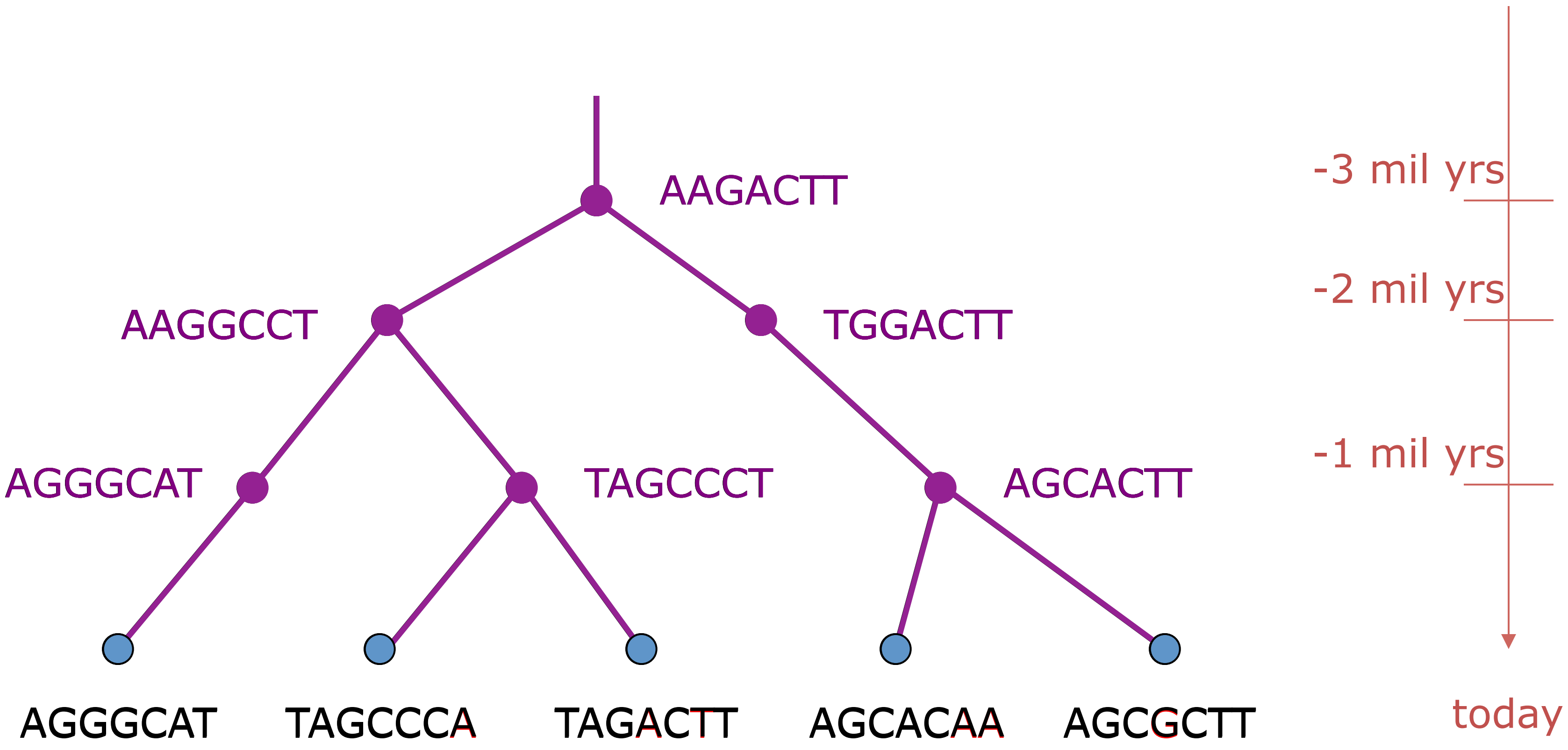
Data mining techniques to explore multiple optima

This talk

This talk

- Statistical gene tree estimation
  - Models of evolution
  - Identifiability and statistical consistency
  - Absolute fast converging methods
- Statistical species tree estimation
  - Gene tree conflict due to incomplete lineage sorting
  - The multi-species coalescent model
  - Identifiability and statistical consistency
  - New methods for species tree estimation
    - Statistical Binning (in press)
    - ASTRAL (Bioinformatics 2014)
  - The challenge of gene tree estimation error

DNA Sequence Evolution (Idealized)



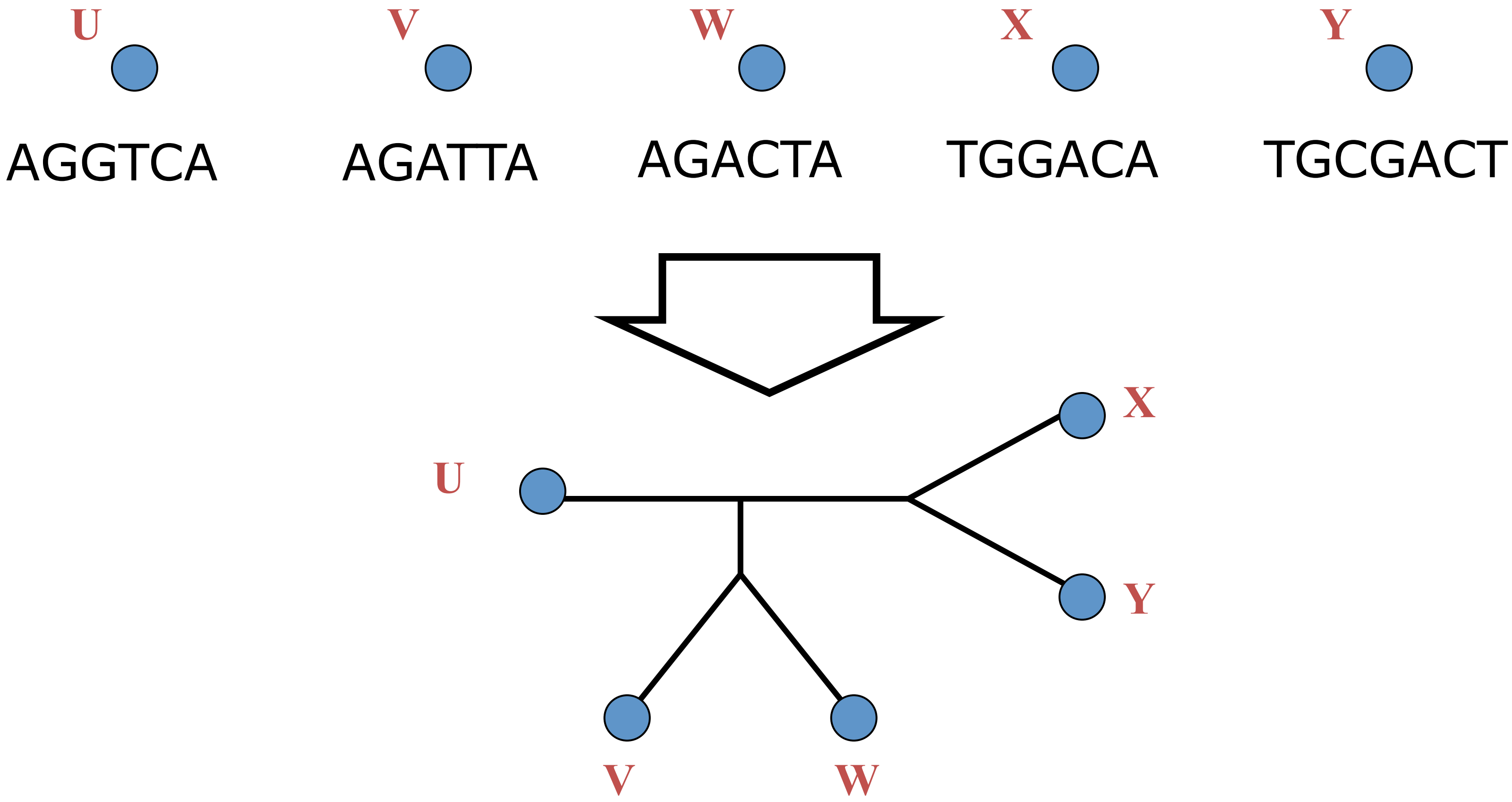
Markov Model of Site Evolution

Simplest (Jukes-Cantor, 1969):

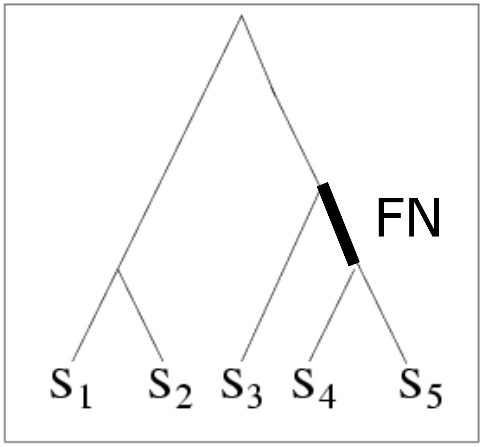
- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge*, it changes with equal probability to each of the remaining states.*
- The evolutionary process is Markovian.

The different sites are assumed to evolve independently and identically down the tree (with rates that are drawn from a gamma distribution).

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



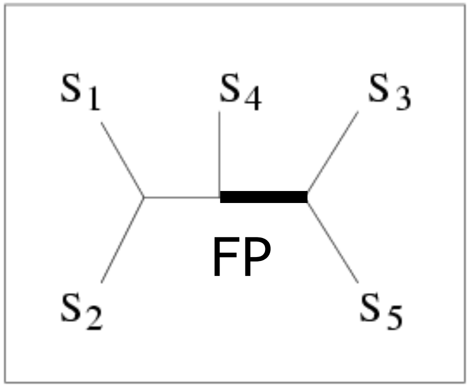
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

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

50% error rate

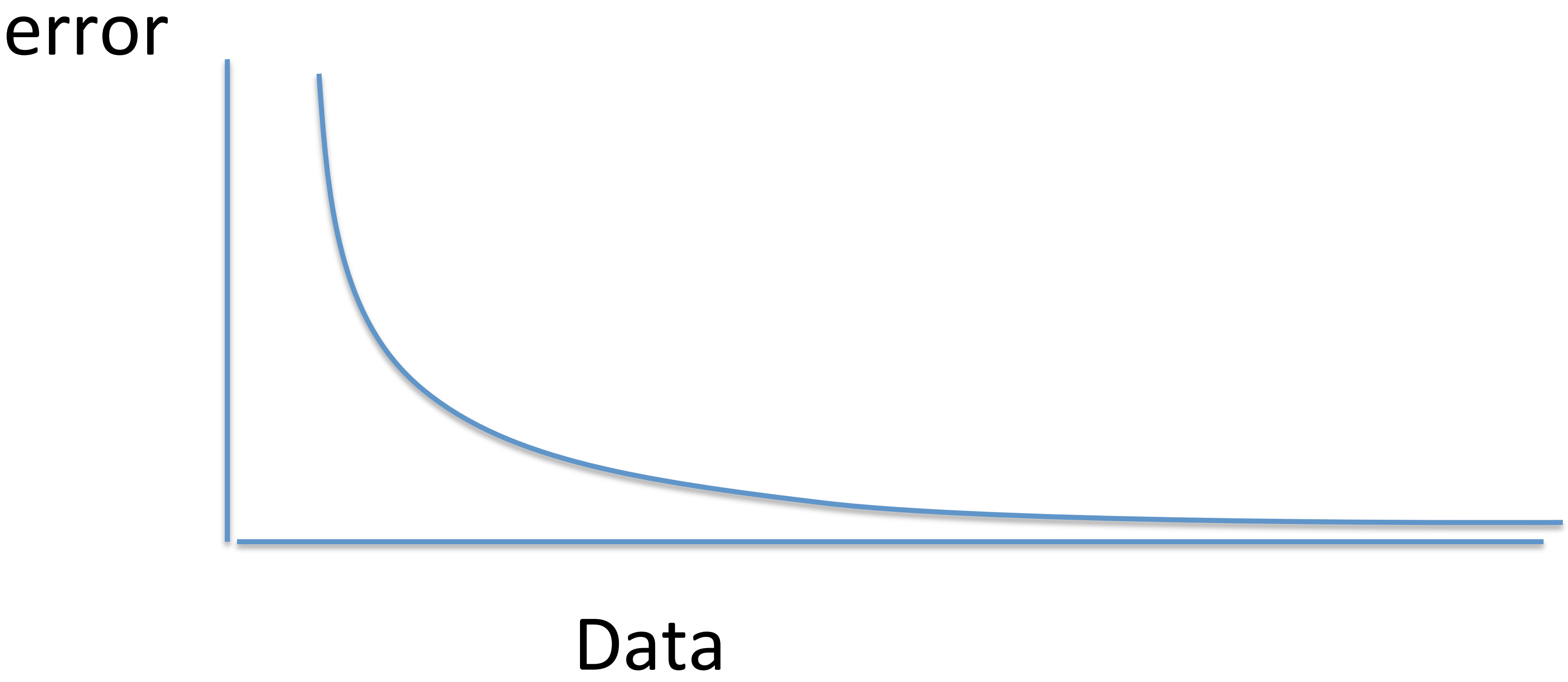


INFERRED TREE

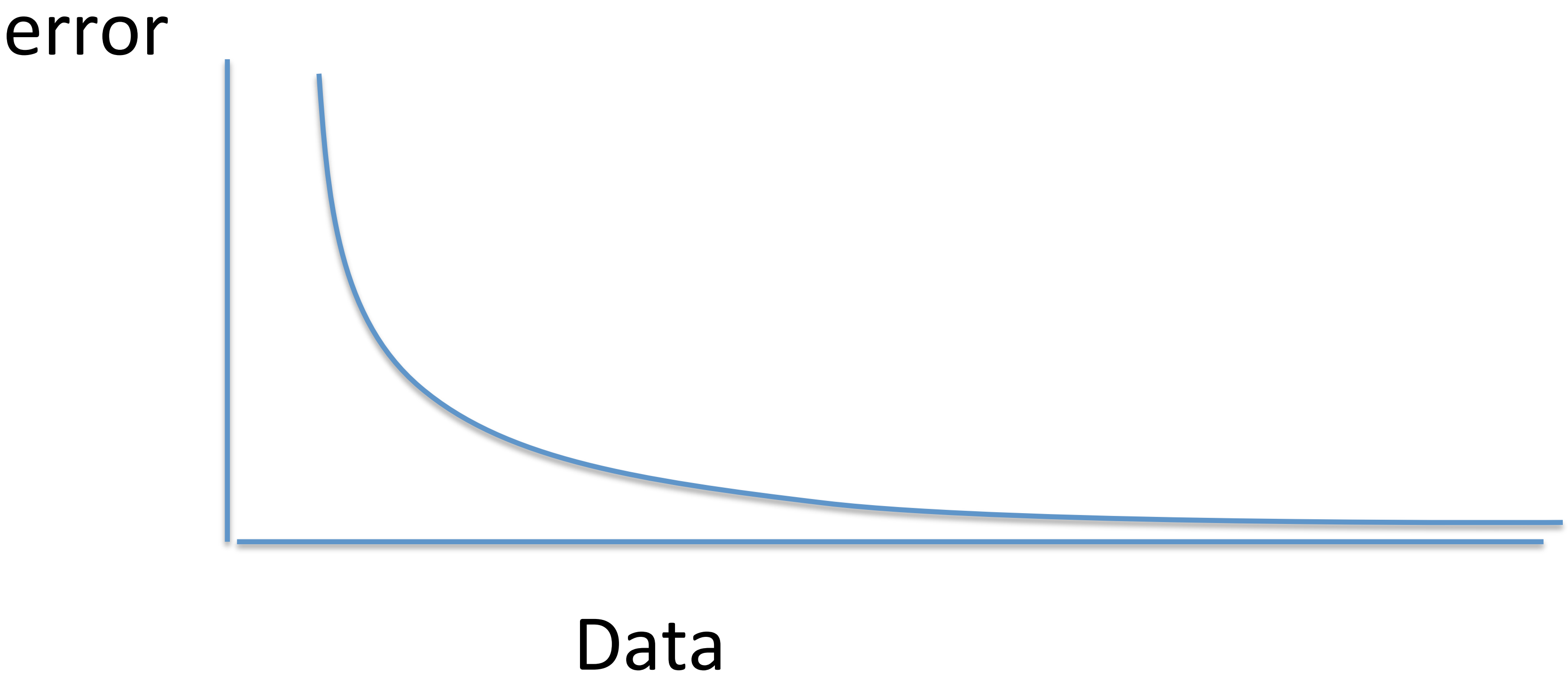
Questions

- 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?

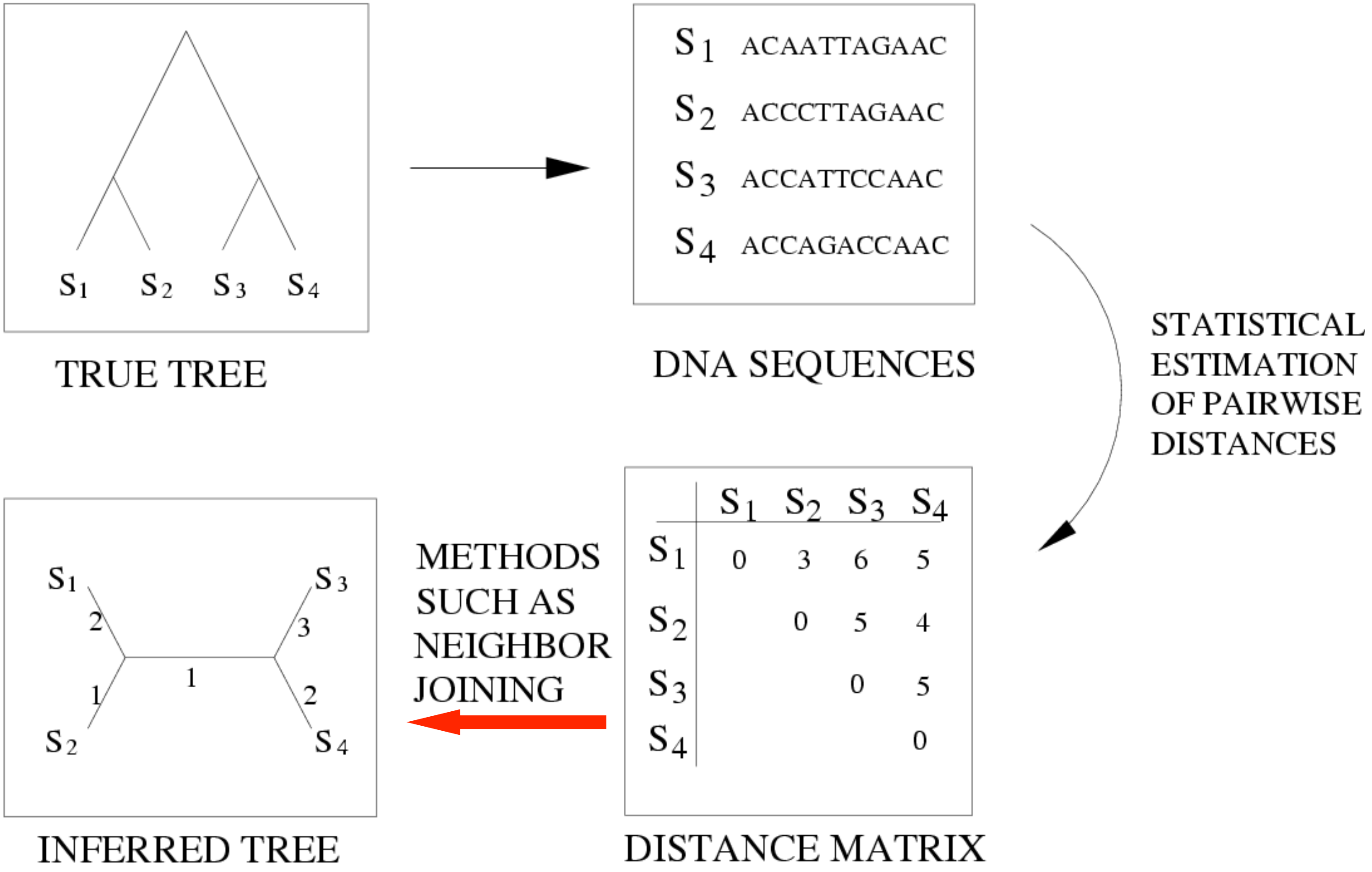
Statistical Consistency



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

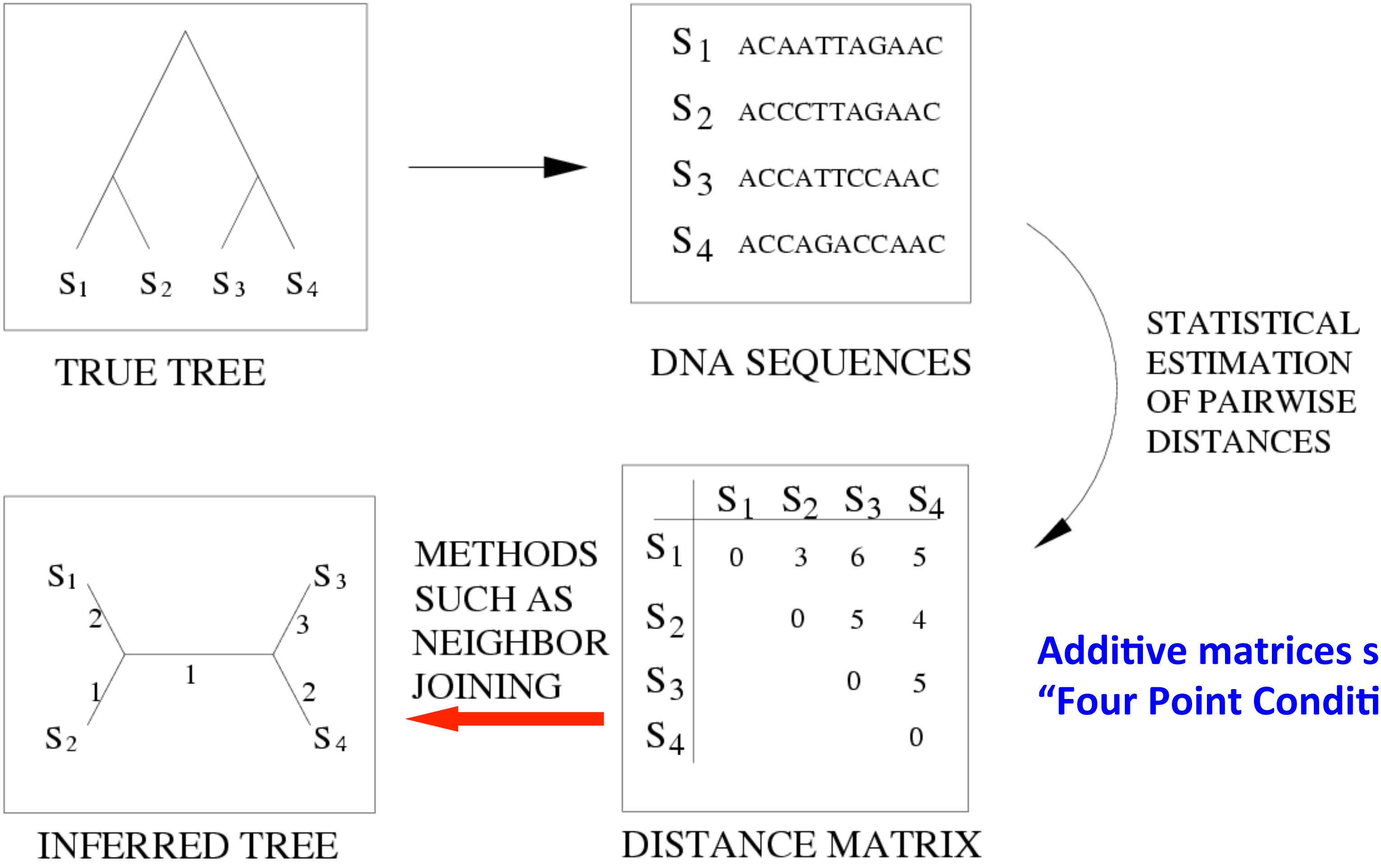
	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS
SUCH AS
NEIGHBOR
JOINING

INFERRED TREE

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



TRUE TREE

S1 ACAATTAGAAC

S2 ACCCTTAGAAC

S3 ACCATTCCAAC

S4 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

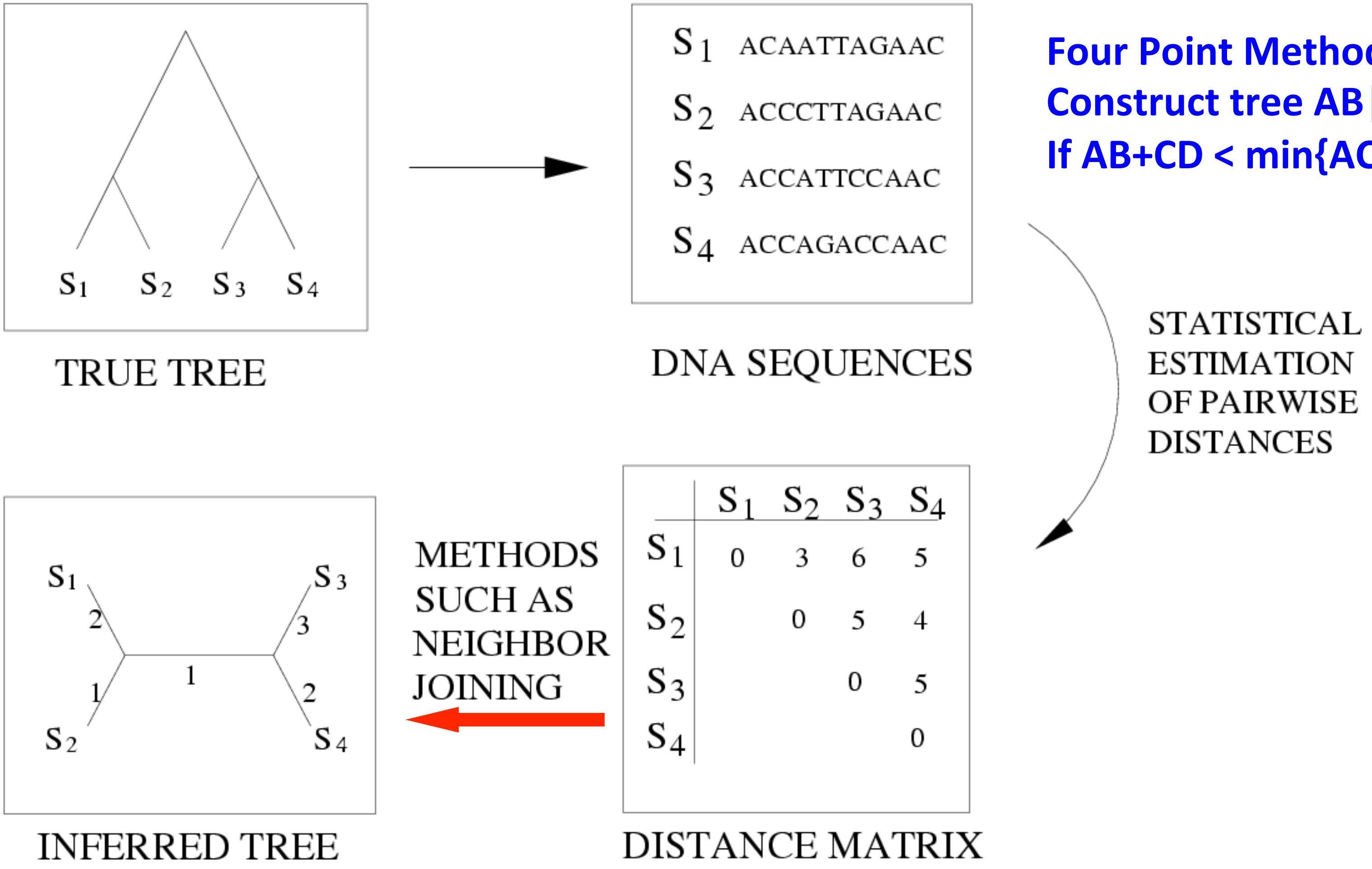
METHODS SUCH AS NEIGHBOR JOINING

INFERRED TREE

Additive matrices satisfy the "Four Point Condition"

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

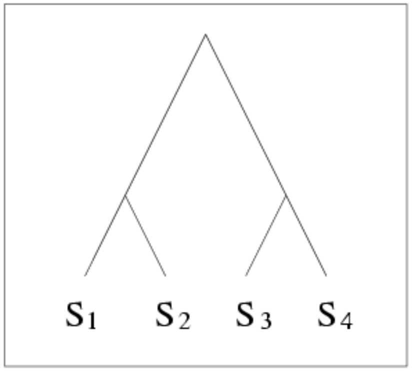
Four Point Method:
Construct tree AB|CD
If AB+CD < min{AC+BD,AD+BC}



S_1 ACAATTAGAAC
S_2 ACCCTTAGAAC
S_3 ACCATTCCAAC
S_4 ACCAGACCAAC

	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

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



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

DNA SEQUENCES

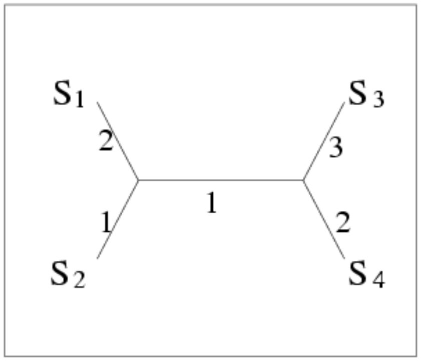
Four Point Method:
Construct tree AB|CD
If AB+CD < min{AC+BD,AD+BC}

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

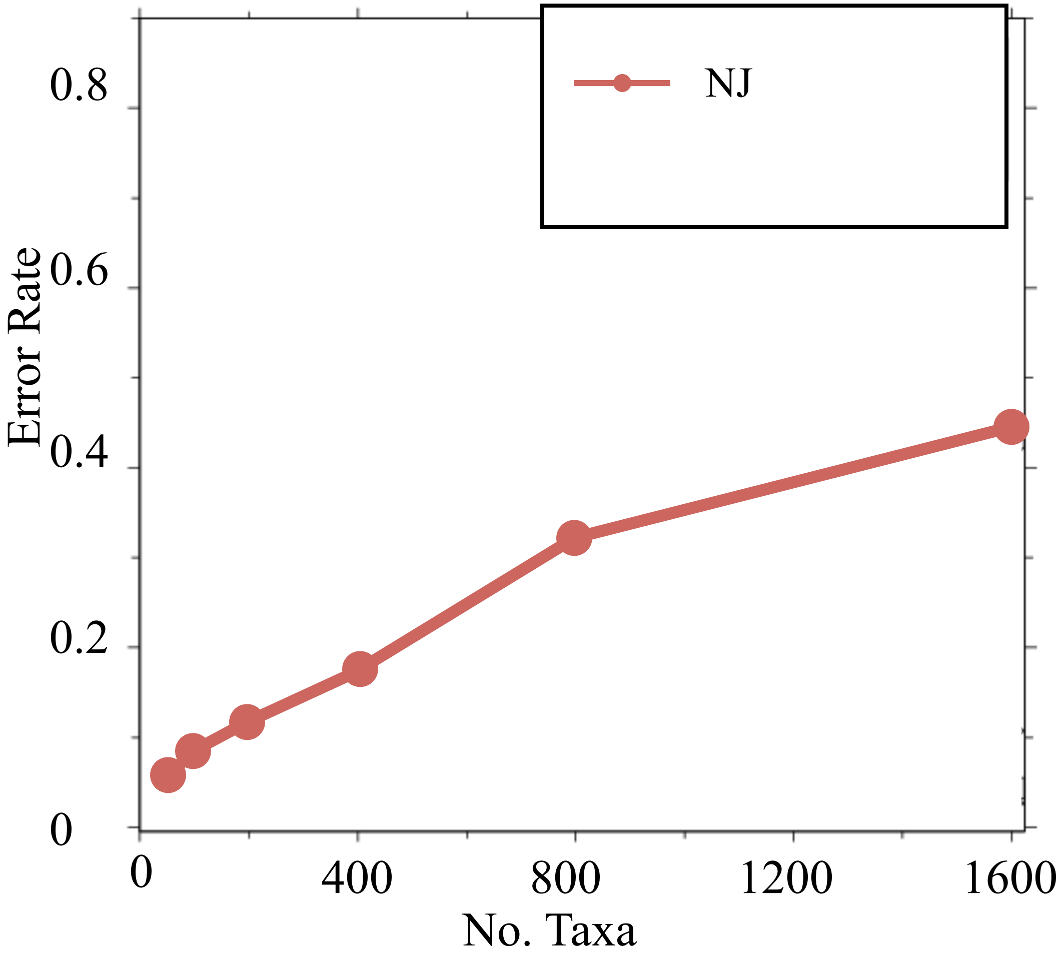


INFERRED TREE

Constructing larger trees:
(1) Compute quartet trees using FPM
(2) Determine if the quartet trees are "compatible"; if so, return the tree on which they agree. Else Return FAIL.

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

Neighbor Joining on large diameter trees

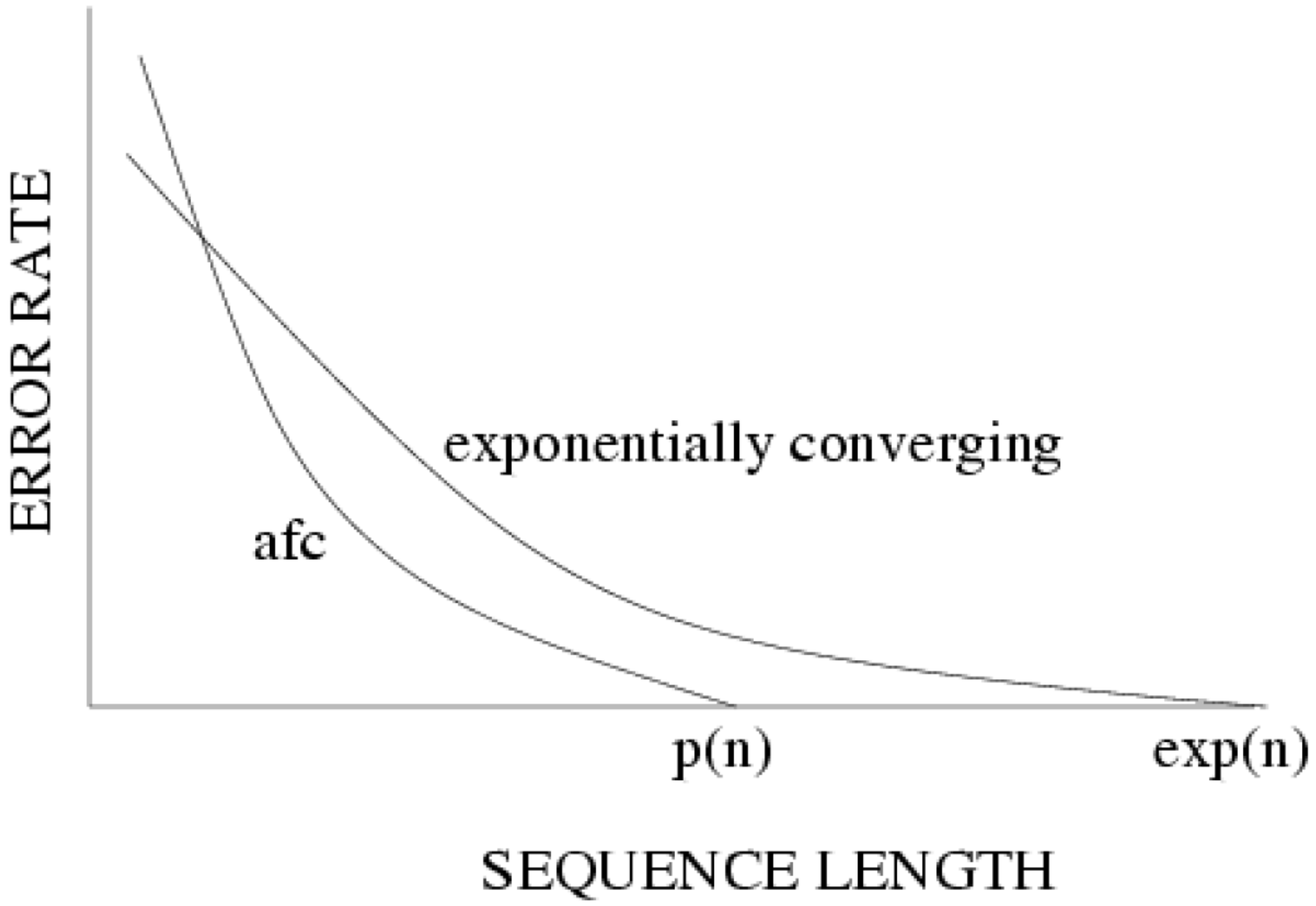


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

[Nakhleh et al. ISMB 2001]

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

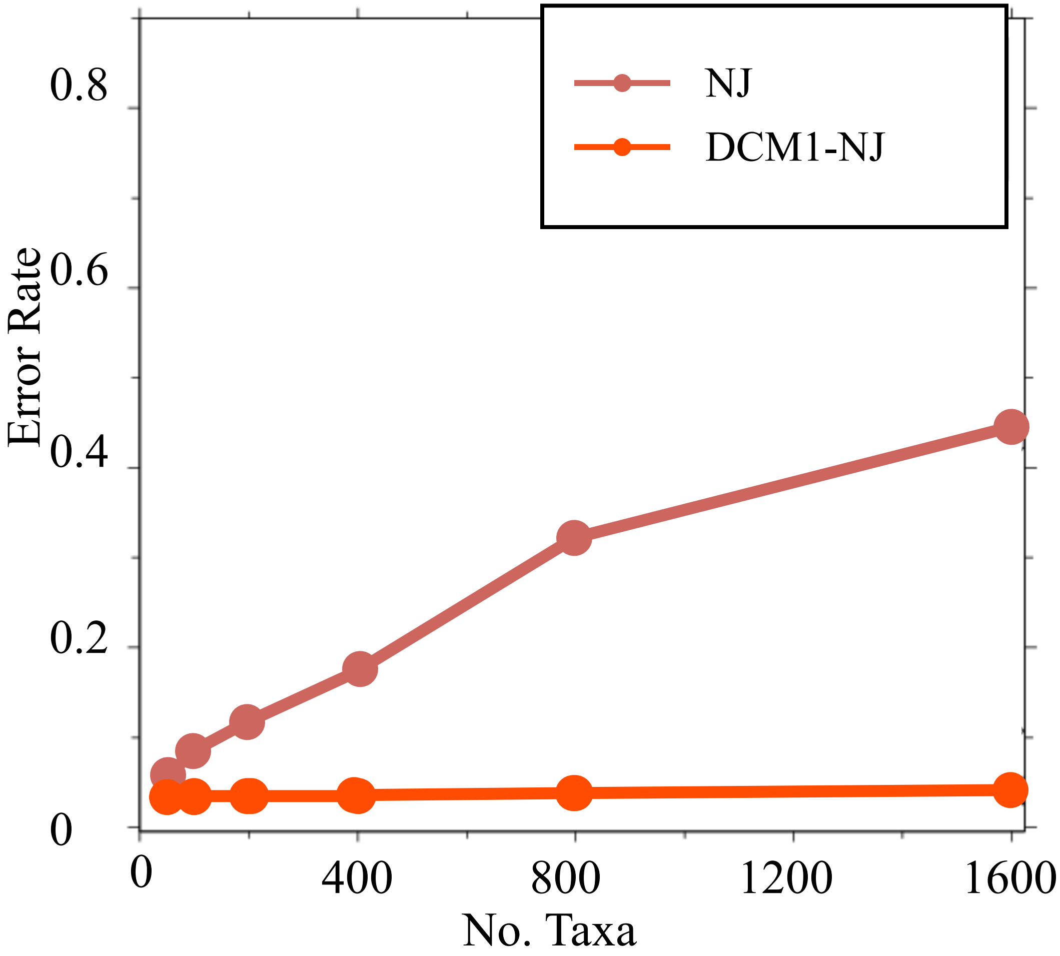


Fast-converging methods (and related work)

- 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)

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.

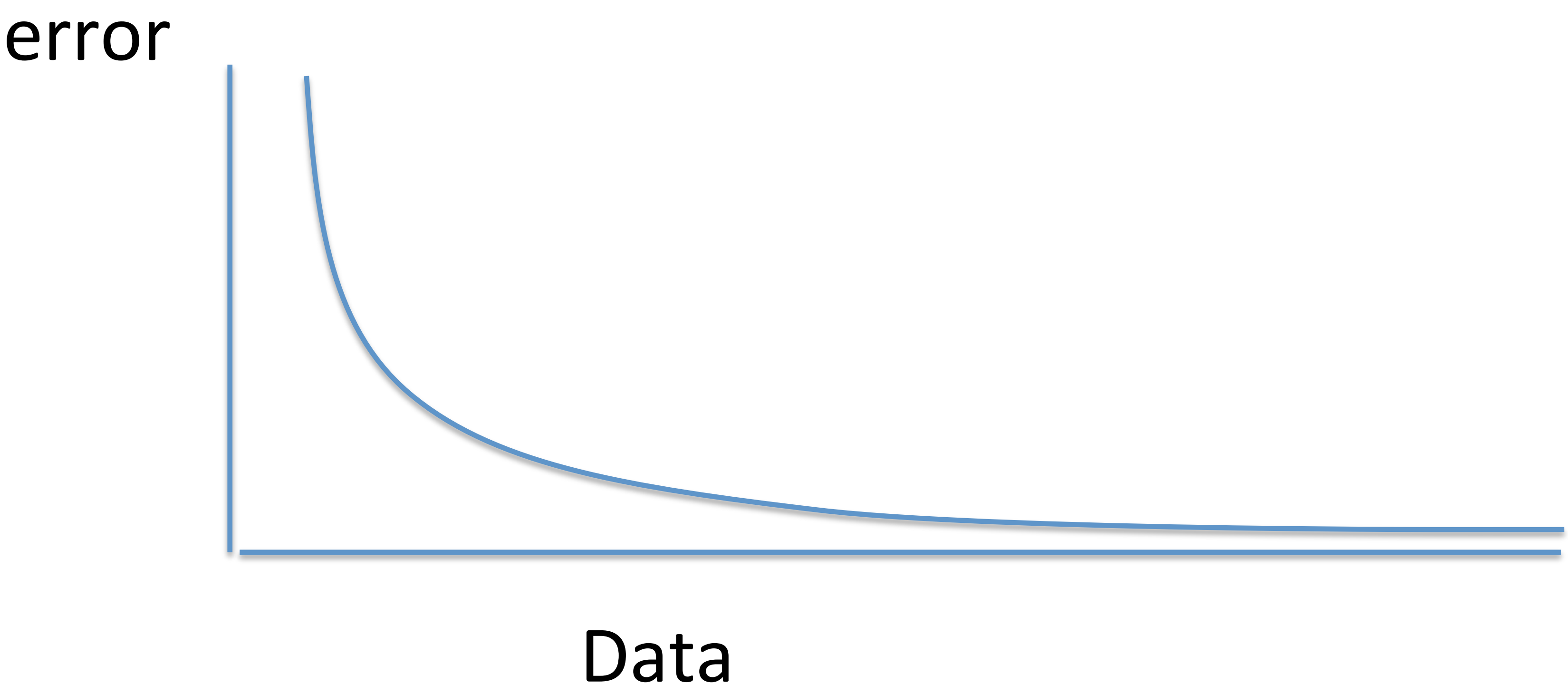
Answers?

- 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?

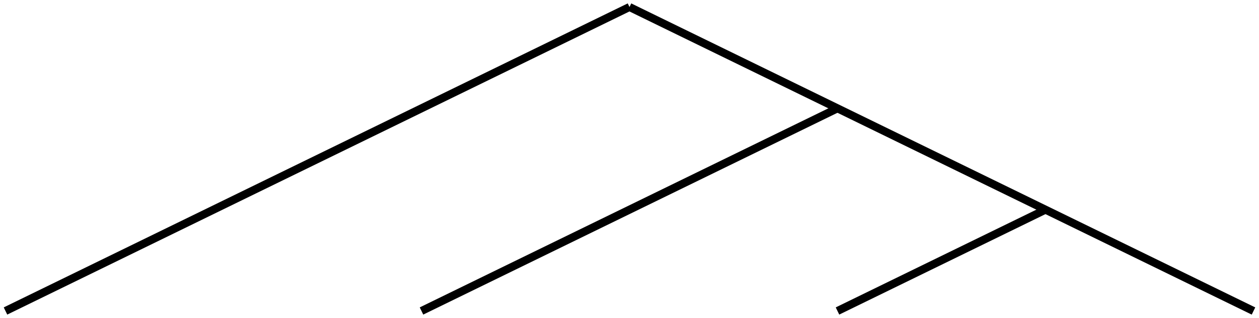
- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.
- Just about everything is NP-hard, and the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

In other words…

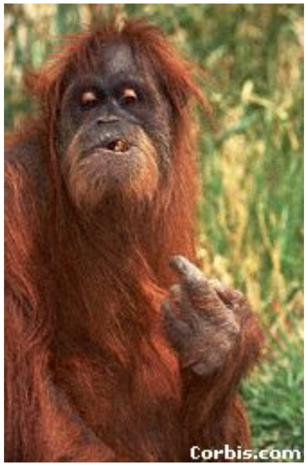


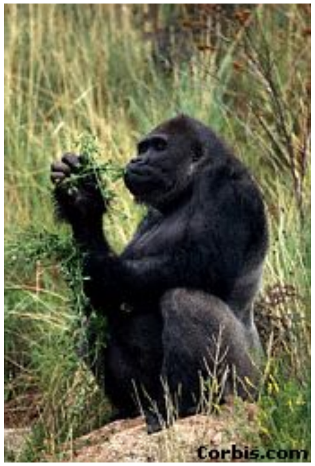
Statistical consistency doesn’t guarantee accuracy w.h.p. unless the sequences ***are long enough.***

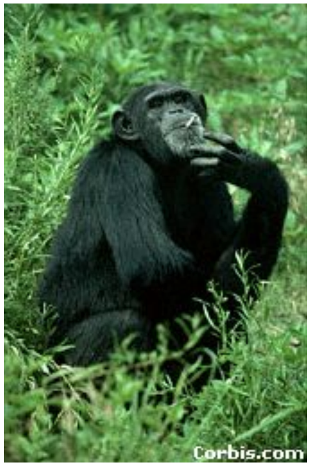
Phylogeny
(evolutionary tree)



Orangutan Gorilla Chimpanzee Human

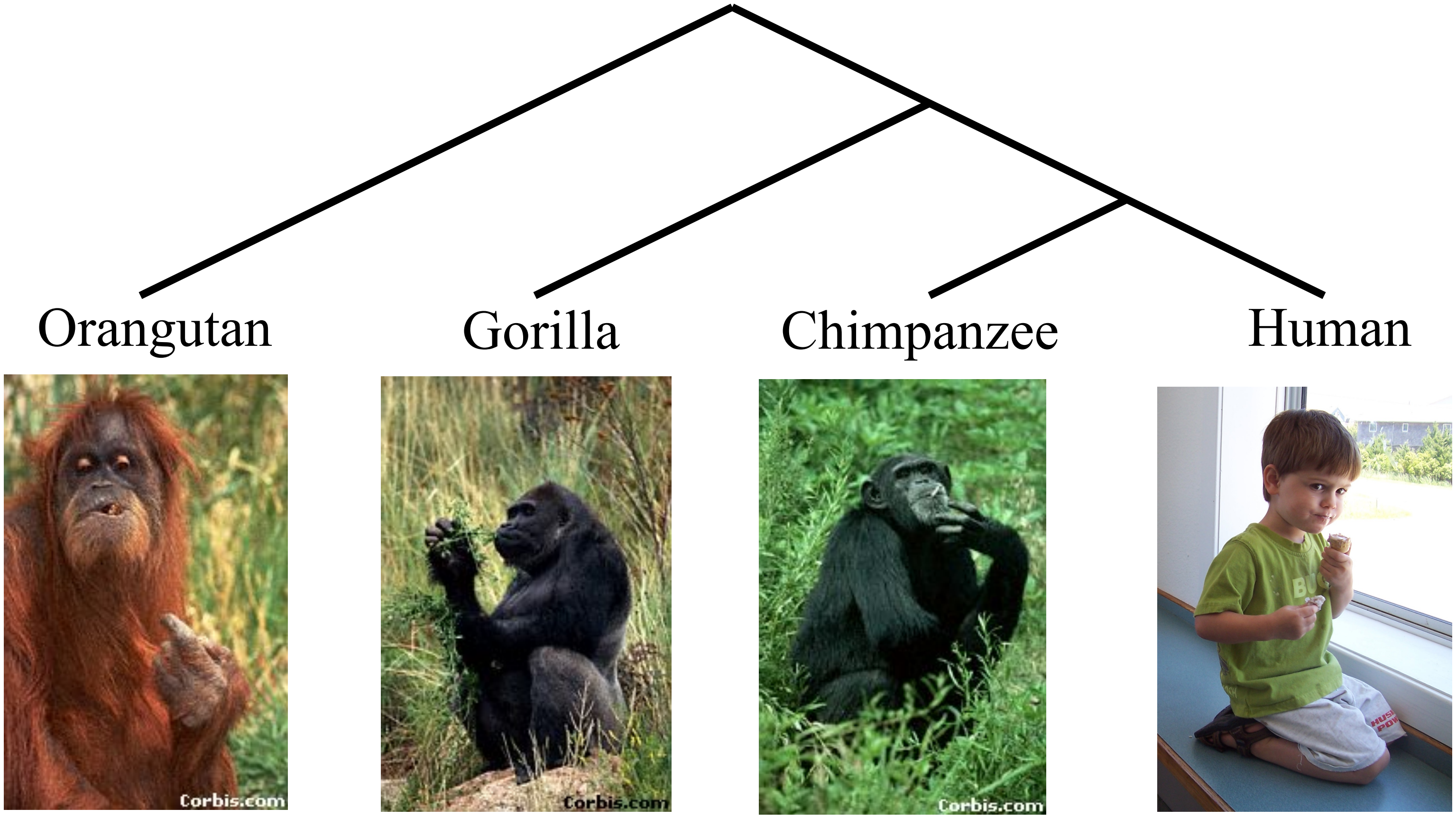






From the Tree of the Life Website,
University of Arizona

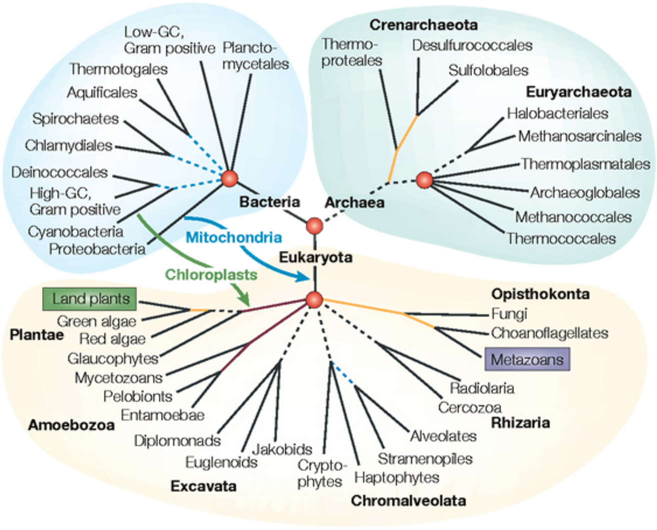
Sampling multiple genes from multiple species



From the Tree of the Life Website,
University of Arizona

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Using multiple genes

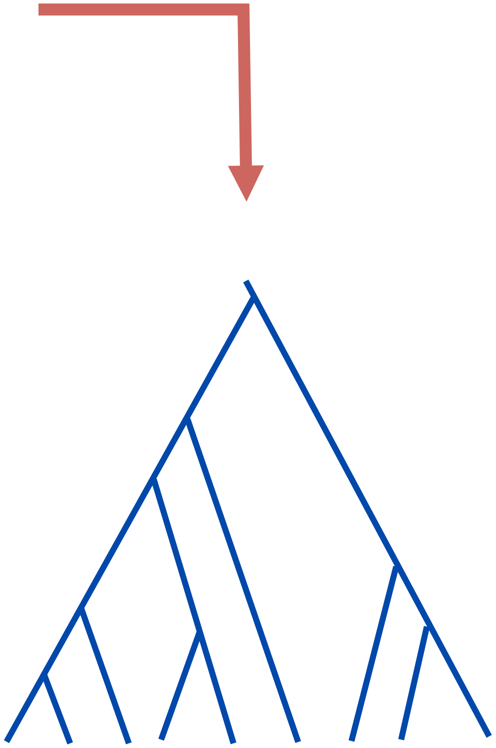
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

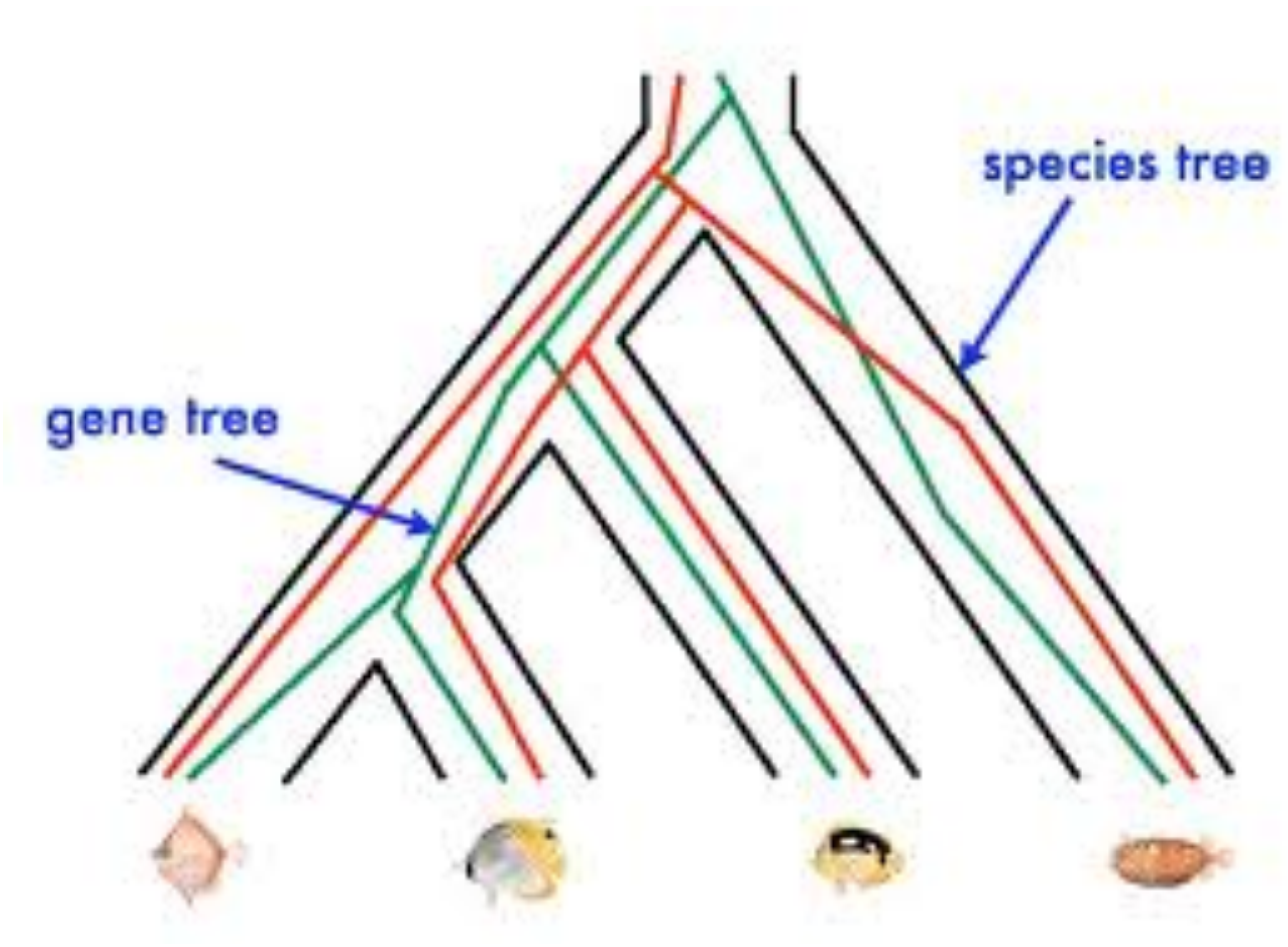
	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

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)



Avian Phylogenomics Project

E Jarvis,
HHMI

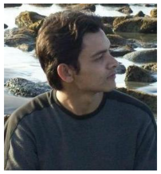
MTP Gilbert,
Copenhagen

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)

1KP: Thousand Transcriptome Project

Gene Tree Incongruence

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

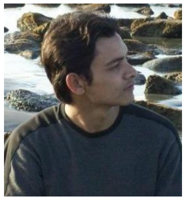
N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin



Md. S.Bayzid
UT-Austin

- 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)

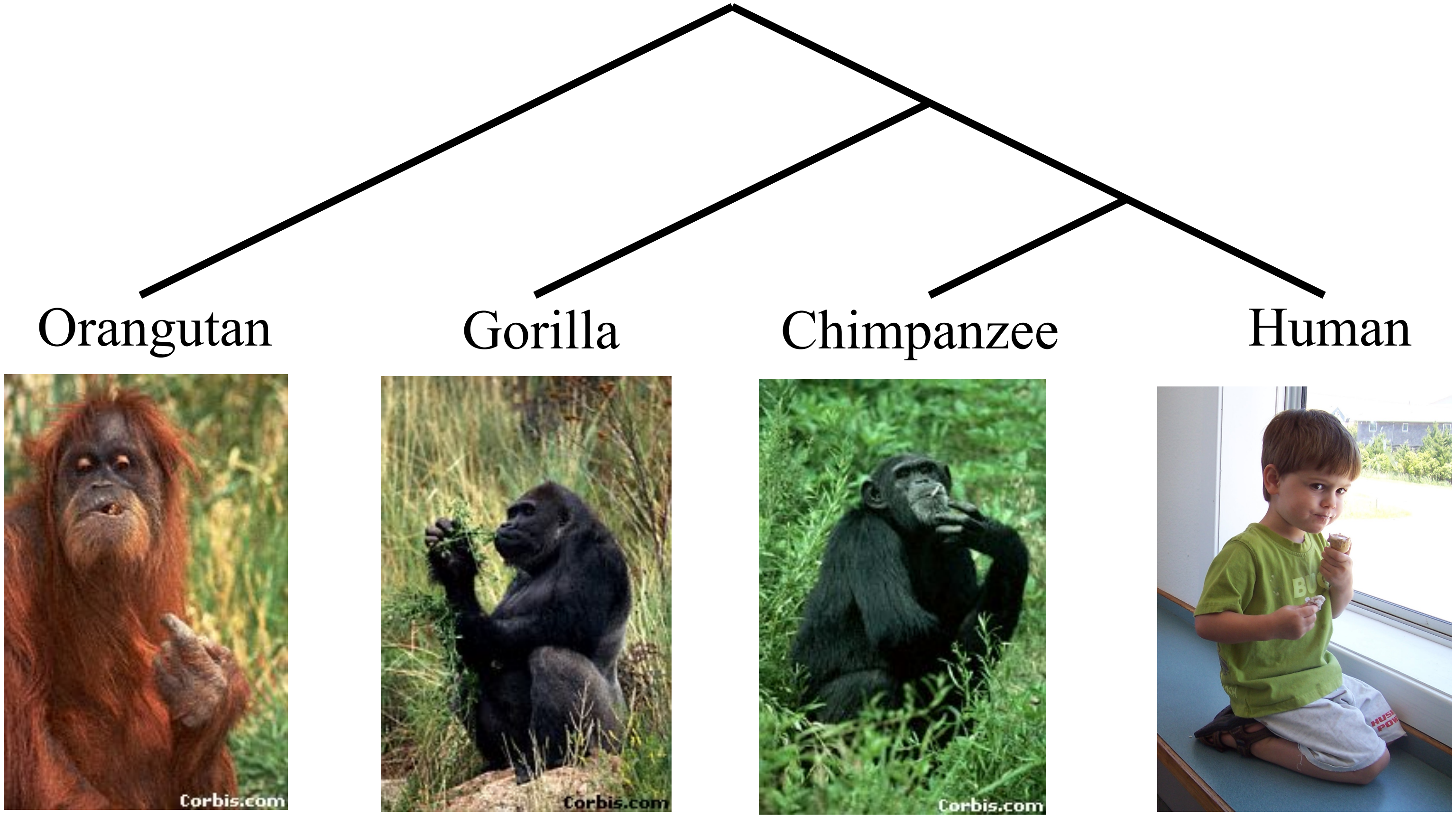
Gene Tree Incongruence

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

Incomplete Lineage Sorting (ILS)

- 1000+ 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.

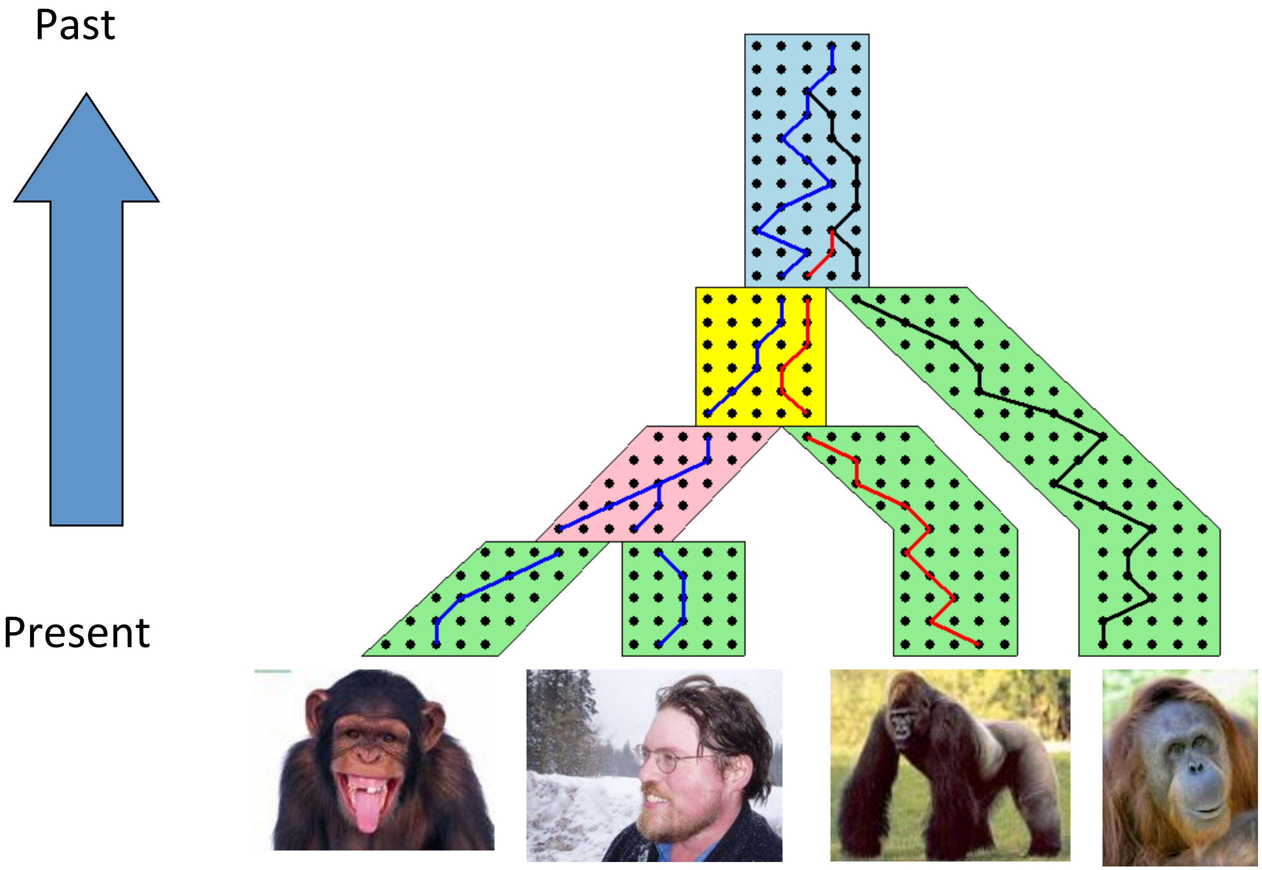
Species tree estimation: difficult, even for small datasets!



From the Tree of the Life Website,
University of Arizona

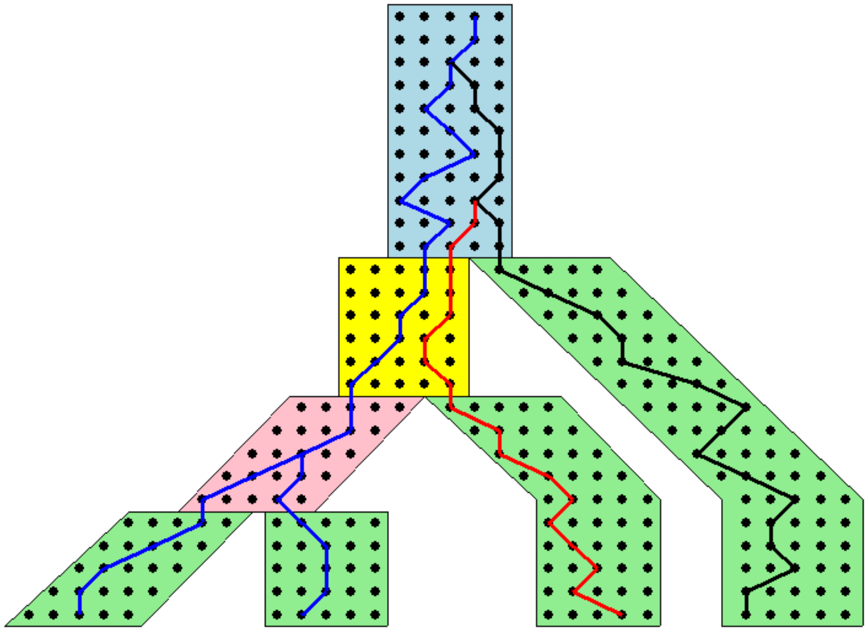
The Coalescent

Courtesy James Degnan



Gene tree in a species tree

Courtesy James Degnan

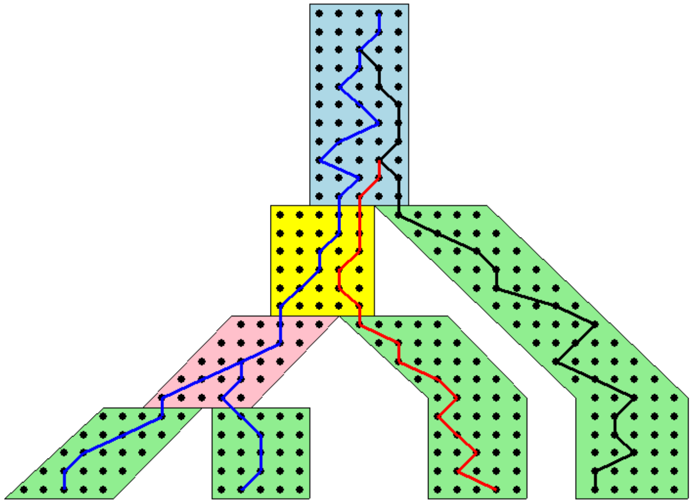


Lineage Sorting

- Population-level process, also called the "Multi-species coalescent" (Kingman, 1982)
- Gene trees can differ from species trees due to short times between speciation events or large population size; this is called "Incomplete Lineage Sorting" or "Deep Coalescence".

Key observation:

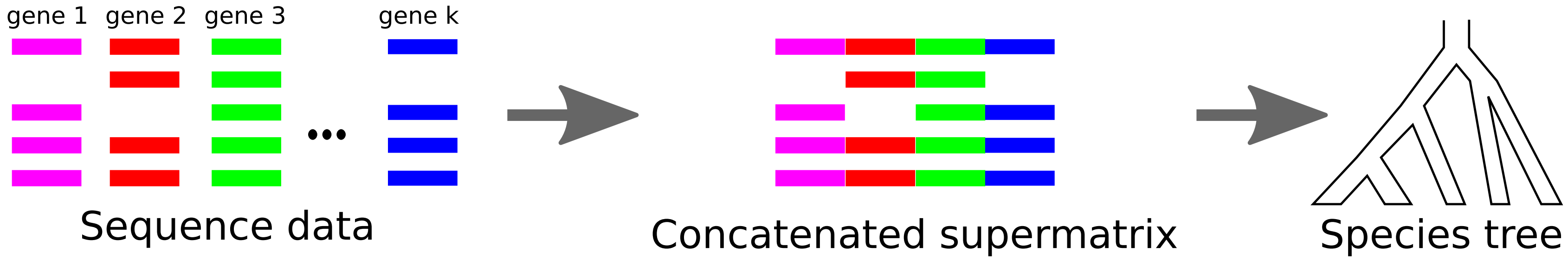
Under the multi-species coalescent model, the species tree defines a *probability distribution on the gene trees, and is identifiable from the distribution on gene trees*



Courtesy James Degnan

Species tree estimation

1- Concatenation: statistically inconsistent (Roch & Steel 2014)



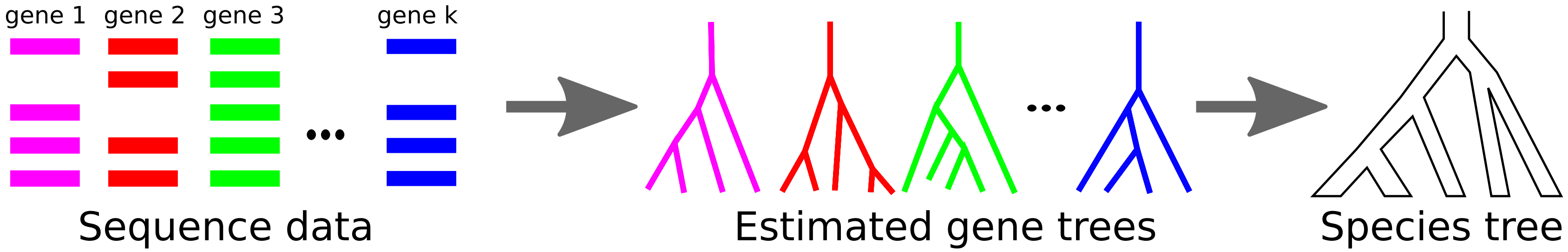
gene 1 gene 2 gene 3 … gene k

Sequence data

Concatenated supermatrix

Species tree

2- Summary methods: can be statistically consistent



gene 1 gene 2 gene 3 … gene k

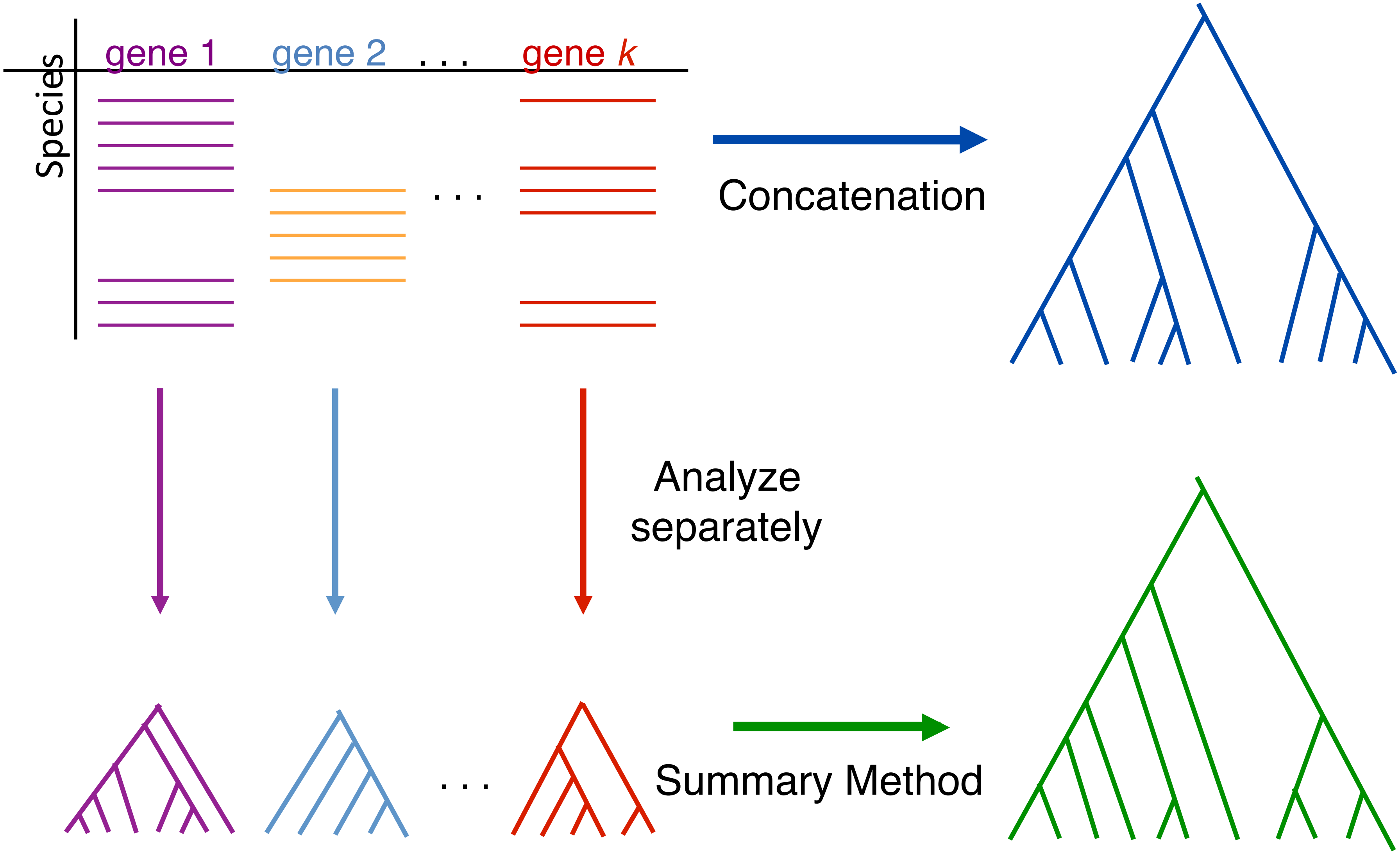
Sequence data

Estimated gene trees

Species tree

3- Co-estimation methods: too slow for large datasets

Two competing approaches



Species

gene 1 gene 2 . . . gene *k*

. . .

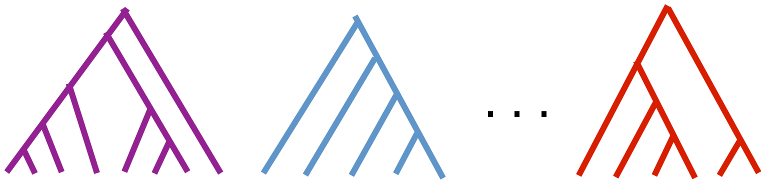
Concatenation

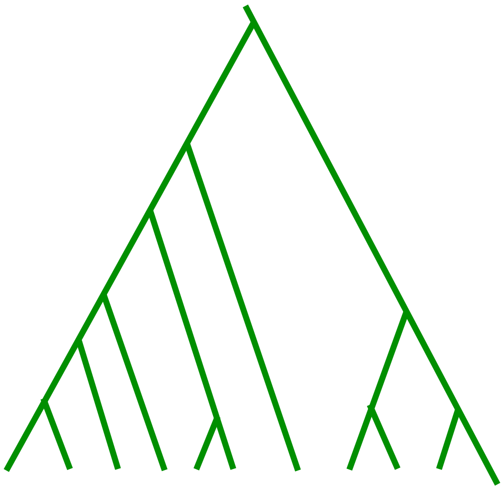
Analyze separately

. . .

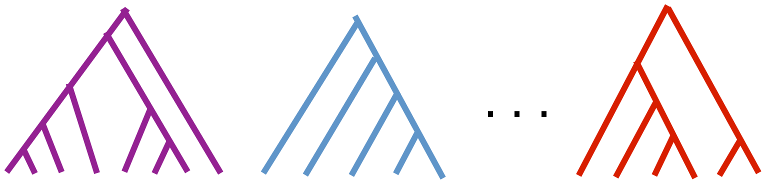
Summary Method

How to compute a species tree?



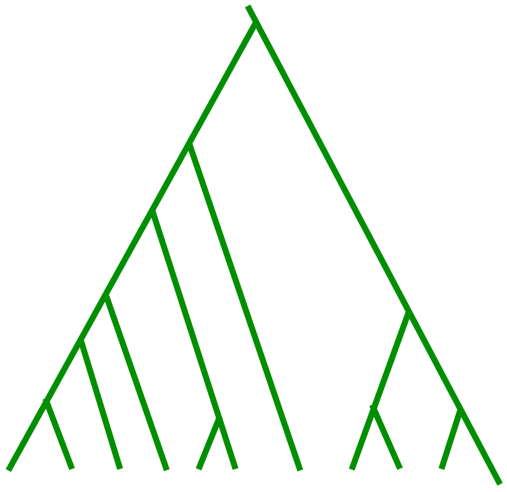


How to compute a species tree?



Techniques:

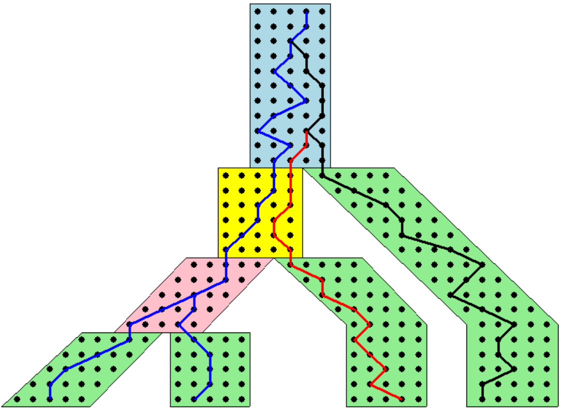
- Most frequent gene tree?
- Consensus of gene trees?
- Other?



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

Courtesy James Degnan

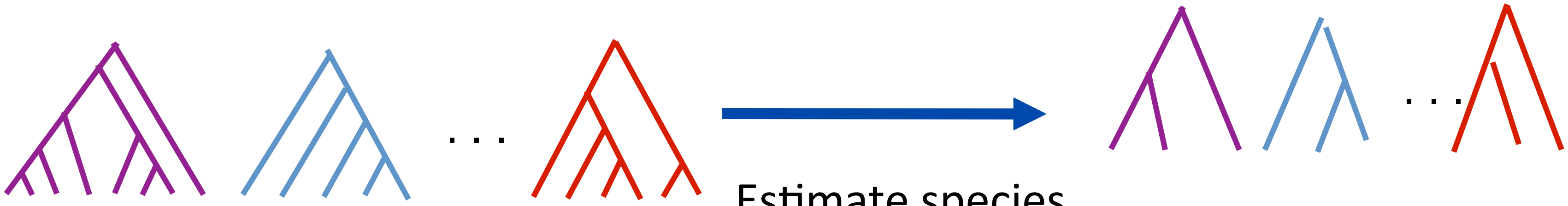
Theorem (Degnan et al., 2006, 2009): Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on {A,B,C} is identical to the rooted species tree induced on {A,B,C}.



How to compute a species tree?

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}.

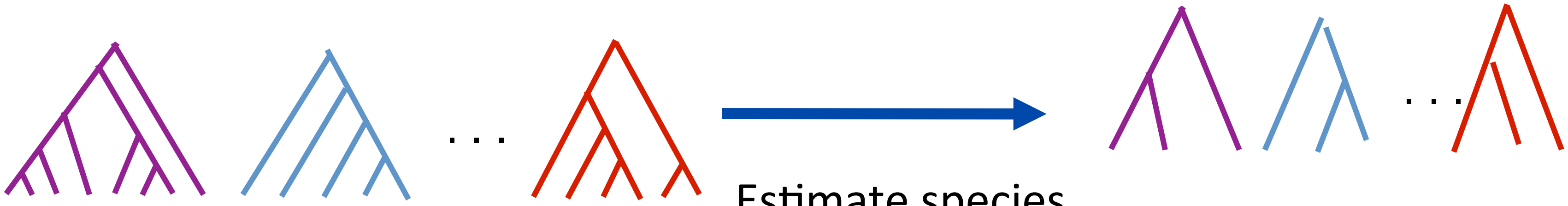
How to compute a species tree?



Estimate species tree for every 3 species

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}.

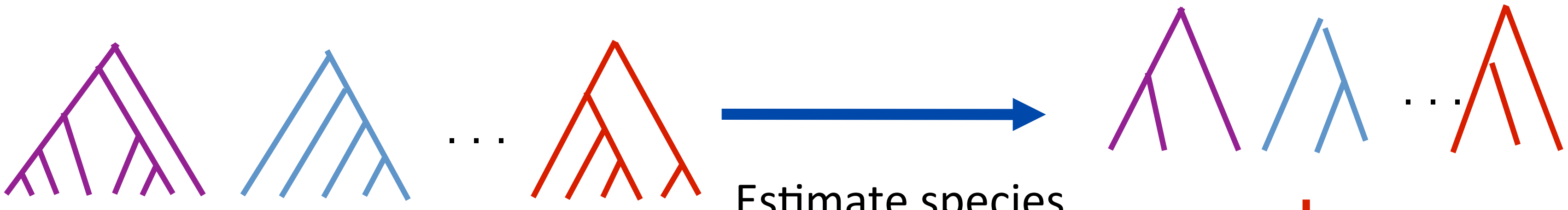
How to compute a species tree?



Estimate species
tree for every
3 species

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

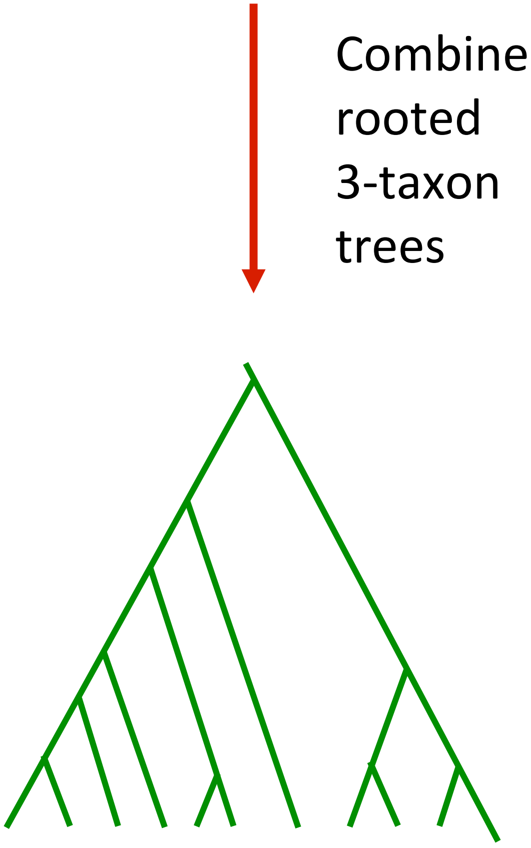
How to compute a species tree?



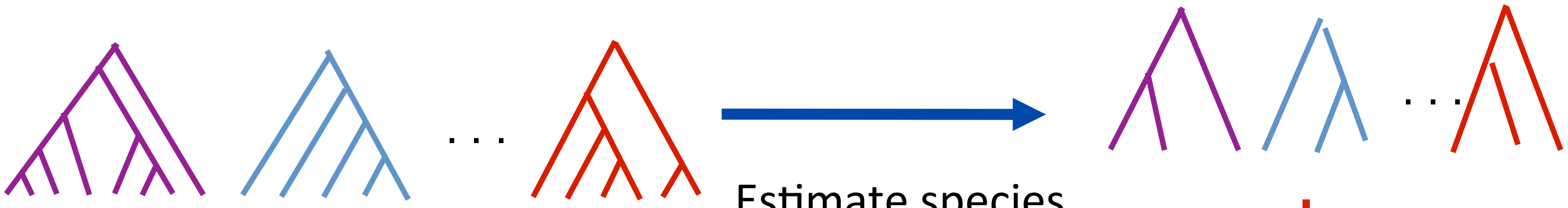
Estimate species tree for every 3 species

Combine rooted 3-taxon trees

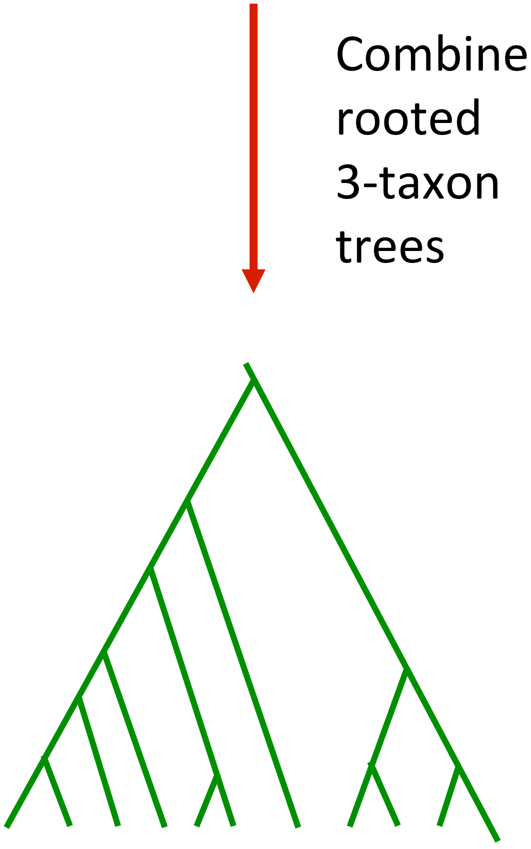
Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.



How to compute a species tree?



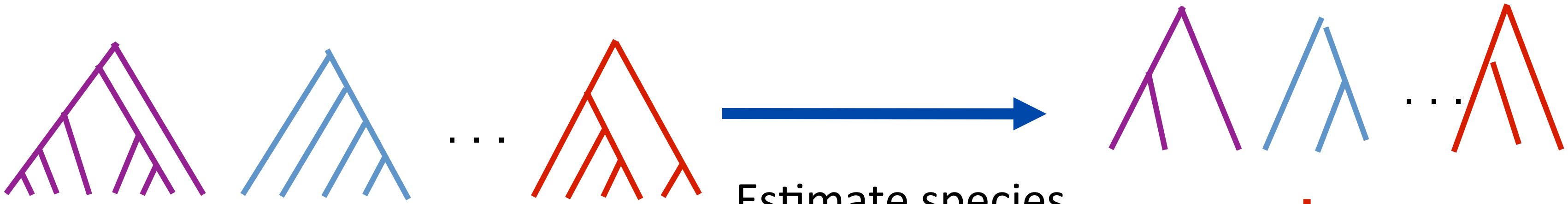
Estimate species tree for every 3 species



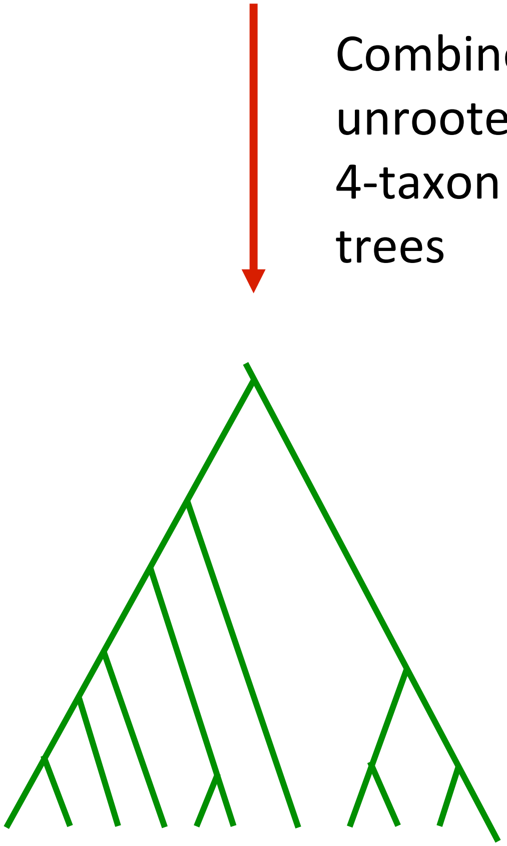
Combine rooted 3-taxon trees

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?



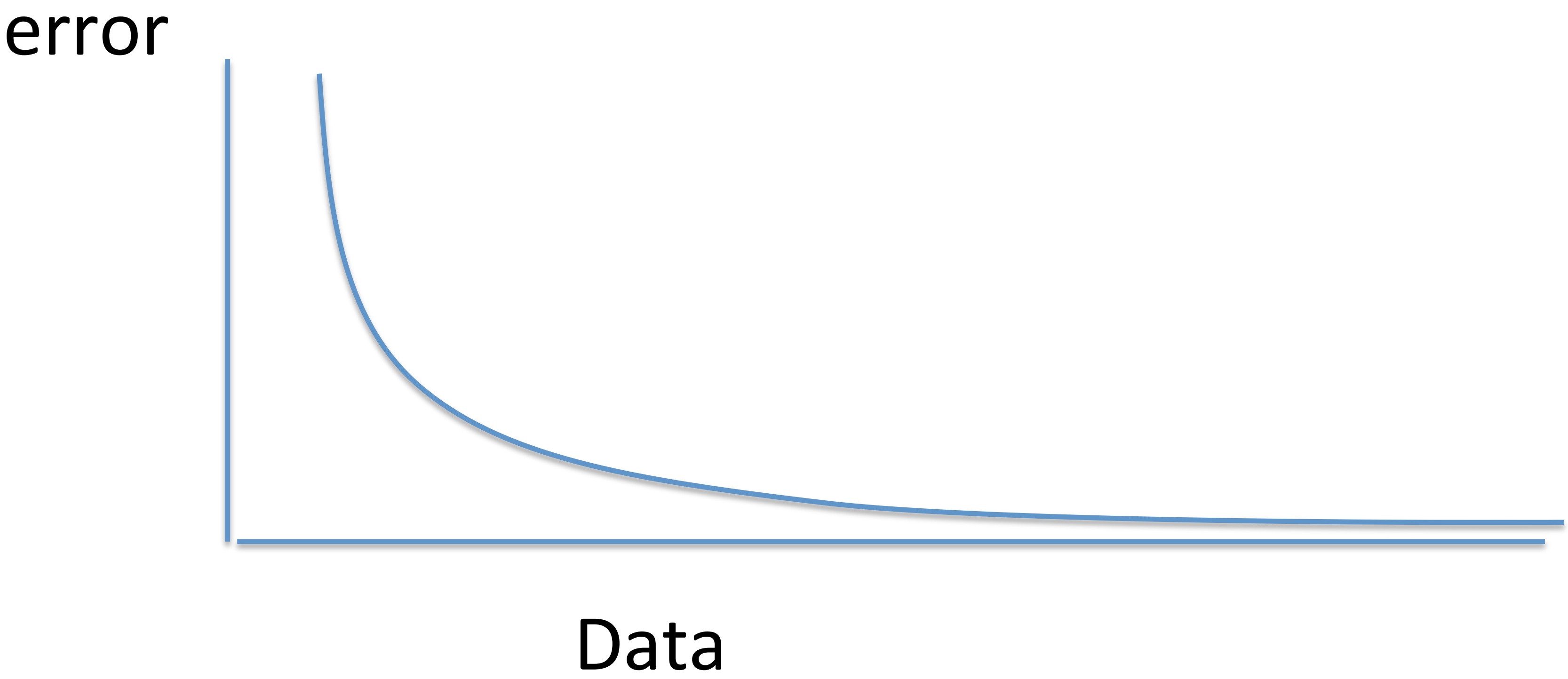
Estimate species tree for every 4 species



Combine unrooted 4-taxon trees

Theorem (Allman et al., 2011, and others): For every four leaves {a,b,c,d}, the most probable unrooted quartet tree on {a,b,c,d} is the true species tree. Hence, the unrooted species tree can be estimated from a large enough number of true unrooted gene trees.

Statistical Consistency

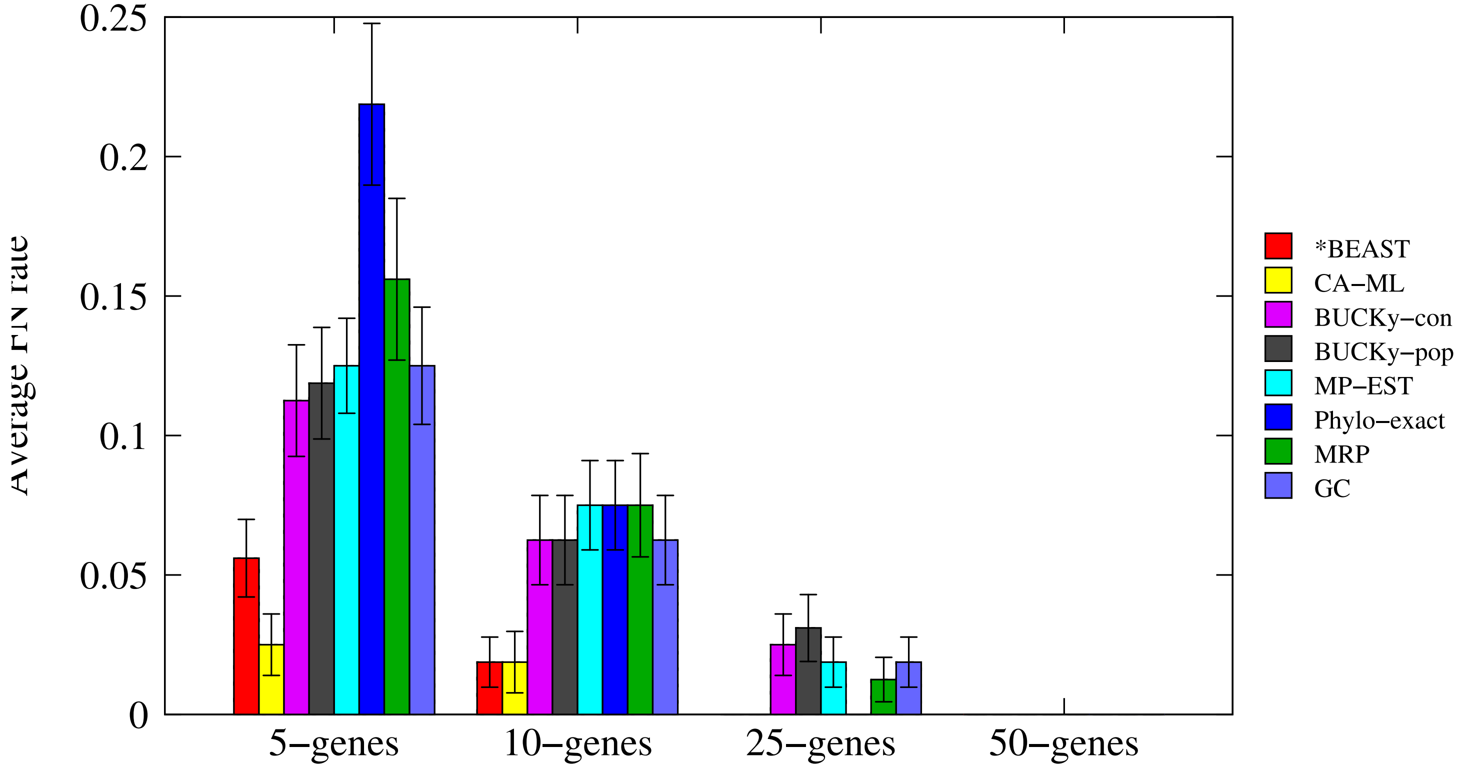


Data are gene trees, presumed to be randomly sampled true gene trees.

Statistically consistent under ILS?

- **MP-EST** (Liu et al. 2010): maximum likelihood estimation of rooted species tree – YES
- BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation –YES
- MDC – NO
- Greedy – NO
- Concatenation under maximum likelihood - NO
- MRP (supertree method) – open

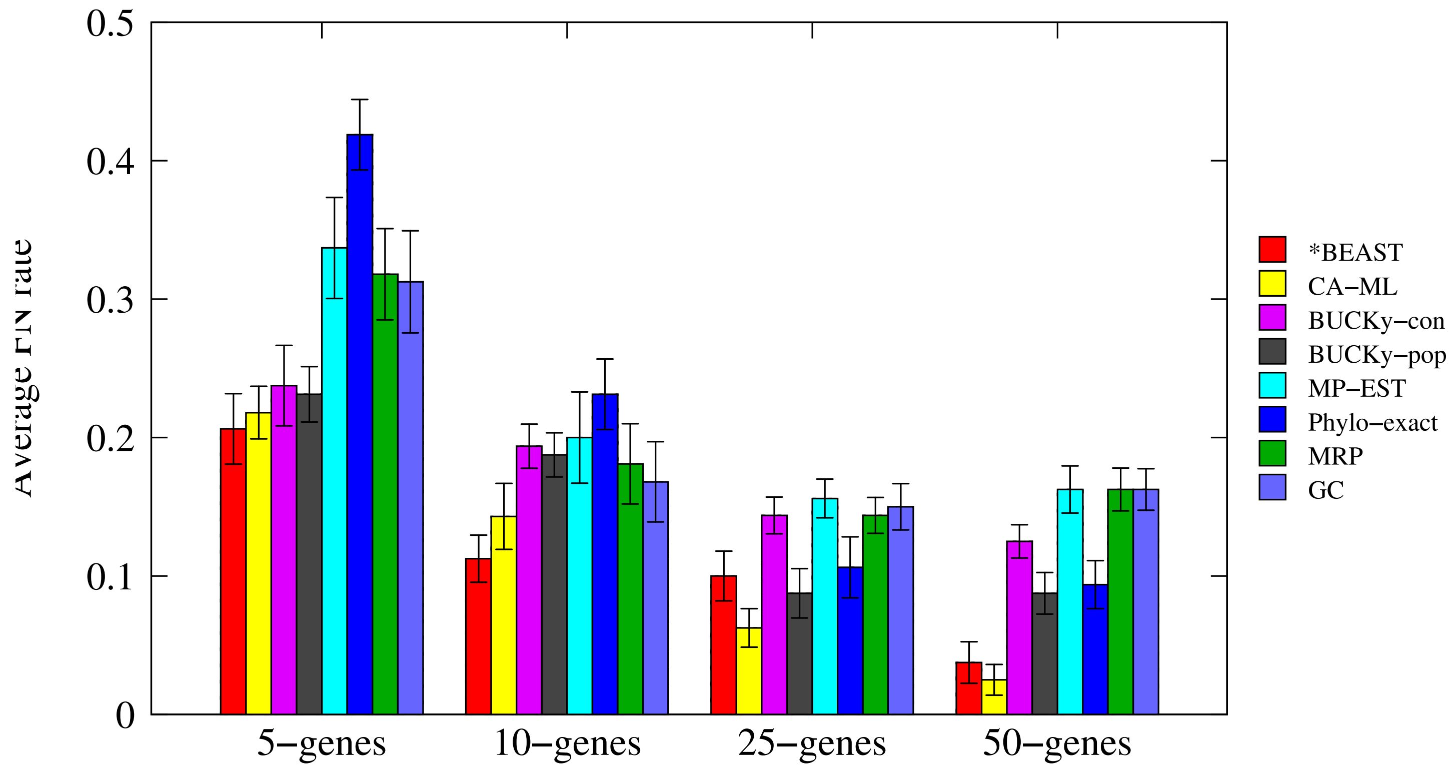
Results on 11-taxon datasets with weak ILS



*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

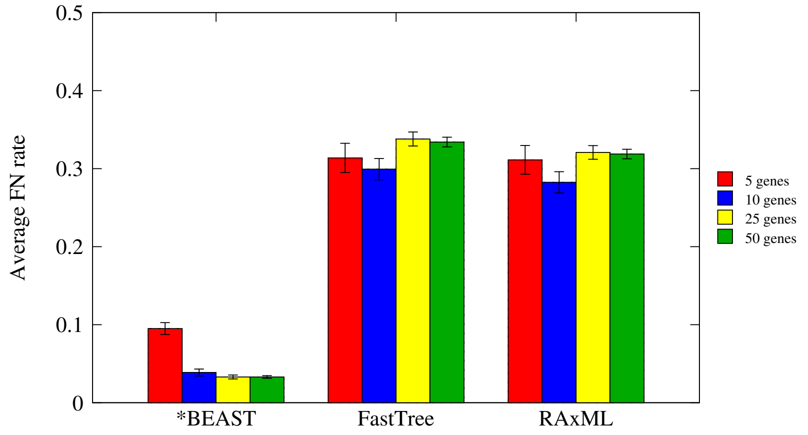
Results on 11-taxon datasets with strongILS



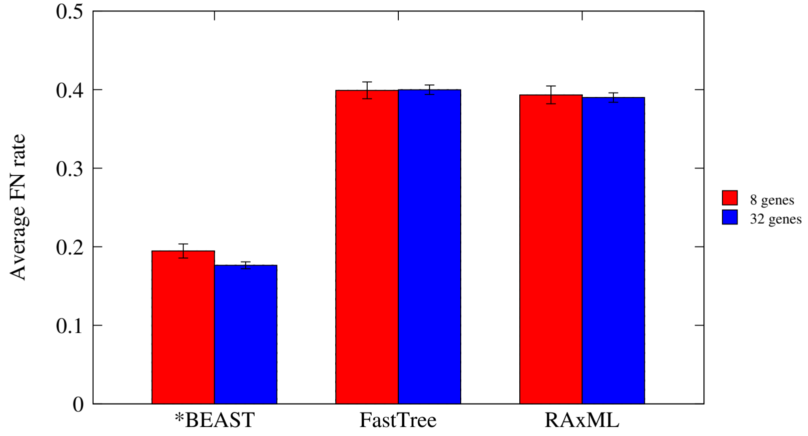
*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: (concatenated analysis) also very accurate

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

*BEAST co-estimation produces more accurate gene trees than Maximum Likelihood



11-taxon weakILS datasets

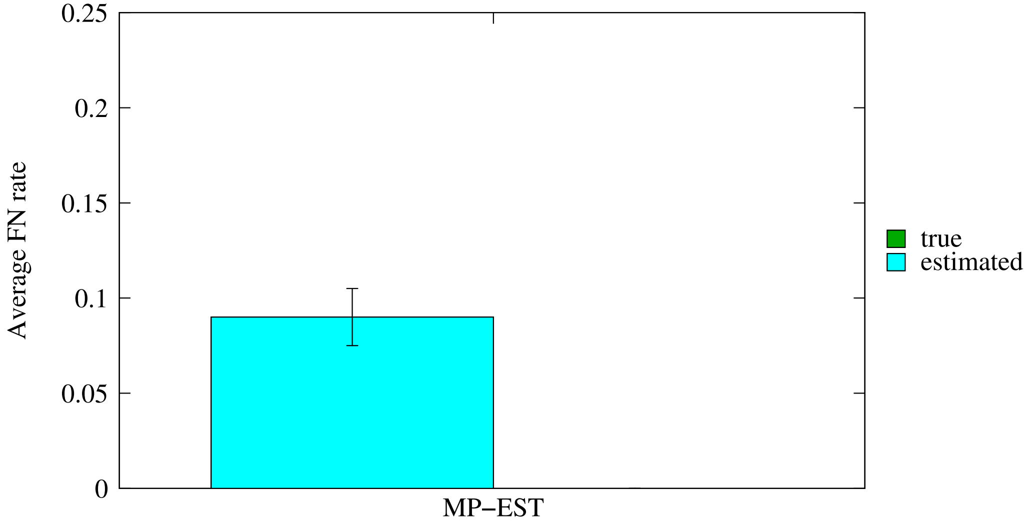


17-taxon (very high ILS) datasets

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

Bayzid & Warnow, Bioinformatics 2013

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.

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.

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
- Develop methods with greater robustness to gene tree error

Addressing gene tree estimation error

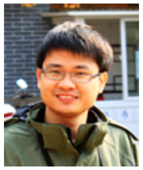
- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- Develop methods with greater robustness to gene tree error
  - ASTRAL. Bioinformatics 2014 (Mirarab et al.)
  - Statistical binning. In press (Mirarab et al.)

Avian Phylogenomics Project

E Jarvis,
HHMI

MTP Gilbert,
Copenhagen

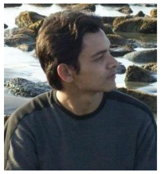
G Zhang,
BGI



T. Warnow
UT-Austin

S. Mirarab
UT-Austin

Md. S. Bayzid,
UT-Austin



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)

Species tree estimated using Statistical Binning with MP-EST
In press

1KP: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

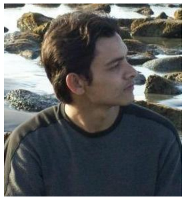
N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin



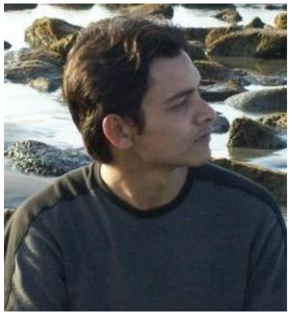
Md. S.Bayzid
UT-Austin

Plus many other people…

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Species tree estimated using ASTRAL (Bioinformatics, 2014) In press

Statistical binning



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

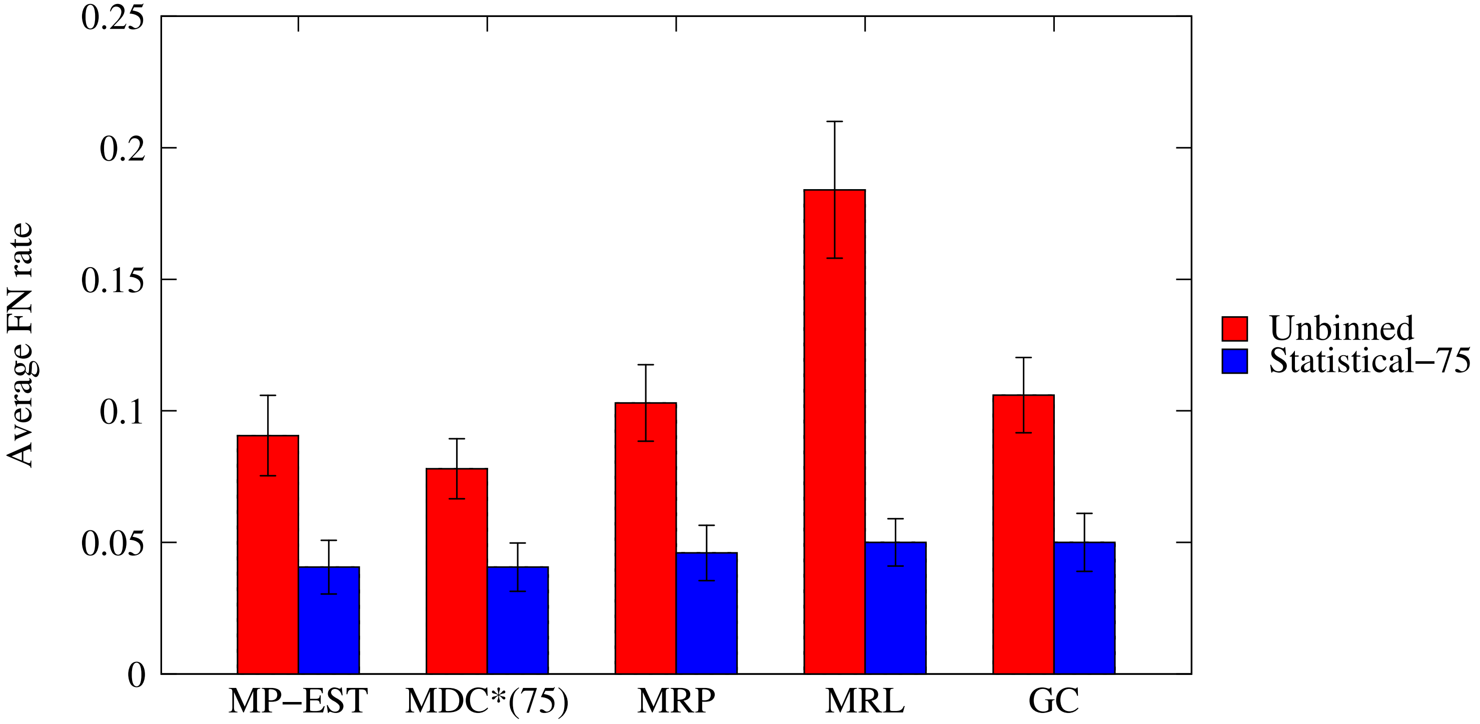
Step 1: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible, and the sets have approximately the same size.

Step 2: estimate "supergene" trees on each set using concatenation (maximum likelihood)

Step 3: combine supergene trees using coalescent-based method

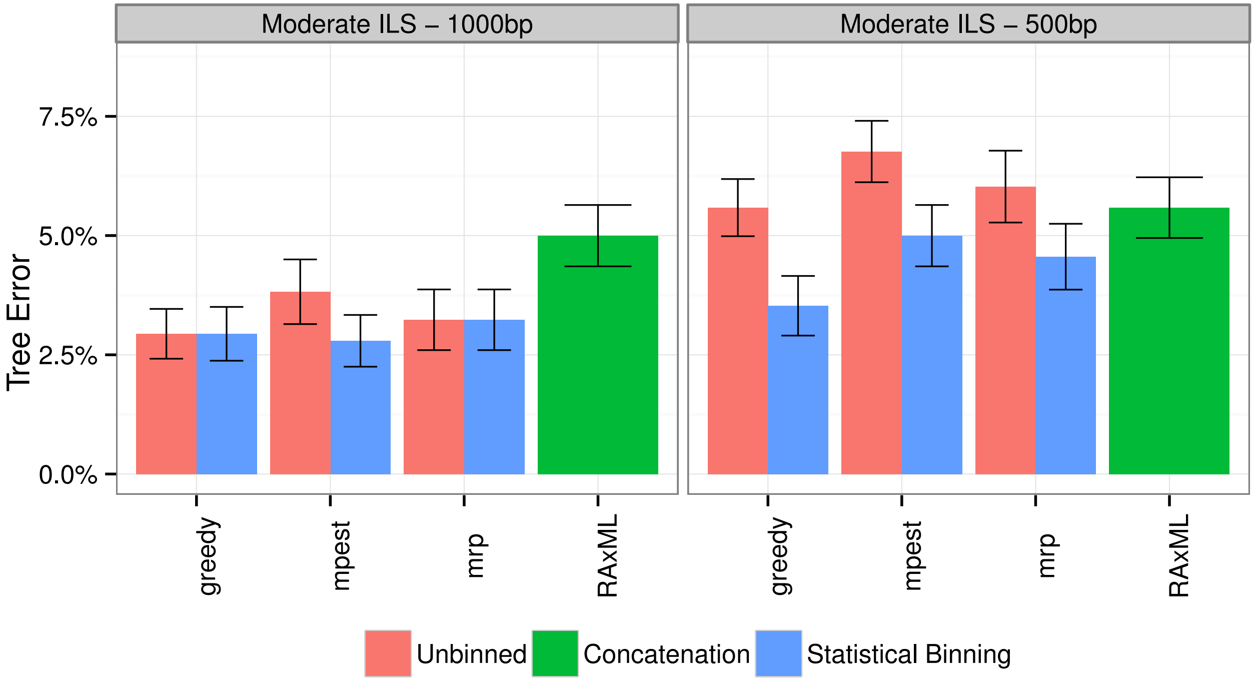
Note: Step 1 requires solving the NP-hard "balanced vertex coloring problem", for which we developed a good heuristic (modified 1979 Brelaz algorithm)

Statistical binning vs. unbinned



Mirarab, et al., to appear (Science 2014)
Binning produces bins with approximate 5 to 7 genes each
Datasets: 11-taxon strongILS datasets with 50 genes, Chung and Ané, Systematic Biology

Mammalian Simulation Study



Observations:

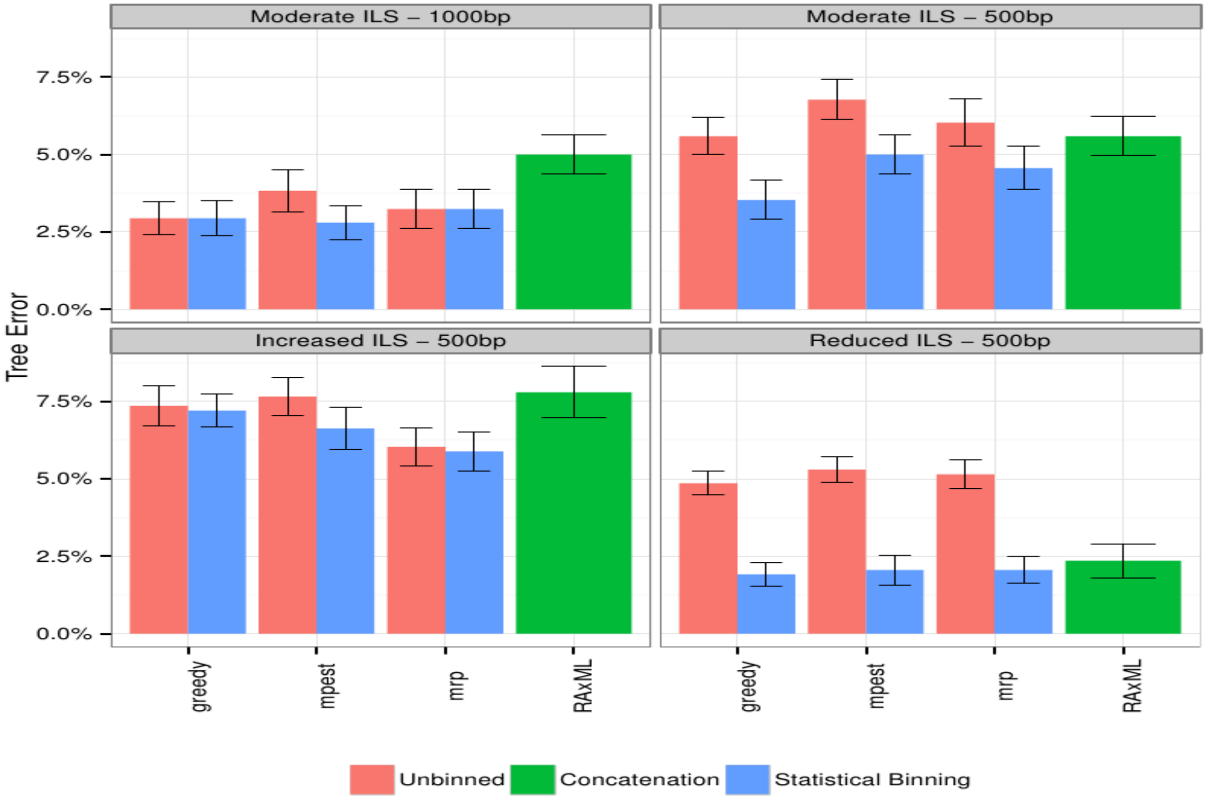
Binning can improve accuracy, but impact depends on accuracy of estimated gene trees and phylogenetic estimation method.

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., to appear

Mammalian simulation



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

MP-EST is statistically consistent; Greedy and Maximum Likelihood are not; unknown for MRP.
Data (200 genes, 20 replicate datasets) based on Song et al. PNAS 2012

ASTRAL

- Accurate Species Trees Algorithm
- Mirarab et al., ECCB 2014 and Bioinformatics 2014
- Statistically-consistent estimation of the species tree from unrooted gene trees

ASTRAL's approach

- Input: set of unrooted gene trees T_1, T_2, …, T_k
- Output: Tree T^* maximizing the total quartet-similarity score to the unrooted gene trees

Theorem:

- An exact solution to this problem would be a statistically consistent algorithm in the presence of ILS

ASTRAL's approach

- Input: set of unrooted gene trees T_1, T_2, …, T_k
- Output: Tree T^* maximizing the total quartet-similarity score to the unrooted gene trees

Theorem:

- An exact solution to this problem is NP-hard

Comment: unknown computational complexity if all trees T_i are on the same leaf set

ASTRAL's approach

- Input: set of unrooted gene trees T_1, T_2, …, T_k and set X of bipartitions on species set S
- Output: Tree T^* maximizing the total quartet-similarity score to the unrooted gene trees, <u>subject to Bipartitions(T^*) drawn from X</u>

Theorem:

- An exact solution to this problem is achievable in polynomial time!

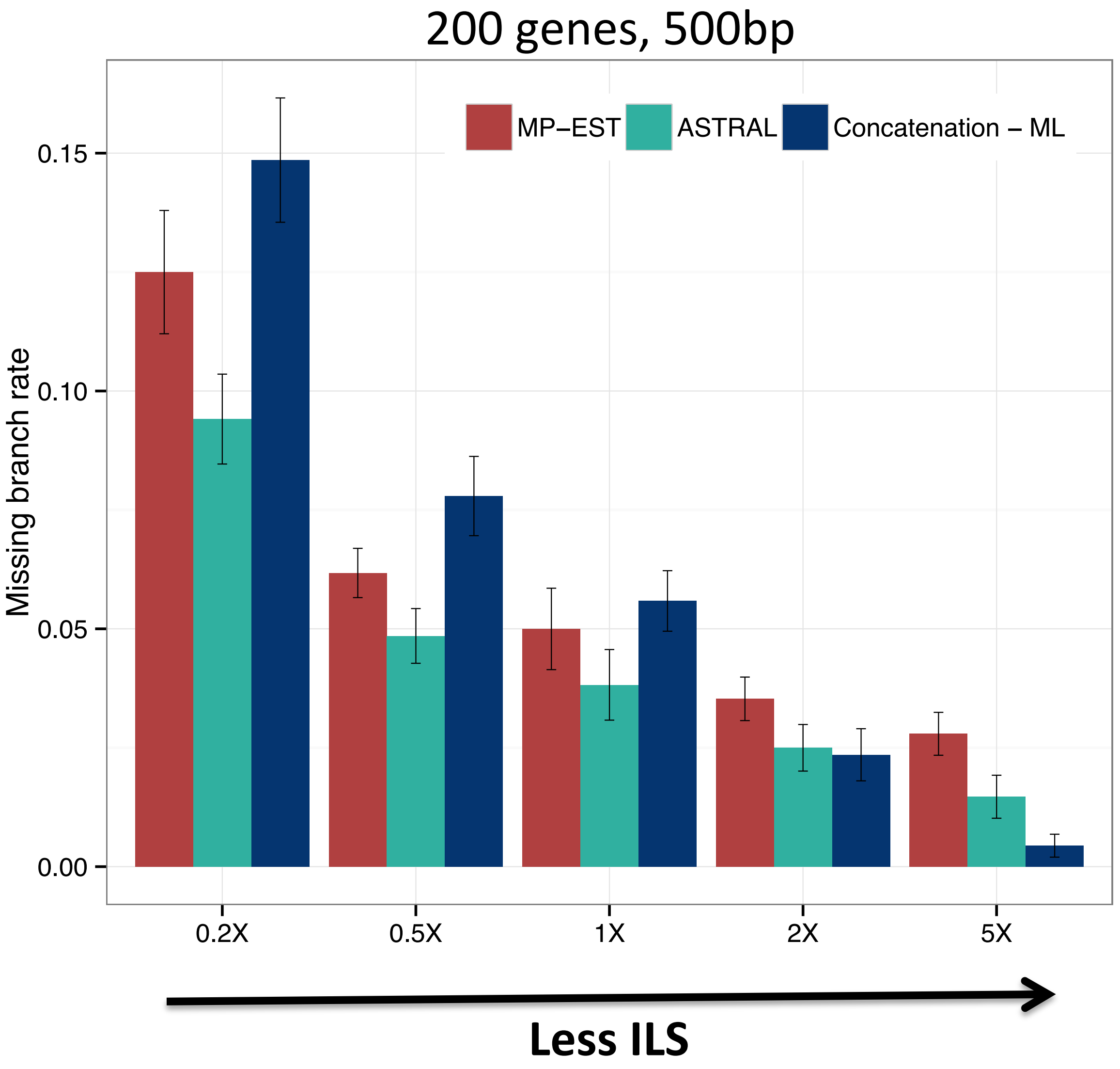
ASTRAL's approach

- Input: set of unrooted gene trees T_1, T_2, …, T_k and set X of bipartitions on species set S
- Output: Tree T^* maximizing the total quartet-similarity score to the unrooted gene trees, <u>subject to Bipartitions(T^*) drawn from X</u>

Theorem:

- Letting X be the set of bipartitions from the input gene trees is statistically consistent and polynomial time.

ASTRAL vs. Concatenation



Basic Question

- Is it possible to estimate the species tree with high probability given a large enough set of estimated gene trees, each with some non-zero probability of error?

Partial answers

Theorem (Roch & Warnow, in preparation): If gene sequence evolution obeys the strong molecular clock, then statistically consistent estimation is possible – even where all gene trees are estimated based on a single site.

Partial answers

Theorem (Roch & Warnow, in preparation): If gene sequence evolution obeys the strong molecular clock, then statistically consistent estimation is possible – even where all gene trees are estimated based on a single site.

Proof (sketch): Under the multi-species coalescent model, the most probable rooted triplet gene tree on {a,b,c} is the true species tree for {a,b,c}, and this remains true (when the molecular clock holds) *even for triplet gene trees estimated on a single site*.

When molecular clock fails

- Without the molecular clock, the estimation of the species tree is based on quartet trees.

When molecular clock fails

- Without the molecular clock, the estimation of the species tree is based on quartet trees.
- Although the most probable quartet tree is still the true species tree, this is *no longer true for estimated quartet trees – except for very long sequences.*

When molecular clock fails

- Without the molecular clock, the estimation of the species tree is based on quartet trees.
- Although the most probable quartet tree is still the true species tree, this is *no longer true for estimated quartet trees – except for very long sequences.*
- *No positive results established for any of the current coalescent-based methods in use!*

Summary

Gene tree estimation under Markov models of evolution:

- Absolute fast converging (afc) methods: true trees from polynomial length sequences.

Coalescent-based species tree estimation:

- Gene tree estimation error impacts species tree estimation.
- Statistical binning (in press) improves coalescent-based species tree estimation from multiple genes, used in Avian Tree (in press).
- ASTRAL (Bioinformatics, 2014) more robust to gene tree estimation error, used in Plant Tree (in press).
- Identifiability in the presence of gene tree estimation error? Yes under the strong molecular clock, very limited results otherwise.
- New questions about statistical inference, focusing on the impact of input error.

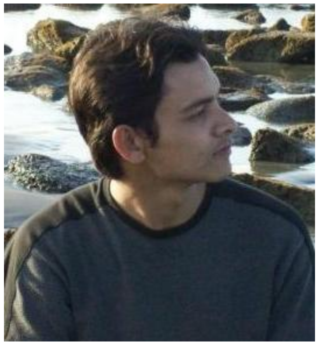
Computational Phylogenetics

Interesting combination of different mathematics and computer science:

- 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

Acknowledgments



PhD students: Siavash Mirarab* and Md. S. Bayzid**
Sebastien Roch (Wisconsin) – work began at IPAM (UCLA)

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

TACC and UTCS computational resources

* Supported by HHMI Predoctoral Fellowship
** Supported by Fulbright Foundation Predoctoral Fellowship

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

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:

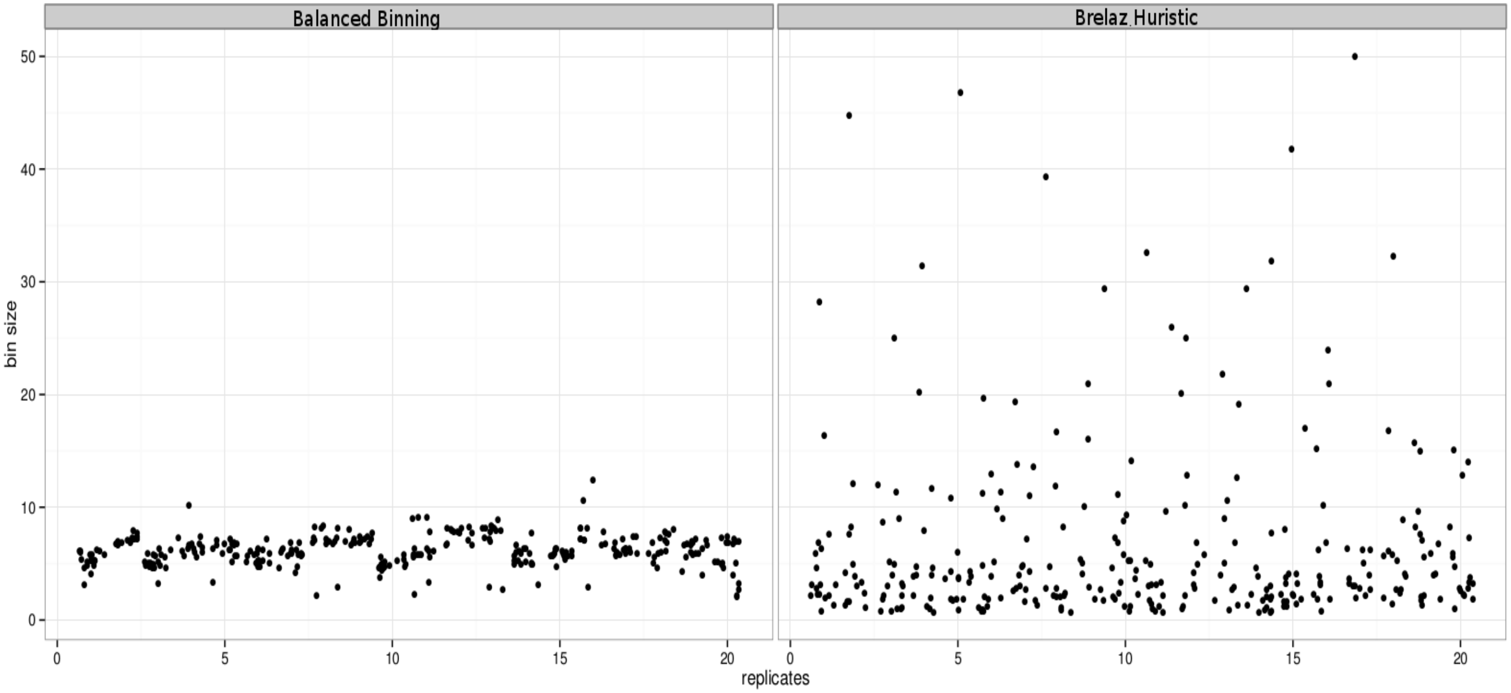
- Naïve binning (Bayzid and Warnow, Bioinformatics 2013)
- Statistical binning (Mirarab, Bayzid, and Warnow, to appear, Science)

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.

Balanced Statistical Binning



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

Avian Phylogenomics Project

E Jarvis,
HHMI

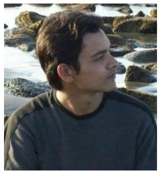
MTP Gilbert,
Copenhagen

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, under review

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, under review

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, under review

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.

Other recent related work

- ASTRAL: statistically consistent method for species tree estimation under the multi-species coalescent (Mirarab et al., Bioinformatics 2014)
- DCM-boosting coalescent-based methods (Bayzid et al., RECOMB-CG and BMC Genomics 2014)
- Weighted Statistical Binning (Bayzid et al., in preparation) – statistically consistent version of statistical binning

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

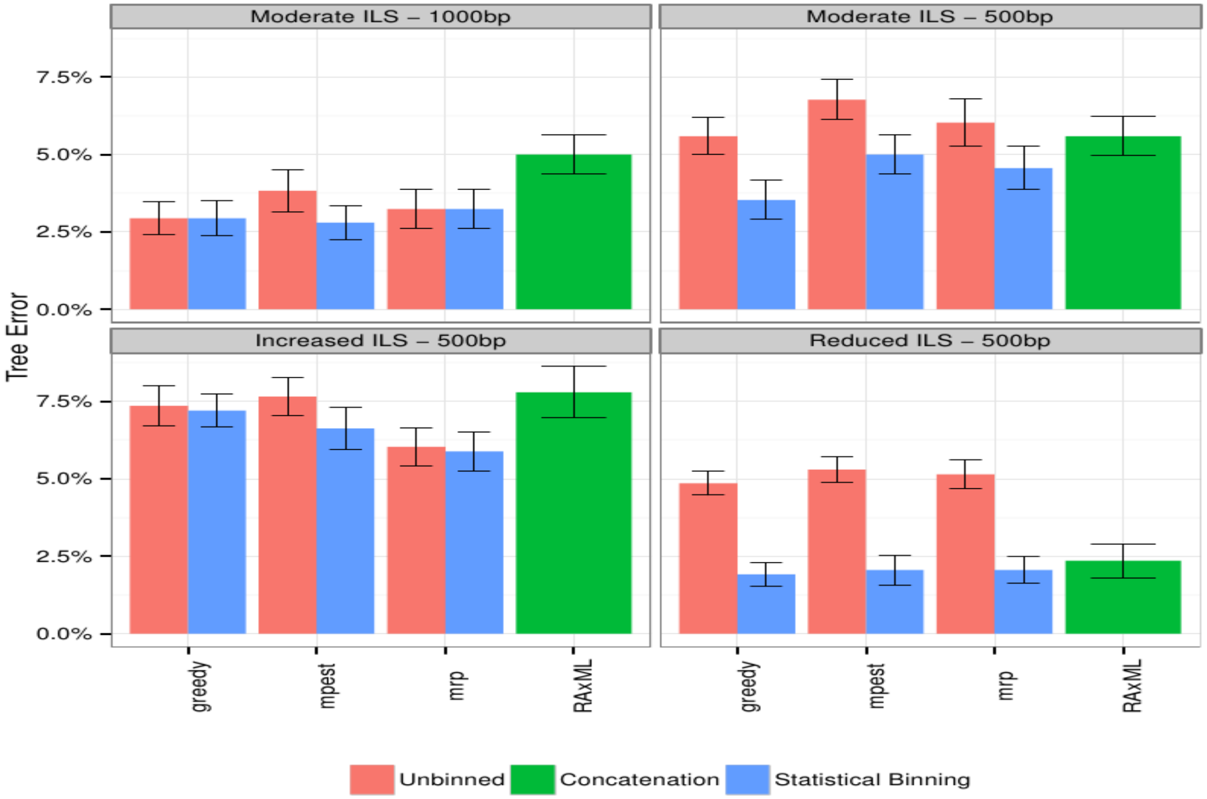
Major scientific goals:

- 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

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

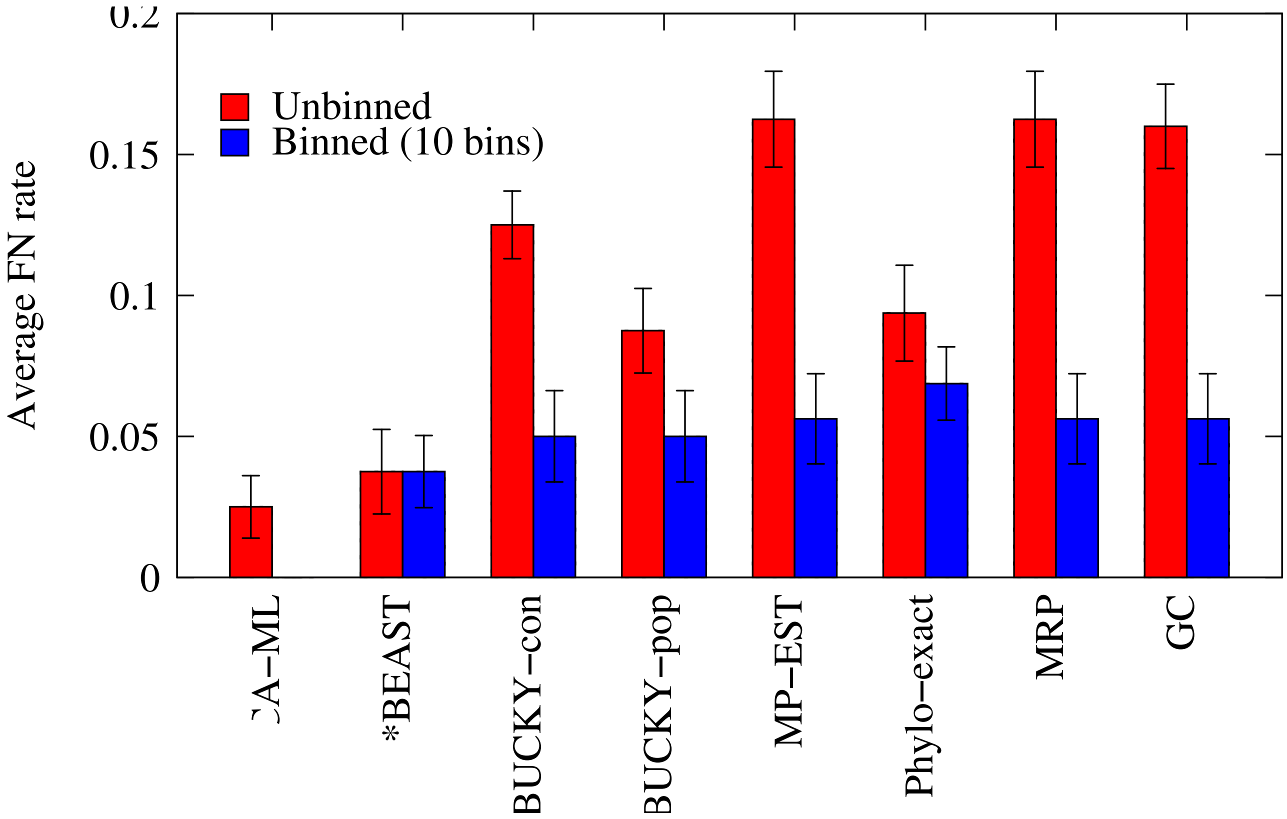
Statistically consistent methods

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

Output: estimated species tree

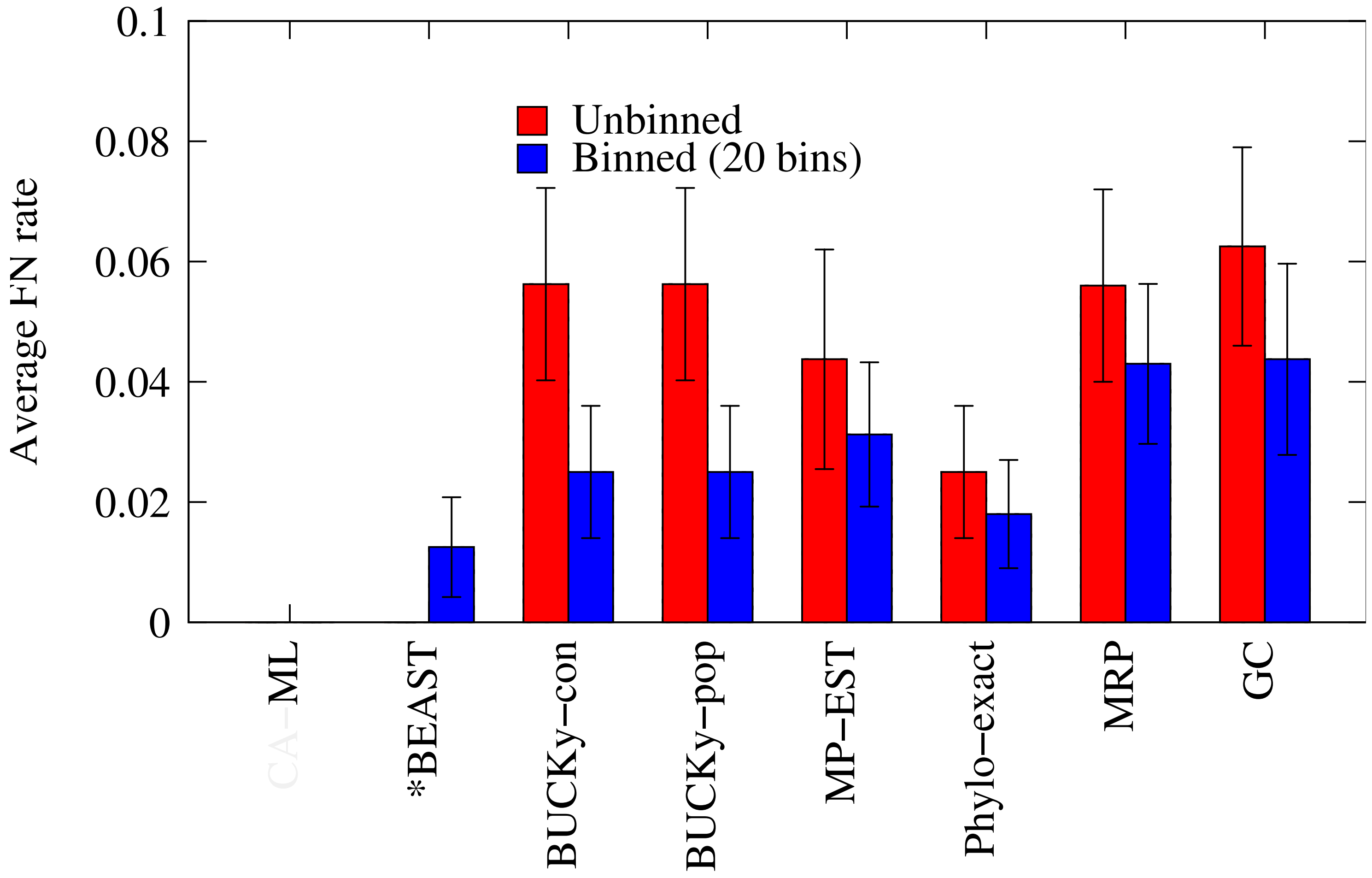
- ***BEAST** (Heled and Drummond 2010): Bayesian co-estimation of gene trees and species trees given sequence alignments
- **MP-EST** (Liu et al. 2010): maximum likelihood estimation of rooted species tree
- **BUCKy-pop** (Ané and Larget 2010): quartet-based Bayesian species tree estimation

Naïve binning vs. unbinned: 50 genes



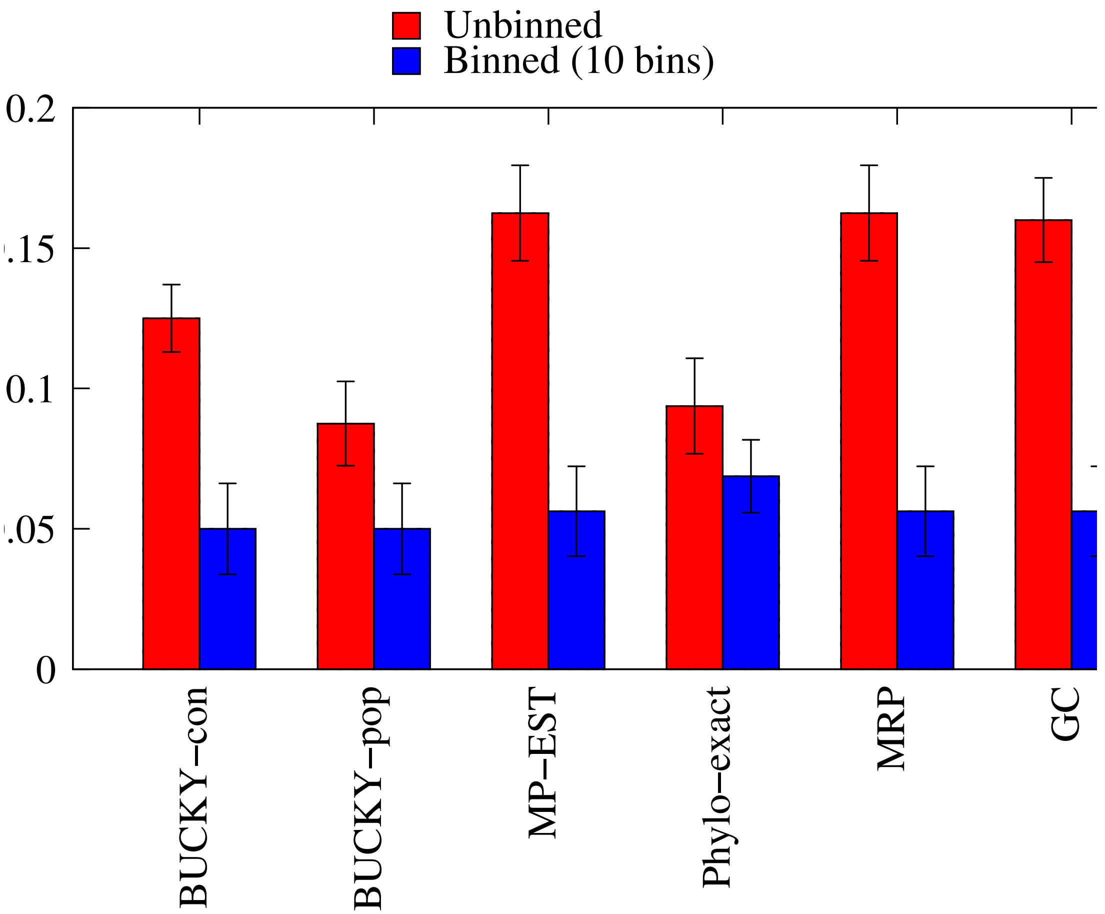
Bayzid and Warnow, Bioinformatics 2013
11-taxon strongILS datasets with 50 genes, 5 genes per bin

Naïve binning vs. unbinned, 100 genes

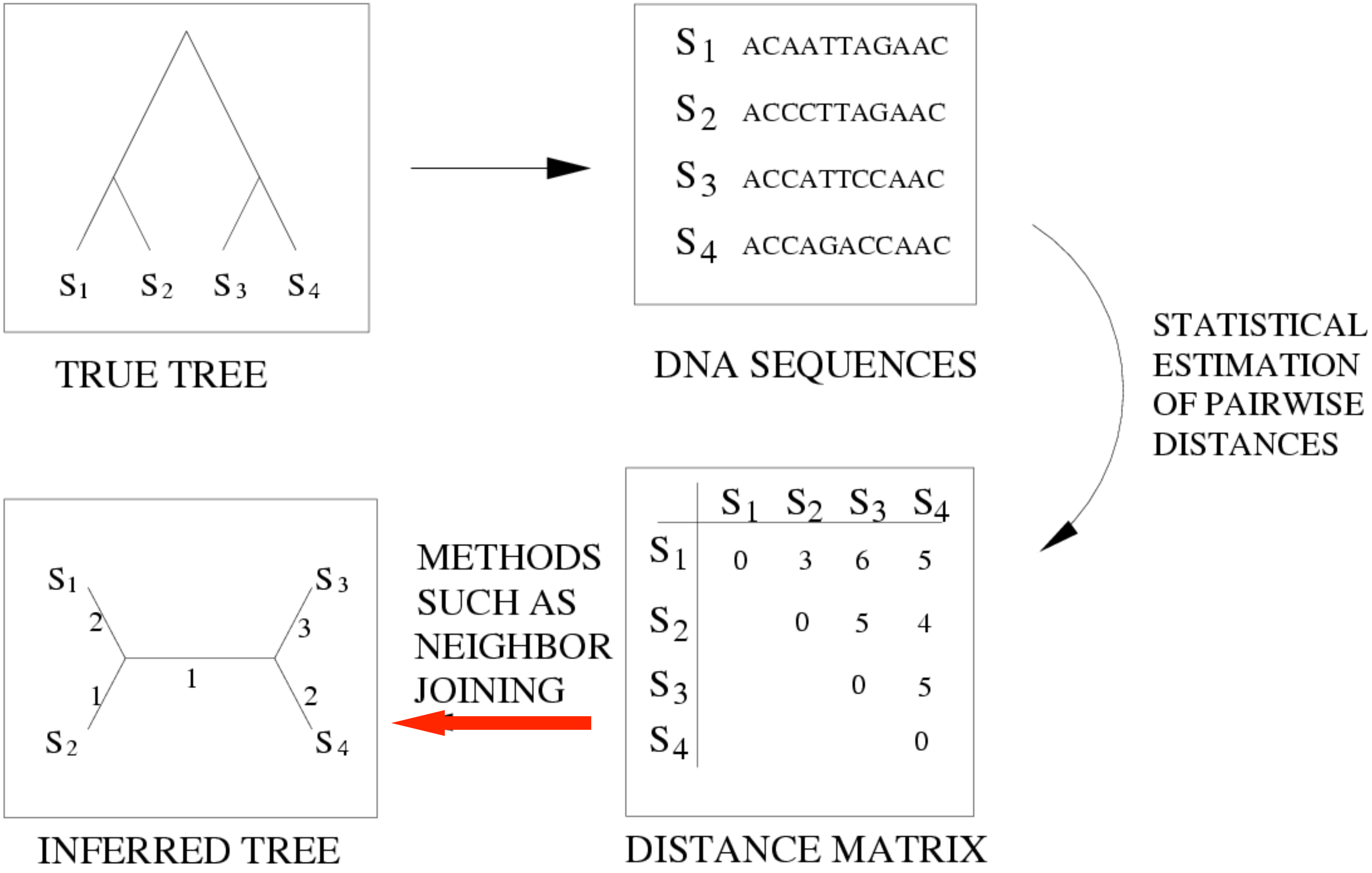


*BEAST did not converge on these datasets, even with 150 hours.
With binning, it converged in 10 hours.

Naïve binning vs. unbinned: 50 genes



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



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