394C, October 2, 2013

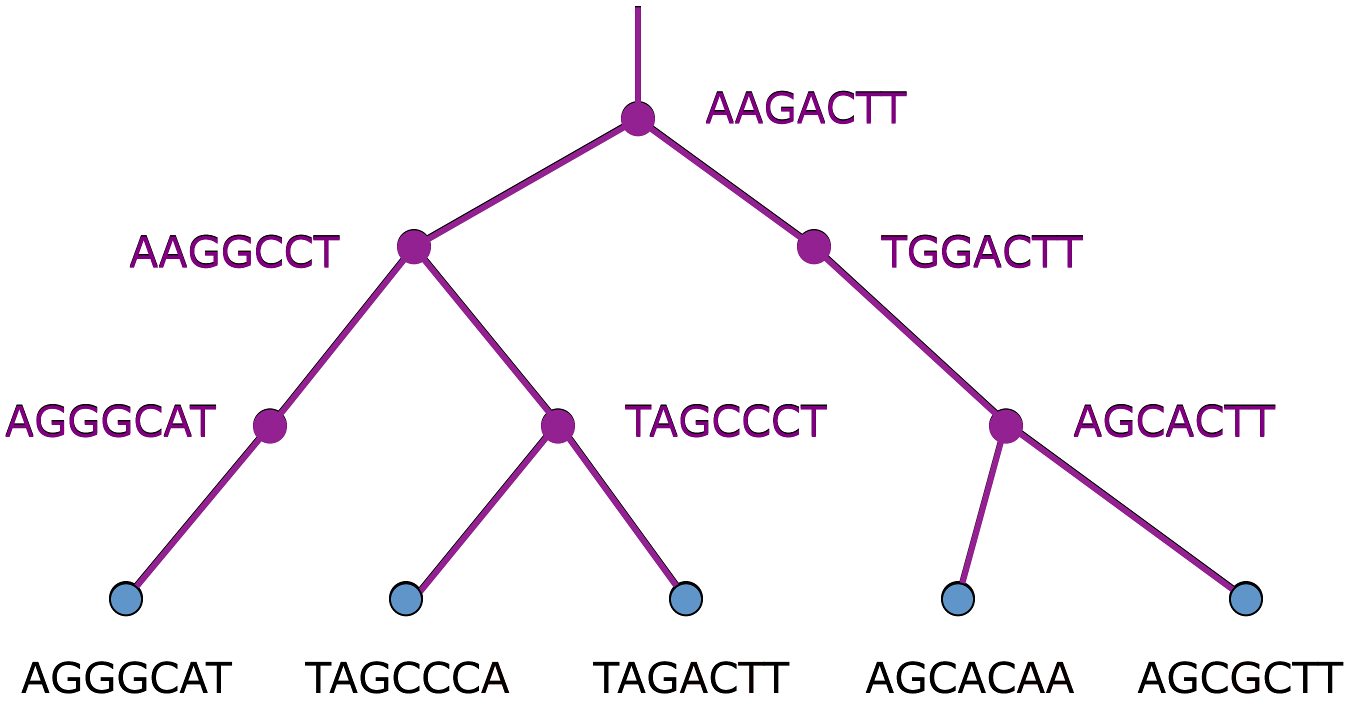
Topics:

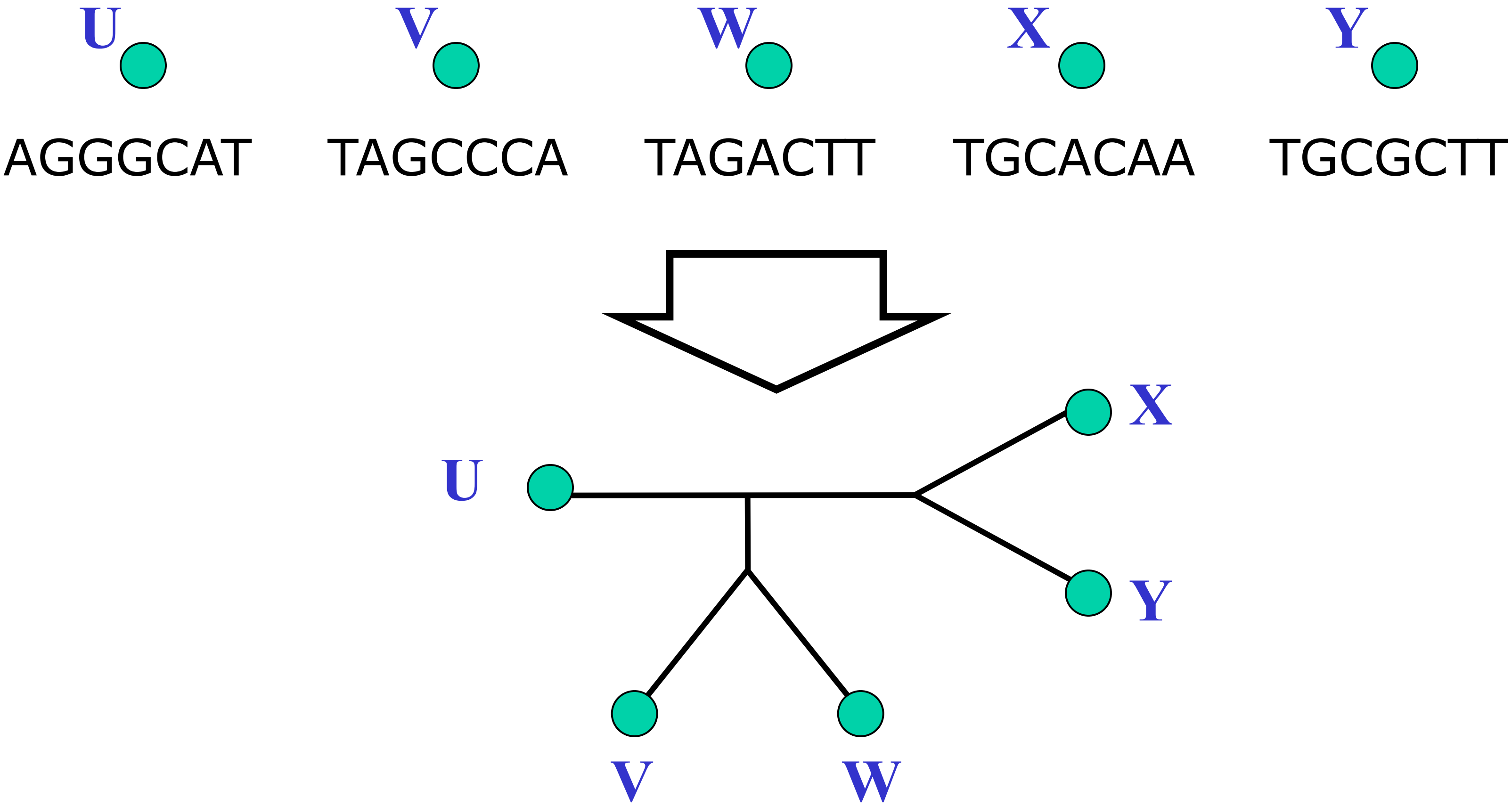
- Multiple Sequence Alignment
- Estimating Species Trees from Gene Trees

Multiple Sequence Alignment

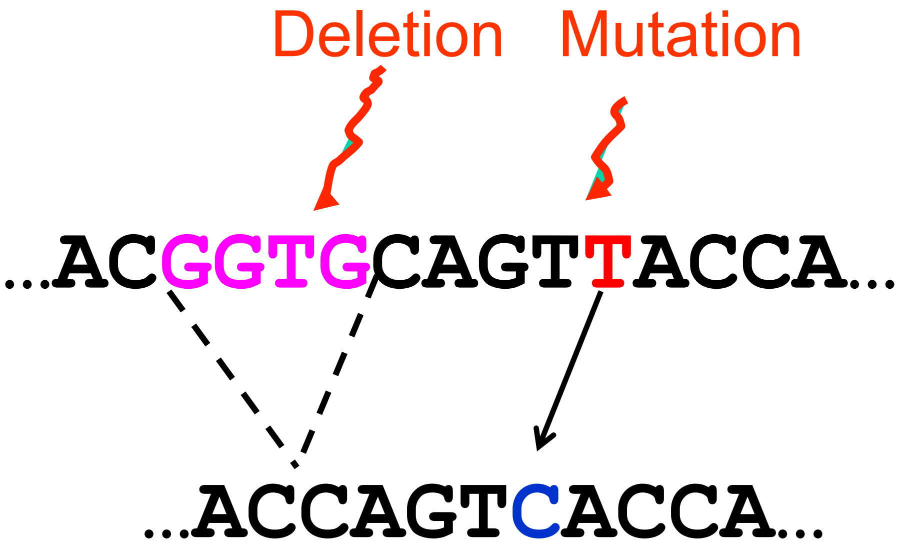
- Multiple Sequence Alignments and Evolutionary Histories (the meaning of “homologous”)
- How to define error rates in multiple sequence alignments
- Minimum edit transformations and pairwise alignments
- Dynamic Programming for calculating a pairwise alignment (or minimum edit transformation)
- Co-estimating alignments and trees

DNA Sequence Evolution





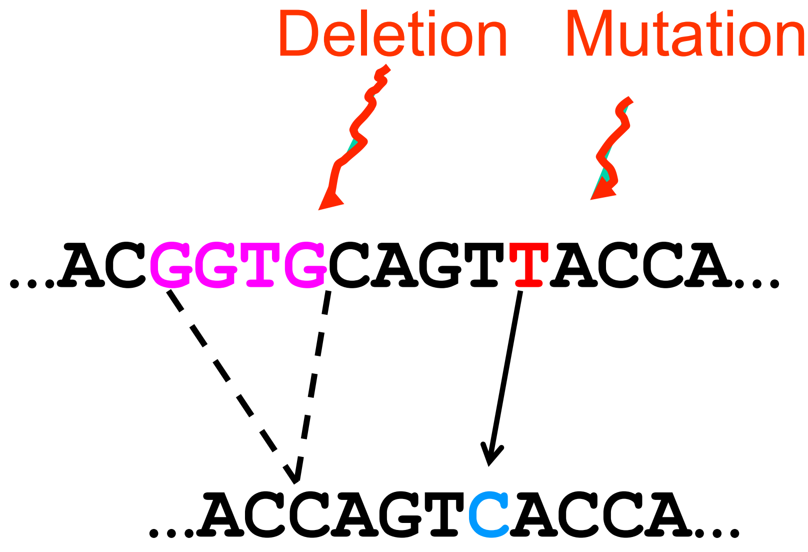
U	V	W	X	Y
AGGGCAT	TAGCCCA	TAGACTT	TGCACAA	TGCGCTT



Deletion Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...



Deletion Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...

```
...ACGGTGCAGTTACCA...
...AC----CAGTCACCA...
```

The true multiple alignment

- Reflects historical substitution, insertion, and deletion events in the true phylogeny

Input: unaligned sequences

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

Phase 1: Multiple Sequence Alignment

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

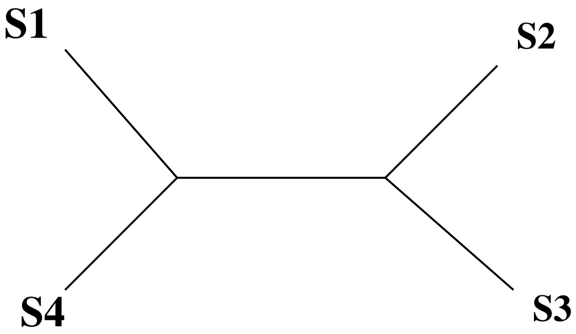
→

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



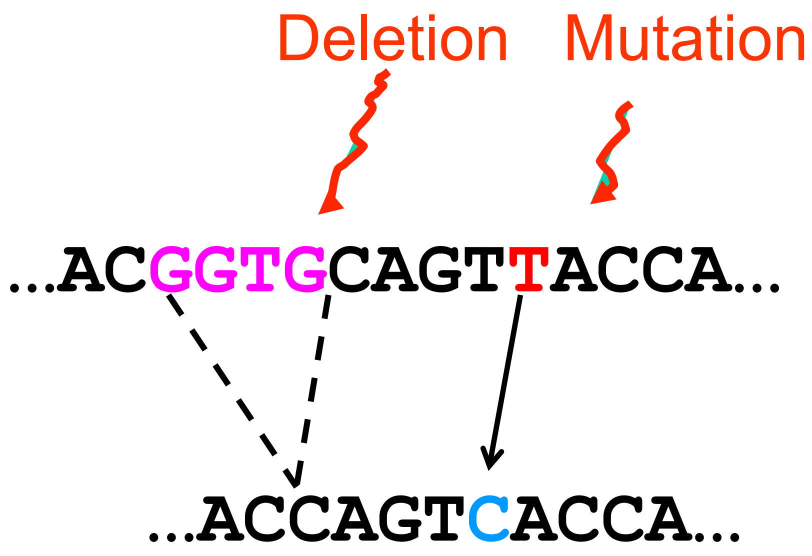
Many methods

Alignment methods

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Opal
- Etc.

Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.



Deletion Mutation

...ACGGTGCAGTTACCA...

...ACCAGTCACCA...

```
...ACGGTGCAGTTACCA...
...AC-----CAGTCACCA...
```

The true multiple alignment

- Reflects historical substitution, insertion, and deletion events in the true phylogeny

But how do we try to estimate this?

Pairwise alignments and edit transformations

- Each pairwise alignment implies one or more edit transformations
- Each edit transformation implies one or more pairwise alignments
- So calculating the edit distance (and hence minimum cost edit transformation) *is the same* as calculating the optimal pairwise alignment

Edit distances

- Substitution costs may depend upon which nucleotides are involved (e.g, transition/ transversion differences)

- Gap costs
  - Linear (aka "simple"): gapcost(L) = cL
  - Affine: gapcost(L) = c+c'L
  - Other: gapcost(L) = c+c' log(L)

Computing optimal pairwise alignments

- The cost of a pairwise alignment (*under a simple gap model*) is just the sum of the costs of the columns

- Under affine gap models, it's a bit more complicated (but not much)

Computing edit distance

- Given two sequences and the edit distance function F(.,.), how do we compute the edit distance between two sequences?

- Simple algorithm for standard gap cost functions (e.g., affine) based upon dynamic programming

DP alg for simple gap costs

- Given two sequences A[1…n] and B[1…m], and an edit distance function F(.,.) with unit substitution costs and gap cost C,
- Let
  - $A = A_1,A_2,\ldots,A_n$
  - $B = B_1,B_2,\ldots,B_m$
- Let M(i,j)=F(A[1…i],B[1…j]) (i.e., the edit distance between these two prefixes)

Dynamic programming algorithm

Let M(i,j)=F(A[1…i],B[1…j])

- M(0,0)=0
- M(n,m) stores our answer
- How do we compute M(i,j) from other entries of the matrix?

Calculating M(i,j)

- Examine final column in some optimal pairwise alignment of A[1…i] to B[1…j]
- Possibilities:
  - Nucleotide over nucleotide: previous columns align A[1…i-1] to B[1…j-1]:
  - Indel (-) over nucleotide: previous columns align A[1…i] to B[1…j-1]:
  - Nucleotide over indel: previous columns align A[1…i-1] to B[1…j]:

Calculating M(i,j)

- Examine final column in some optimal pairwise alignment of A[1…i] to B[1…j]
- Possibilities:
  - Nucleotide over nucleotide: previous columns align A[1…i-1] to B[1…j-1]: $M(i,j)=M(i-1,j-1)+subcost(A_i,B_j)$
  - Indel (-) over nucleotide: previous columns align A[1…i] to B[1…j-1]: $M(i,j)=M(i,j-1)+indelcost$
  - Nucleotide over indel: previous columns align A[1…i-1] to B[1…j]: $M(i,j)=M(i-1,j)+indelcost$

Calculating M(i,j)

- M(i,j) = min {
  M(i-1,j-1)+subcost(A_i,B_j),
  M(i,j-1)+indelcost, M(i-1,j)+indelcost }

O(nm) DP algorithm for pairwise alignment using simple gap costs

- Initialize M(0,j) = M(j,0) = j*indelcost

- For i=1…n
  - For j = 1…m
    - M(i,j) = min {

      M(i-1,j-1)+subcost(A_i,B_j),
      M(i,j-1)+indelcost,
      M(i-1,j)+indelcost

      }

- Return M(n,m)
- Add arrows for backtracking (to construct an optimal alignment and edit transformation rather than just the cost)

Modification for other gap cost functions is straightforward but leads to an increase in running time

Sum-of-pairs optimal multiple alignment

- Given set S of sequences and edit cost function F(.,.),
- Find multiple alignment that minimizes the sum of the implied pairwise alignments (Sum-of-Pairs criterion)

- NP-hard, but can be approximated
- Is this useful?

Other approaches to MSA

- Many of the methods used in practice do not try to optimize the sum-of-pairs
- Instead they use probabilistic models (HMMs)
- Often they do a progressive alignment on an estimated tree (aligning alignments)
- Performance of these methods can be assessed using real and simulated data

Many methods

Alignment methods

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Opal
- Etc.

Phylogeny methods

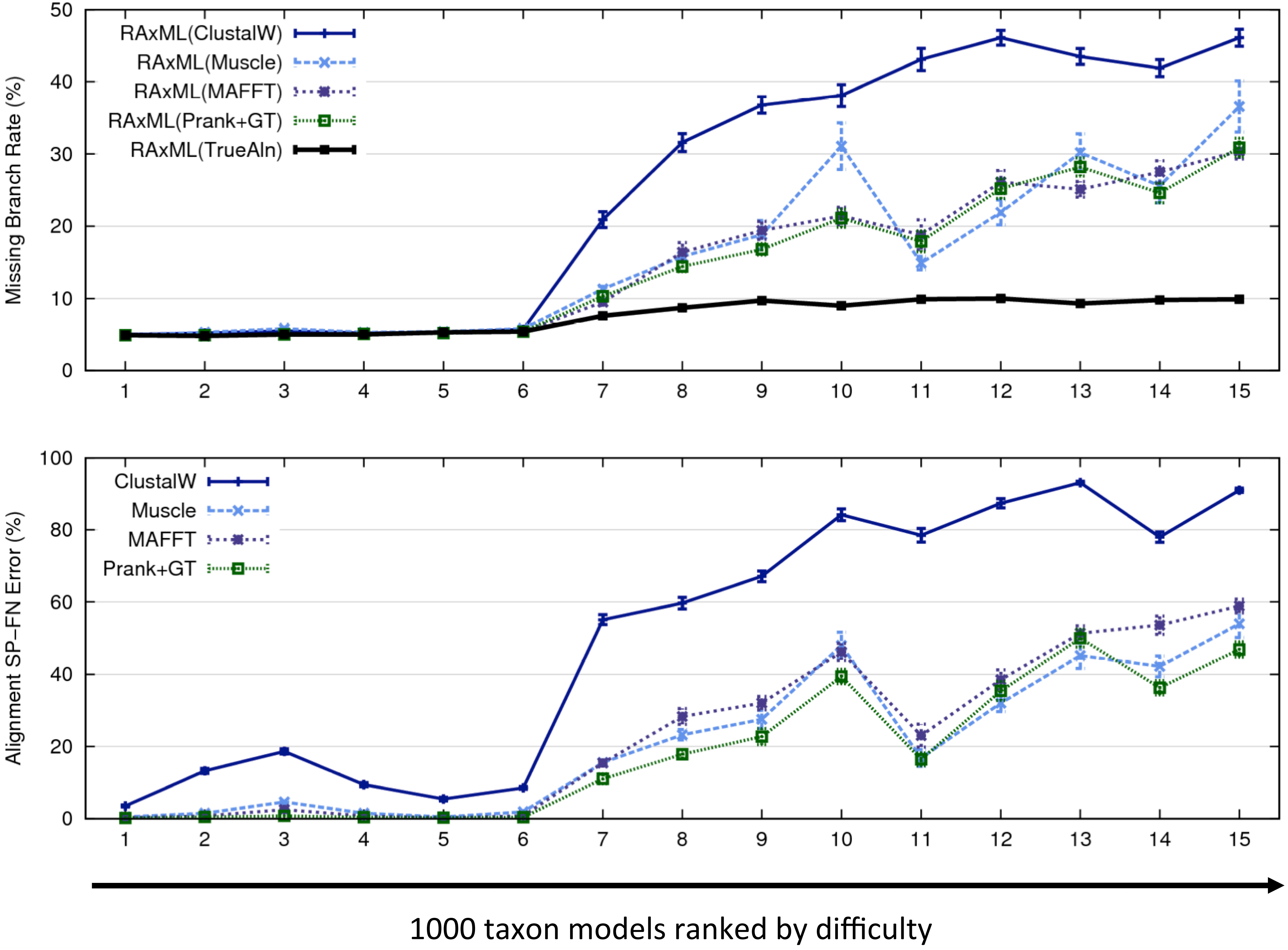
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

Simulation study

- ROSE simulation:
  - 1000, 500, and 100 sequences
  - Evolution with substitutions and indels
  - Varied gap lengths, rates of evolution
- Computed alignments
- Used RAxML to compute trees
- Recorded tree error (missing branch rate)
- Recorded alignment error (SP-FN)

Alignment Error

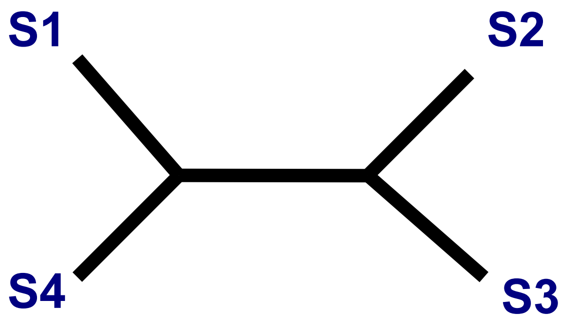
- Given a multiple sequence alignment, we represent it as a set of pairwise homologies.
- To compare two alignments, we compare their sets of pairwise homologies.
- The SP-FN (sum-of-pairs false negative rate) is the percentage of the true homologies (those present in the true alignment) that are missing in the estimated alignment.
- The SP-FP (sum-of-pairs false positive rate) is the percentage of the homologies in the estimated alignment that are not in the true alignment.


Missing Branch Rate (%)
50
40
30
20
10
0
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Alignment SP−FN Error (%)
100
80
60
40
20
0
ClustalW
Muscle
MAFFT
Prank+GT
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1000 taxon models ranked by difficulty

Problems with the two phase approach

- Manual alignment can have a high level of subjectivity (and can take a long time).
- Current alignment methods fail to return reasonable alignments on markers that evolve with high rates of indels and substitutions, especially if these are large datasets.
- We discard potentially useful markers if they are difficult to align.

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



and

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Simultaneous estimation of trees and alignments

Simultaneous Estimation Methods

- Likelihood-based (under model of evolution including insertion/deletion events)
  - ALIFRITZ, BAli-Phy, BEAST, StatAlign, others
  - Computationally intensive
  - Most are limited to small datasets (< 30 sequences)

Treelength-based

- Input: Set S of unaligned sequences over an alphabet ∑, and an edit distance function F(.,.) (must account for gaps and substitutions)

- Output: Tree T with sequences S at the leaves and other sequences at the internal nodes so as to minimize

$$\sum_e F(s_v,s_w),$$

where the sum is taken over all edges $e=(s_v,s_w)$ in the tree

Minimizing treelength

- Given set S of sequences and edit distance function F(.,.),
- Find tree T with S at the leaves and sequences at the internal nodes so as to minimize the treelength (sum of edit distances)
- NP-hard but can be approximated
- NP-hard even if the tree is known!

Minimizing treelength

- The problem of finding sequences at the internal nodes of a fixed tree was introduced by Sankoff.
- Several algorithmic results related to this problem, with pretty theory
- Most popular software is POY, which tries to optimize tree length.
- The accuracy of any tree or alignment depends upon the edit distance function F(.,.), but so far even good affine distances don't produce very good trees or alignments.

More

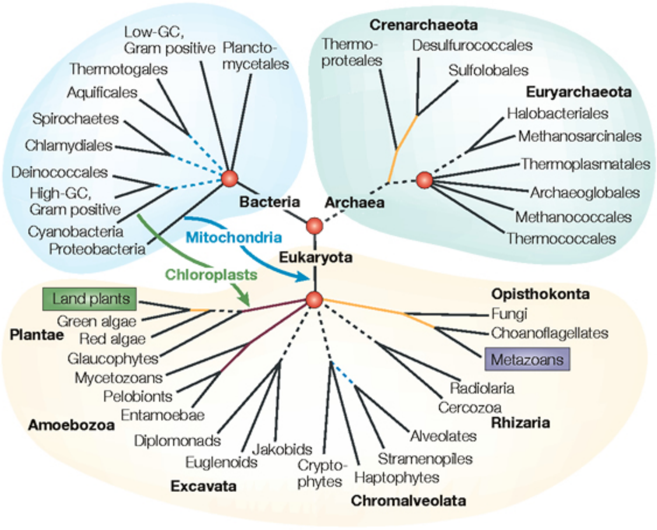
- SATé: a heuristic method for simultaneous estimation and tree alignment
- POY, POY*, and BeeTLe: results of how changing the gap penalty from simple to affine impacts the alignment and tree
- Impact of guide tree on MSA
- Statistical co-estimation using models that include indel events (Statalign, Alifritz, BAliPhy)
- UPP (ultra-large alignments using SEPP)
- Alignment estimation in the presence of duplications and rearrangements
- Visualizing large alignments
- The differences between amino-acid alignments and nucleotide alignments (especially for non-coding data)

Research Projects

- How to use indel information in an alignment?
- Do the statistical estimation methods (Bali-Phy, StatAlign, etc.) produce more accurate alignments than standard methods (e.g., MAFFT)? Do they result in better trees?
- What benefit do we get from an improved alignment? (What biological problem does the alignment method help us solve, besides tree estimation?)

Phylogenomics

(Phylogenetic estimation from whole genomes)

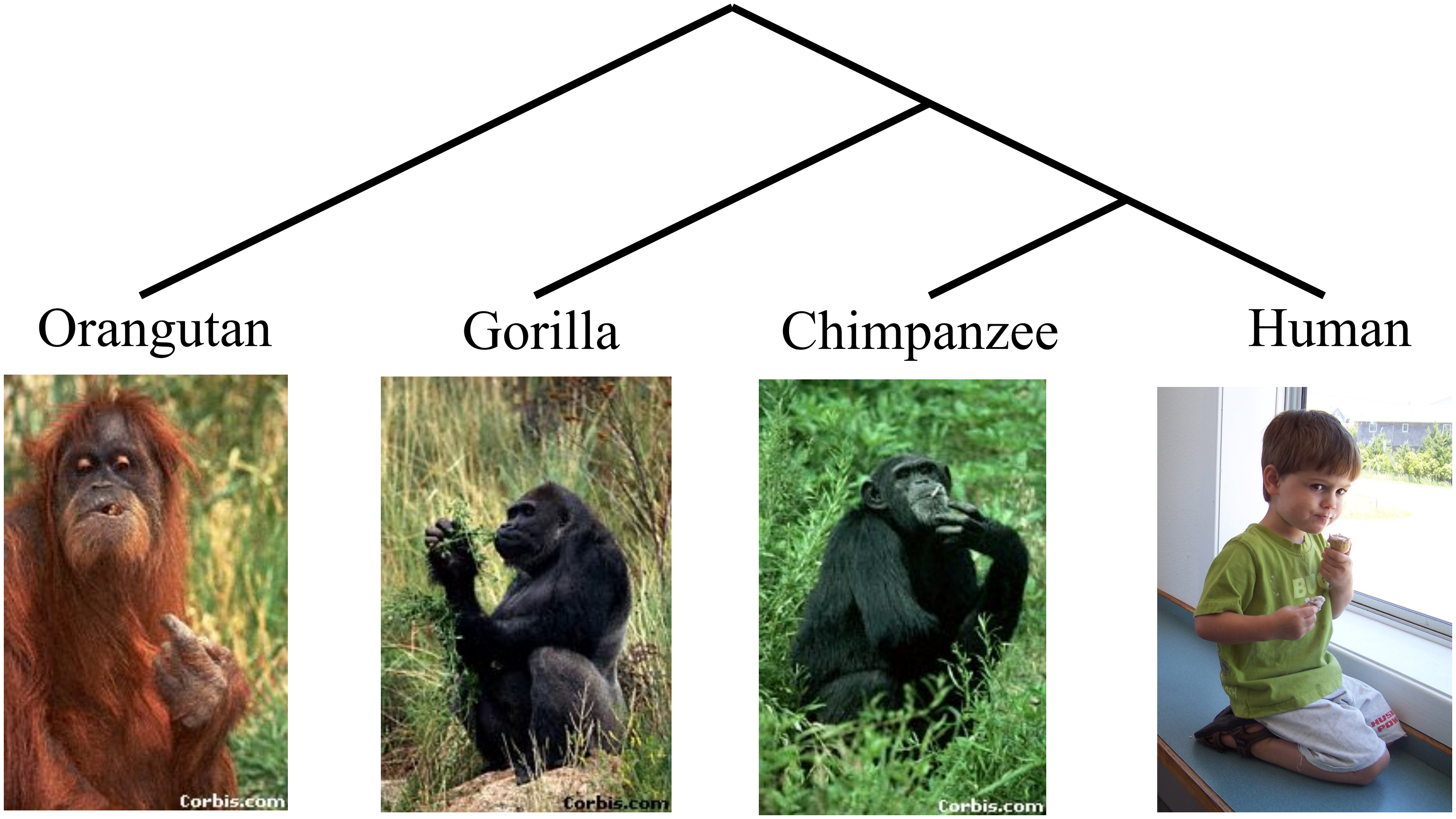


Nature Reviews | Genetics

Gene Trees to Species Trees

- Gene trees are “inside” species trees
- Causes of gene tree discord
- Incomplete lineage sorting
- Methods for estimating species trees from gene trees

Sampling multiple genes from multiple species



Orangutan Gorilla Chimpanzee Human

From the Tree of the Life Website,
University of Arizona

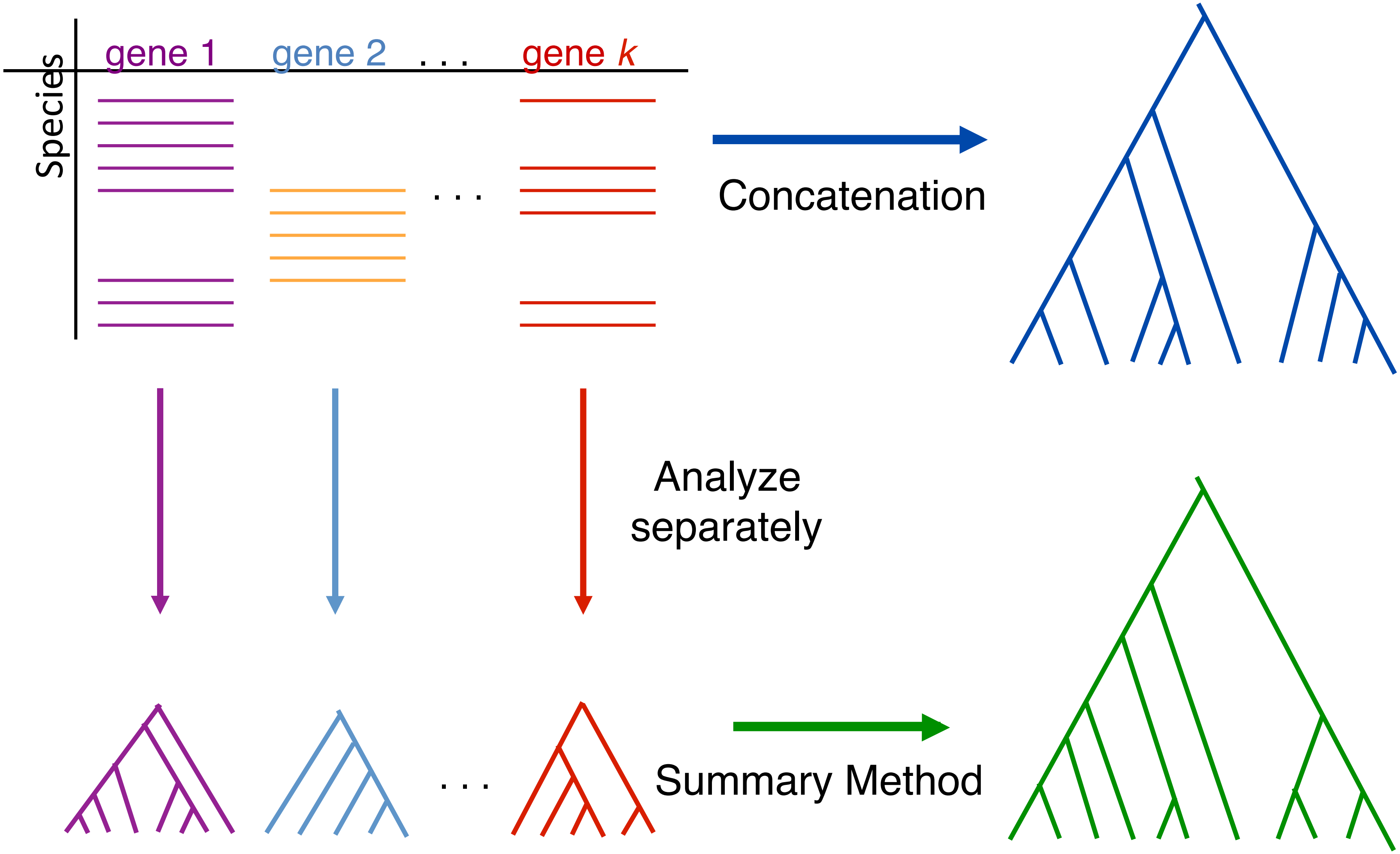
Using multiple genes

	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

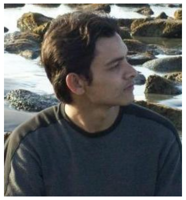
N. Matasci
iPlant

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)

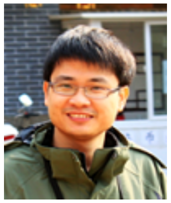
Challenges:
Multiple sequence alignments of > 100,000 sequences
Gene tree incongruence

Avian Phylogenomics Project

Erich Jarvis,
HHMI

MTP Gilbert,
Copenhagen

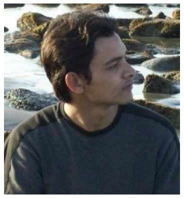
G Zhang,
BGI



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:

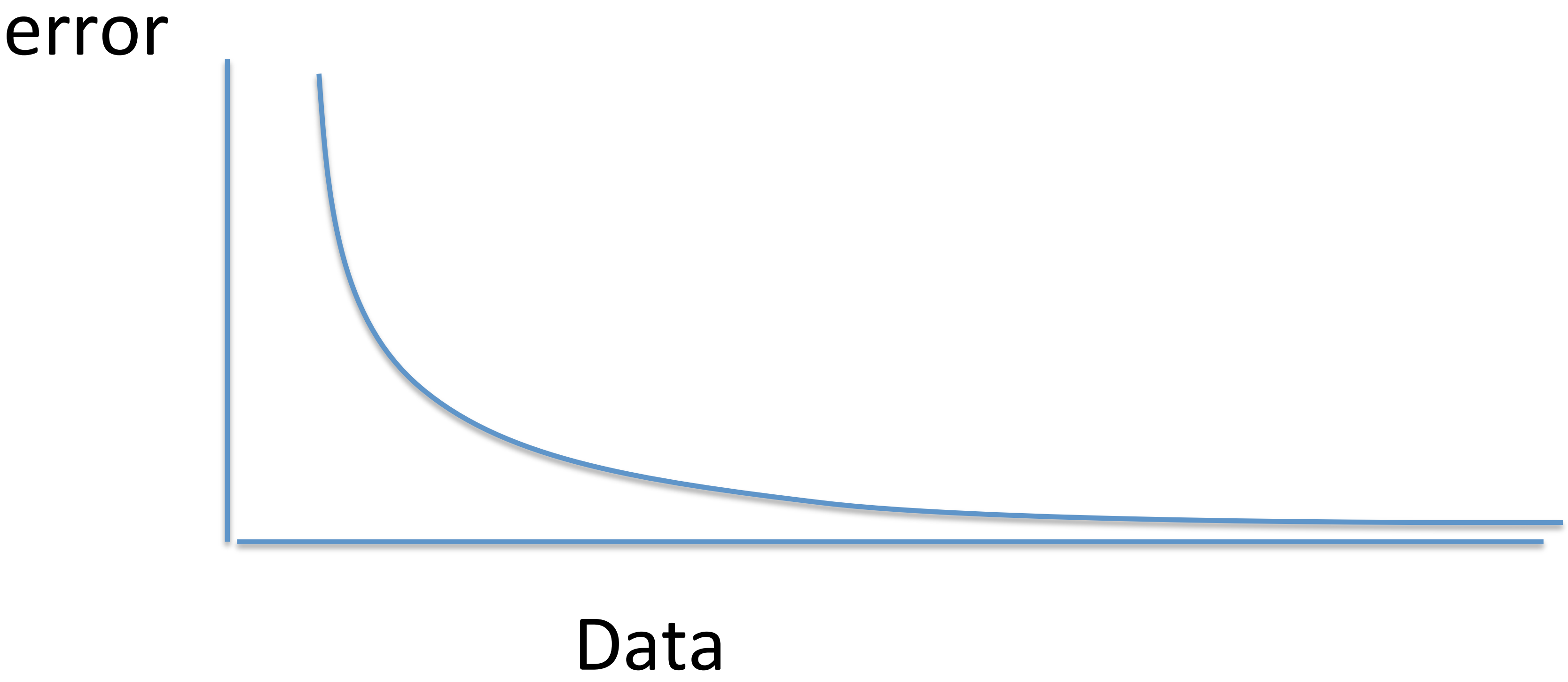
Maximum likelihood on multi-million-site sequence alignments

Massive gene tree incongruence

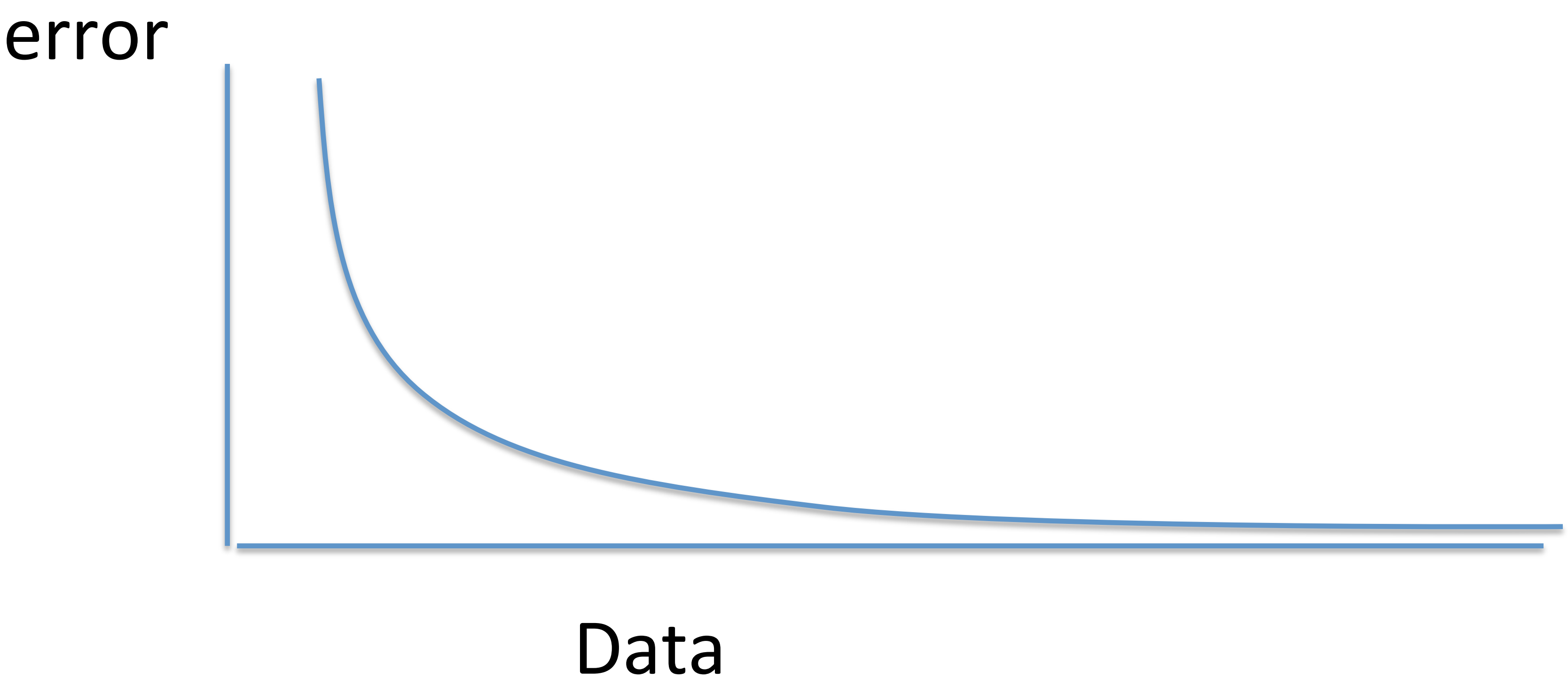
Questions

- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- What is the computational complexity of an estimation problem?

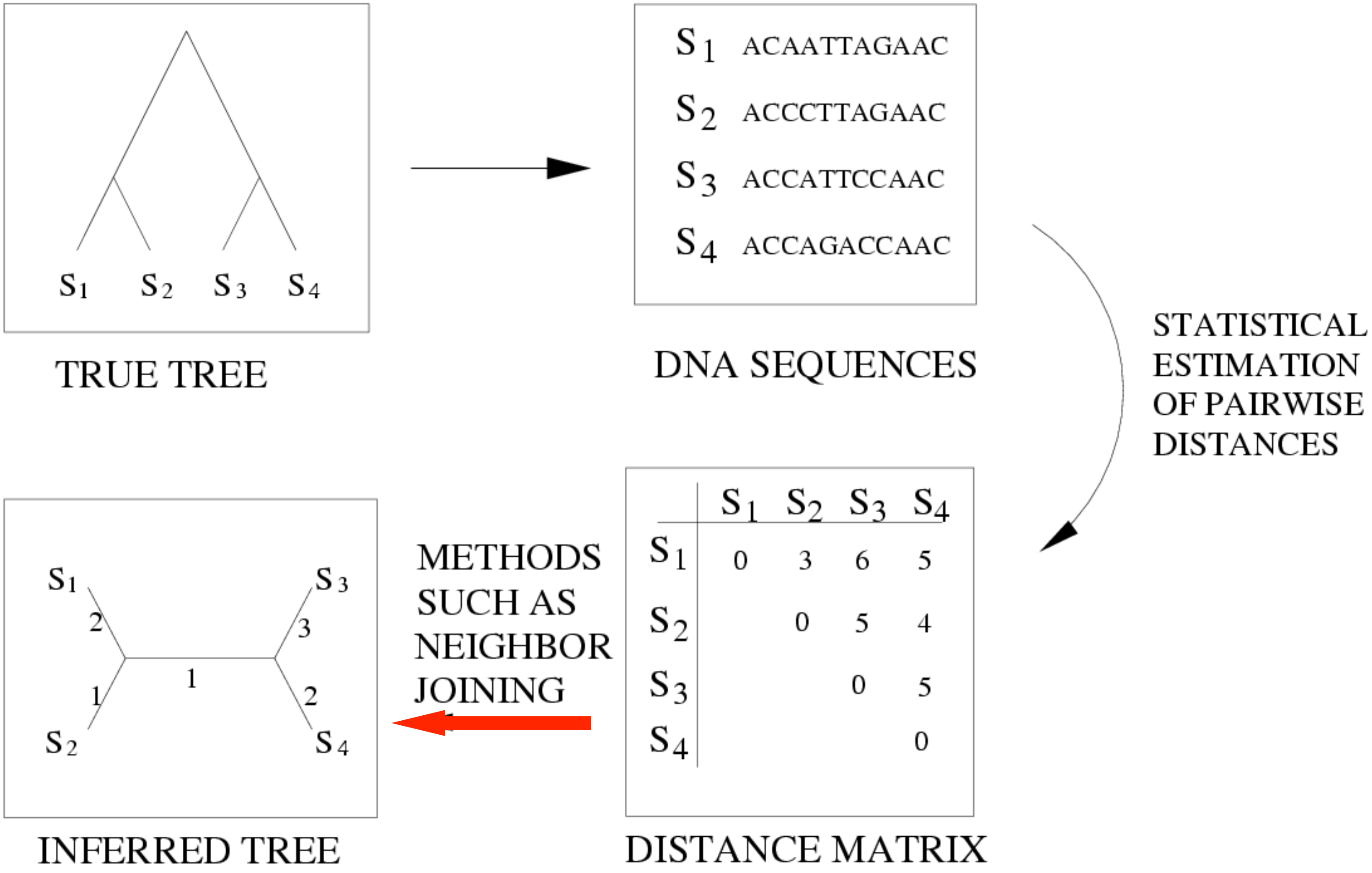
Statistical Consistency



Statistical Consistency



Data are sites in an alignment



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

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

Questions

- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- 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.

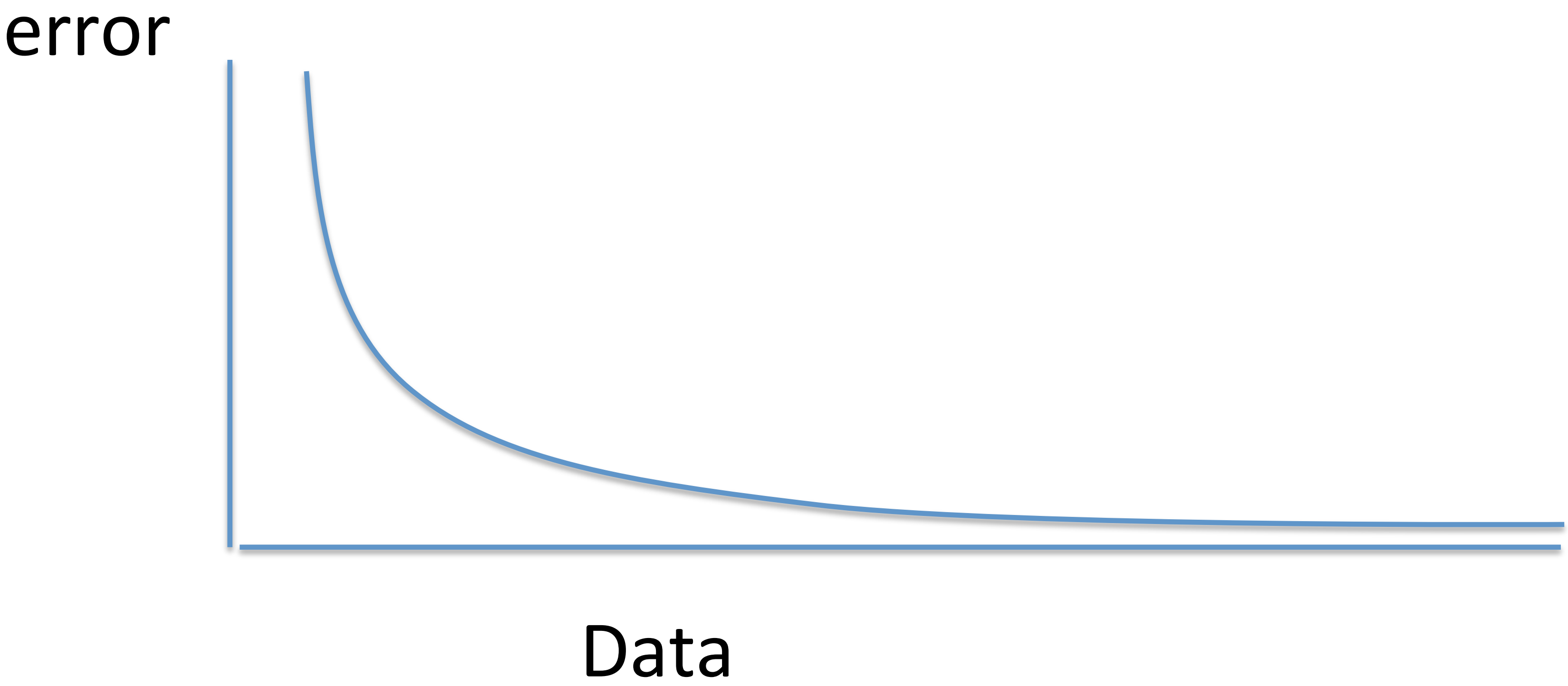
Answers?

- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Just about everything is NP-hard, and the datasets are big.

Answers?

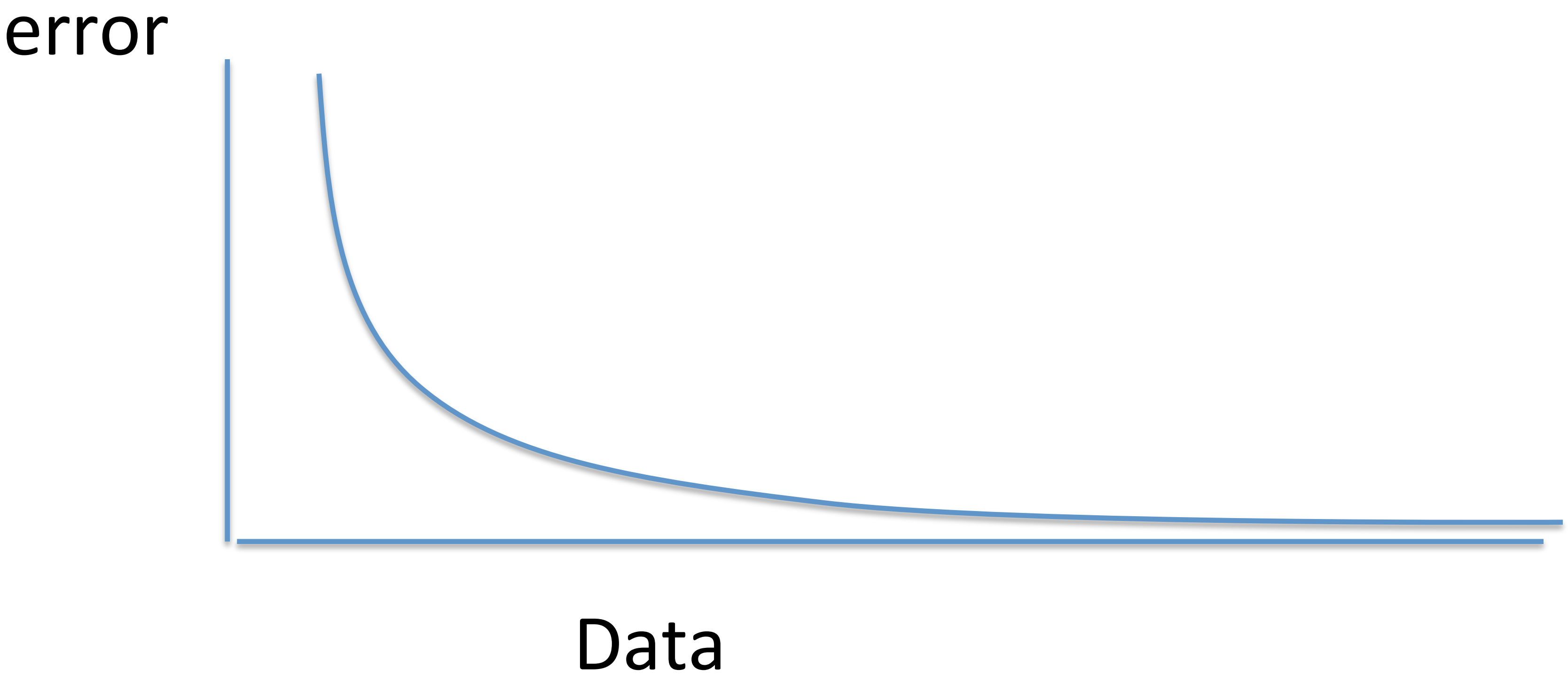
- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Just about everything is NP-hard, and the datasets are big.
- Extensive studies show that even the best methods produce gene trees with some error.

In other words…



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

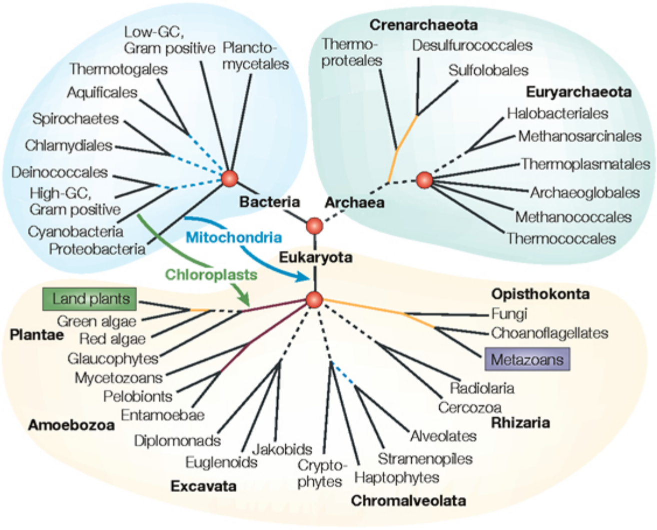
Species Tree Estimation from Gene Trees



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

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Using multiple genes

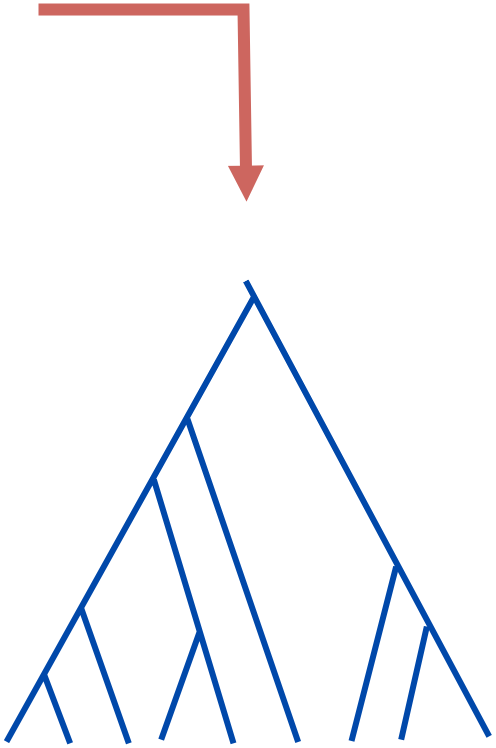
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

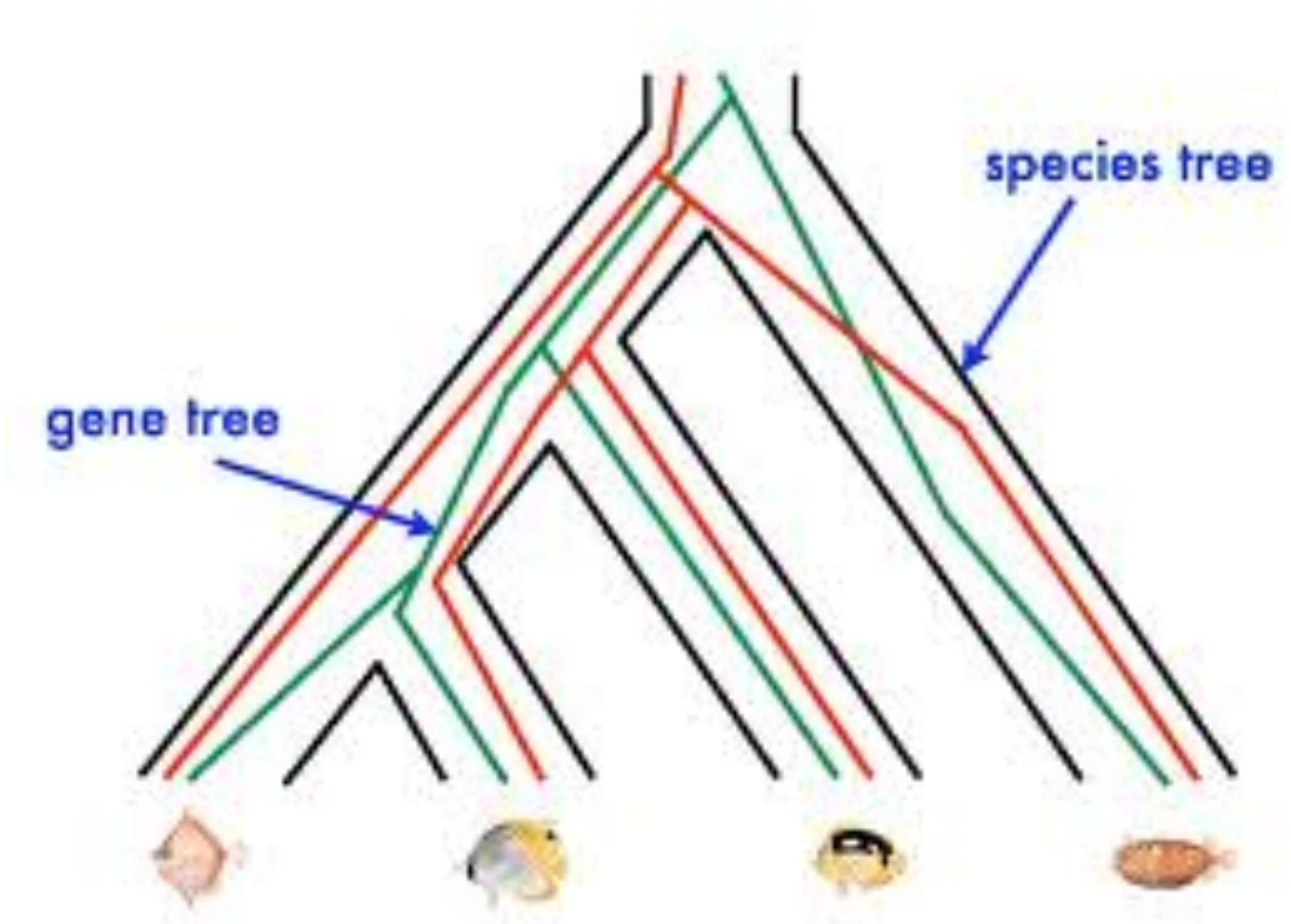
	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

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

Gene Tree Incongruence

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

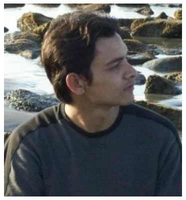
N. Wickett
Northwestern

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)

Avian Phylogenomics Project

E Jarvis,
HHMI

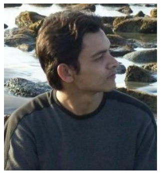
MTP Gilbert,
Copenhagen

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)

Gene Tree Incongruence

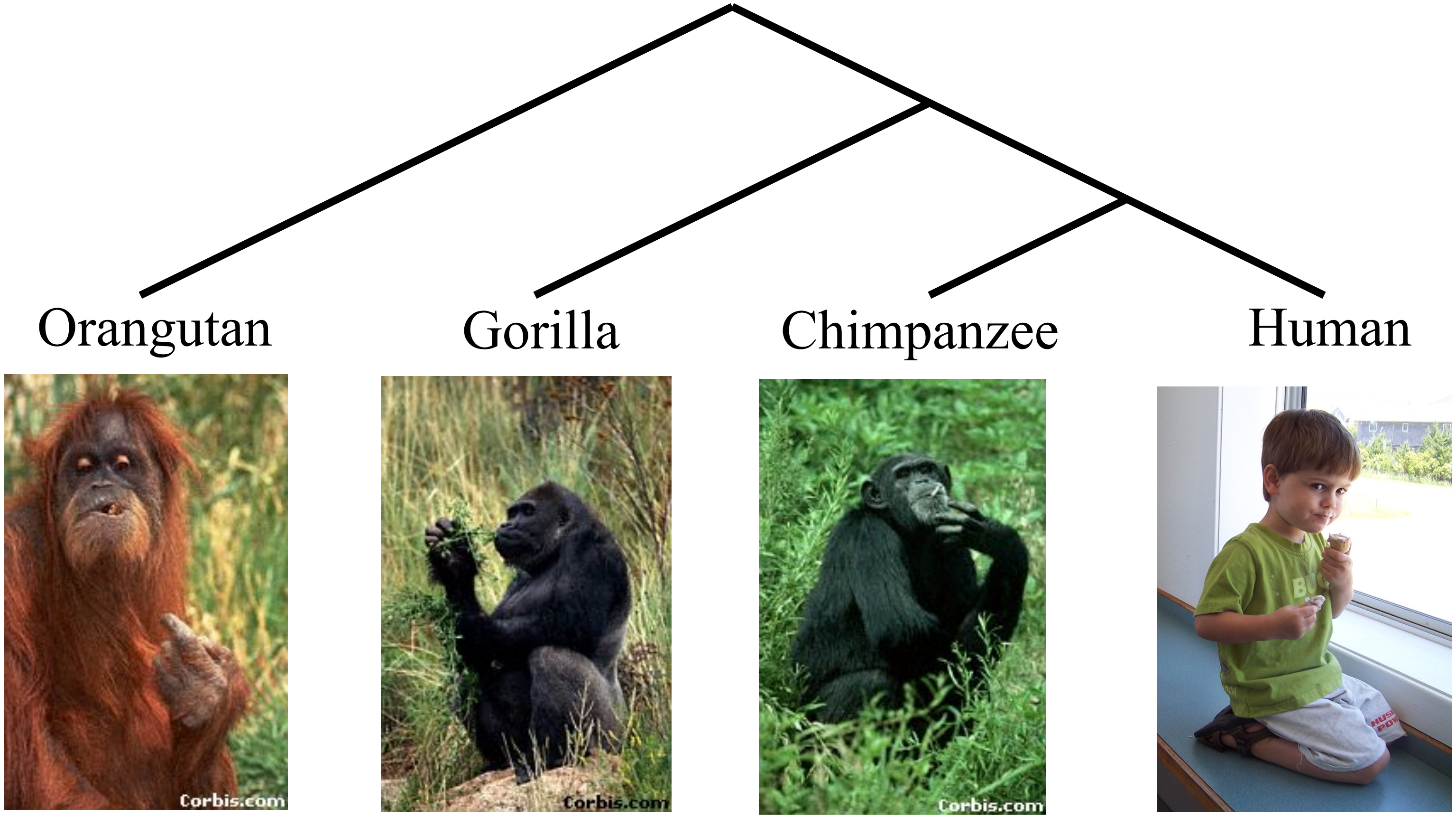
Gene Tree Incongruence

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

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!

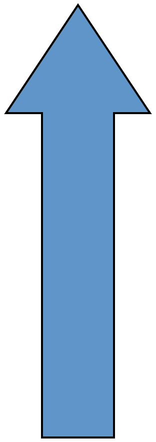


Orangutan Gorilla Chimpanzee Human

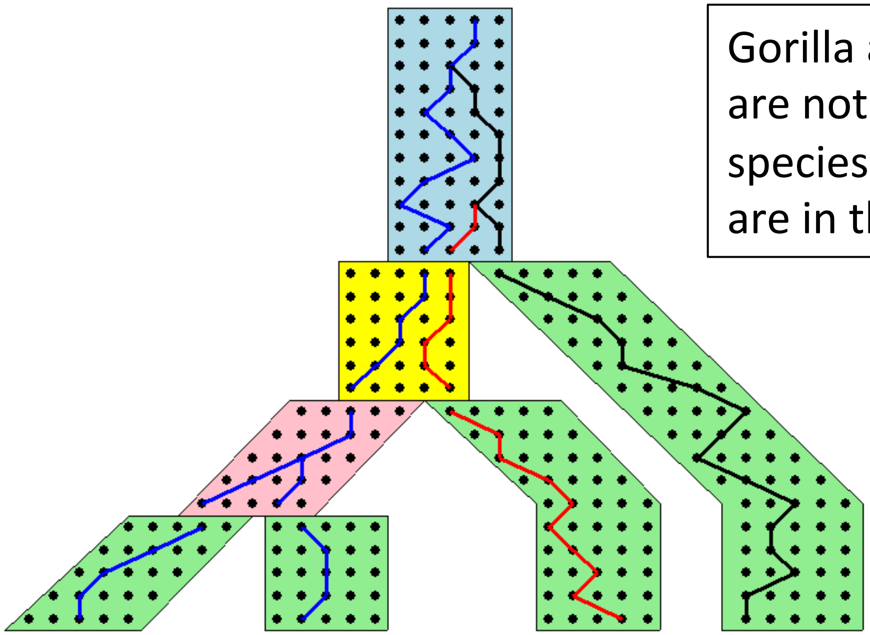
From the Tree of the Life Website,
University of Arizona

The Coalescent

Past



Present

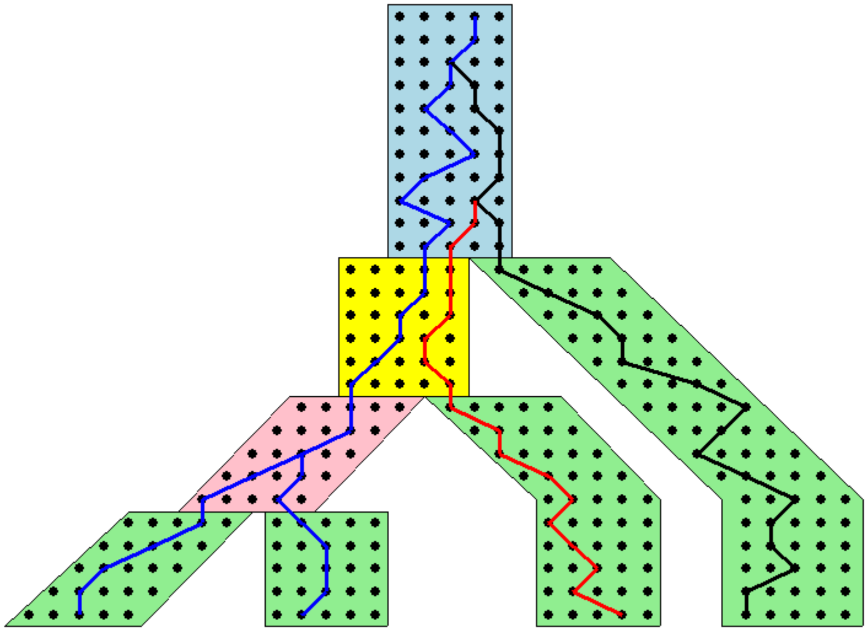


Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Courtesy James Degnan

Gene tree in a species tree

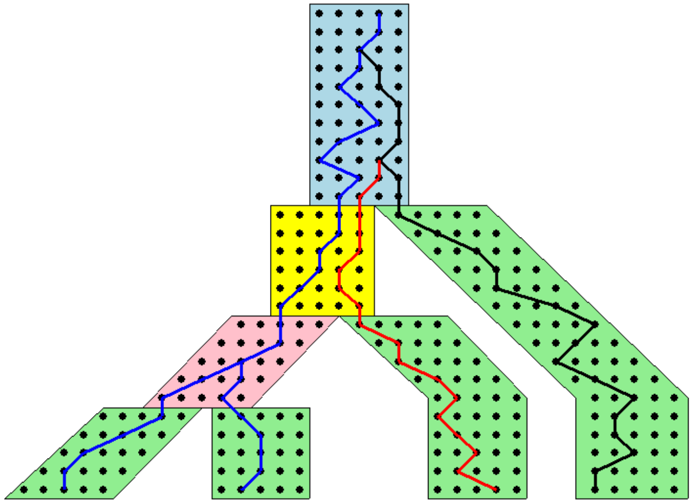
Courtesy James Degnan



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*

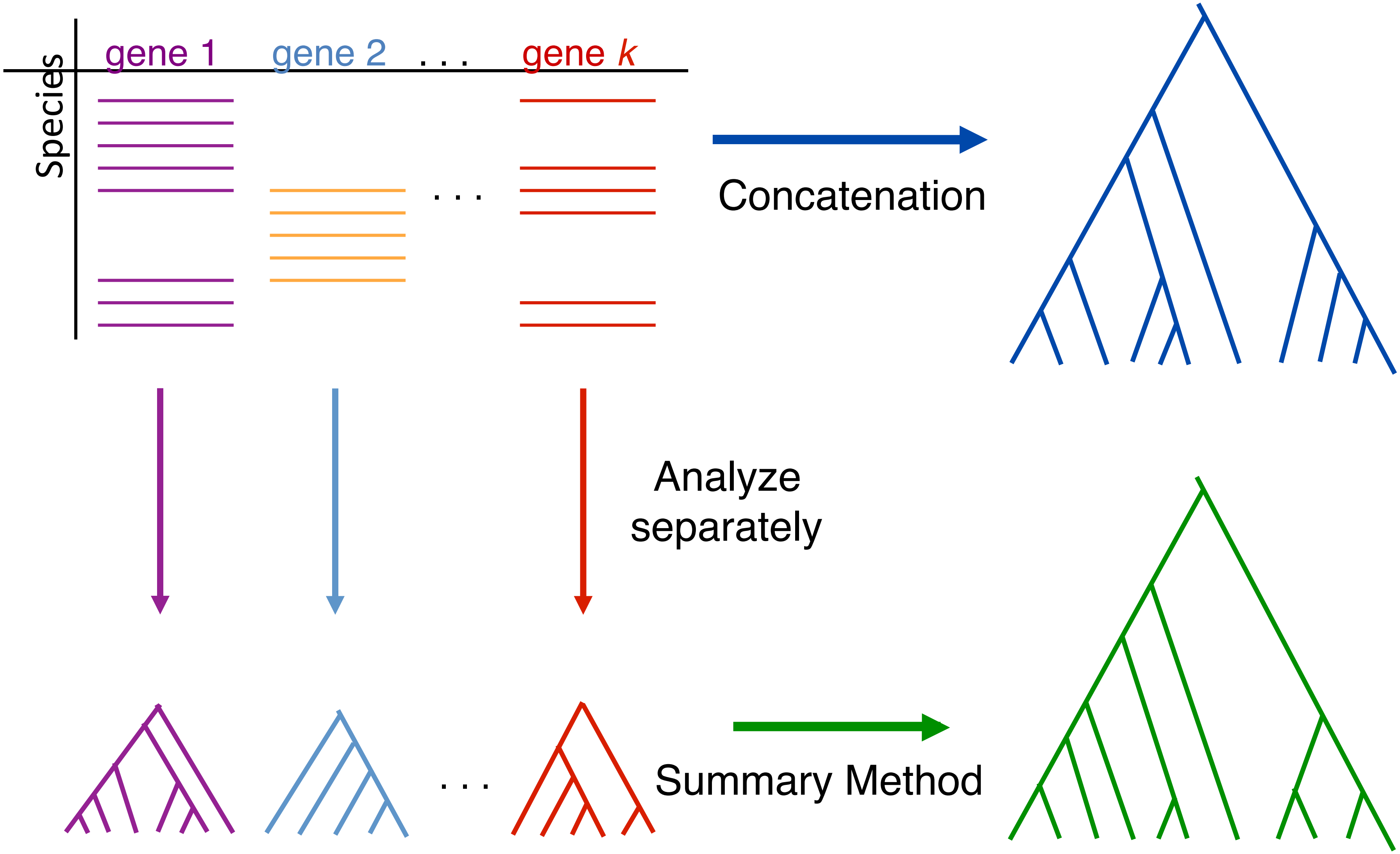


Courtesy James Degnan

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



gene 1

gene 2

. . .

gene *k*

Species

. . .

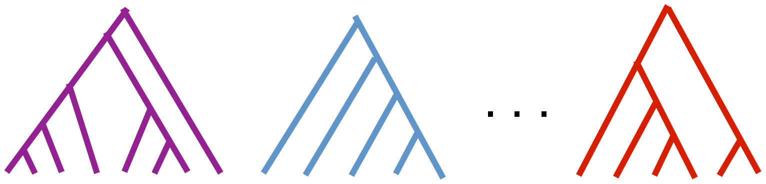
Concatenation

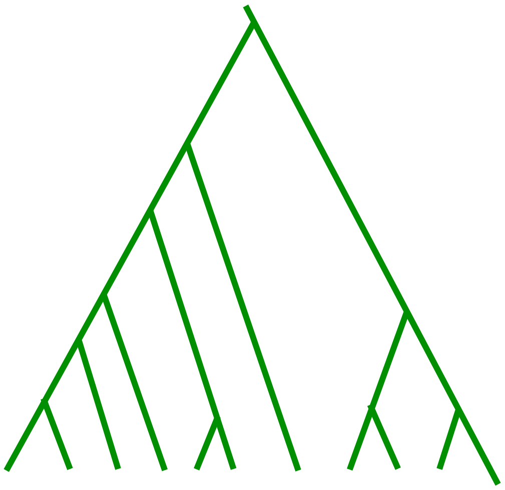
Analyze separately

. . .

Summary Method

How to compute a species tree?

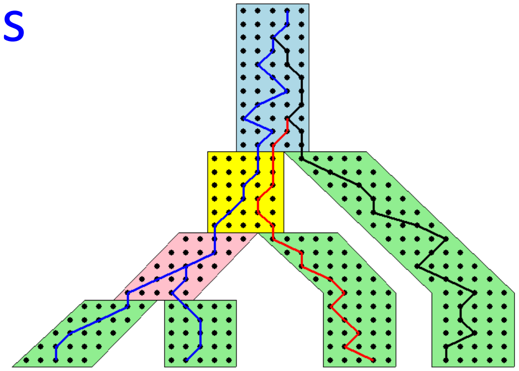




MDC Problem (Maddison 1997)

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.,

$$\text{minimize } MDC(T, X) = \mathbf{\Sigma}_i\ XL(T,t_i)$$

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!

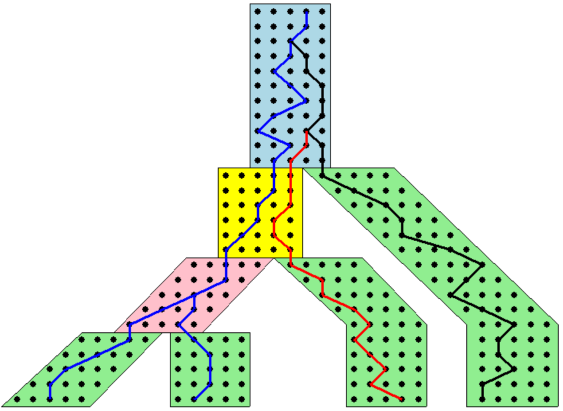
Statistically consistent under ILS?

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

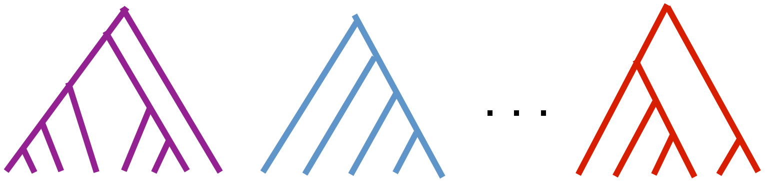
Under the multi-species coalescent model, the species tree defines a probability distribution on the gene trees

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}.

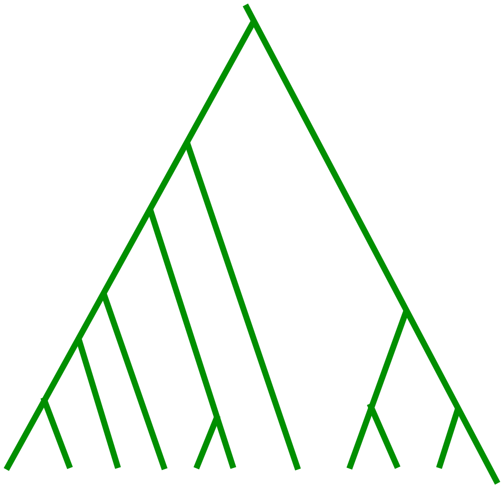


How to compute a species tree?



Techniques:

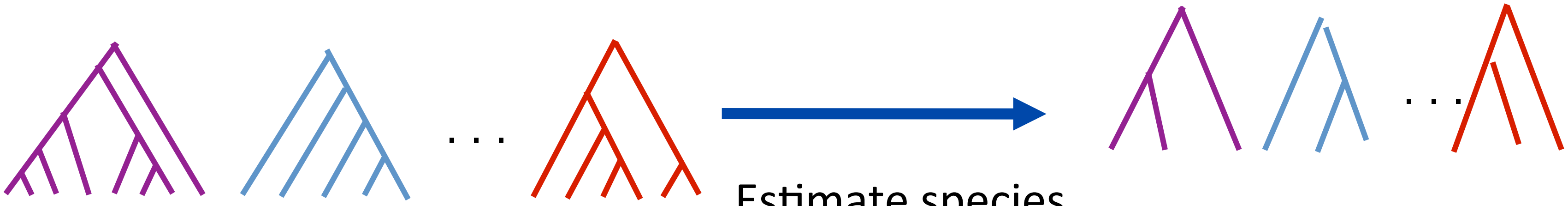
- MDC?
- Most frequent gene tree?
- Consensus of gene trees?
- Other?



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}.

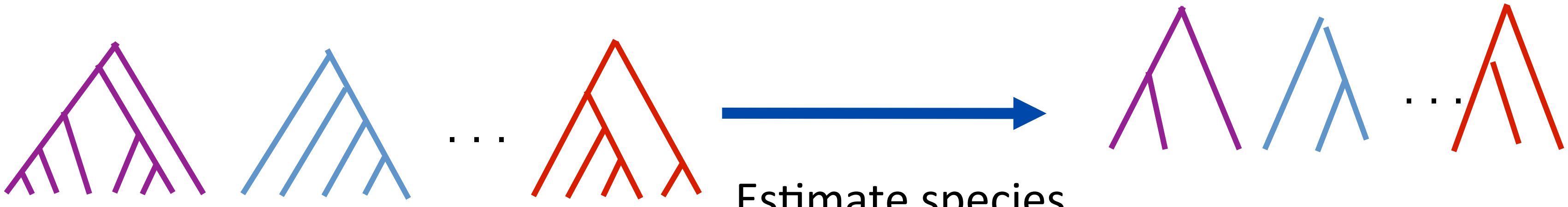
How to compute a species tree?



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}.

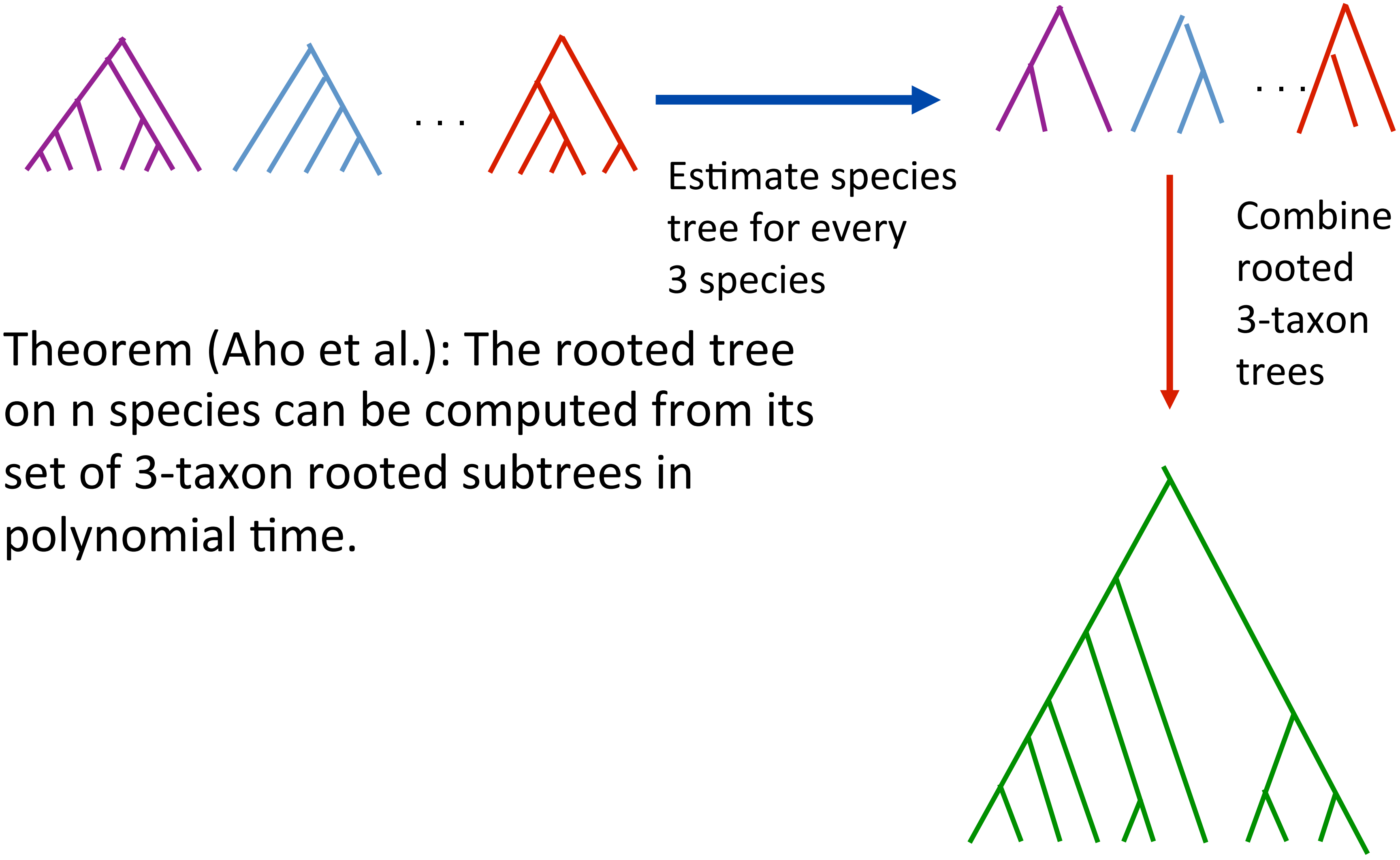
How to compute a species tree?



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?



. . .

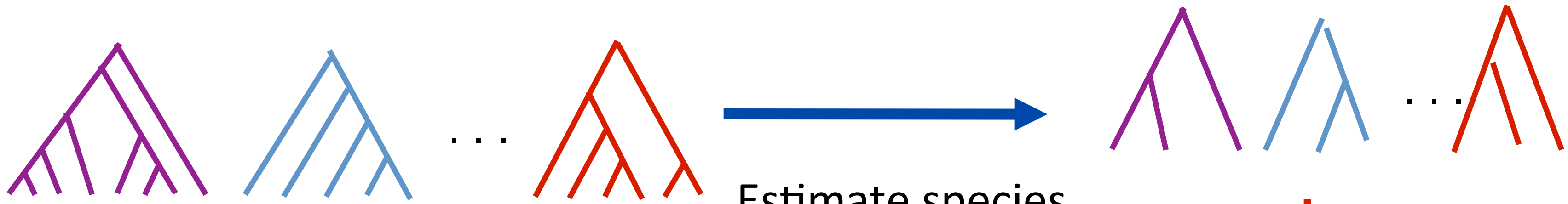
Estimate species tree for every 3 species

. . .

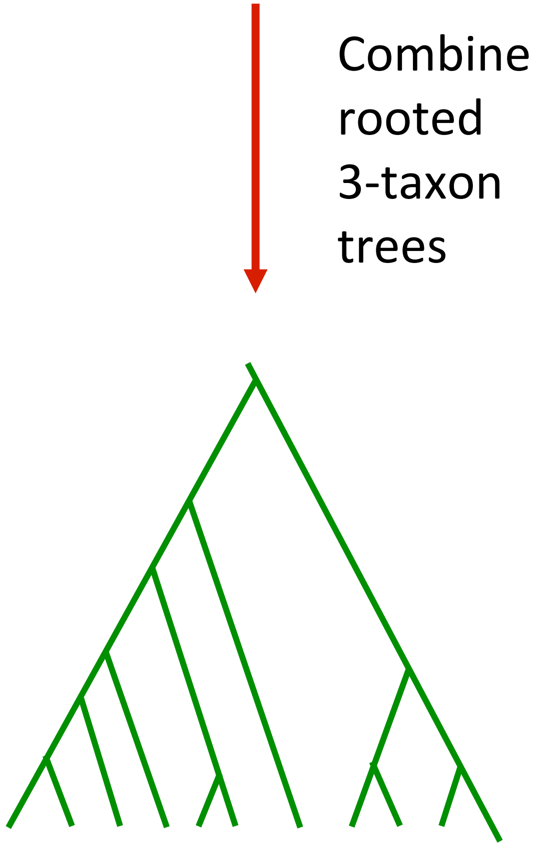
Combine rooted 3-taxon trees

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?



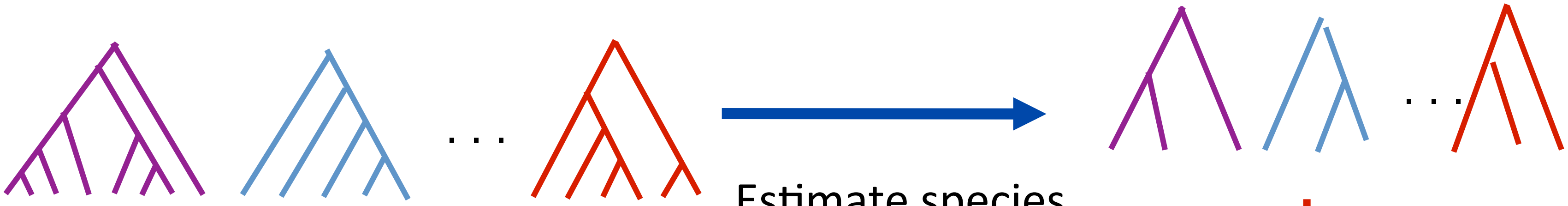
Estimate species tree for every 3 species



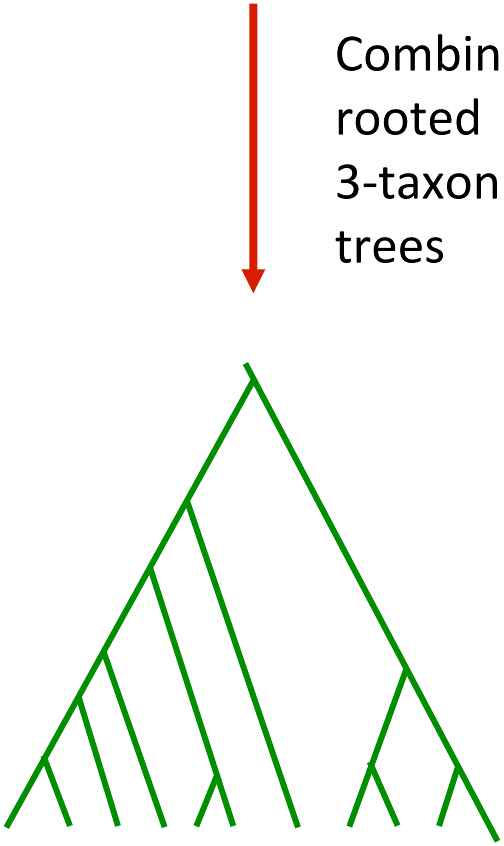
Combine rooted 3-taxon trees

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 3 species

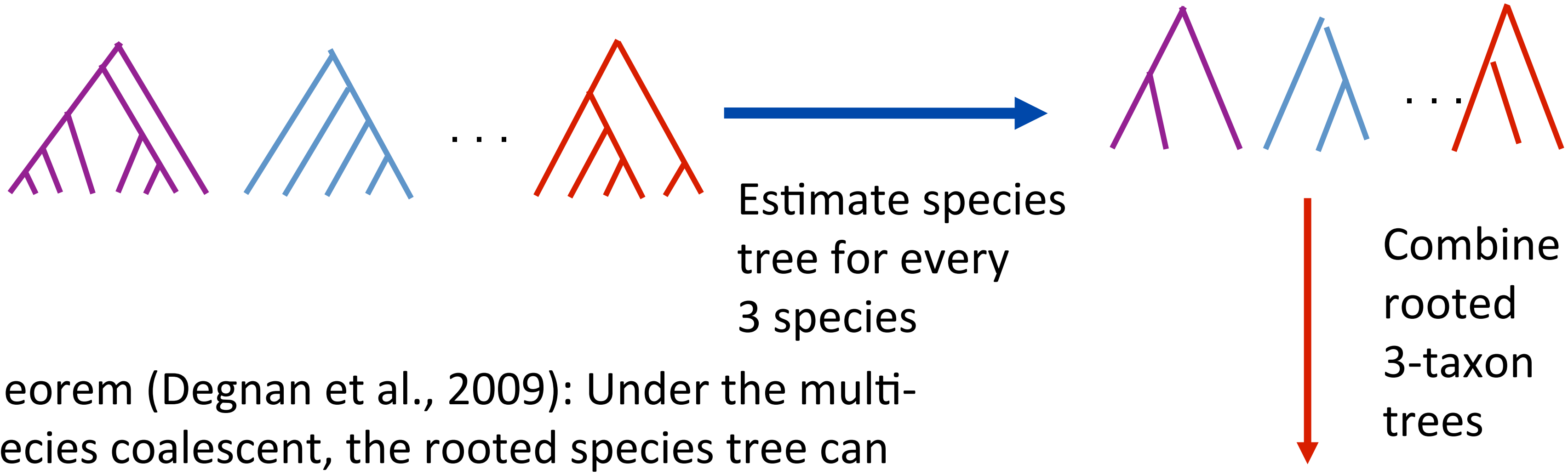


Combine rooted 3-taxon trees

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.

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?

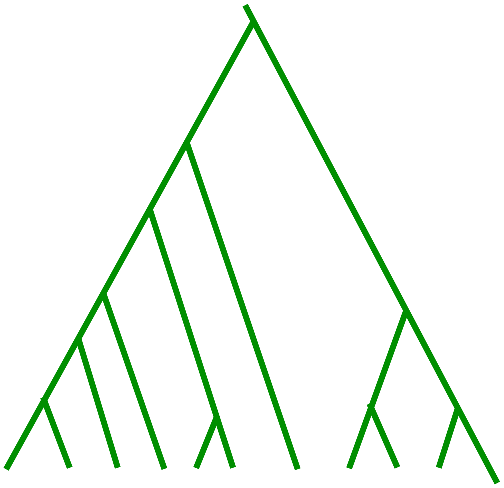


Estimate species tree for every 3 species

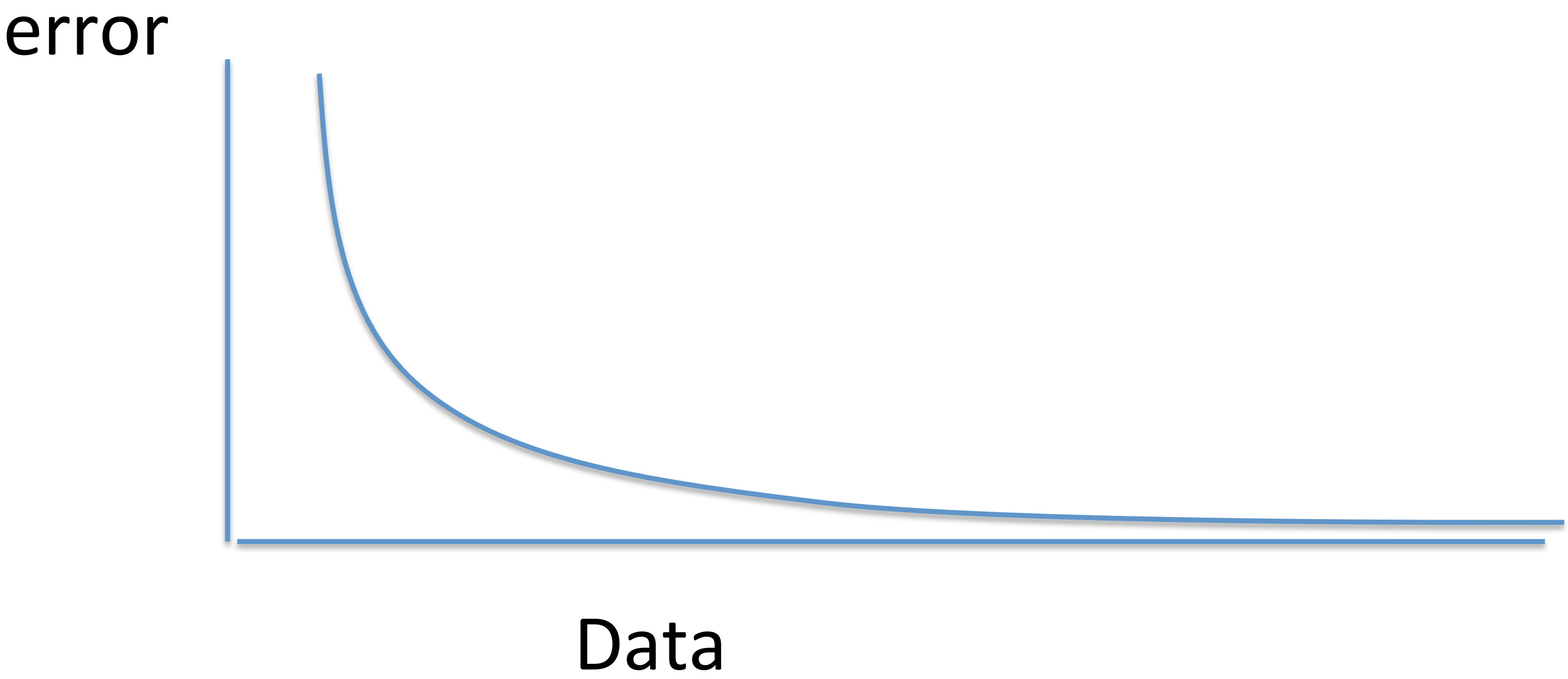
Combine rooted 3-taxon trees

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.

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 true gene trees.

Statistically consistent methods under ILS

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

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

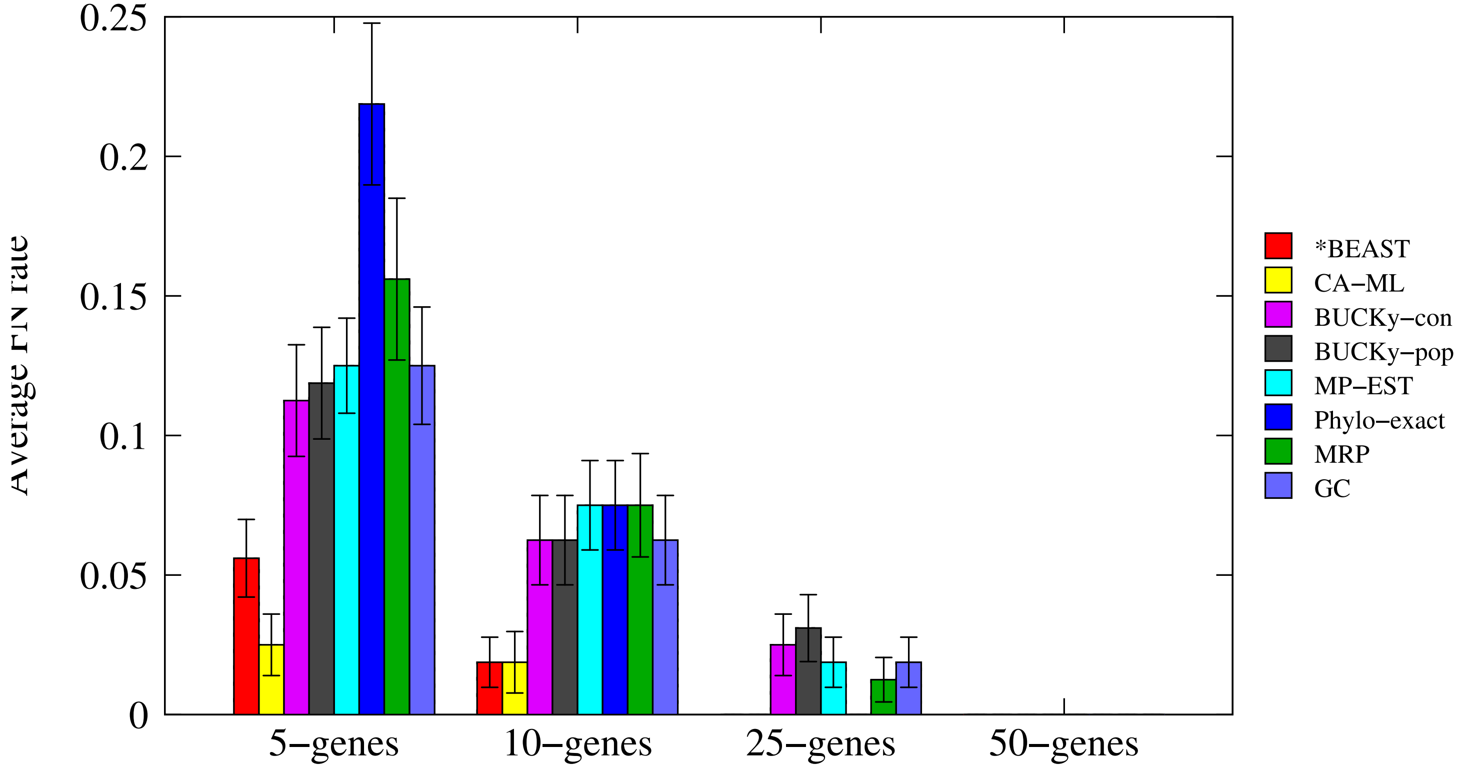
*BEAST (Heled and Drummond, 2011), co-estimates gene trees and species trees

(and some others)

Questions

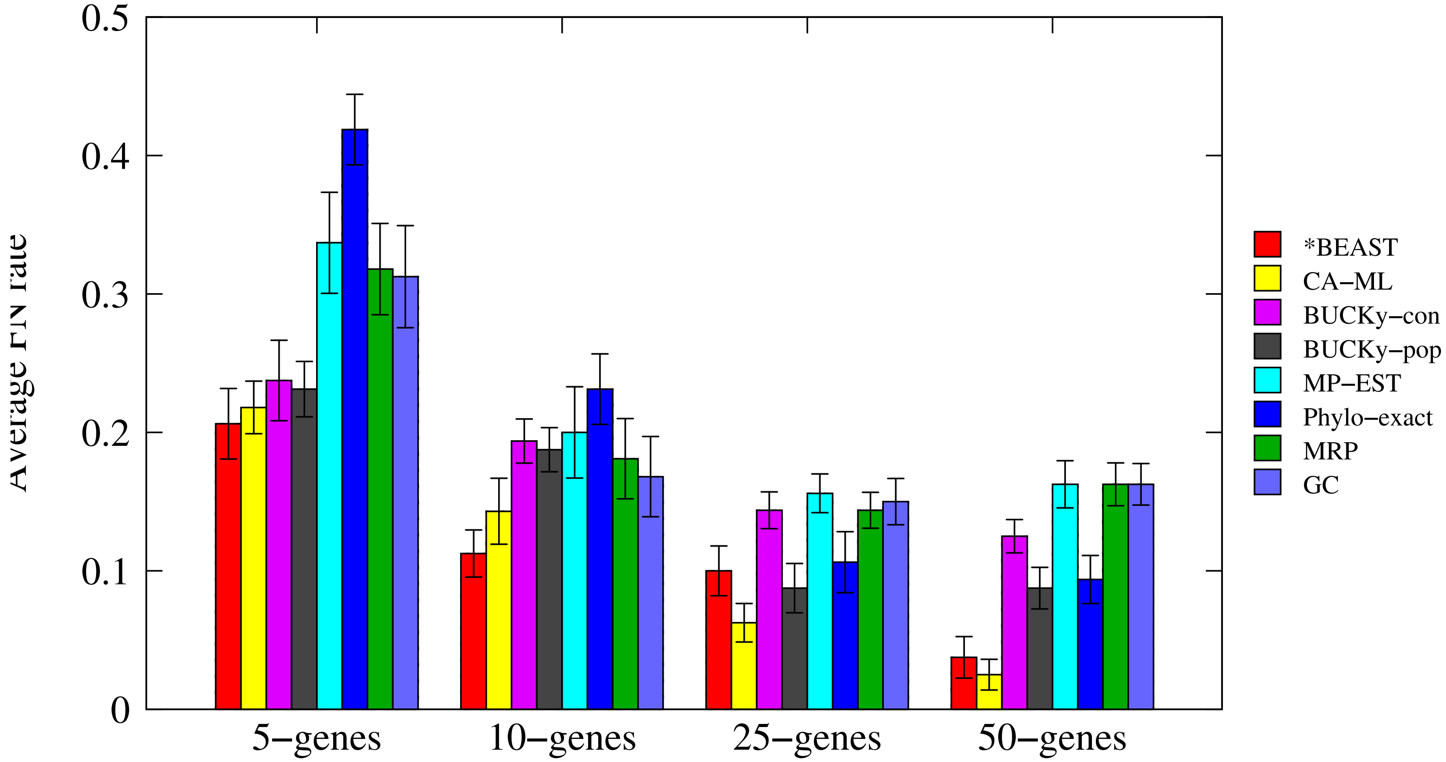
- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- What is the computational complexity of an estimation problem?
- What is the performance in practice?

Results on 11-taxon weakILS



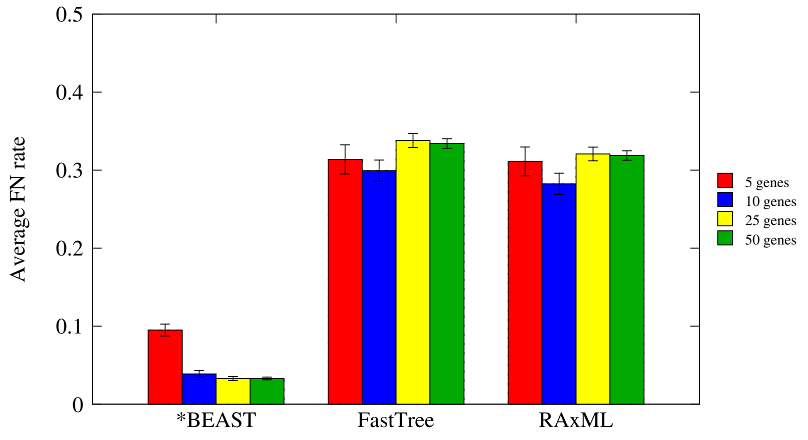
20 replicates studied, due to computational challenge of running *BEAST and BUCKy

Results on 11-taxon strongILS

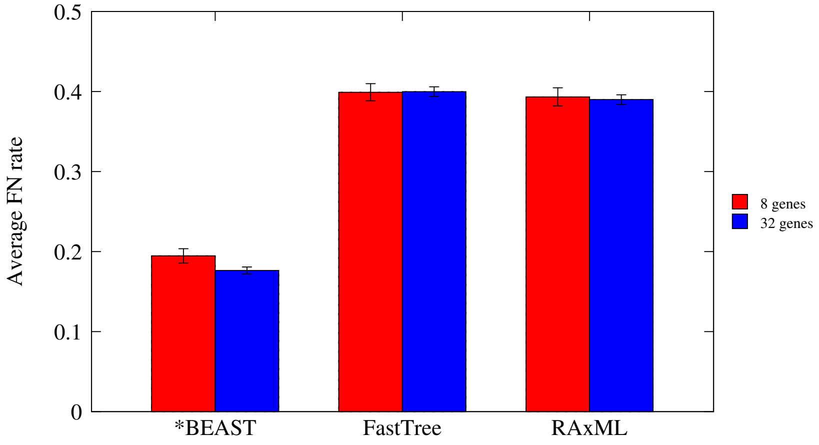


20 replicates studied, due to computational challenge of running *BEAST and BUCKy

*BEAST is better than ML at estimating gene trees



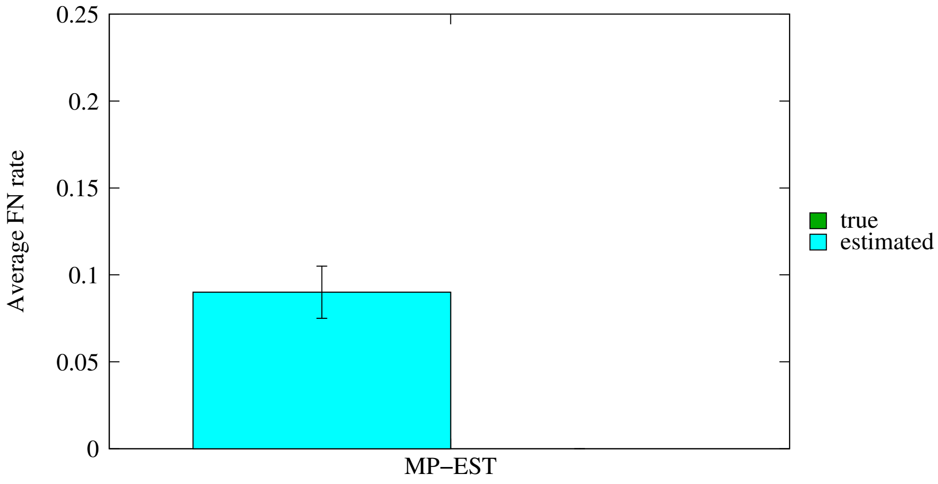
11-taxon weakILS datasets



17-taxon (very high ILS) datasets

- FastTree-2 and RAxML very close in accuracy
- *BEAST much more accurate than both ML methods
- *BEAST gives <u>biggest improvement under low-ILS conditions</u>

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 phylogenetic signal

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

Controversies/Open Problems

- Concatenation may (or may not be) statistically consistent under ILS – but some simulation studies suggest it can be positively misleading.
- Coalescent-based methods have not in general given strong results on biological data – can give poor bootstrap support, or produce strange trees, compared to concatenation.

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.

Research Projects

- Coalescent-based methods: analyze a biological dataset using different coalescent-based methods and compare to concatenation
- Evaluation impact of choice of gene trees (e.g., removing gene trees with low support)
- Evaluate impact of missing taxa in gene trees
- Develop new coalescent-based method (e.g., combine quartet trees)
- Evaluate scalability of coalescent-based methods