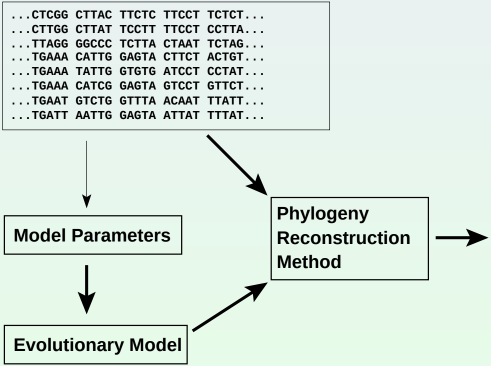
Comparison of commonly used methods for combining multiple phylogenetic data sets

Anne Kupczok, Heiko A. Schmidt and Arndt von Haeseler

Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

SMBE, June, 2009

Phylogeny reconstruction scheme



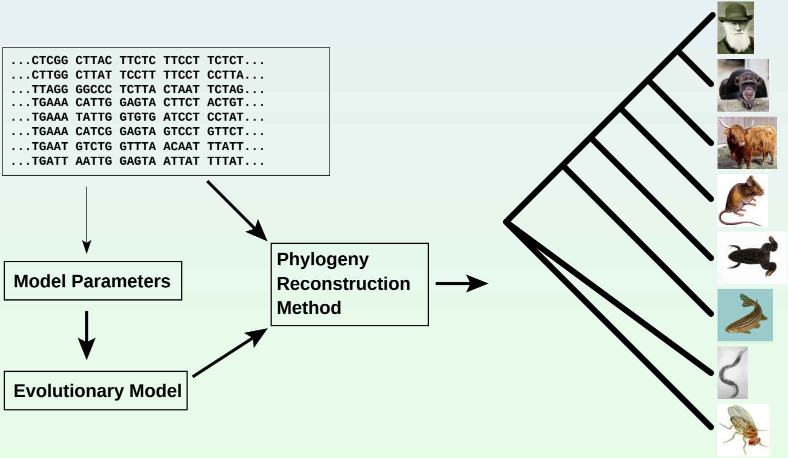
```
...CTCGG CTTAC TTCTC TTCCT TCTCT...
...CTTGG CTTAT TCCTT TTCCT CCTTA...
...TTAGG GGCCC TCTTA CTAAT TCTAG...
...TGAAA CATTG GAGTA CTTCT ACTGT...
...TGAAA TATTG GTGTG ATCCT CCTAT...
...TGAAA CATCG GAGTA GTCCT GTTCT...
...TGAAT GTCTG GTTTA ACAAT TTATT...
...TGATT AATTG GAGTA ATTAT TTTAT...
```

Model Parameters

Evolutionary Model

Phylogeny Reconstruction Method

Phylogeny reconstruction scheme



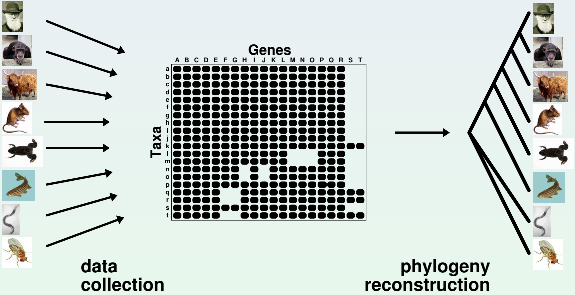
```
...CTCGG CTTAC TTCTC TTCCT TCTCT...
...CTTGG CTTAT TCCTT TTCCT CCTTA...
...TTAGG GGCCC TCTTA CTAAT TCTAG...
...TGAAA CATTG GAGTA CTTCT ACTGT...
...TGAAA TATTG GTGTG ATCCT CCTAT...
...TGAAA CATCG GAGTA GTCCT GTTCT...
...TGAAT GTCTG GTTTA ACAAT TTATT...
...TGATT AATTG GAGTA ATTAT TTTAT...
```

Model Parameters

Evolutionary Model

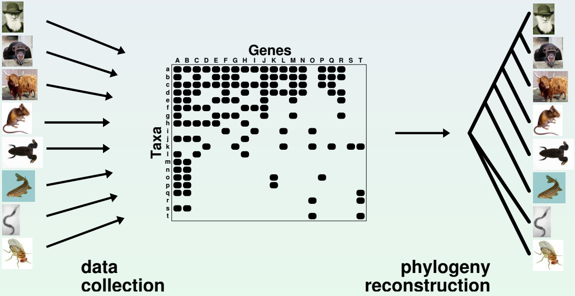
Phylogeny Reconstruction Method

Multi-Locus Datasets



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
a	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
b	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
c	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
d	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
e	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
f	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
g	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
h	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
i	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
j	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
k	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
l	●	●	●	●	●	●	●	●	●	●	●	●				●	●	●		
m	●	●	●	●	●	●	●	●	●	●	●	●				●	●	●		
n	●	●	●	●	●	●	●		●		●	●	●	●	●	●	●	●		
o	●	●	●	●	●	●	●		●		●	●	●	●	●	●	●	●		
p	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
q	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●	●
r	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●	●
s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
t	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●	●

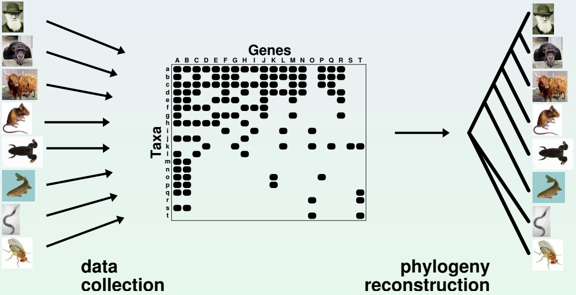
Multi-Locus Datasets



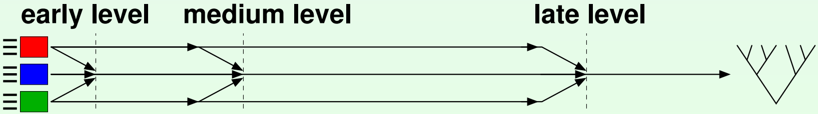
Genes

Taxa	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
a	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●		
b	●		●		●	●	●			●	●	●	●	●		●	●	●		
c	●	●	●	●	●	●	●	●	●	●	●		●	●		●	●			
d	●	●	●			●		●		●	●	●	●	●			●	●		
e	●	●			●	●	●			●			●					●		
f	●	●	●	●				●	●	●										
g	●				●	●	●			●			●					●		
h	●	●	●	●	●			●												
i						●			●			●			●					
j	●	●	●					●												
k				●			●					●			●		●		●	●
l	●		●					●												
m	●	●																		
n	●	●																		
o	●	●									●					●				
p	●	●									●									
q	●	●																		●
r															●					●
s	●	●																		
t															●					●

Multi-Locus Datasets

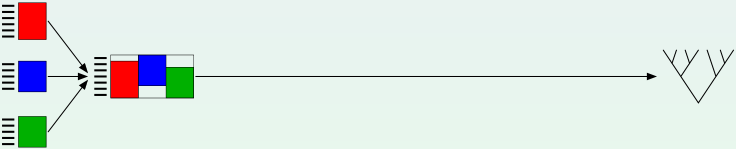


Approaches:


early level **medium level** **late level**

Early-level combination: Superalignment (Supermatrix)

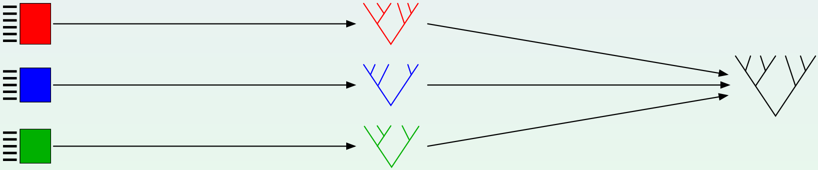
- Combination by concatenating data sets:



- Any tree reconstruction method can be applied to the large alignment
- here: Maximum Likelihood with IQPNNI (Vinh and von Haeseler, 2004)

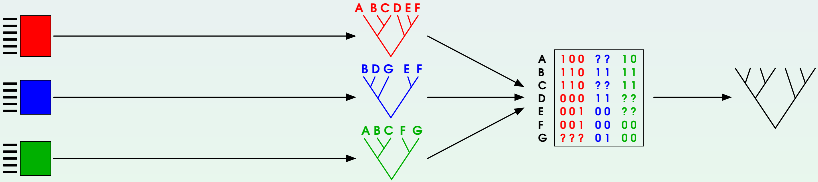
Late-level combination: Supertree

Construct separate trees for each gene and combine them to a supertree:



Late-level combination: Supertree

Construct separate trees for each gene and combine them to a supertree:



A	100	??	10
B	110	11	11
C	110	??	11
D	000	11	??
E	001	00	??
F	001	00	00
G	???	01	00

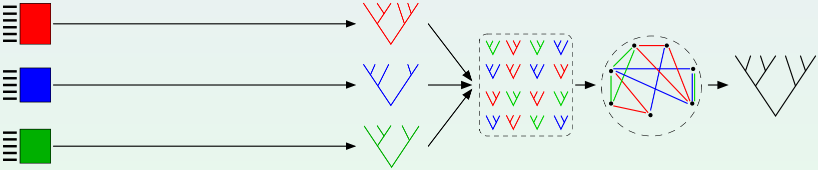
Supertree methods combine different kinds of information from the gene trees:

Split information → Matrix Representation:

- MR with Parsimony (MRP, Baum, 1992; Ragan, 1992)
- MR with Flipping (MRF, e.g. Chen et al., 2003)

Late-level combination: Supertree

Construct separate trees for each gene and combine them to a supertree:



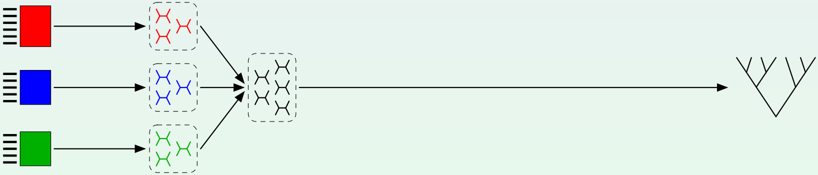
Supertree methods combine different kinds of information from the gene trees:

Triplet information → Rooted triplets:

- MinCut (Semple and Steel, 2000)
- Modified MinCut (Page, 2002)
- MaxCut (Snir and Rao, 2006)

Medium-level combination

Intermediate data (not final trees) are computed from every gene alignment and subsequently combined to a tree.



SuperQP: Combination of quartet likelihoods (Schmidt, 2003)

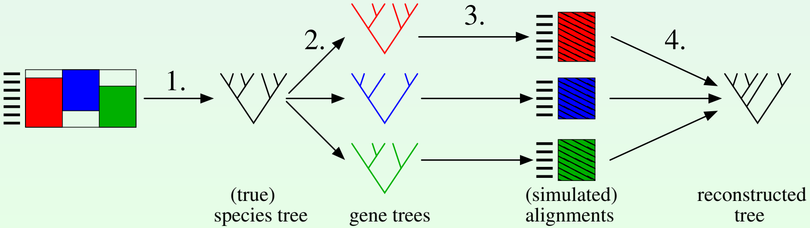
Medium-level combination

Intermediate data (not final trees) are computed from every gene alignment and subsequently combined to a tree.

Average Consensus: Average over distance matrix for each gene (Lapointe and Cucumel, 1997)
Super Distance Matrix (SDM): Additional weights estimated (Criscuolo et al., 2006)

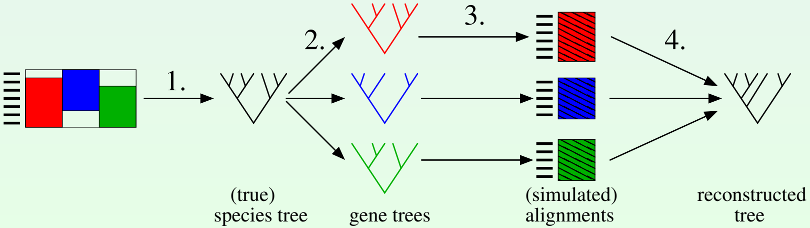
Simulation setting

1. Estimate an ML tree with branch lengths and model parameters from a data superalignment → species tree



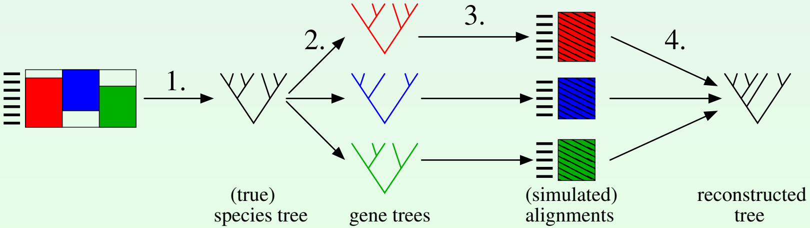
Simulation setting

1. Estimate an ML tree with branch lengths and model parameters from a data superalignment → species tree
2. Generate gene trees



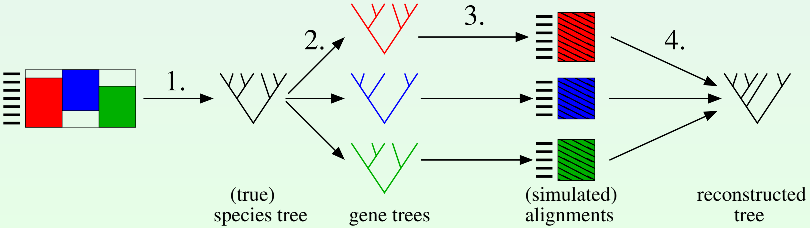
Simulation setting

1. Estimate an ML tree with branch lengths and model parameters from a data superalignment → species tree
2. Generate gene trees
3. Simulate alignments along the gene trees



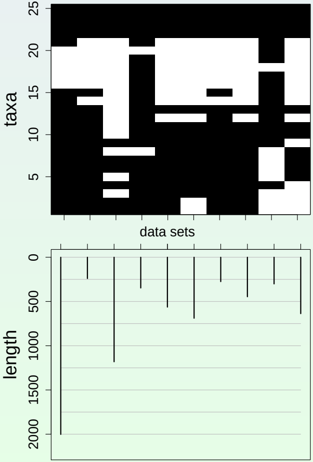
Simulation setting

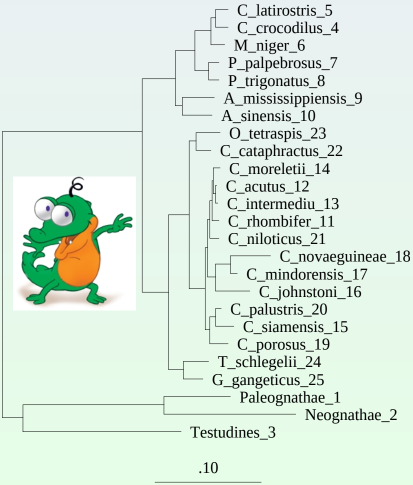
1. Estimate an ML tree with branch lengths and model parameters from a data superalignment → species tree
2. Generate gene trees
3. Simulate alignments along the gene trees
4. Apply the reconstruction methods to each data set and compare the result with the model tree



Species tree

10 genes of 25 Crocodylia species (Gatesy et al., 2004)

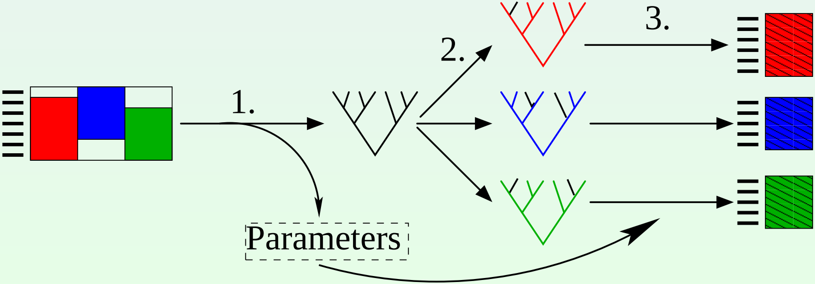


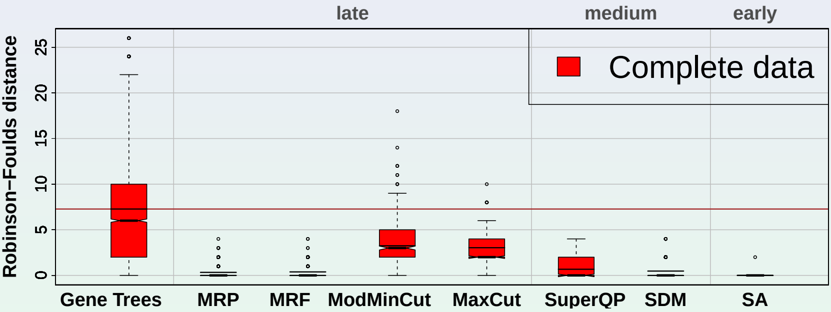


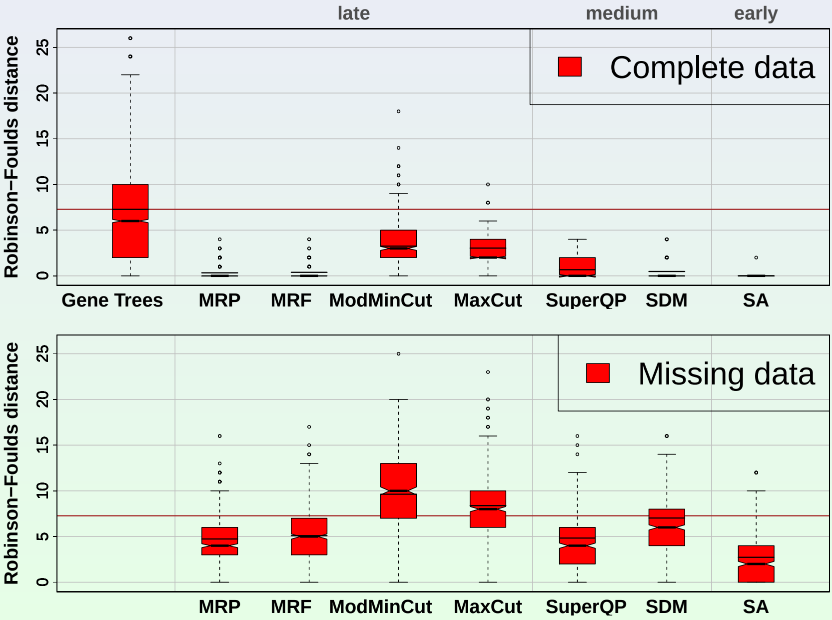
Complete and missing data

Step 2: Gene trees are the complete model tree (complete data) or the pruned model tree (missing data)

Step 3: Simulation with the parameters estimated from the superalignment



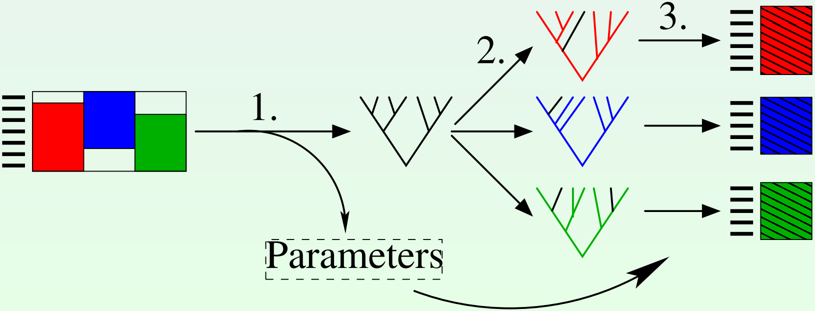

late
medium
early
Complete data
Robinson−Foulds distance
0
5
10
15
20
25
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA

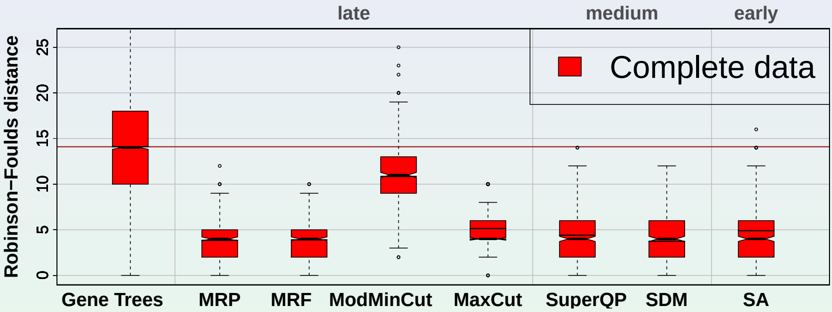

late
medium
early
Robinson−Foulds distance
0
5
10
15
20
25
Complete data
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA
Robinson−Foulds distance
0
5
10
15
20
25
Missing data
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA

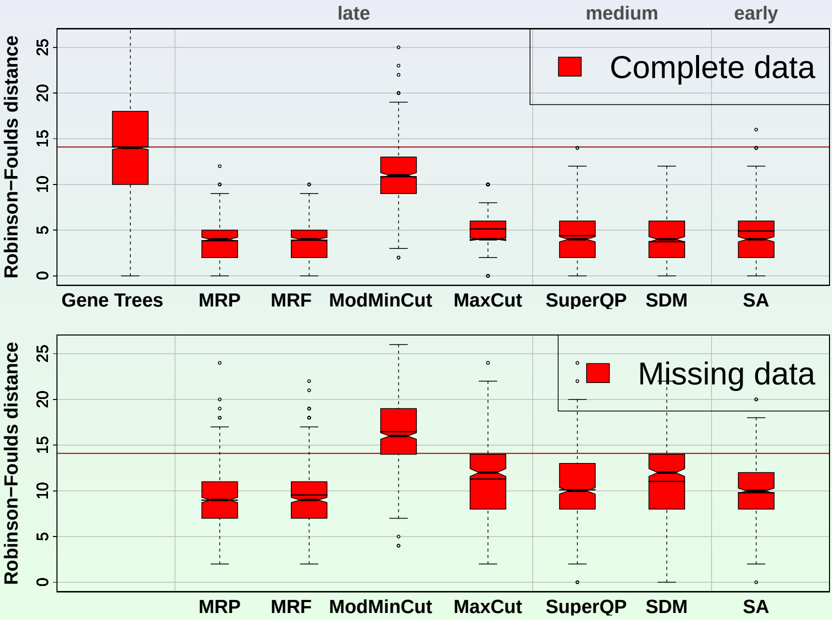
Incomplete lineage sorting

Step 2: For every simulation, a gene tree is generated from the species tree with a coalescent process ($\theta = 0.005$, changes $\approx$ 22 % of the branches)

Step 3: Simulation with the parameters estimated from the superalignment




late
medium
early
🟥 Complete data
Robinson–Foulds distance
0
5
10
15
20
25
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA


late
medium
early
Robinson−Foulds distance
0
5
10
15
20
25
🟥 Complete data
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA
Robinson−Foulds distance
0
5
10
15
20
25
🟥 Missing data
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA

Summary

- Simulation of sequence-based phylogenetic analysis for multiple data sets
- With the assumption of congruent gene trees, superalignment yields the highest accuracy
- In case of high incongruency among (true) gene trees other methods may outperform superalignment
- Matrix representation methods are the best choice for supertree reconstruction

Summary

- Simulation of sequence-based phylogenetic analysis for multiple data sets
- With the assumption of congruent gene trees, superalignment yields the highest accuracy
- In case of high incongruency among (true) gene trees other methods may outperform superalignment
- Matrix representation methods are the best choice for supertree reconstruction

Acknowledgements:

- Gregory Ewing
- WWTF for funding

CIBIV MFPL
Center for Integrative
Bioinformatics Vienna