Multi-Locus Phylogenies Reconstructed from Incomplete Gene-Sets

Heiko A. Schmidt

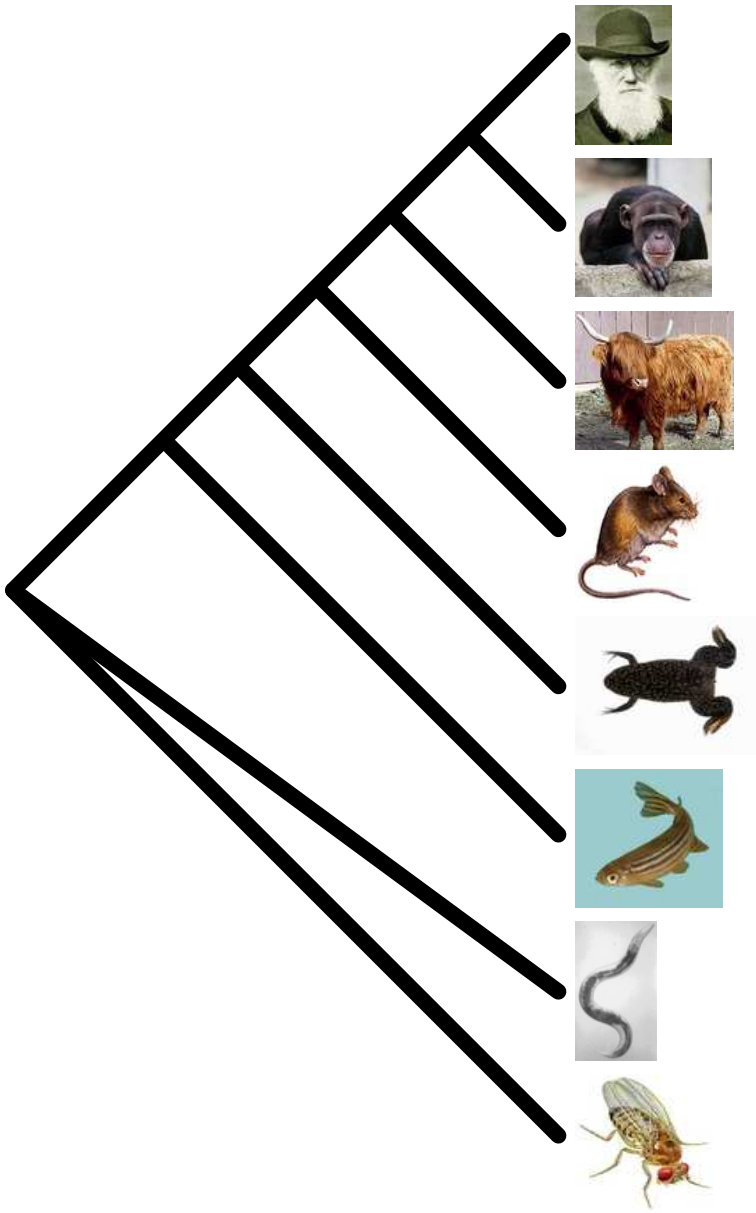
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories
Vienna, Austria
heiko.schmidt@univie.ac.at

CIBIV MFPL

universität wien

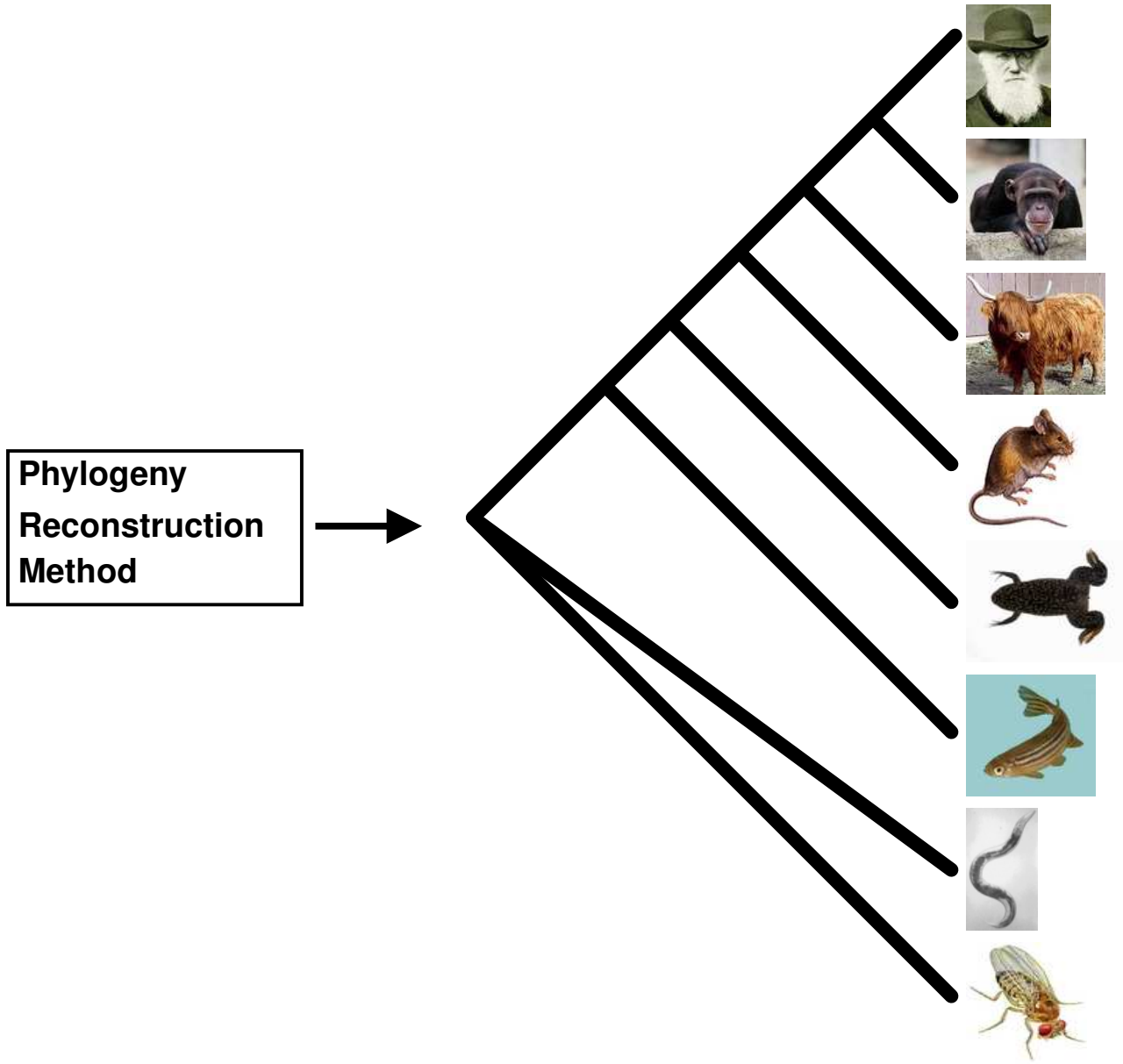
MEDICAL UNIVERSITY OF VIENNA

Phylogeny Reconstruction Scheme

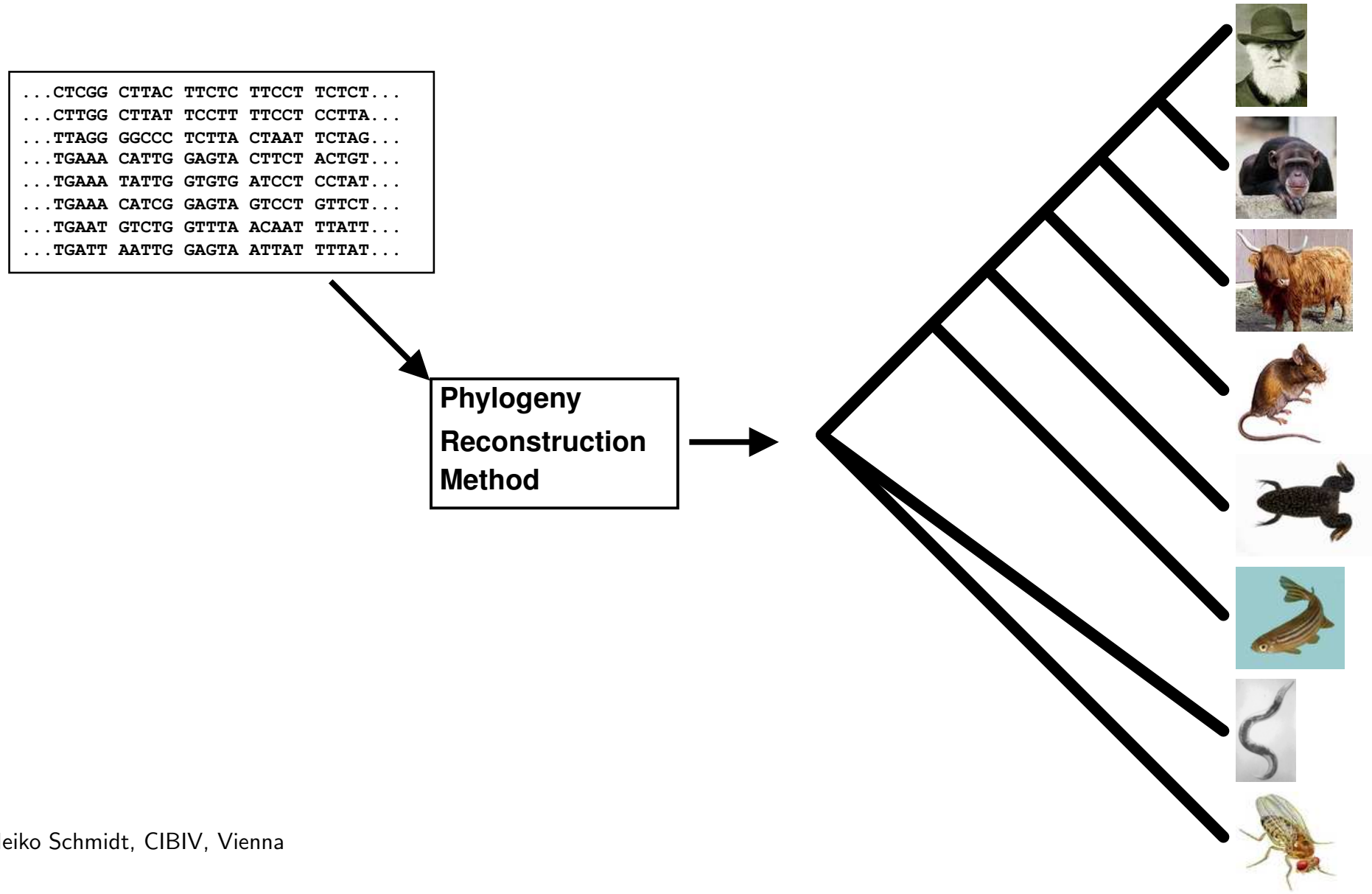


Heiko Schmidt, CIBIV, Vienna

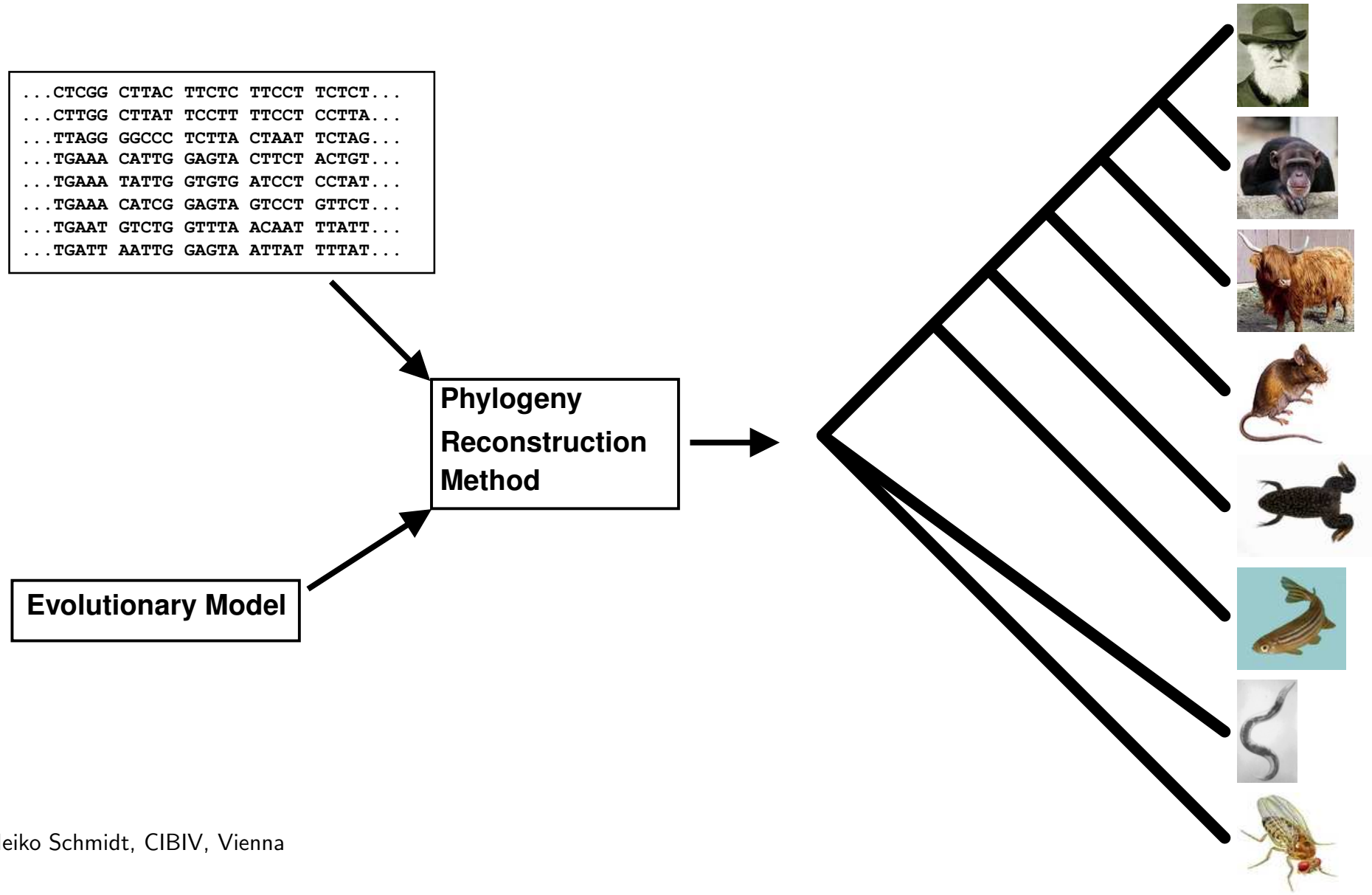
Phylogeny Reconstruction Scheme



Phylogeny Reconstruction Scheme

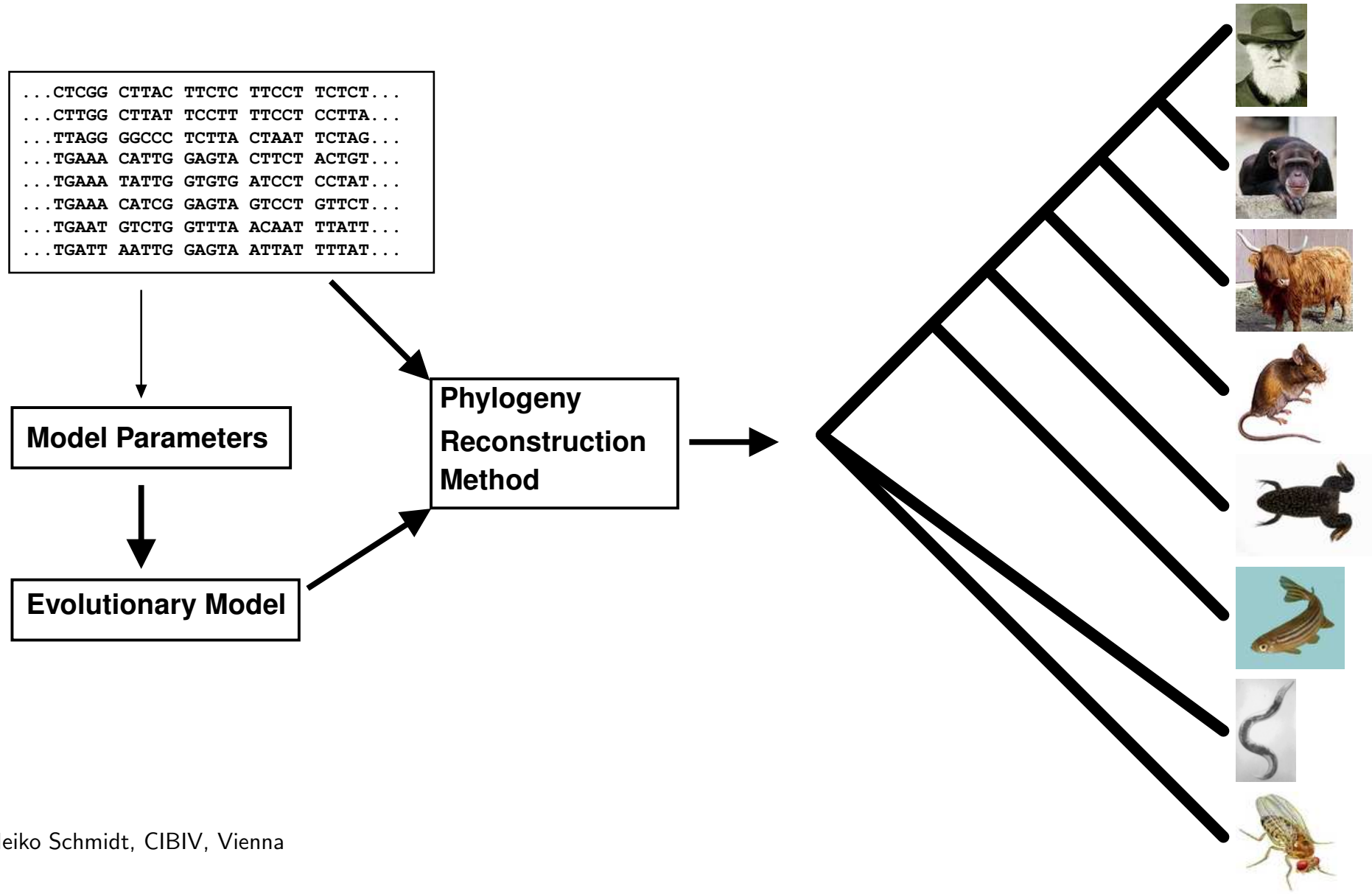


Phylogeny Reconstruction Scheme



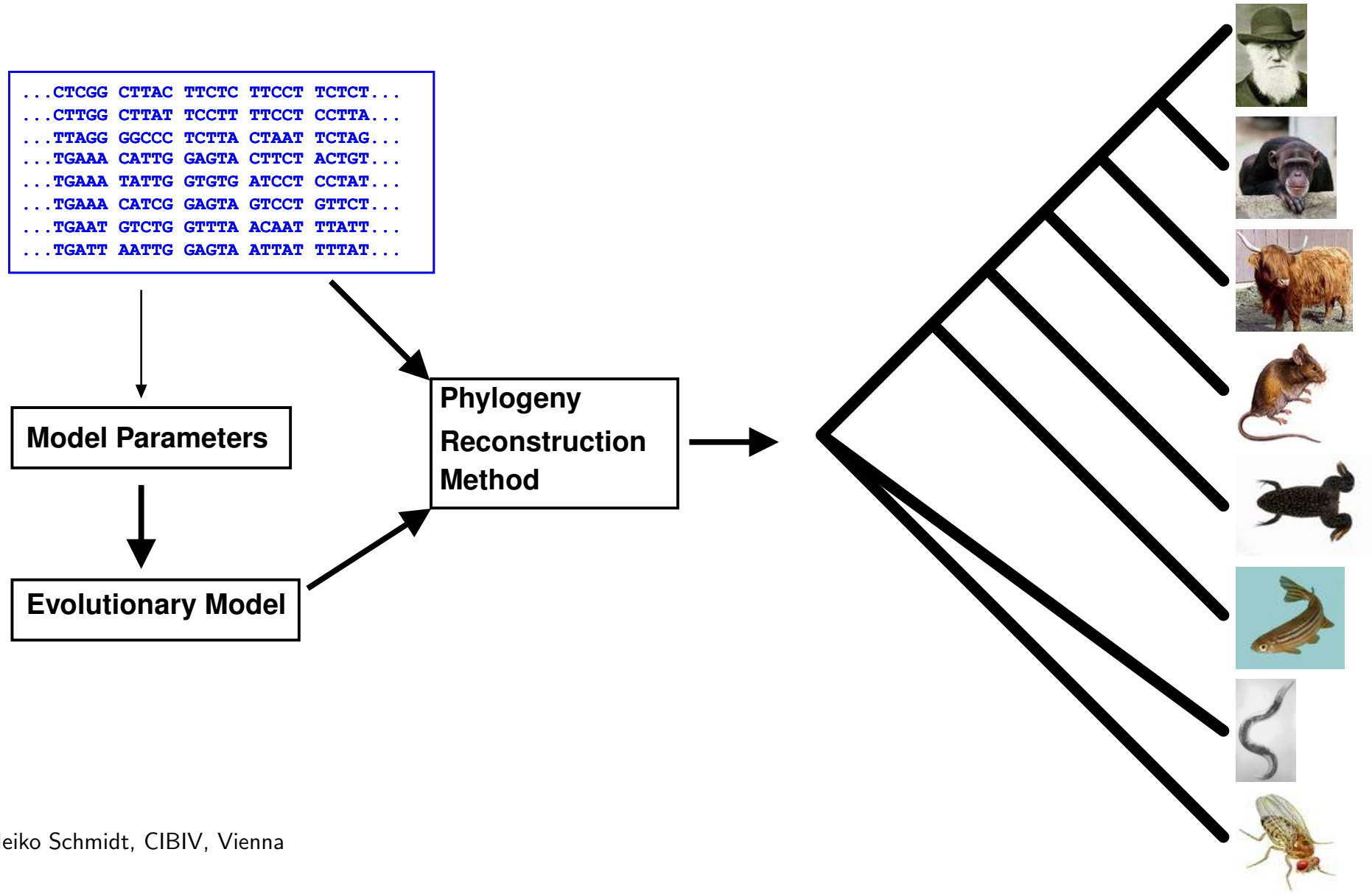
Heiko Schmidt, CIBIV, Vienna

Phylogeny Reconstruction Scheme



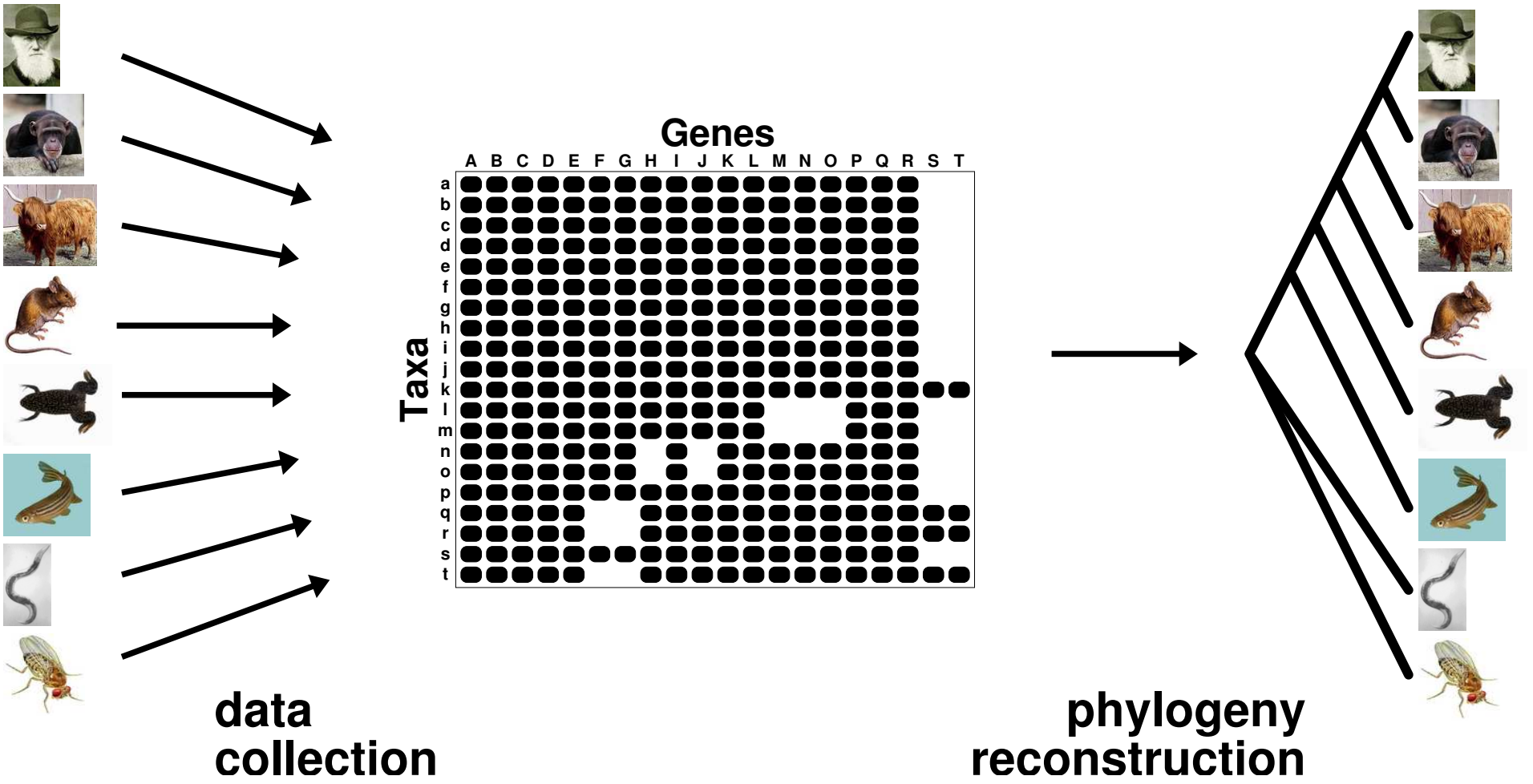
Heiko Schmidt, CIBIV, Vienna

Phylogeny Reconstruction Scheme



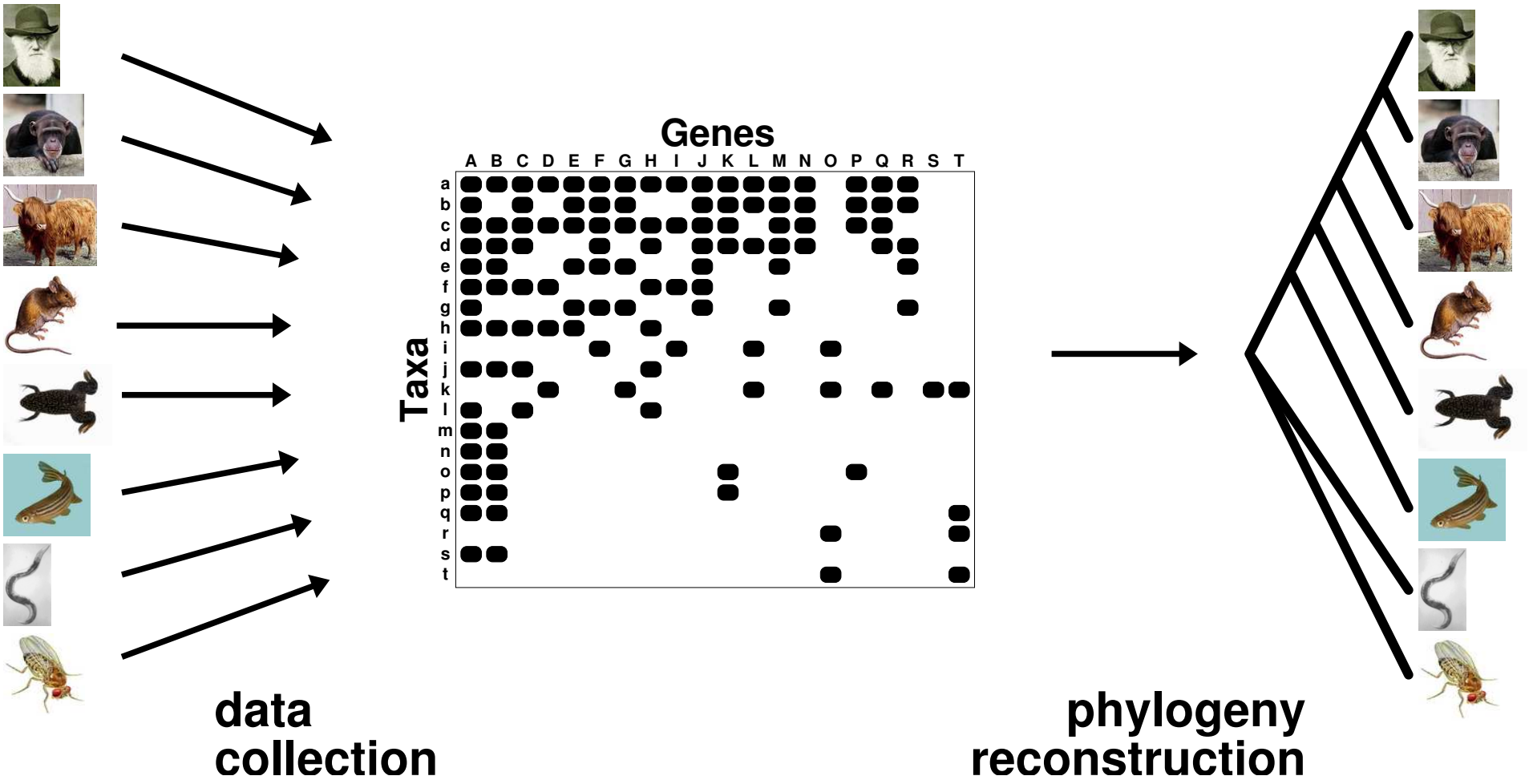
Heiko Schmidt, CIBIV, Vienna

Multi-Locus Datasets



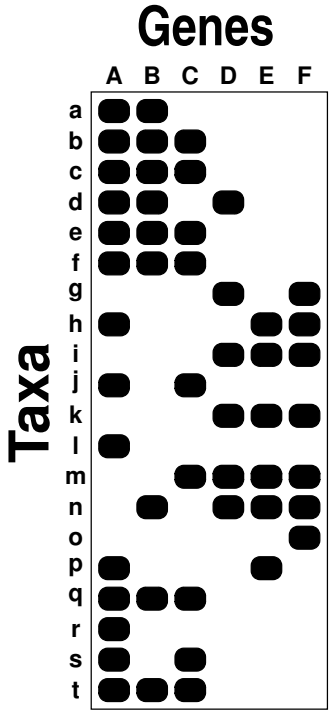
Heiko Schmidt, CIBIV, Vienna

Multi-Locus Datasets



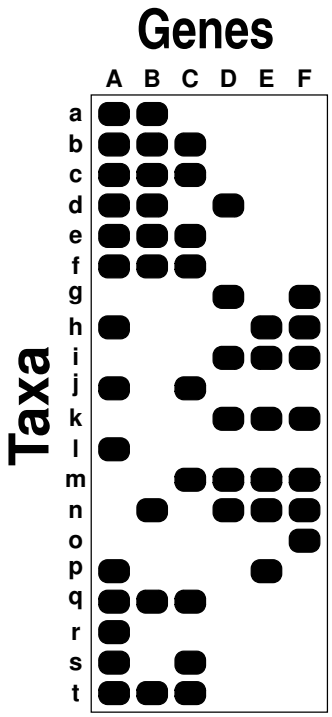
Heiko Schmidt, CIBIV, Vienna

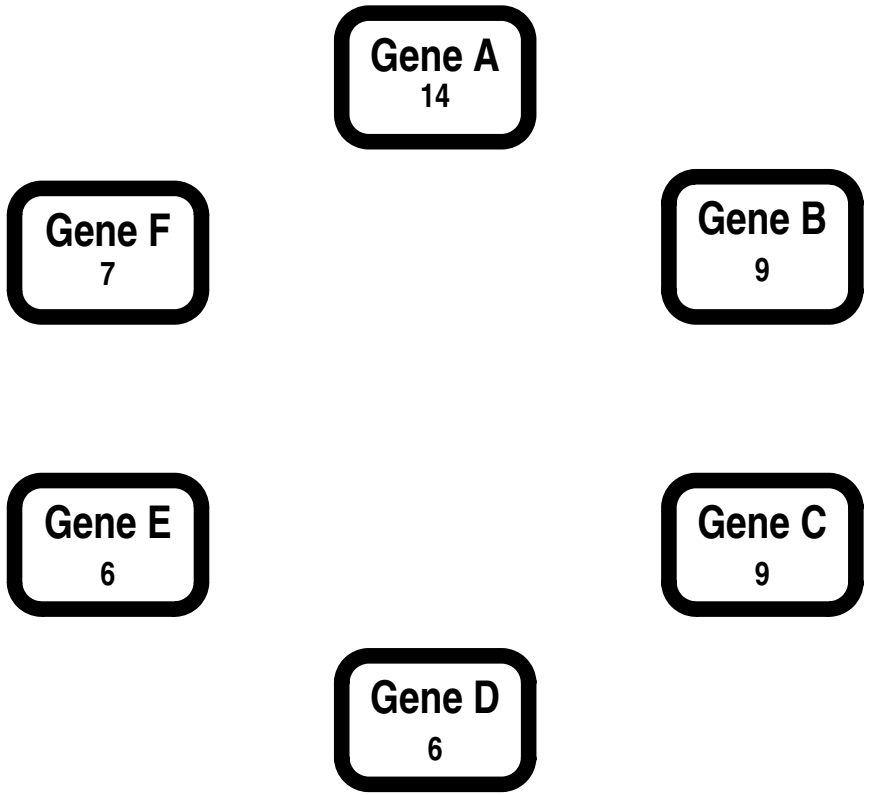
Assessing Combinability of Genesets: The Overlap Graph



Heiko Schmidt, CIBIV, Vienna

Assessing Combinability of Genesets: The Overlap Graph

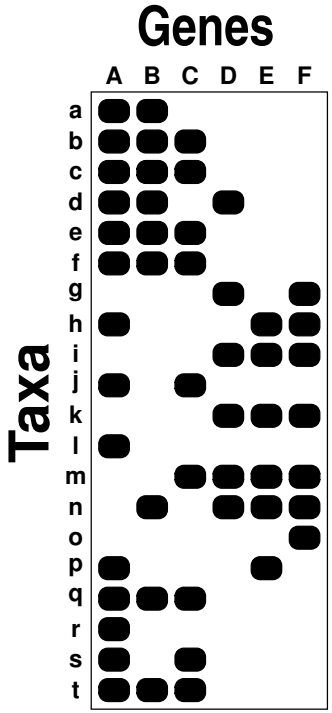


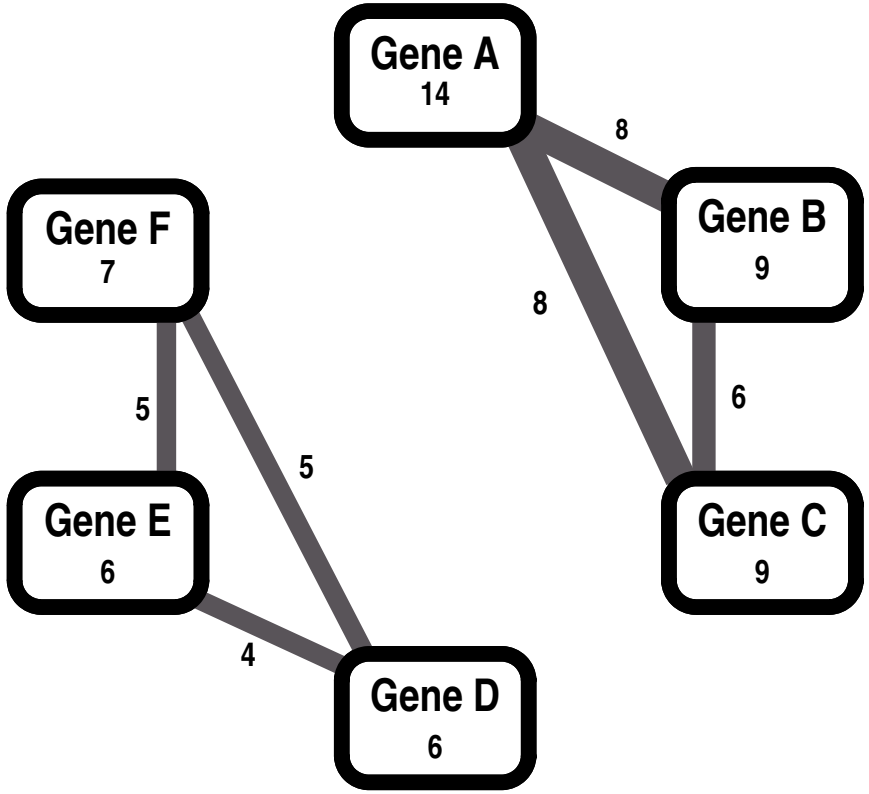


Overlap Graph

nodes = genes

Assessing Combinability of Genesets: The Overlap Graph

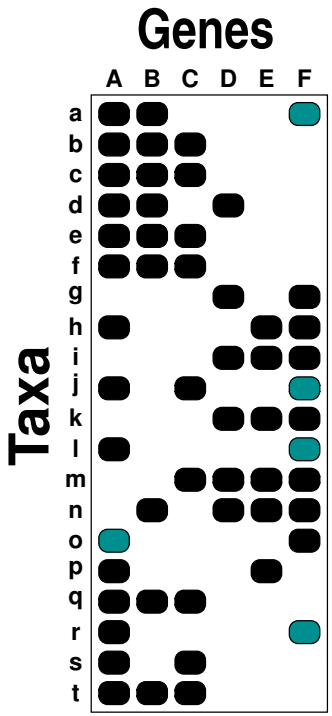


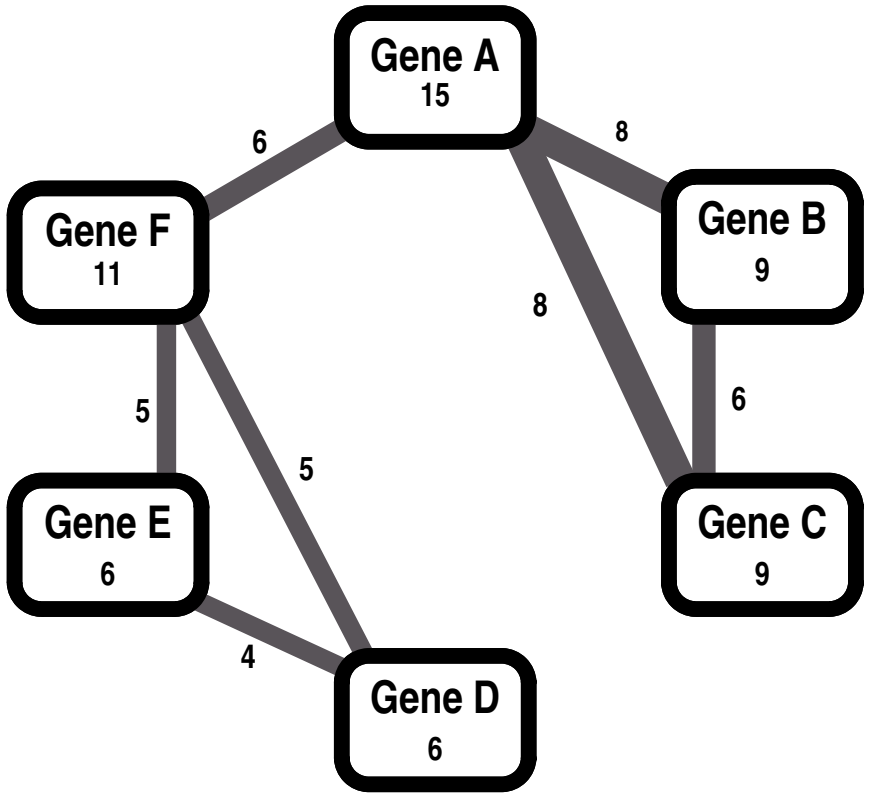


Overlap Graph

nodes = genes
edges = (sufficient) overlap (≥ 3)
edge weights = (pairwise) geneset overlap

Assessing Combinability of Genesets: The Overlap Graph

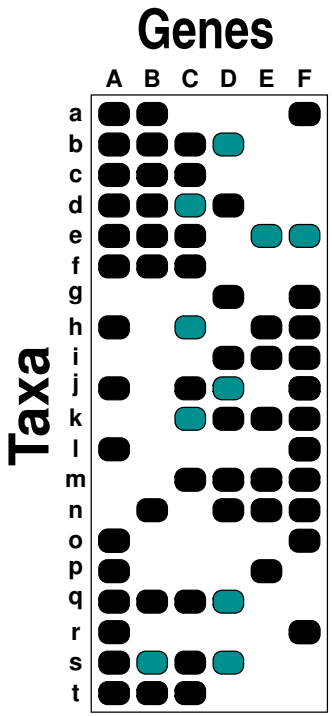


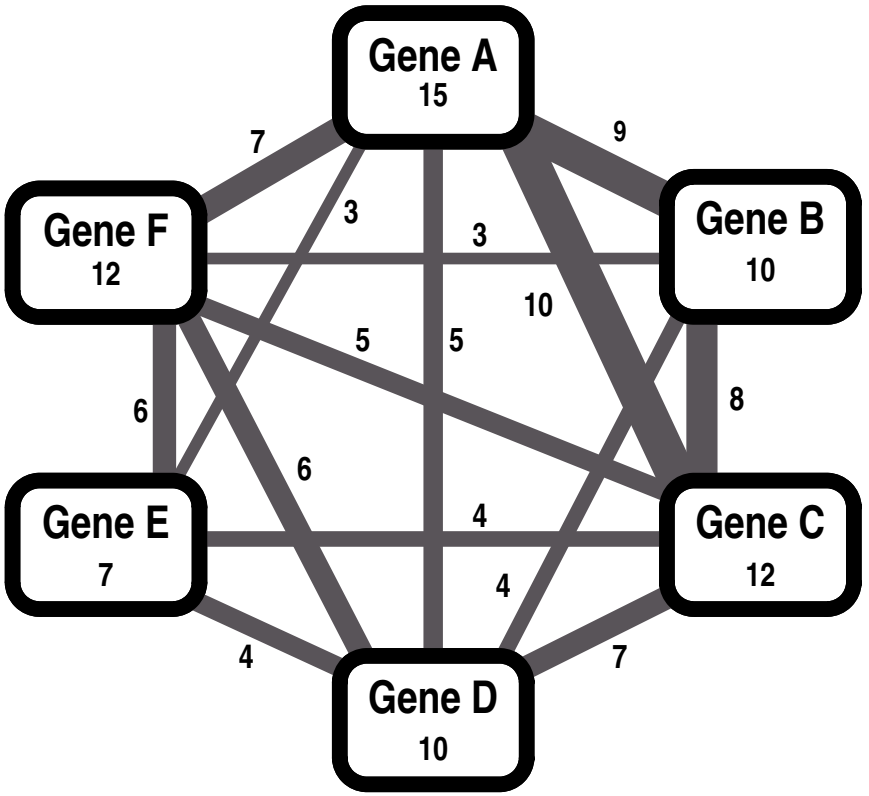


Overlap Graph

nodes = genes
edges = (sufficient) overlap (≥ 3)
edge weights = (pairwise) geneset overlap

Assessing Combinability of Genesets: The Overlap Graph





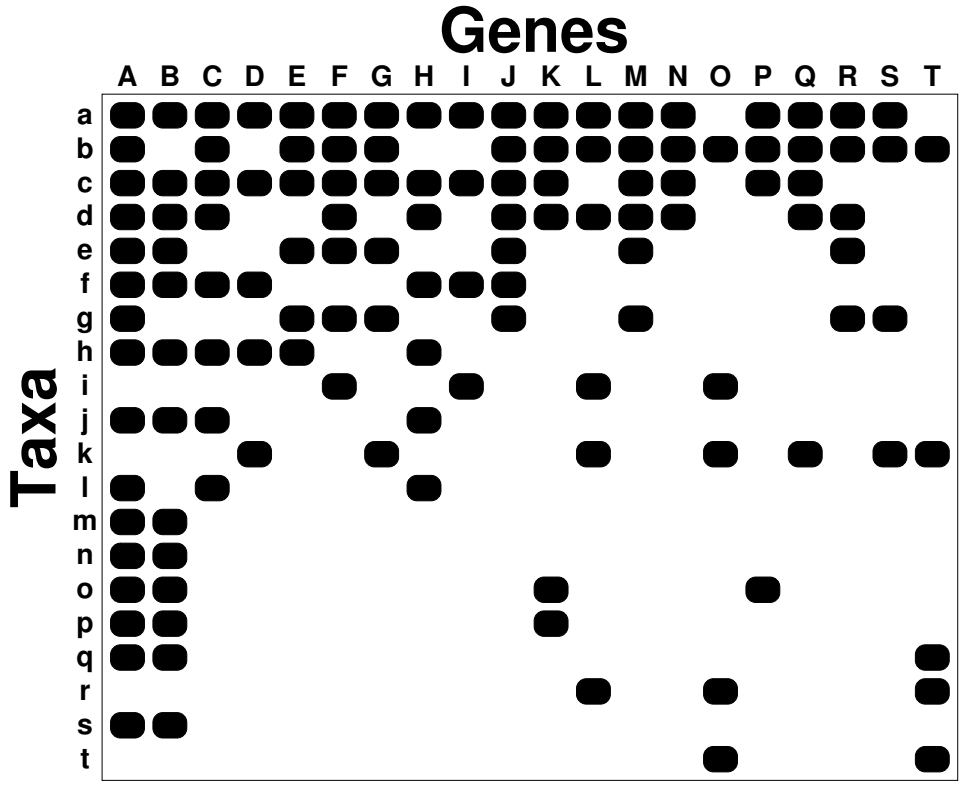
Overlap Graph

nodes = genes
edges = (sufficient) overlap (≥ 3)
edge weights = (pairwise) geneset overlap

Quality Assessments

- Overlap Graph based check for connectivity (connected components)
- Overlap Graph based check for fragility (minimum cut)
- Connectivity within defined groups

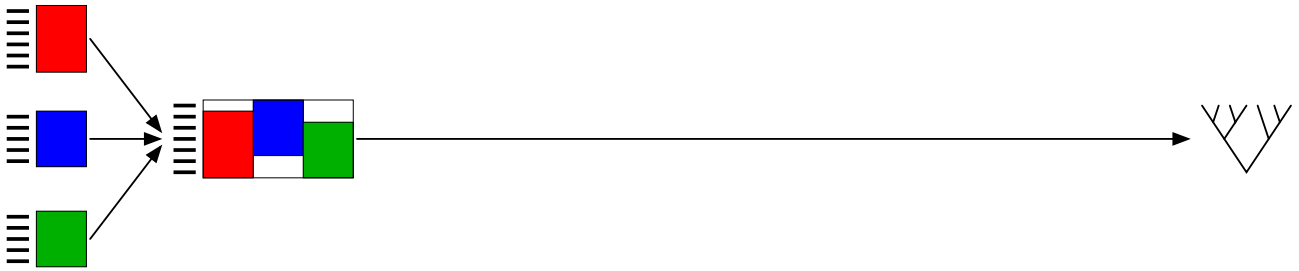
Patchiness of the Data Matrix



How to combine genesets with missing data?

Early-Level Combination: Supermatrix/'Total Evidence'

Combination by concatenating datasets:

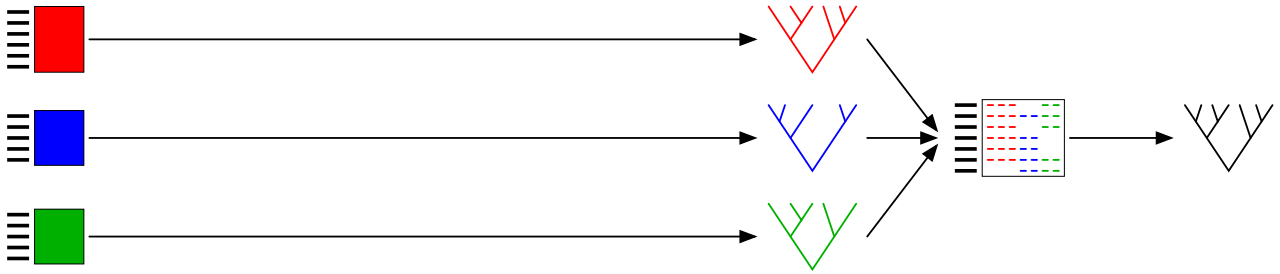


Any method possible for tree reconstruction:

- distance methods
- maximum parsimony
- likelihood methods

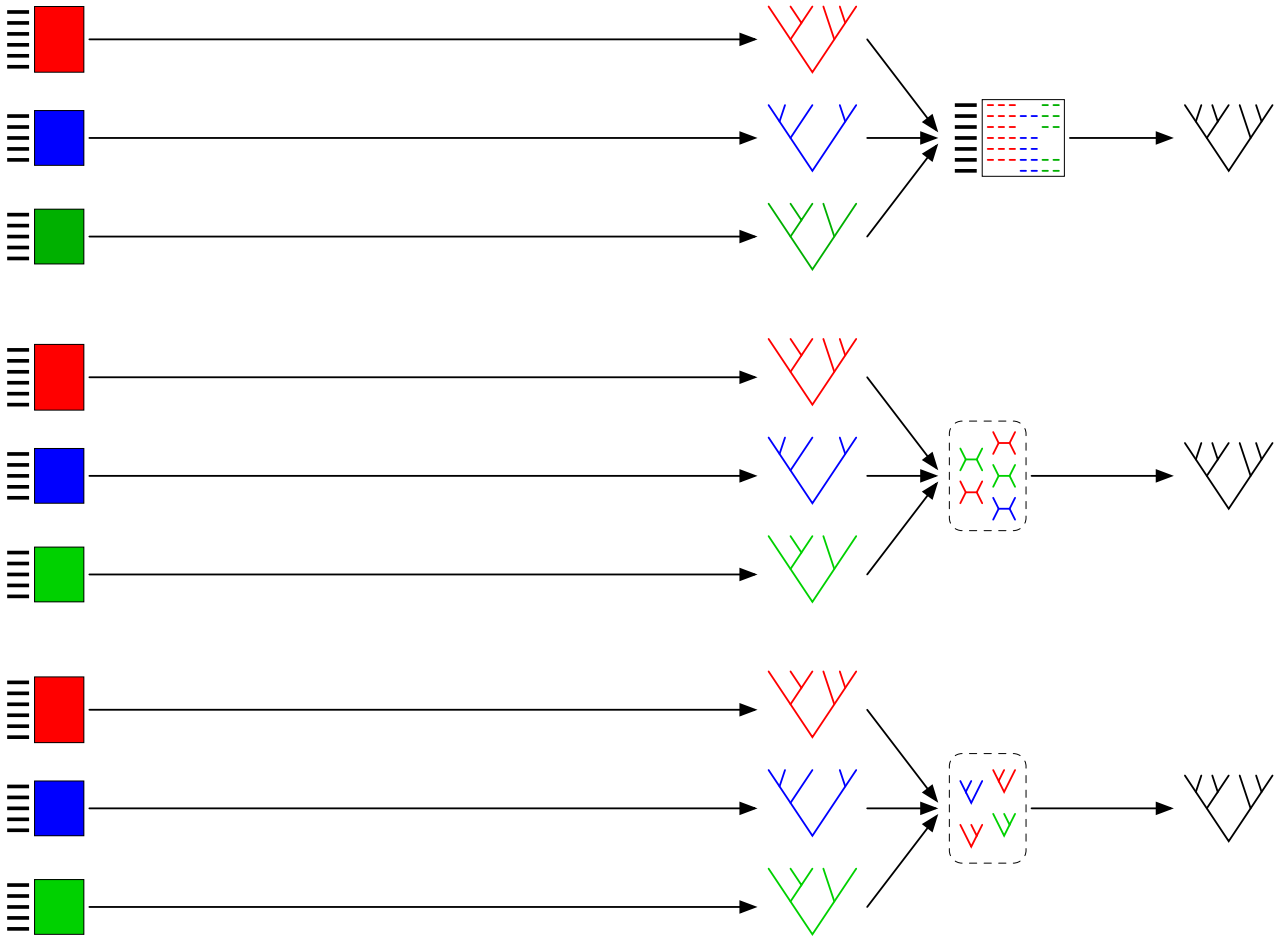
Late-Level Combination: Supertrees

Construct separate trees for each geneset and combine them to supertrees:



Late-Level Combination: Supertrees

Construct separate trees for each geneset and combine them to supertrees:



The 'Total Evidence' vs. Supertree/Consensus Debate

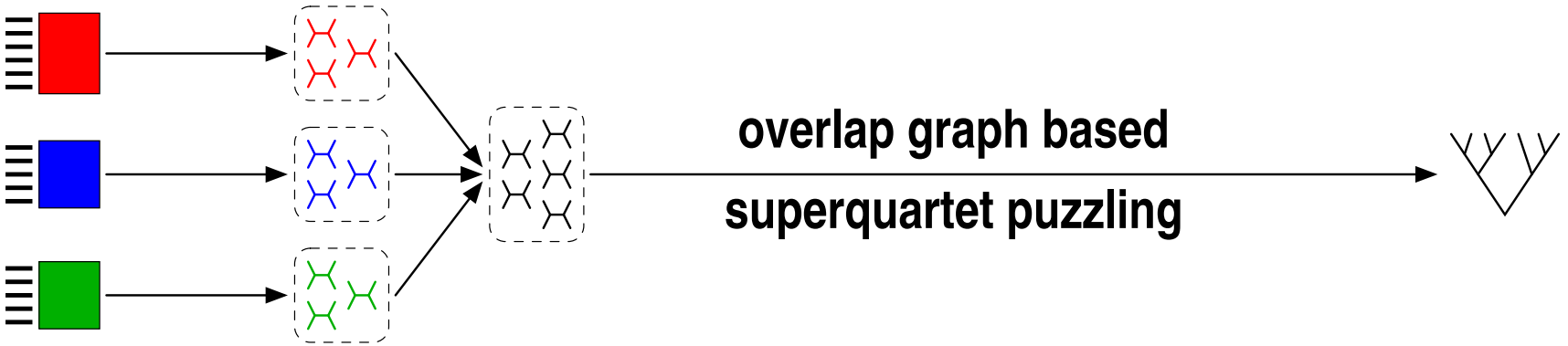
Total Evidence/Supermatrix Methods:

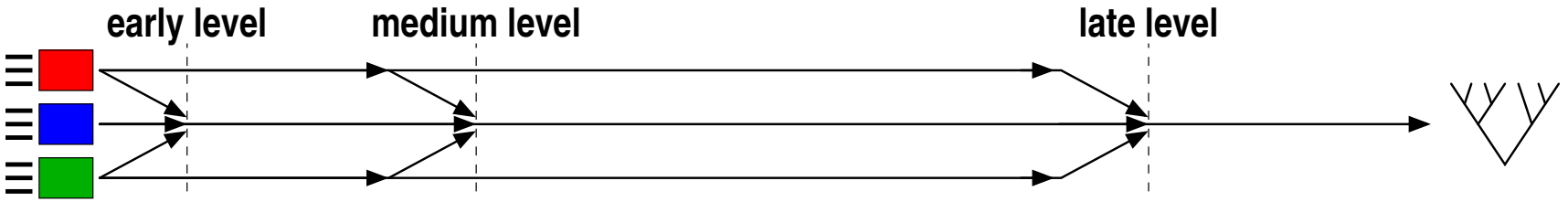
- ⊕ used all sequence information available for tree reconstruction.
- ⊖ Different genomic regions can have significantly different evolutionary parameters (e.g., histones and immunoglobulines)

Supertree/Consensus Methods:

- ⊖ All sequence information lost before combining the genetrees, sequences play no role in the final tree reconstruction.
- ⊕ Evolutionary models can be adjusted for each geneset separately.
- ⊕ (Can be used if only trees are available and for different types of data.)

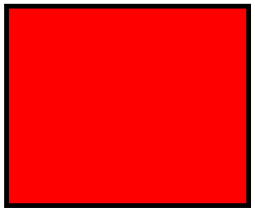
'Medium-Level' Combination – Superquartet Puzzling





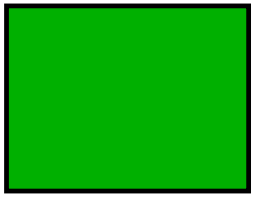
Superquartet Puzzling (I) - Data Combination

Alignment A



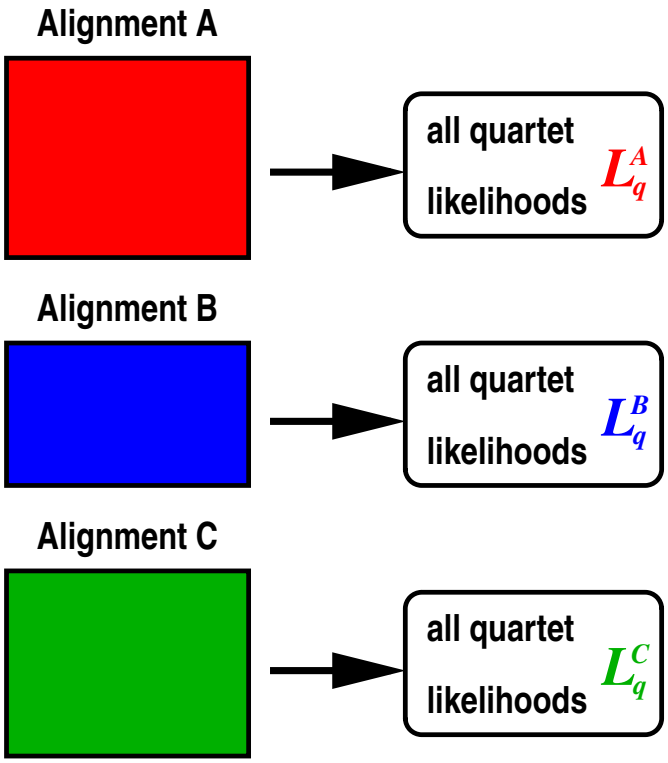
Alignment B

Alignment C

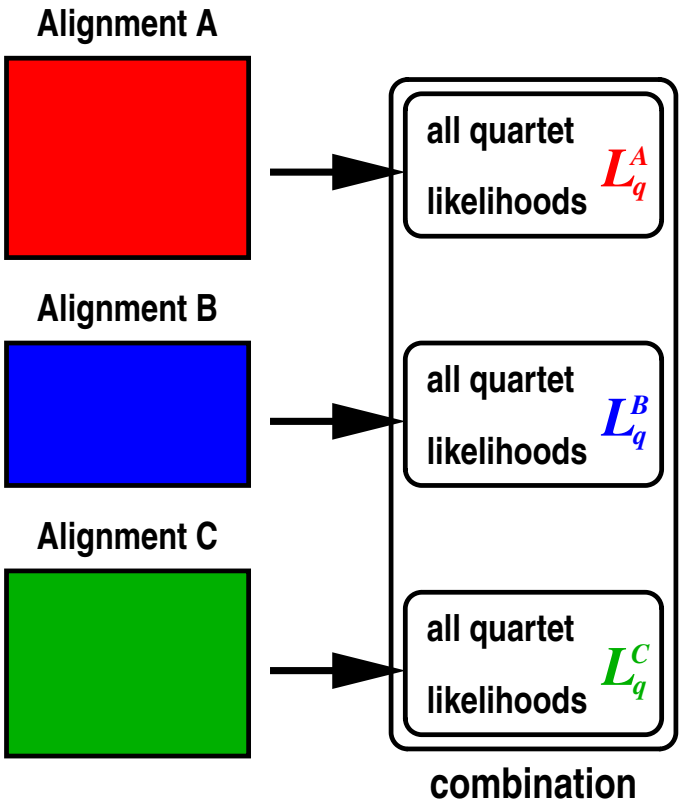


Heiko Schmidt, CIBIV, Vienna

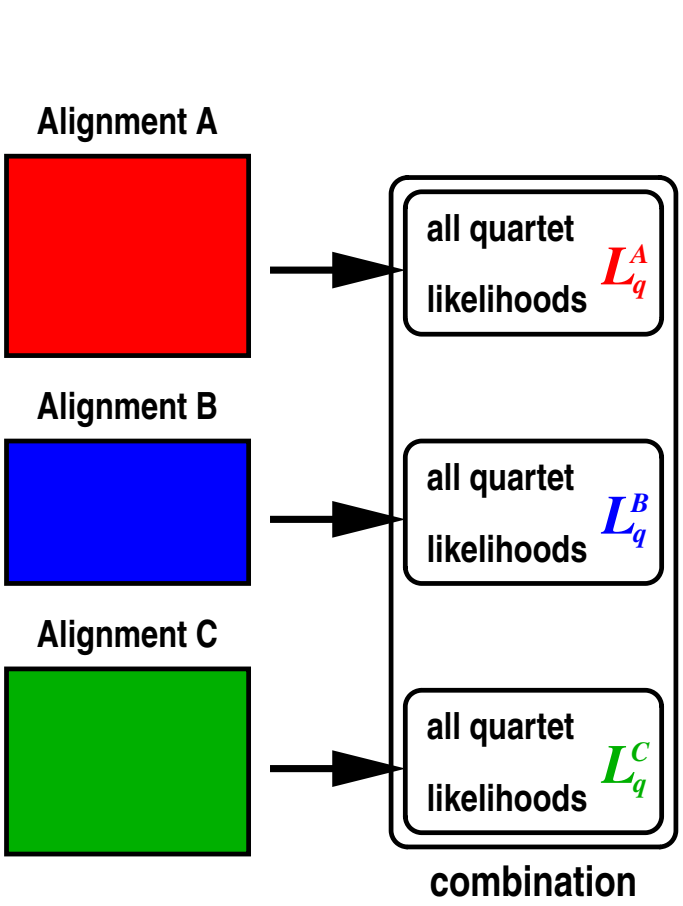
Superquartet Puzzling (I) - Data Combination

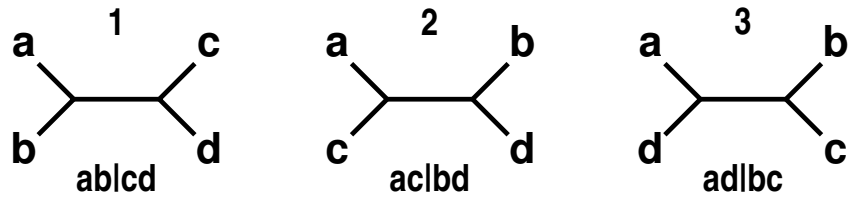


Superquartet Puzzling (I) - Data Combination

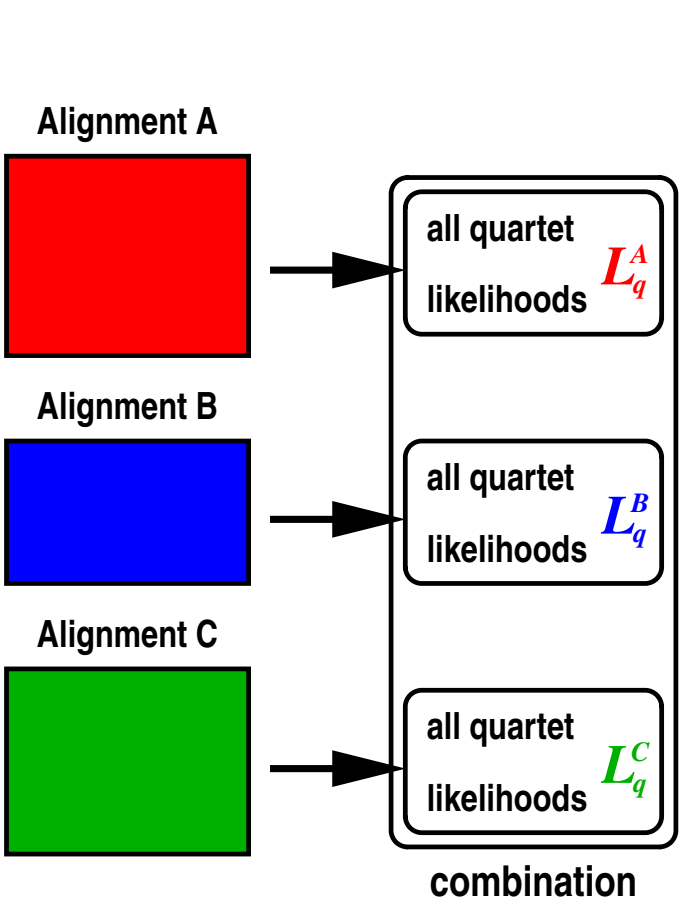


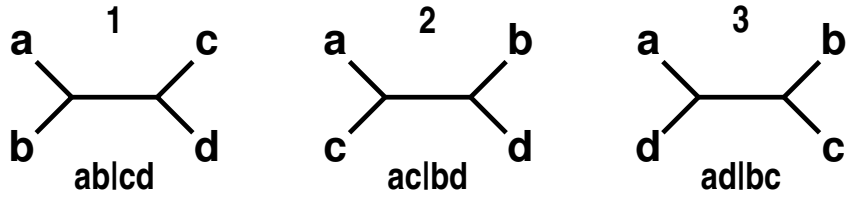
Superquartet Puzzling (I) - Data Combination





Superquartet Puzzling (I) - Data Combination



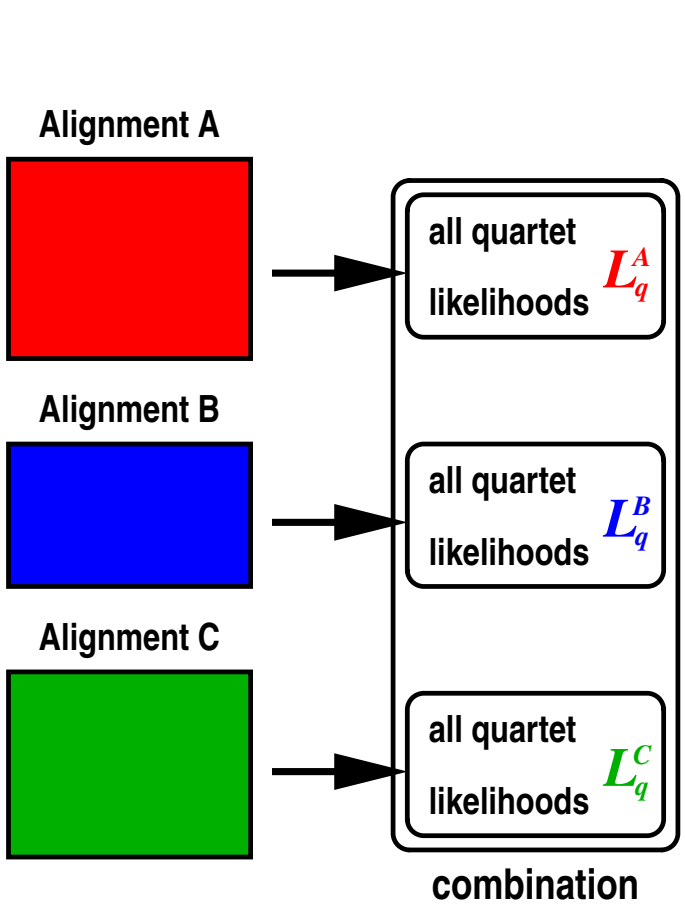


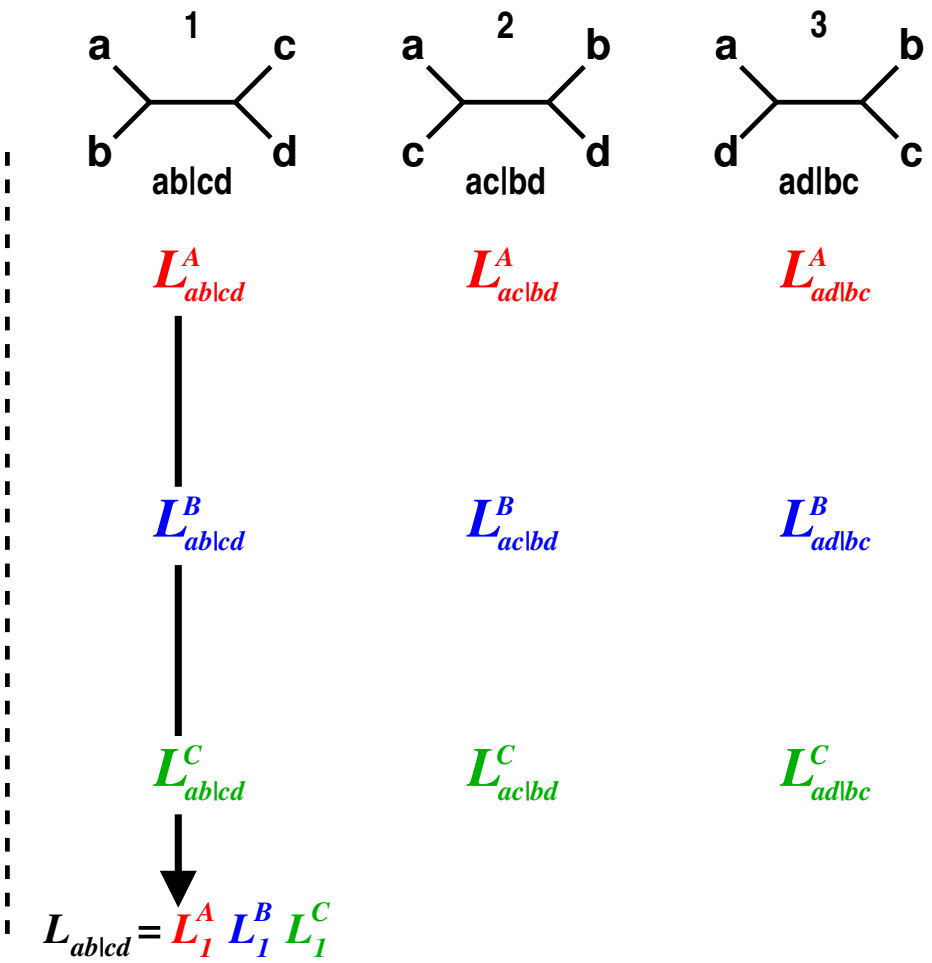
$L^A_{ab|cd}$ $L^A_{ac|bd}$ $L^A_{ad|bc}$

$L^B_{ab|cd}$ $L^B_{ac|bd}$ $L^B_{ad|bc}$

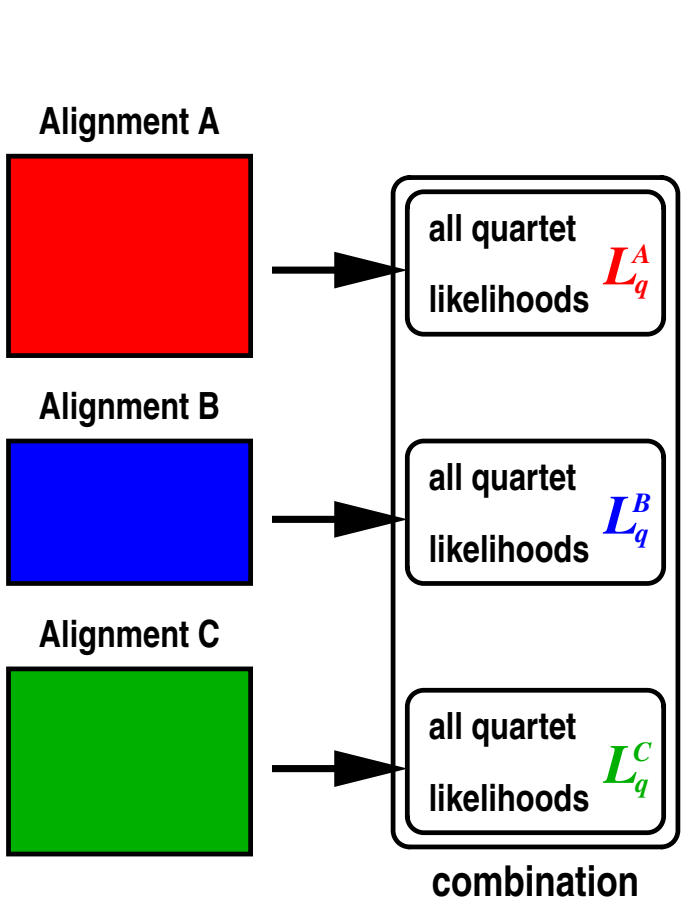
$L^C_{ab|cd}$ $L^C_{ac|bd}$ $L^C_{ad|bc}$

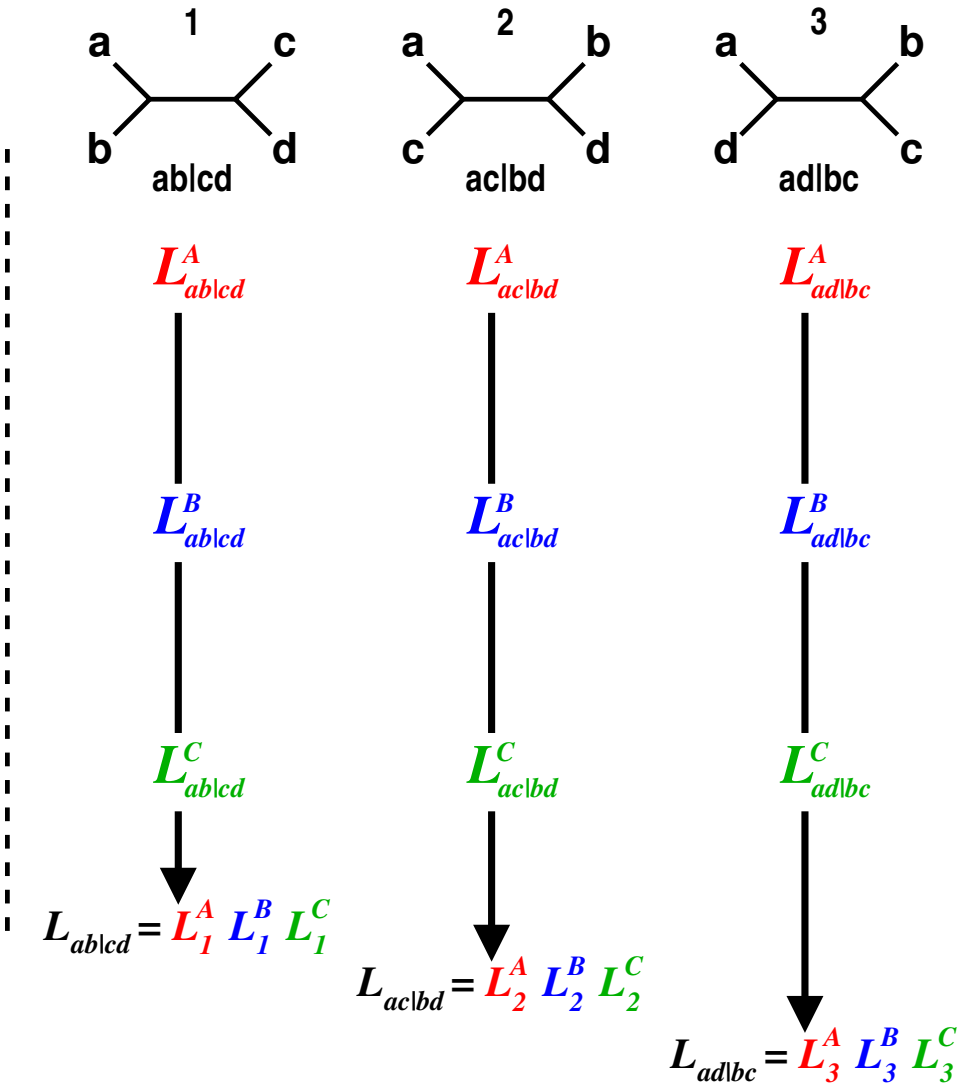
Superquartet Puzzling (I) - Data Combination





Superquartet Puzzling (I) - Data Combination





Superquartet Puzzling (II) – 'Medium-Level' Combination

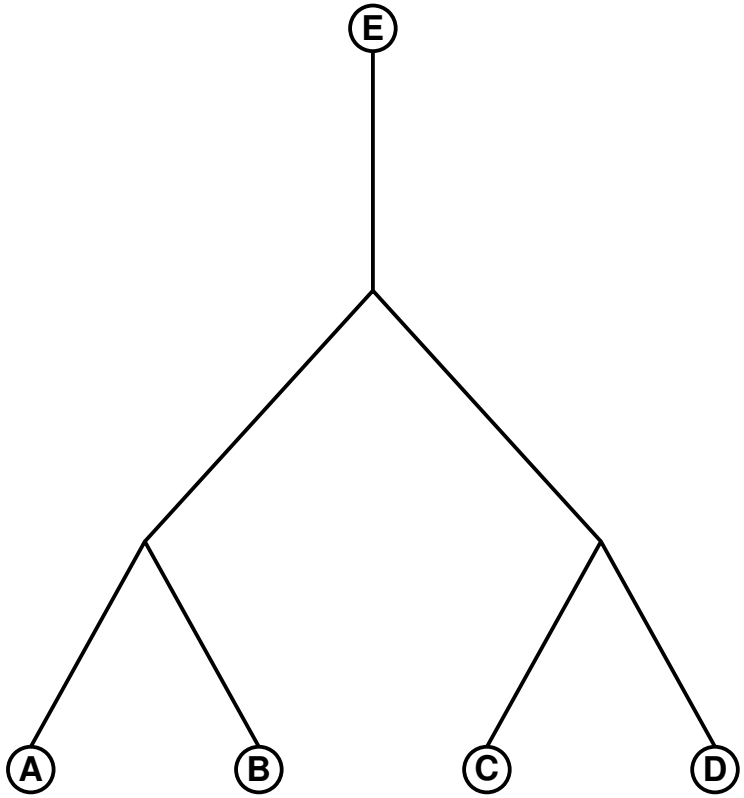
- From the posterior quartet likelihoods

$$p_{ab|cd} = \frac{L_{ab|cd}}{L_{ac|bd} + L_{ab|cd} + L_{ad|bc}}$$

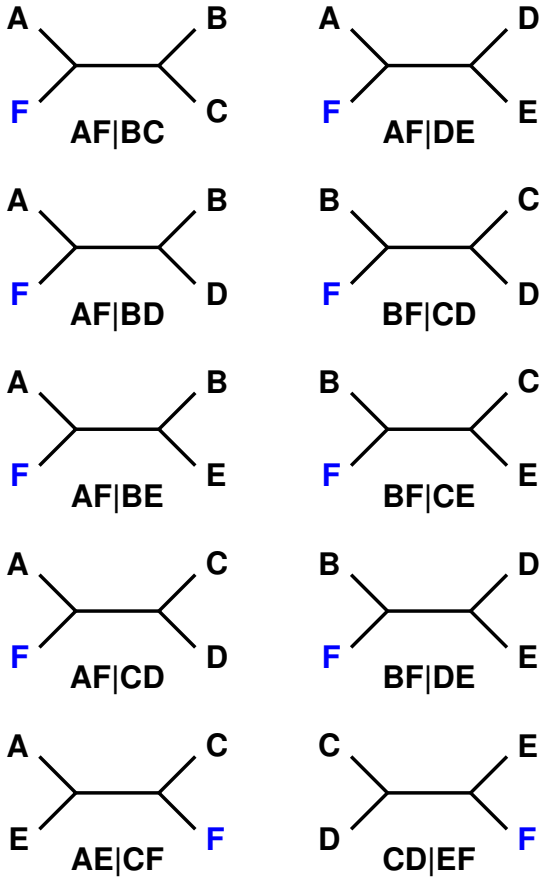
  a set of 'superquartets' is constructed.

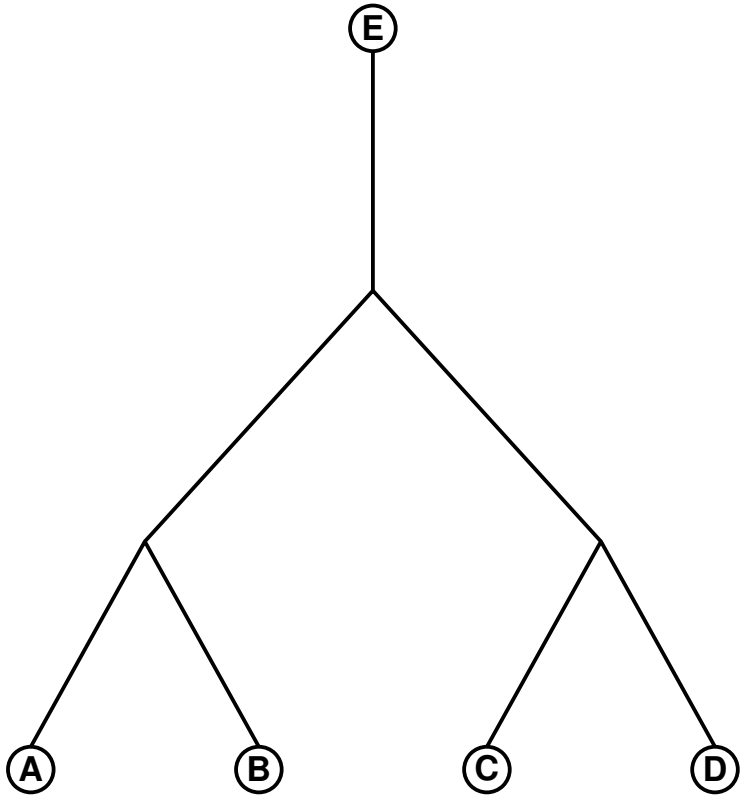
- The superquartets are assembled into a tree applying an algorithm similar to *Quartet Puzzling*, but guided by the overlap graph.

Excursion: Quartet Puzzling

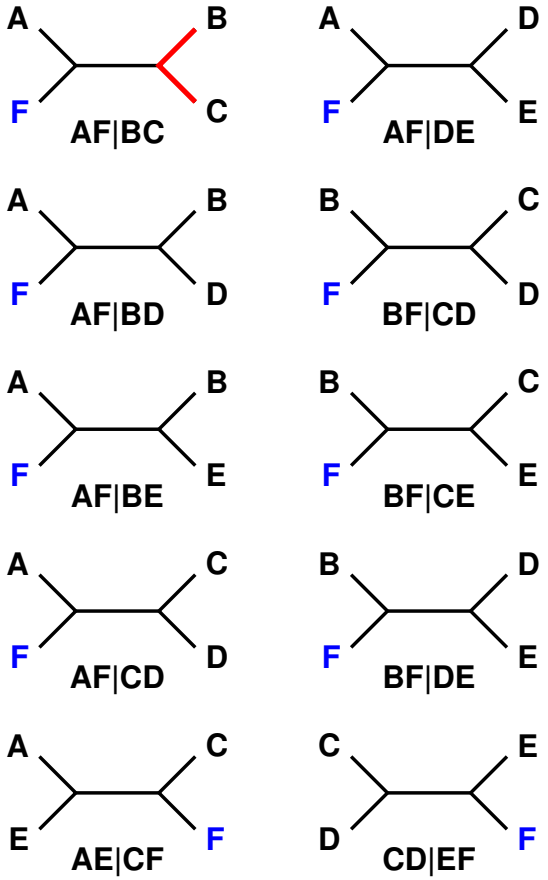


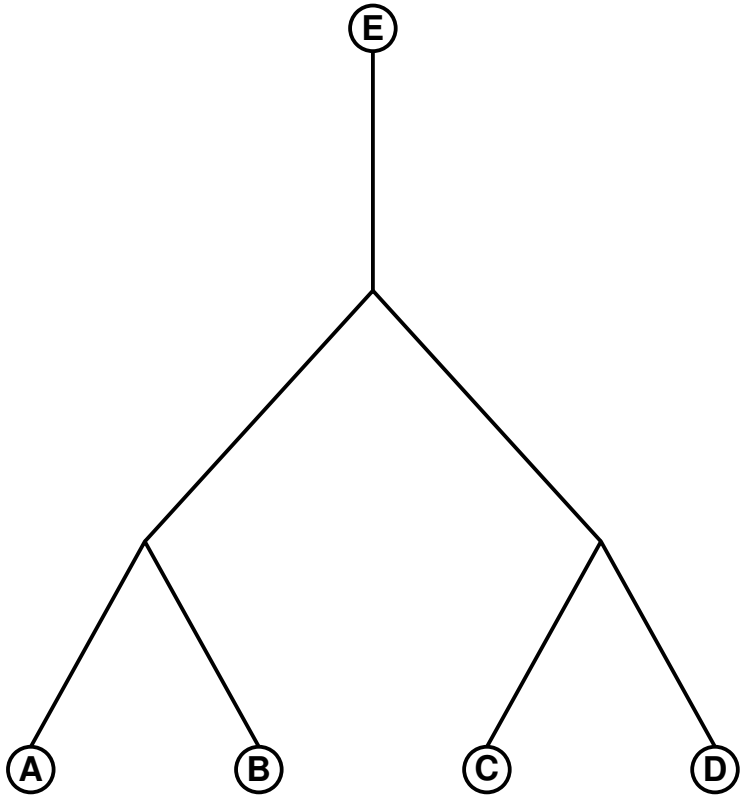
Excursion: Quartet Puzzling



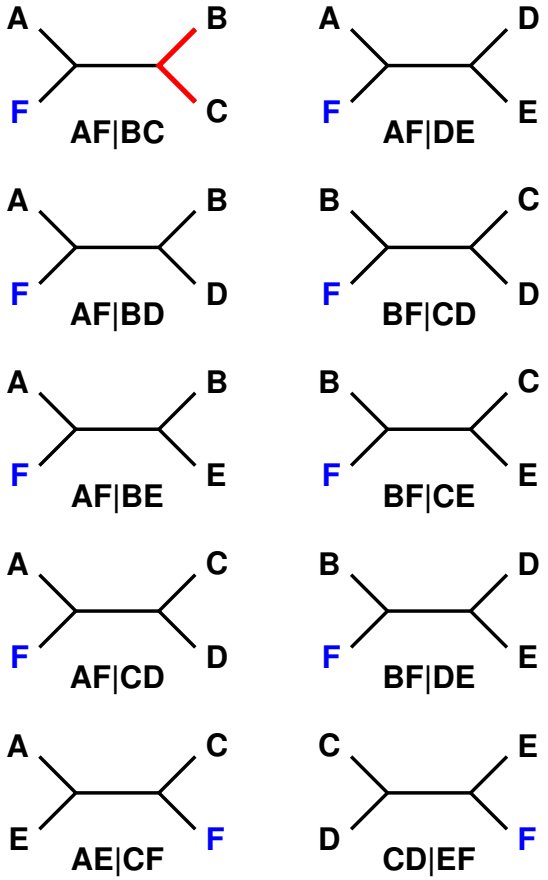


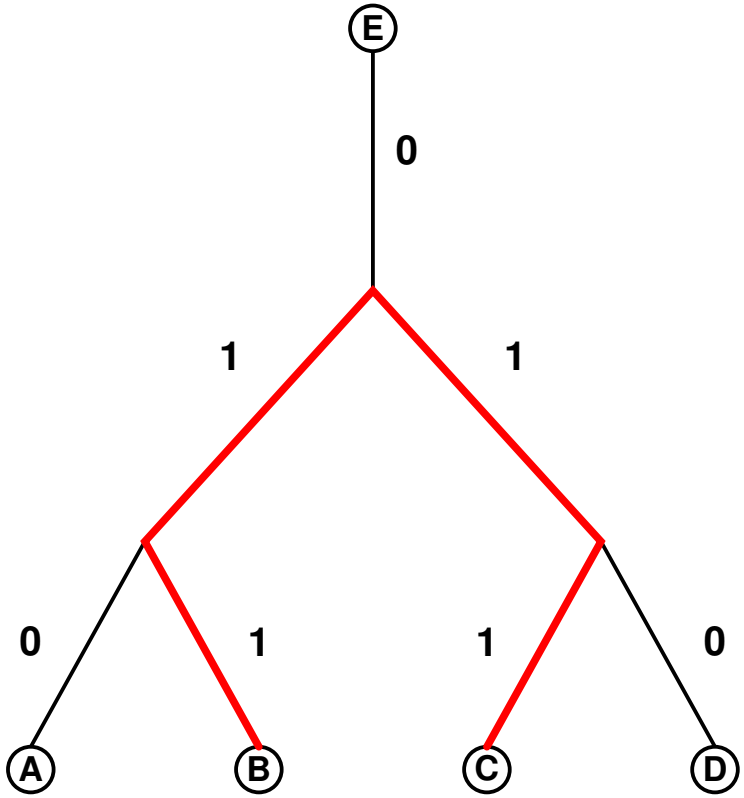
Excursion: Quartet Puzzling



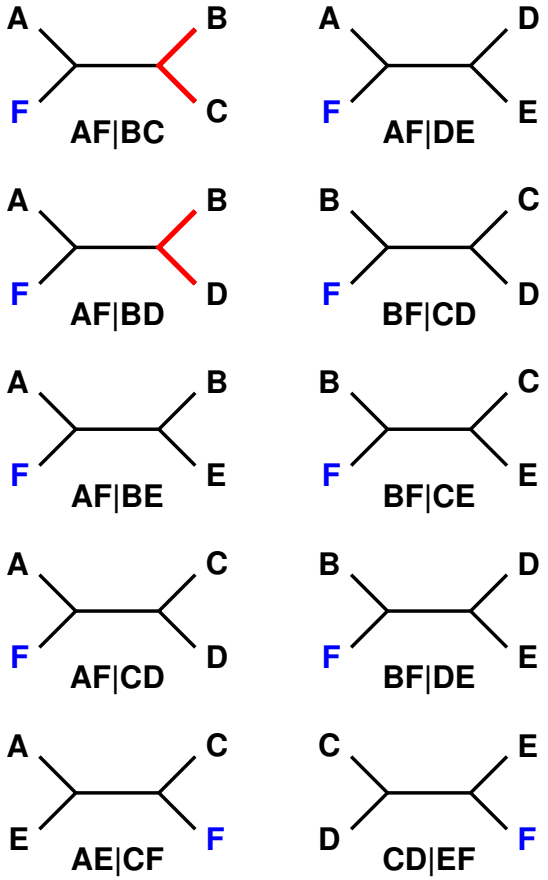


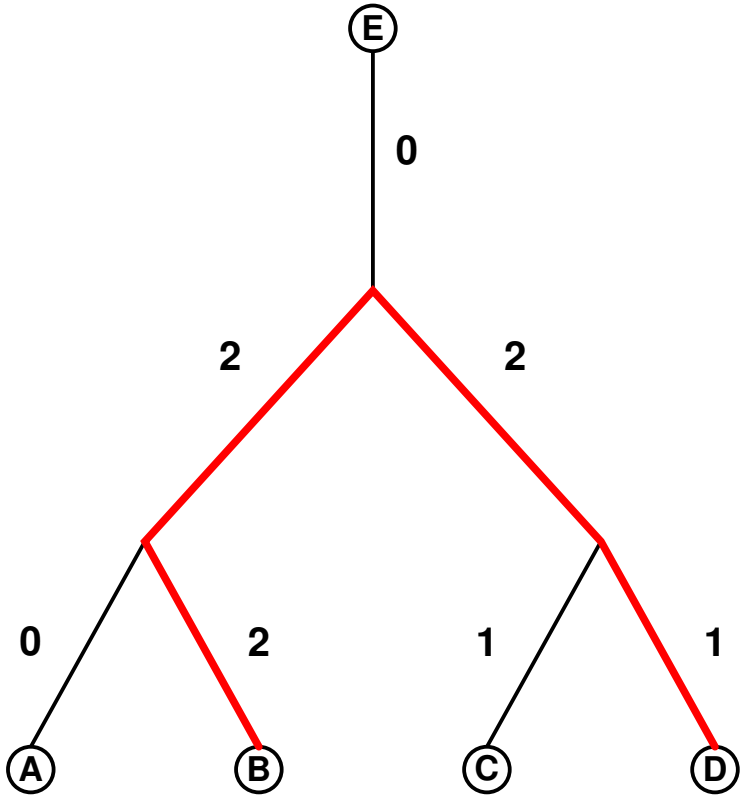
Excursion: Quartet Puzzling



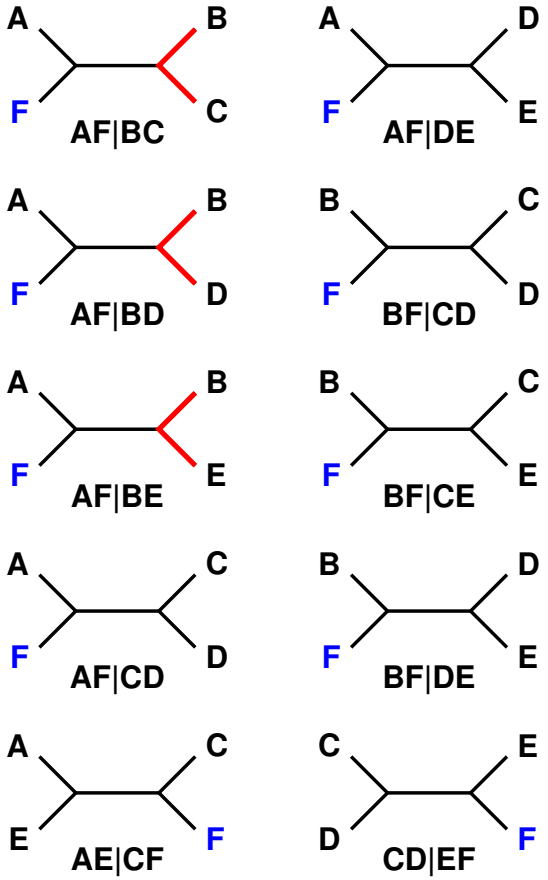


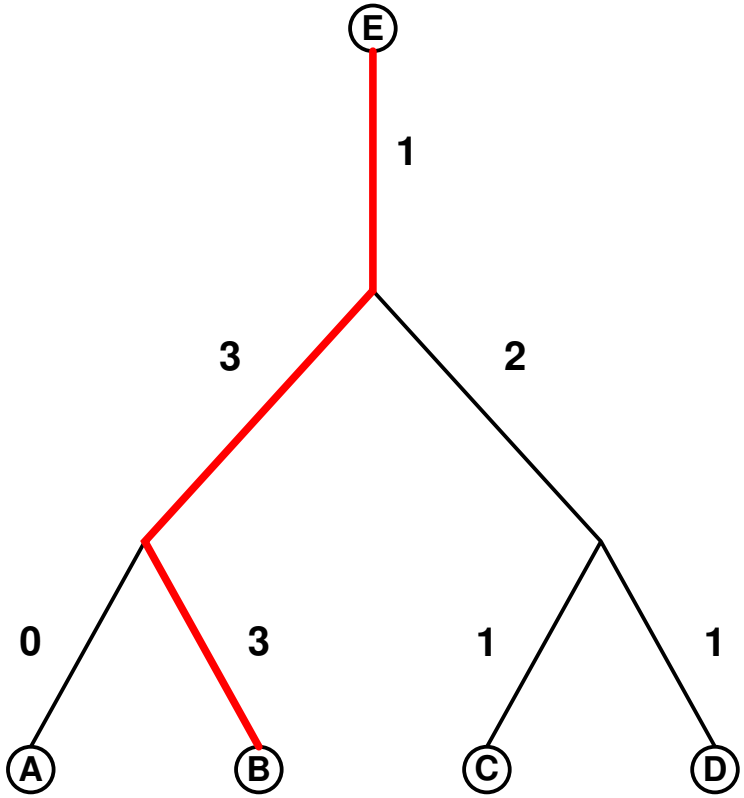
Excursion: Quartet Puzzling



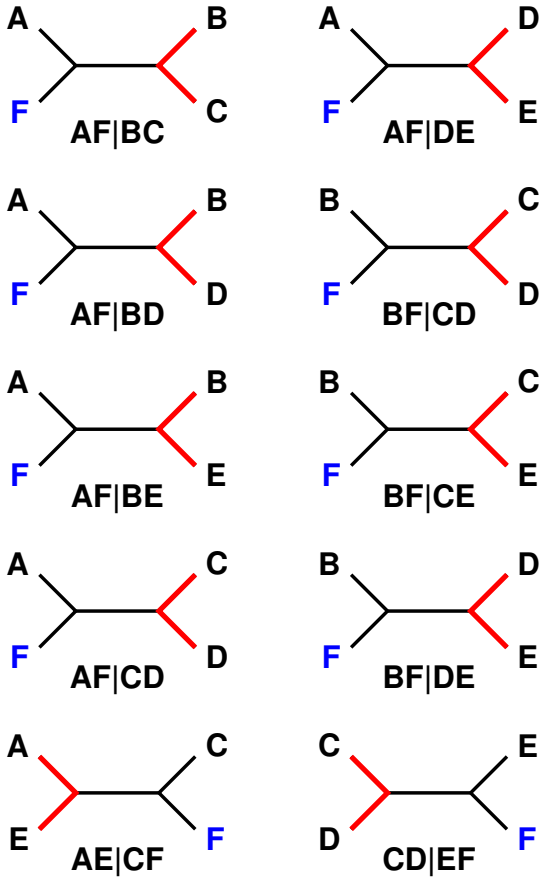


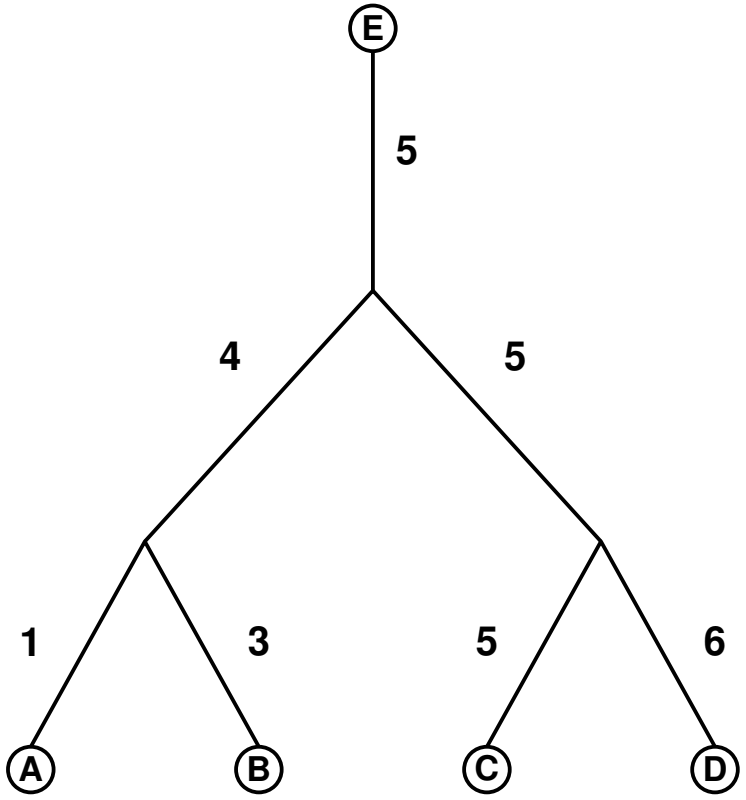
Excursion: Quartet Puzzling



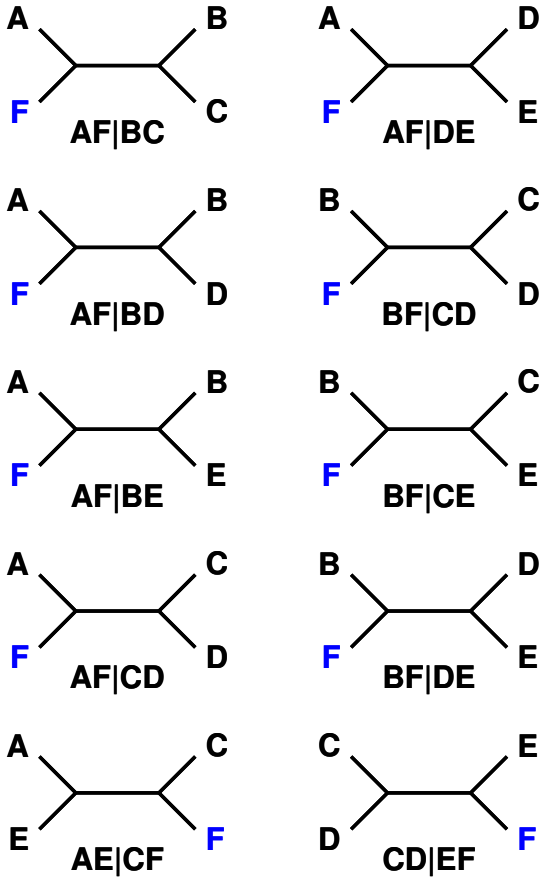


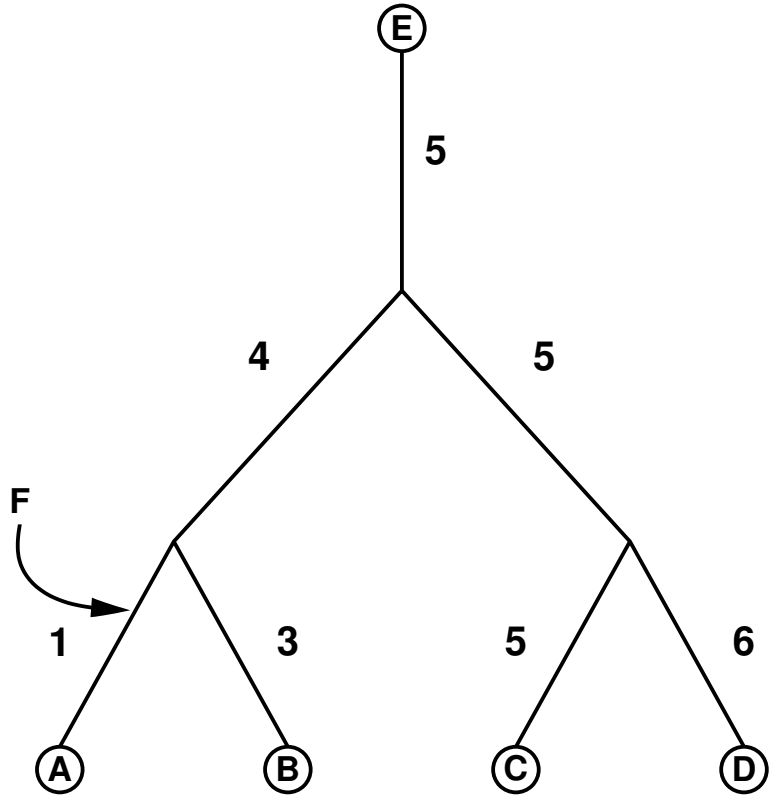
Excursion: Quartet Puzzling



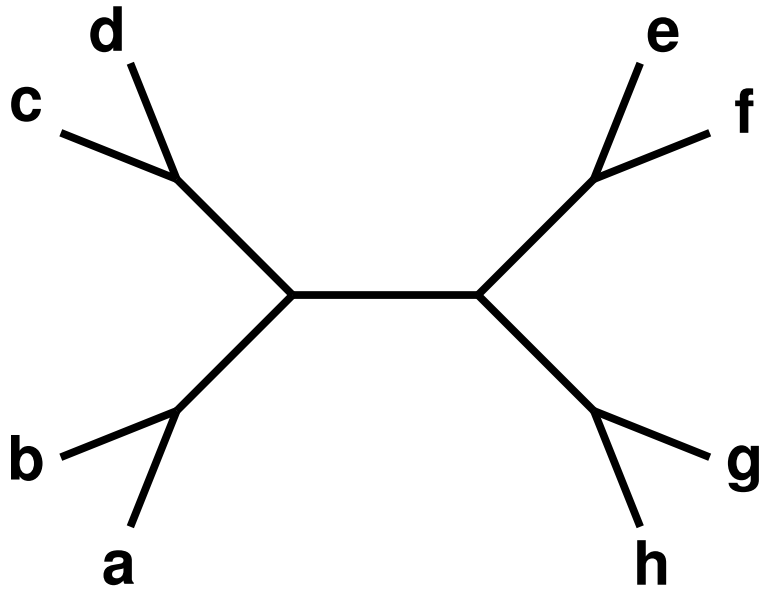


Excursion: Quartet Puzzling

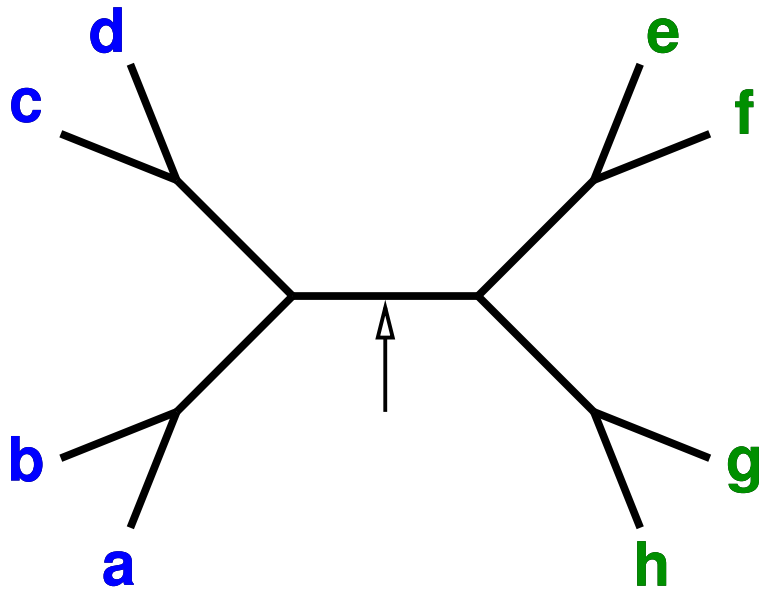




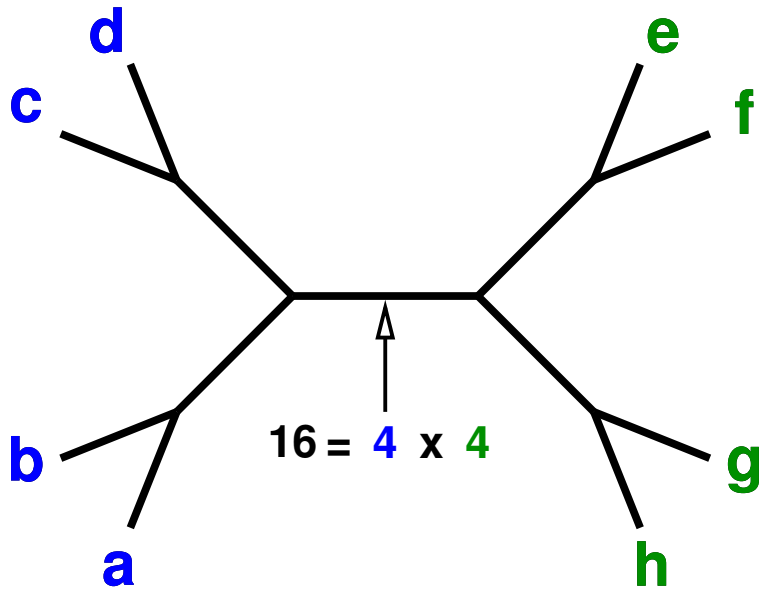
Edge Penalties Biased by Penalty Paths



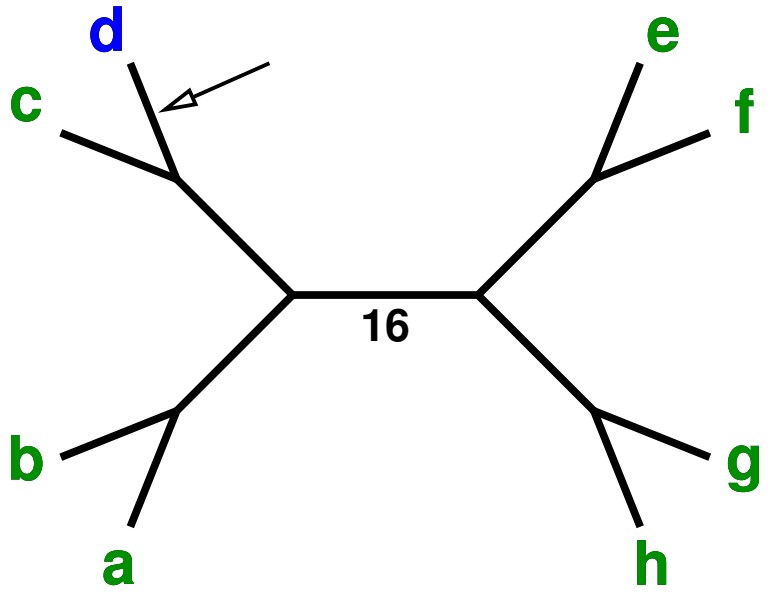
Edge Penalties Biased by Penalty Paths



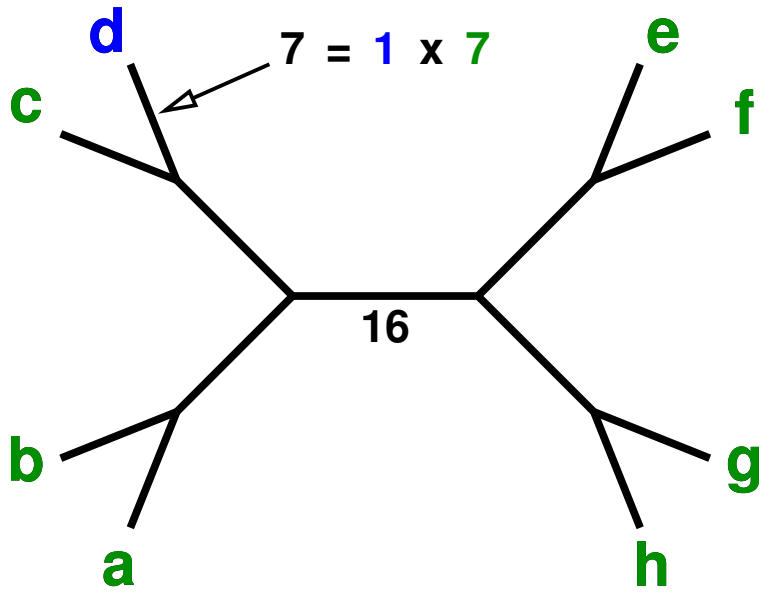
Edge Penalties Biased by Penalty Paths



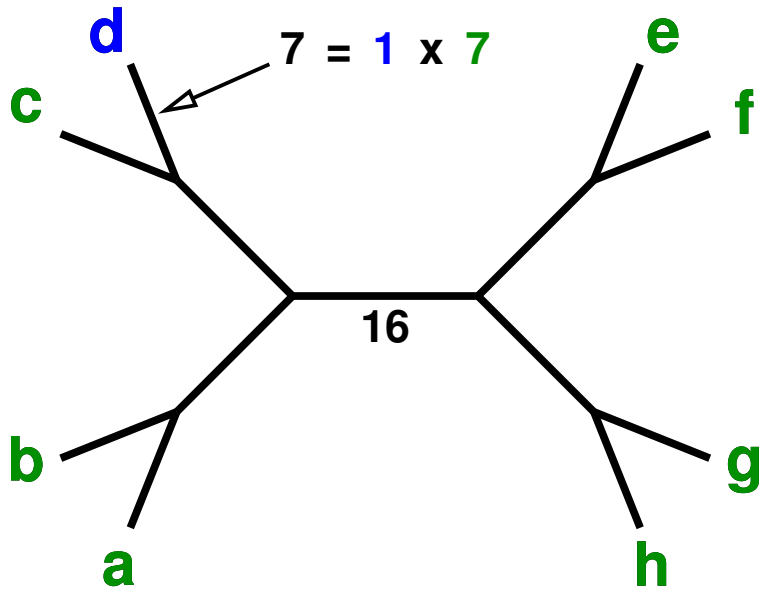
Edge Penalties Biased by Penalty Paths



Edge Penalties Biased by Penalty Paths

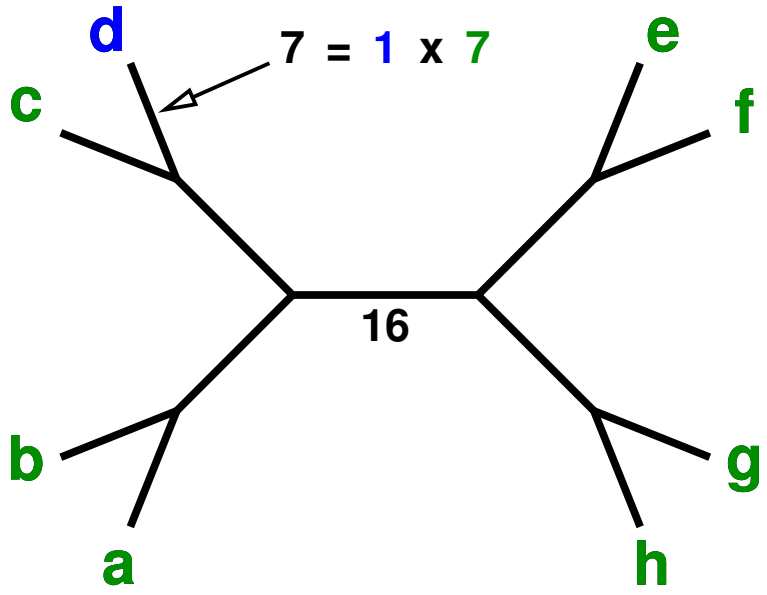


Edge Penalties Biased by Penalty Paths



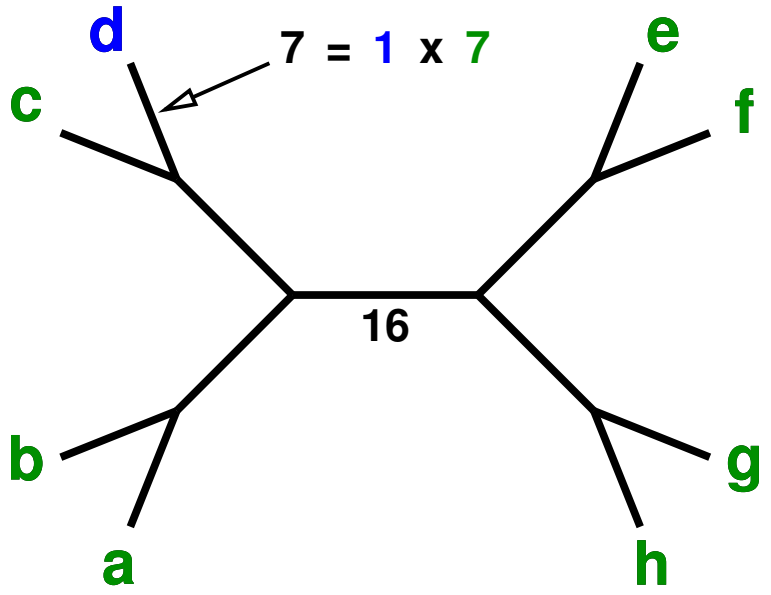
- Bias seems to be no problem, if 'information' is available (A. Führer)

Edge Penalties Biased by Penalty Paths



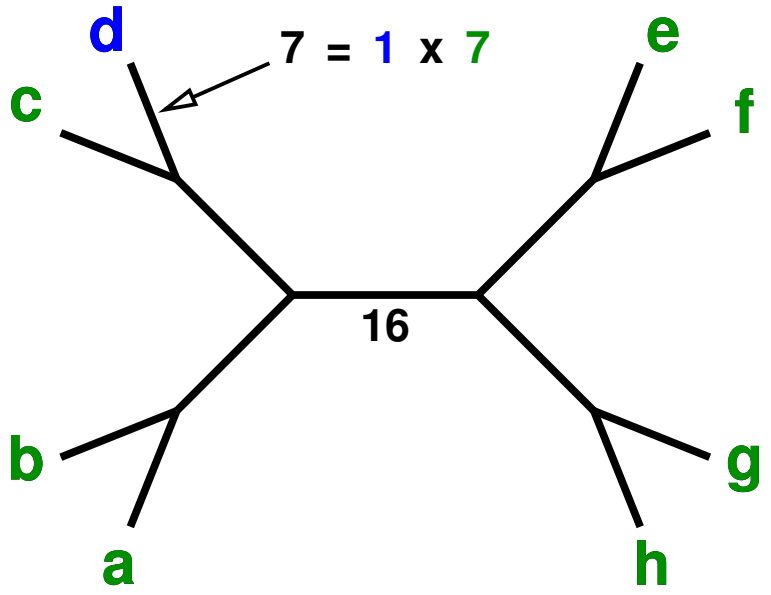
- Bias seems to be no problem, if 'information' is available (A. Führer)
- but we have 'missing information', resolved randomly

Edge Penalties Biased by Penalty Paths



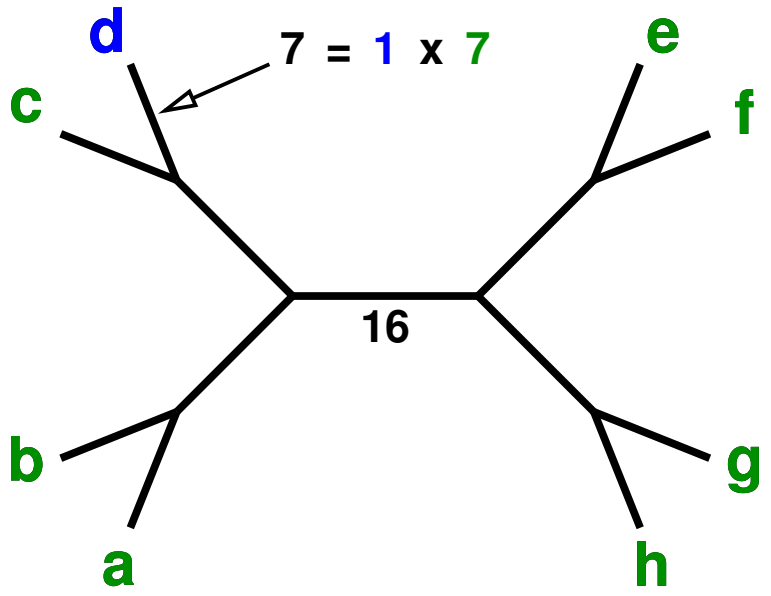
- Bias seems to be no problem, if 'information' is available (A. Führer)
- but we have 'missing information', resolved randomly
- ignoring 'missing data' would lead to un(der)-penalized subtrees

Edge Penalties Biased by Penalty Paths



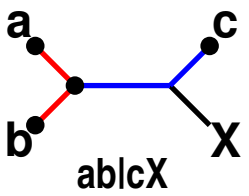
- Bias seems to be no problem, if 'information' is available (A. Führer)
- but we have 'missing information', resolved randomly
- ignoring 'missing data' would lead to un(der)-penalized subtrees: However, not because there is no contradiction

Edge Penalties Biased by Penalty Paths

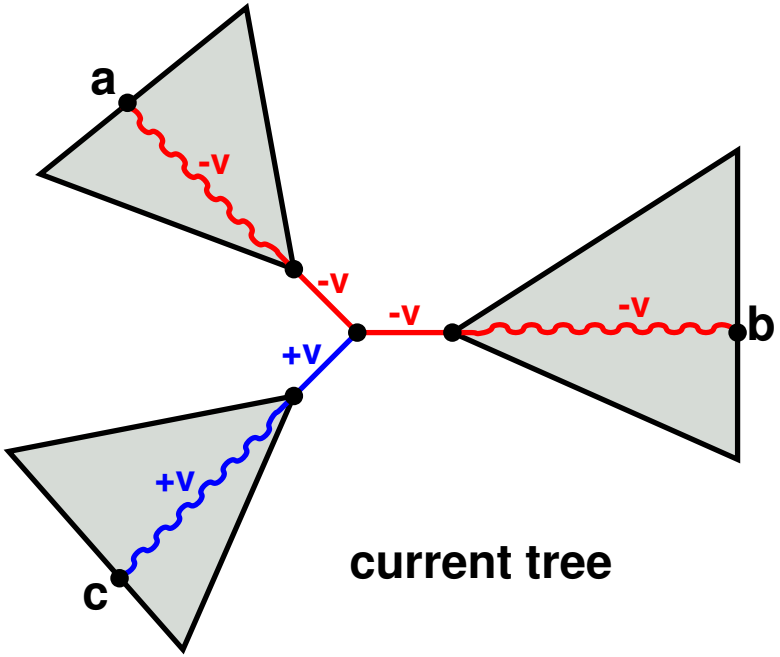


- Bias seems to be no problem, if 'information' is available (A. Führer)
- but we have 'missing information', resolved randomly
- ignoring 'missing data' would lead to un(der)-penalized subtrees:
  However, not because there is no contradiction,
  but merely caused by missing data!

Voting Schemes

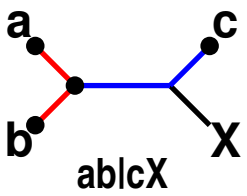


penalty/bonus v=2



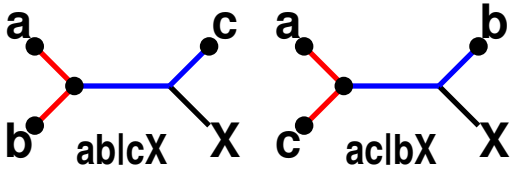
Voting Schemes

Resolved Quartet:

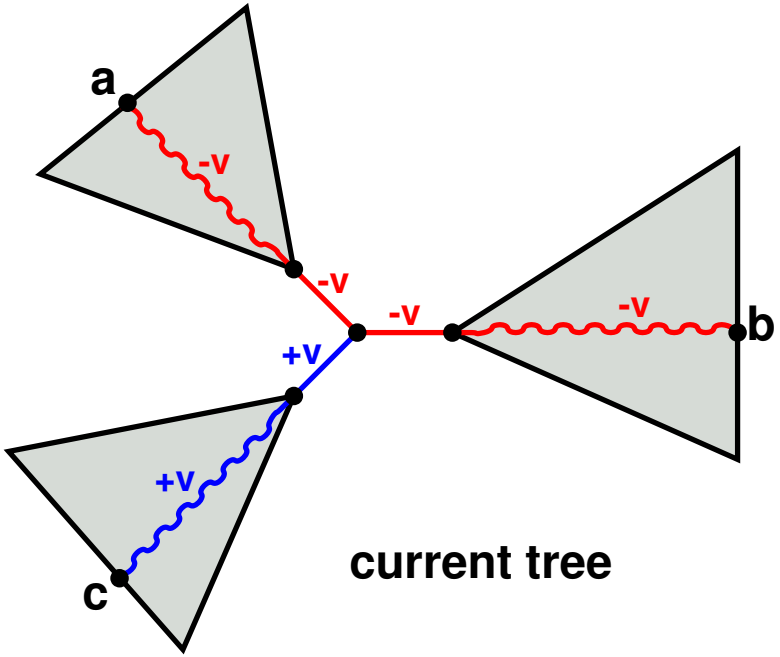


penalty/bonus v=2

Partly Resolved Quartet:



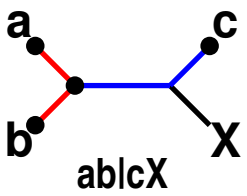
penalty/bonus v=1
for each topology



current tree

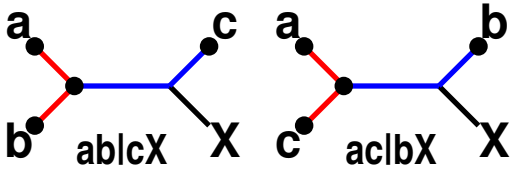
Voting Schemes

Resolved Quartet:

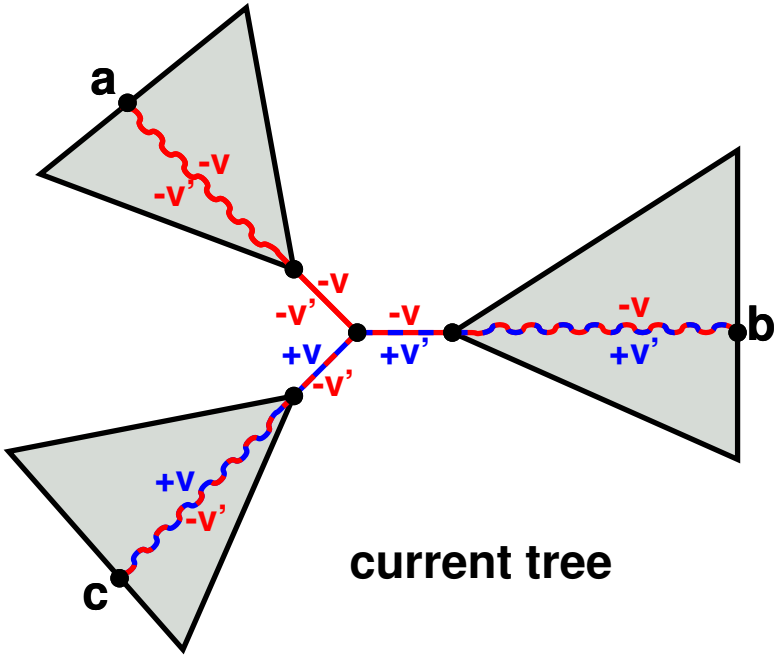


penalty/bonus v=2

Partly Resolved Quartet:



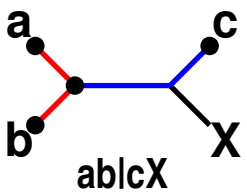
penalty/bonus v=1
for each topology



current tree

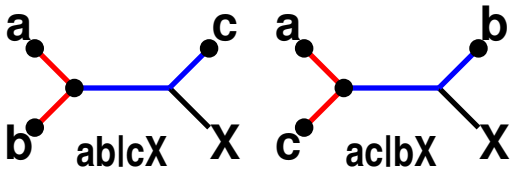
Voting Schemes

Resolved Quartet:

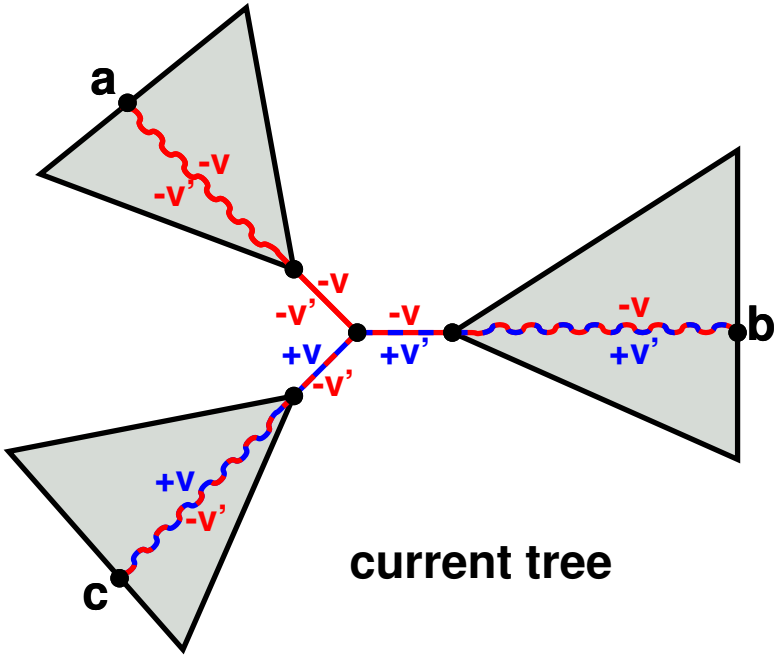


penalty/bonus v=2

Partly Resolved Quartet:



penalty/bonus v=1
for each topology



current tree

$$insertion\ score = \frac{bonus - penalty}{bonus + penalty + missing}$$

Overlap Graph Guided Input Order

A random addition order of taxa is generated as follows:

1. Start with a random geneset g.
2. Insert all taxa of g in random order.
3. Choose a new geneset g that has overlap with some geneset already in the tree.
4. Iterate 2. and 3. this until all taxa are inserted.

Analogous to Prim-MST build path (with equal edge weights).

Complete SQP Algorithm

1. Check combinability with overlap graph
2. Compute quartet likelihoods for each geneset
3. Construct superquartets
4. Construct many trees using the Prim-MST build path on the overlap graph using the SQP voting scheme
5. Construct a consensus

Biological Data: *Poacaea* (GPWG, 2001)

6 genesets

nucleus

- phytochrome B (*phyB*)
- granule bound starch synthase I (*waxy*)
- internal transcribed spacer (ITS)

chloroplast

- NADH dehydrogenase F (*ndhF*)
- ribulose 1,5-bisphosphate carboxylase LSU (*rbcL*)
- RNA polymerase II β" (*rpoC*)

Phylogenetic structure (GPWG, 2001)

- **PACCAD** Clade
  - **P**anicoidea
  - **A**ristidoidea
  - **C**hloridoidea
  - **C**entrothecoidea
  - **A**rundinoidea
  - **D**anthonioidea
- **BEP** Clade
  - **B**ambusoidea
  - **E**rhartoidea
  - **P**ooidea
- early grasses

Biological Data: *Poacaea* (GPWG, 2001)

6 genesets

nucleus

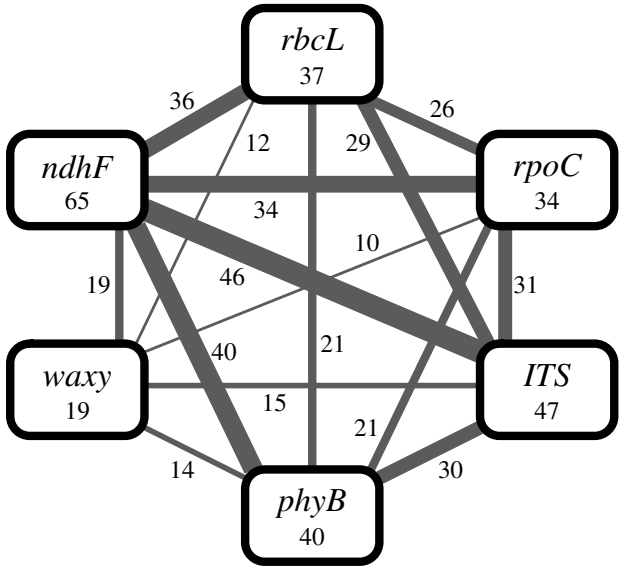
- phytochrome B (*phyB*)
- granule bound starch synthase I (*waxy*)
- internal transcribed spacer (ITS)

chloroplast

- NADH dehydrogenase F (*ndhF*)
- ribulose 1,5-bisphosphate carboxylase LSU (*rbcL*)
- RNA polymerase II β" (*rpoC*)

Phylogenetic structure (GPWG, 2001)

- **PACCAD** Clade
  - **P**anicoidea
  - **A**ristidoidea
  - **C**hloridoidea
  - **C**entrothecoidea
  - **A**rundinoidea
  - **D**anthonioidea
- **BEP** Clade
  - **B**ambusoidea
  - **E**rhartoidea
  - **P**ooidea
- early grasses



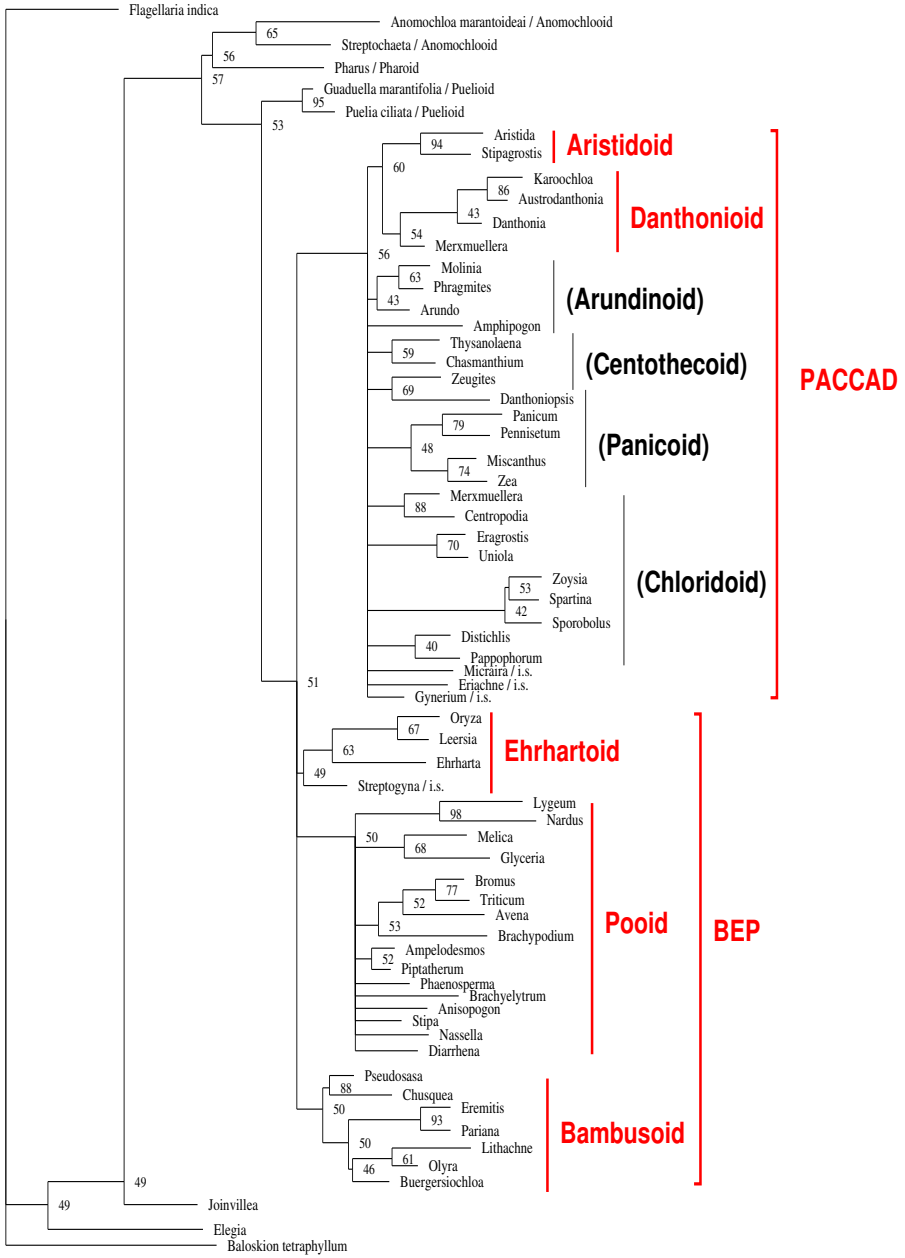
Poacaea-Dataset (GPWG, 2001)

	ndhF	*phyB*	*rbcL*	*rpoC*	*waxy*	ITS	tot. evid.
sequence origin	chloroplast	nucleus	chloroplast	chloroplast	nucleus	nucleus	mixed
# sequences	65	40	37	34	19	47	66
alignment length	2210	1182	1344	777	773	322	6608
constant sites	10.1%	5.6%	45.3%	2.4%	54.2%	48.4%	0%
A content	27.3%	21.5%	27.1%	40.5%	21.3%	18.7%	26.4%
C content	16.5%	26.9%	19.3%	14.6%	29.4%	31.6%	20.0%
G content	17.6%	29.2%	24.9%	28.2%	33.8%	33.1%	23.0%
T content	38.7%	22.3%	28.7%	16.7%	15.5%	16.6%	30.6%
χ^2 test failed	0	4	0	1	2	2	43
ts:tv ratio	1.93	1.85	1.66	2.98	1.05	1.45	1.73
Y:R ts ratio	1.33	0.96	0.82	0.21	0.63	0.85	1.01
average distance	0.077	0.160	0.053	0.107	0.147	0.154	0.087

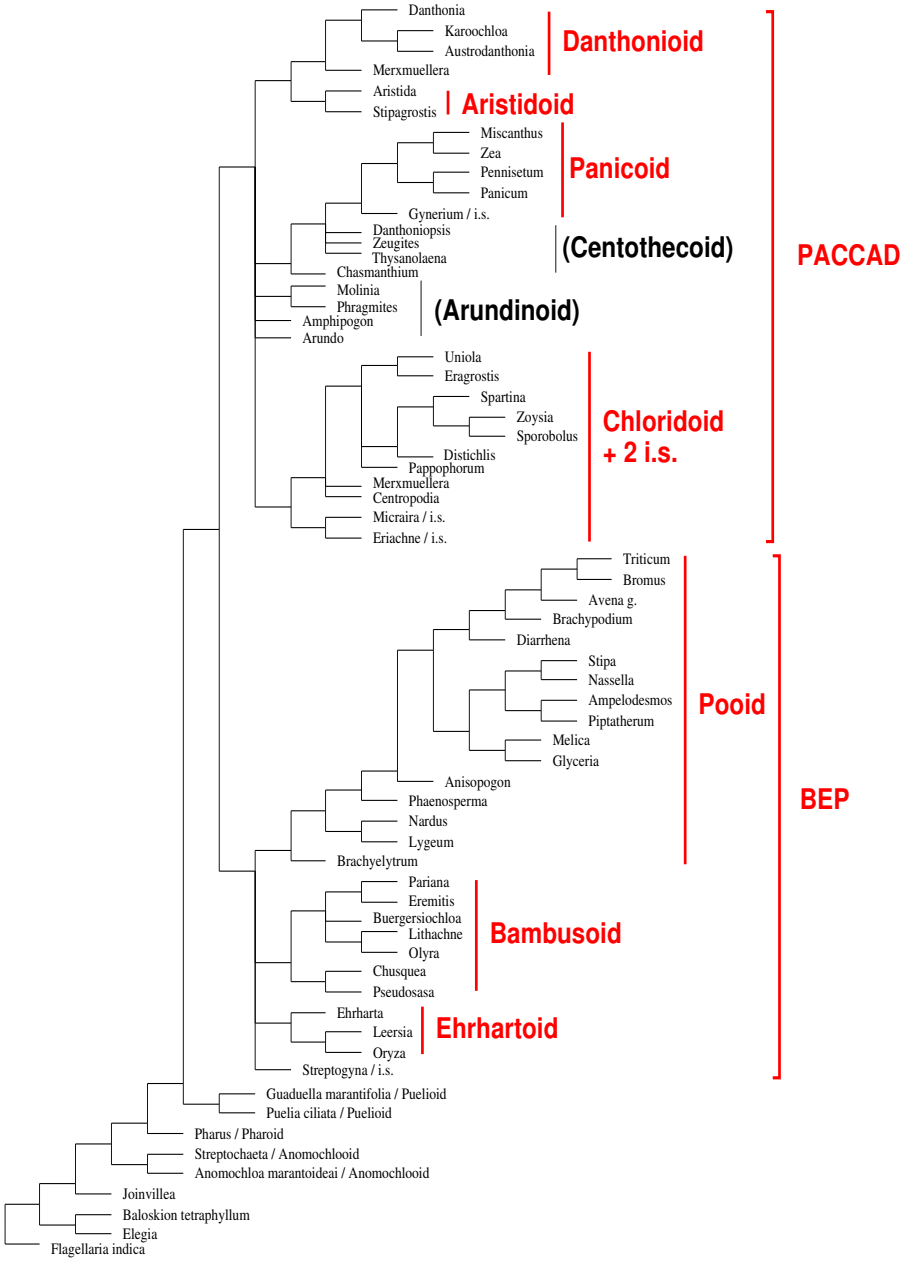
Software Checked

combination	**method**
early level	Total Evidence
late level	MRP + Ragan-Baum coding MRP + Purvis coding MinCut Supertree ModMinCut Supertree
medium level	SuperQuartet Puzzling (SQP)

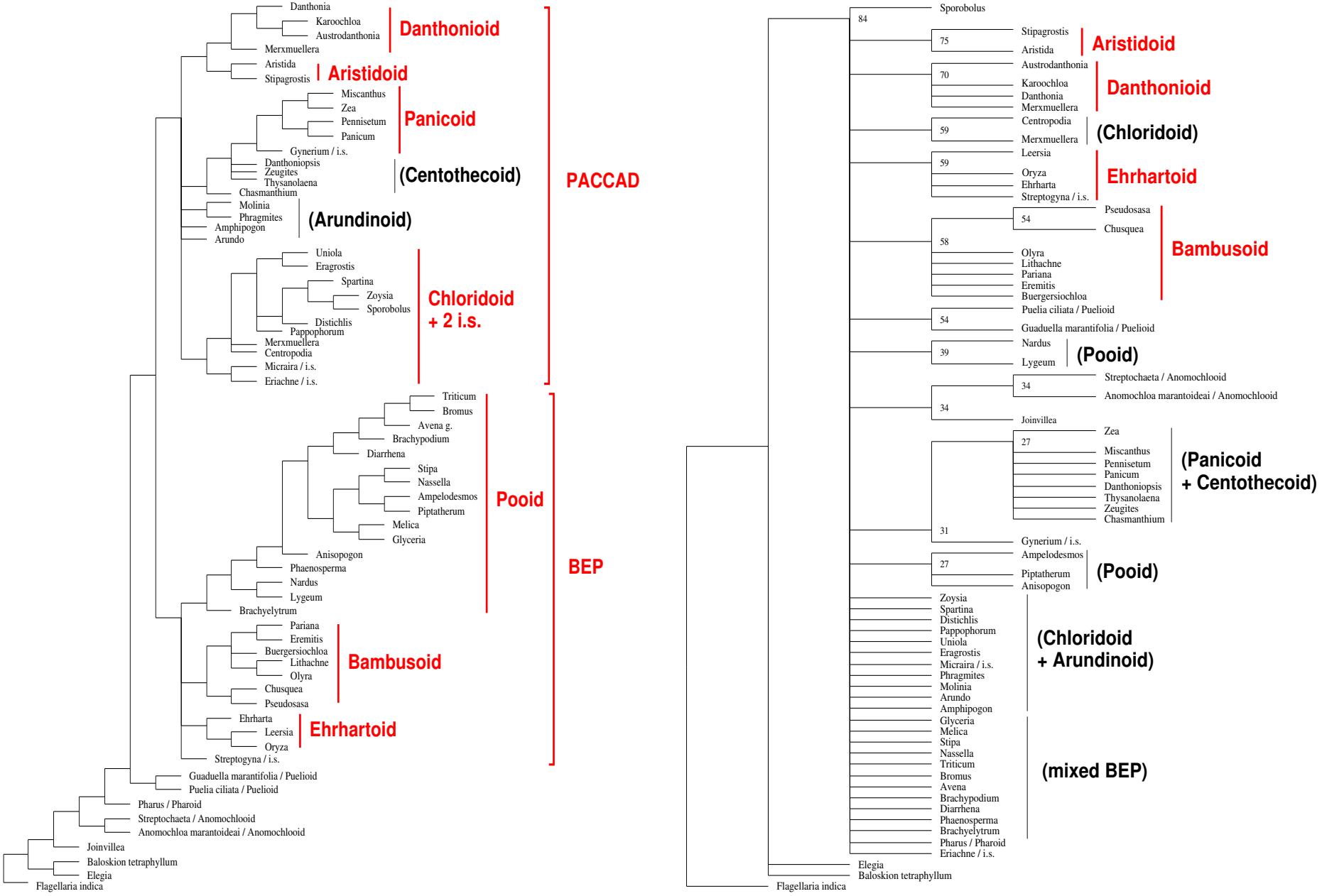
Early Level Approach: Supermatrix



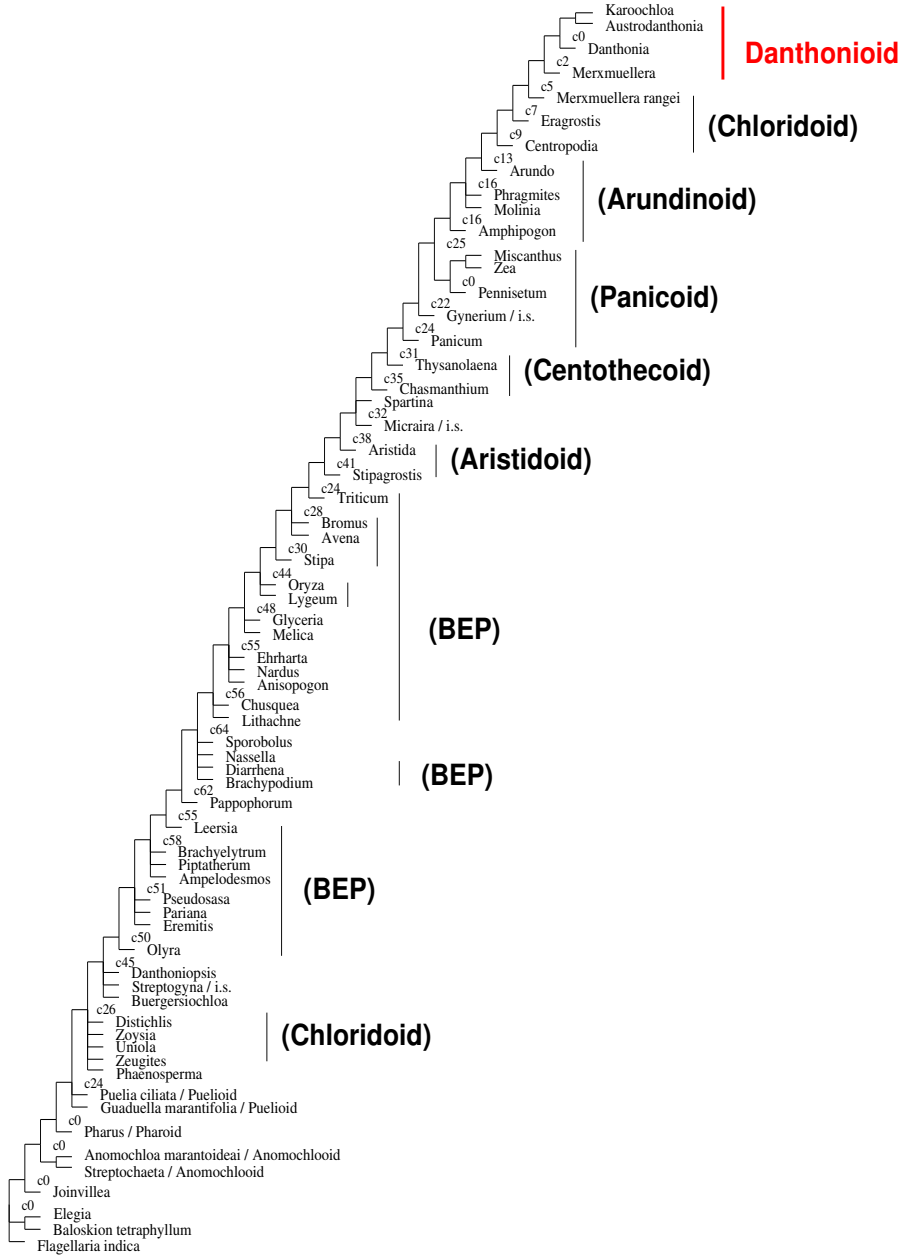
Late Level: MRP Supertree (Baum-Ragan/Purvis coding)



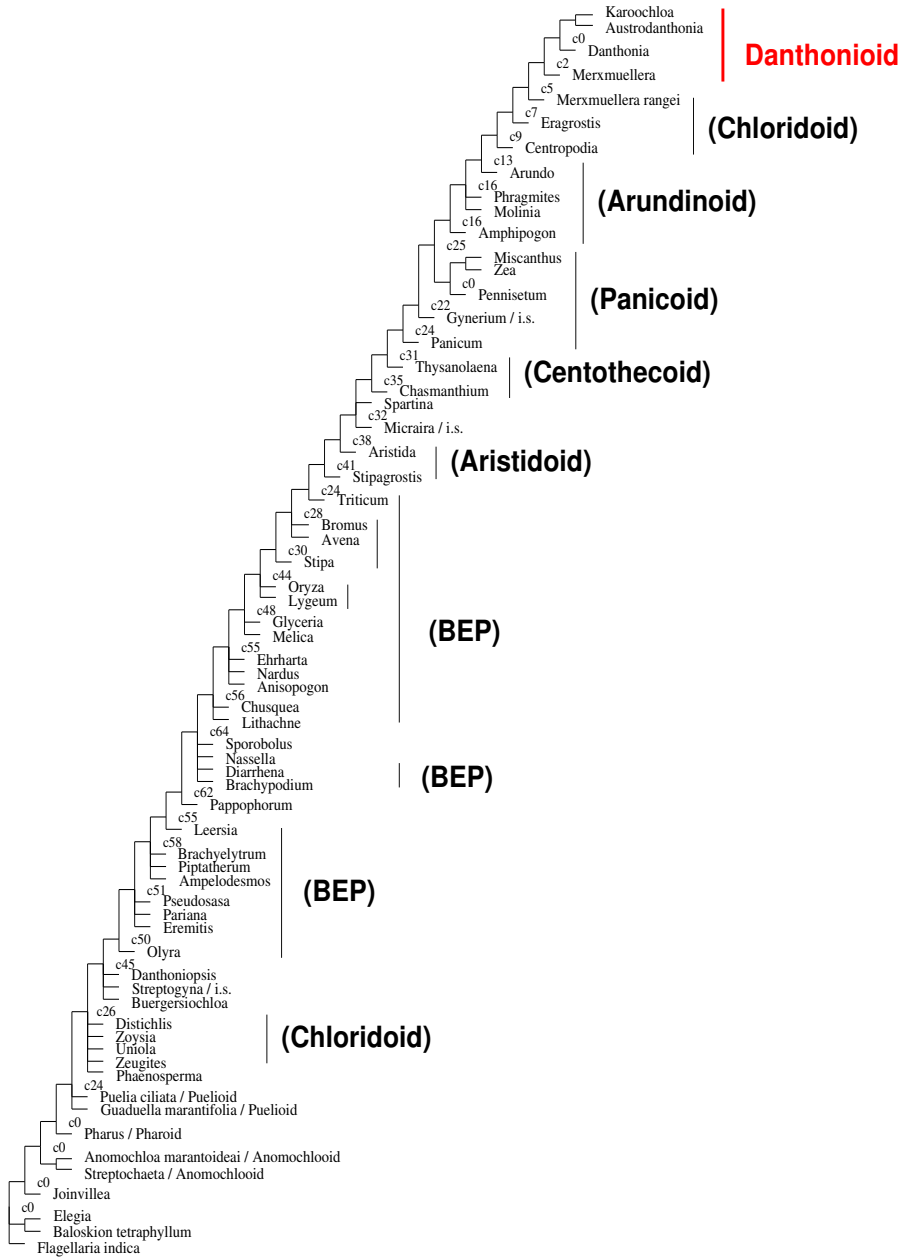
Late Level: MRP Supertree (Baum-Ragan/Purvis coding)

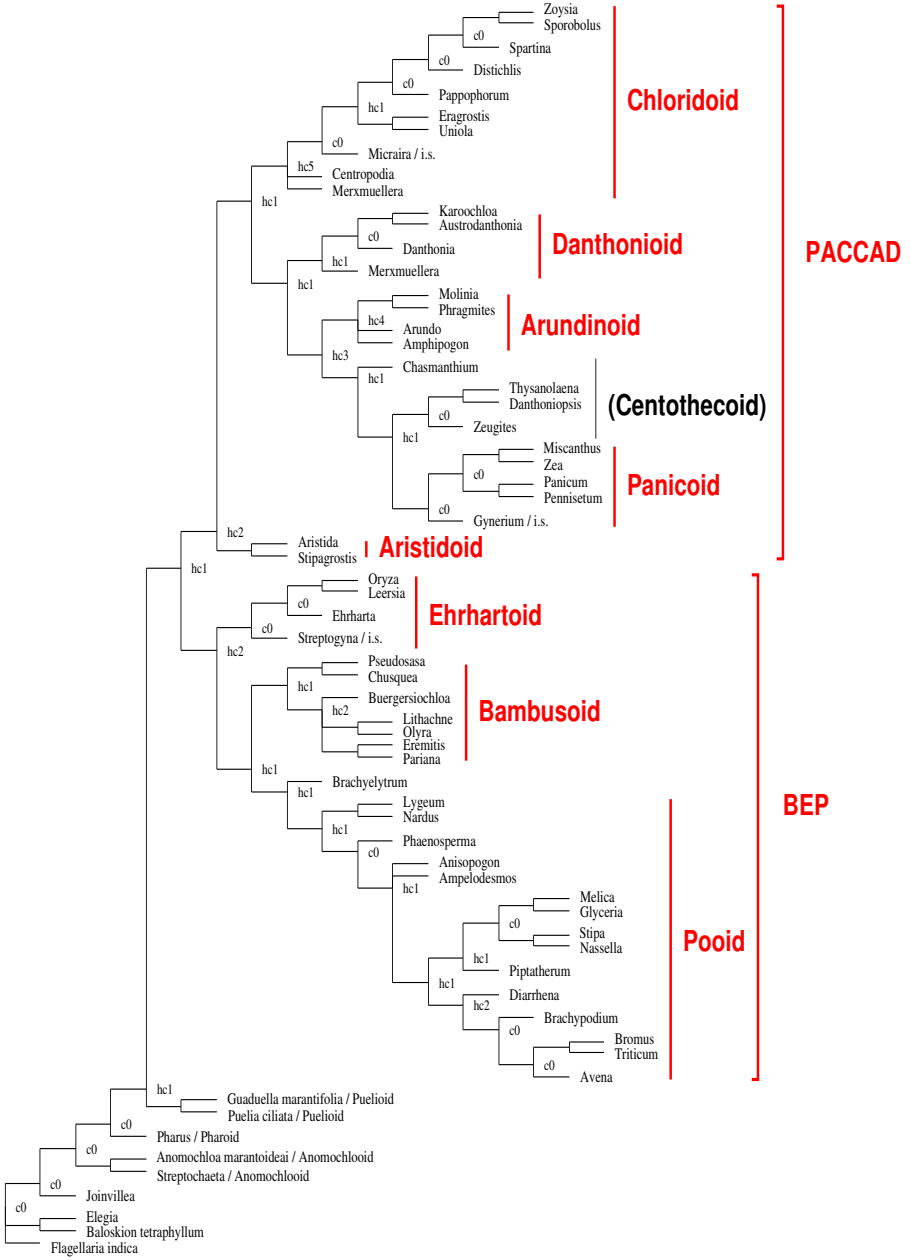


Late Level: Direct Supertrees (MinCut/ModMinCut)

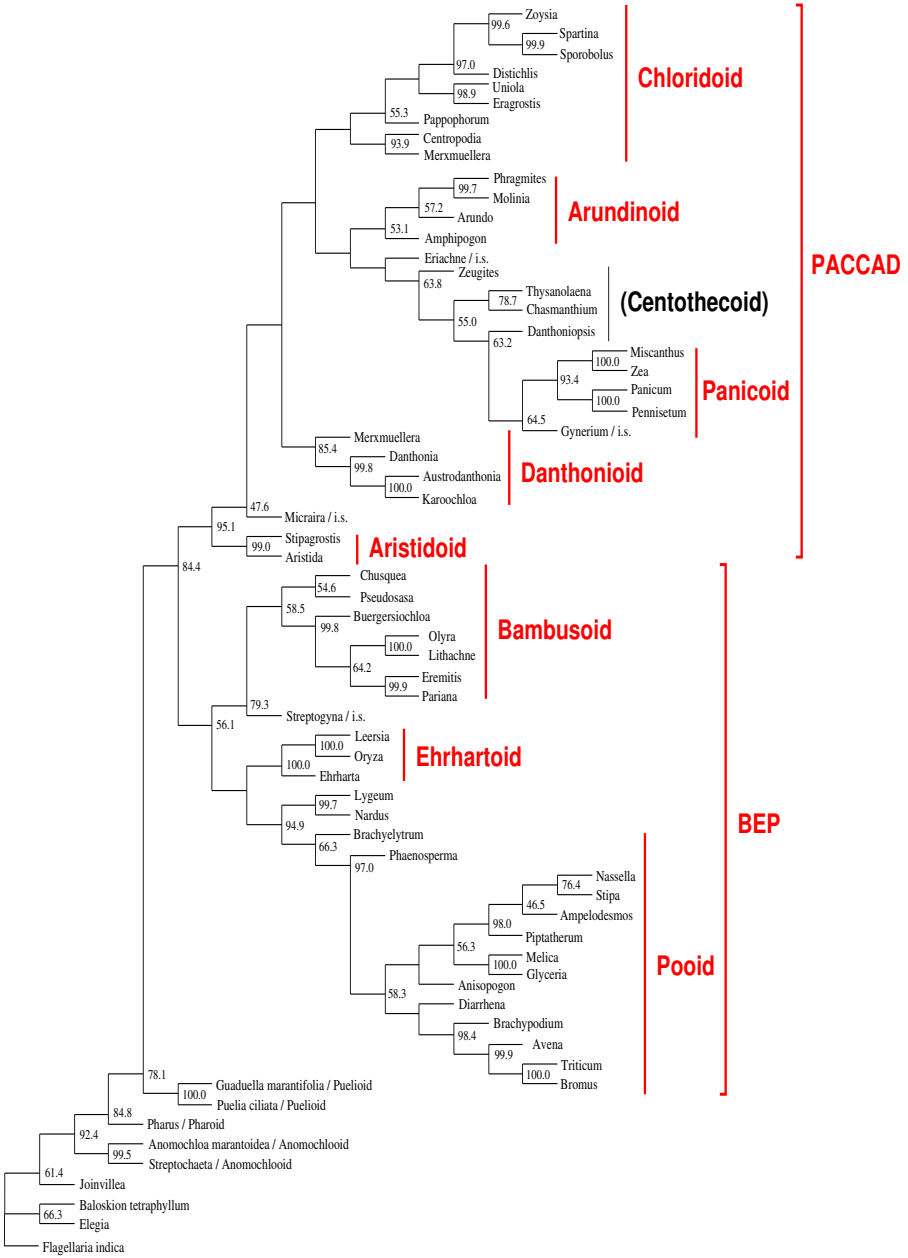


Late Level: Direct Supertrees (MinCut/ModMinCut)





Medium Level Combination: SQP



Summary and Outlook

- The SQP method utilizes data-close information for the combination (like SM).
- SQP models sequence evolution according to each geneset (like ST).
- Voting scheme and overlap graph facilitated better use of available information.
- The new medium level method (SQP) performs well compared to other methods.

Summary and Outlook

- The SQP method utilizes data-close information for the combination (like SM).
- SQP models sequence evolution according to each geneset (like ST).
- Voting scheme and overlap graph facilitated better use of available information.
- The new medium level method (SQP) performs well compared to other methods.

Future work

- Simulations to compare the different combination methods.
- Combination of different data types (e.g., DNA and protein).
- Combination of models with different complexity.