Profile HMMs

Tandy Warnow

BioE/CS 598AGB

Profile Hidden Markov Models

- Basic tool in sequence analysis
- Look more complicated than they really are
- Used to model a family of sequences
- Can be built from a multiple sequence alignment
- Algorithms using profile HMMs are based on dynamic programming (much like Needleman-Wunsch)

Profile

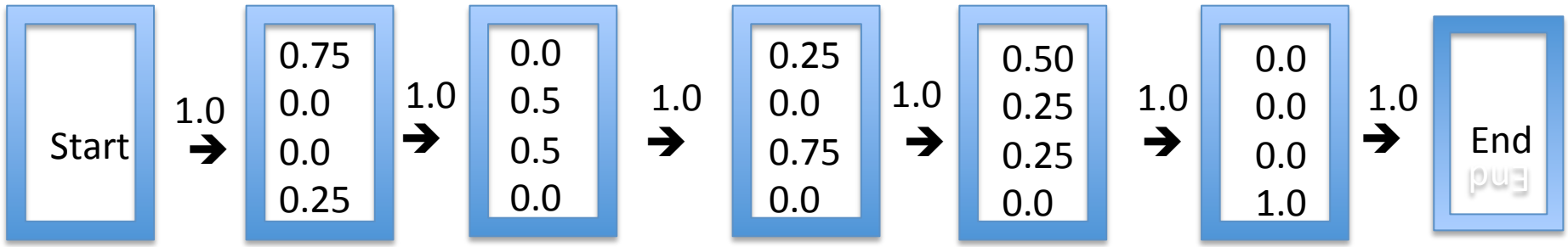
- Given a gap-less multiple sequence alignment, we can build a profile describing what we see

- S1 = A C T A G
- S2 = A C A A G
- S3 = A T T T G
- S4 = G T T C G

	1	2	3	4	5
A	0.75	0.0	0.25	0.50	0.0
C	0.00	0.5	0.00	0.25	0.0
T	0.00	0.5	0.75	0.25	0.0
G	0.25	0.0	0.00	0.00	1.0

Using a profile

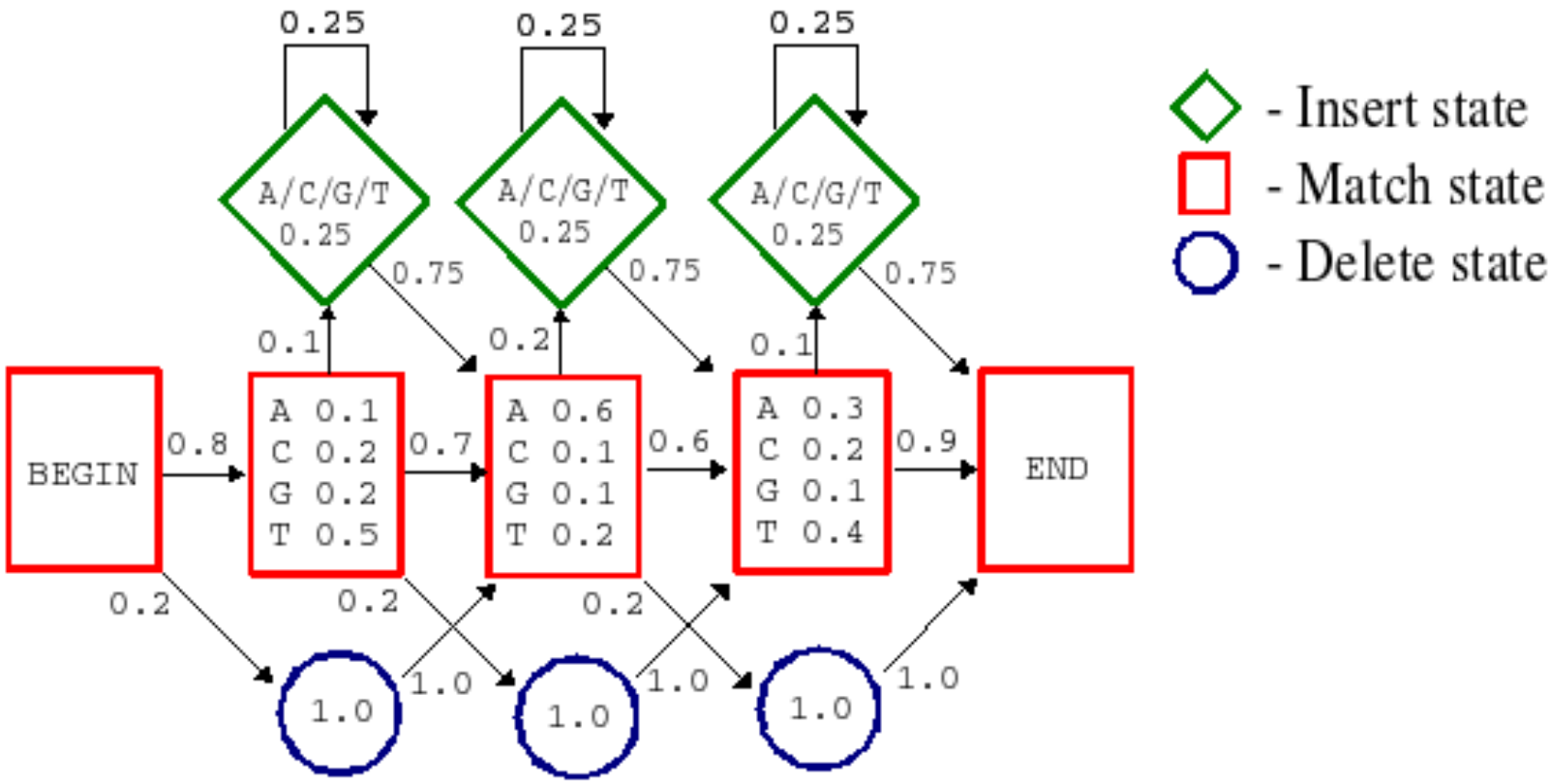
	1	2	3	4	5
A	0.75	0.0	0.25	0.50	0.0
C	0.00	0.5	0.00	0.25	0.0
T	0.00	0.5	0.75	0.25	0.0
G	0.25	0.0	0.00	0.00	1.0



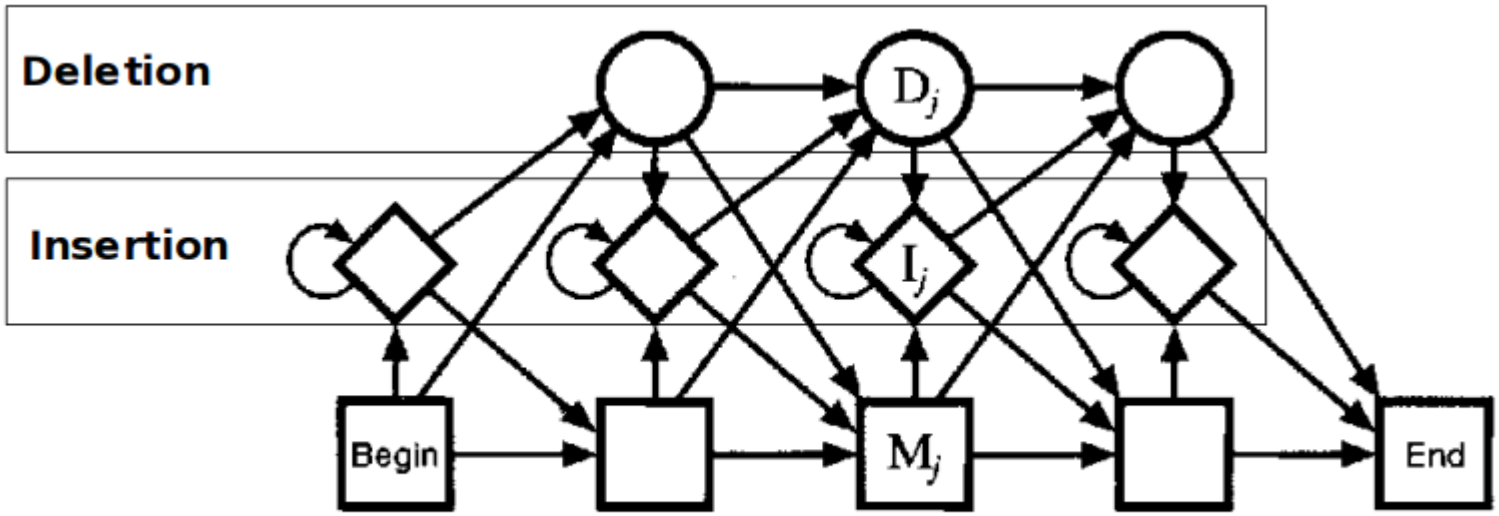
The profile yields a probability distribution of sequences – here, all of the same length.

Adding in insertions

- The profile shown in the previous slide only had *match* states (indicated by rectangles). It doesn't allow any insertions or deletions.
- To model indels, we just have to add additional states to the graphical model.
  - Insertion states: Diamonds (have non-zero emission probabilities)
  - Deletion states: Circles (nothing emitted)



From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html



From http://codecereal.blogspot.com/2011/07/protein-profile-with-hmm.html

Building Profile HMMs

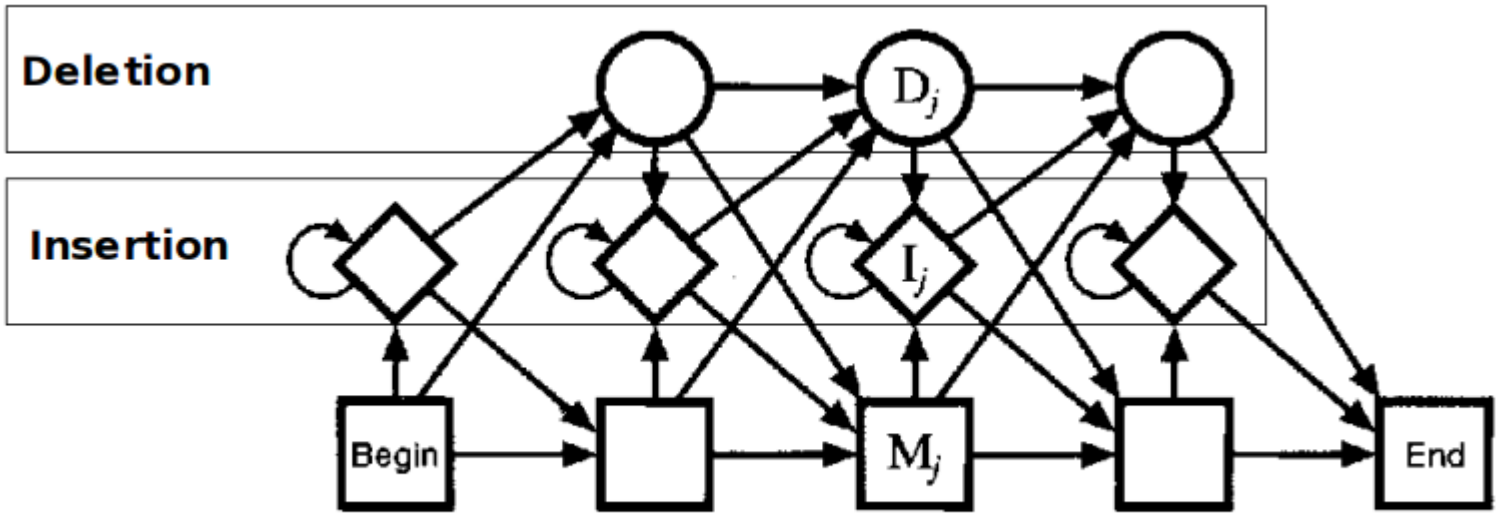
- Profile HMMs can be built from a given multiple sequence alignment – this is not too difficult.
- Profile HMMs can also be built from unaligned sequences. This is a bit complicated.

Using Profile HMMs

- Given a Profile HMM computed for a multiple sequence alignment, you can use it to
  - Recognize related sequences
  - Add related sequences into the multiple sequence alignment
- See PFAM, http://pfam.xfam.org, for how profile HMMs are used to represent groups of functionally and structurally related proteins.

Using profile HMMs to align sequences

- Given an MSA for a set **S** of representative sequences for a gene and set **X** of additional sequences (query sequences)
  - You build a profile HMM for the MSA on S.
  - For each **s in X** you find for the gene, you find the path through the profile HMM that is most likely to generate **s**.
  - The path specifies how to add the sequence into the MSA (only the match states count).
  - Transitivity gives you the final MSA after you add in all the other sequences.



From http://codecereal.blogspot.com/2011/07/protein-profile-with-hmm.html

Other uses of profile HMMs

- Given two profile HMMs (H1 and H2), and a sequence s, you can determine which one is more likely to generate s (again, using dynamic programming).
- Note that computing the probability that a profile HMM generates a sequence requires calculating the probability for *every* path and adding up the probabilities. This can be calculated in polynomial time, using dynamic programming.

Applications of profile HMMs

- Recognizing membership in protein families (see PFAM http://pfam.xfam.org)
- Progressive multiple sequence alignment methods, such as SATCHMO (Edgar and Sjolander, Bioinformatics 2003)
- Phylogenetic placement (e.g., SEPP, Mirarab et al., PSB 2012)
- Taxonomic identification of metagenomic data (e.g., TIPP, Nguyen et al., Bioinformatics 2014)

Another application: UPP

- UPP (Nguyen et al., RECOMB 2015) is a method for ultra-large multiple sequence alignment:
  - Up to 1,000,000 sequences
  - Robust to fragmentary sequences
- Nam-phuong Nguyen (IGB postdoctoral fellow) will present this method on Tuesday.

HMMER

- http://hmmer.janelia.org
- One of the most popular collection of tools to perform analyses based on profile HMMs.
- **HMMER web server: interactive sequence similarity searching, NAR 2011,** http://nar.oxfordjournals.org/content/39/suppl_2/W29

Supertree Estimation

- Purposes:
  - Divide-and-conquer tree estimation
  - Combining analyses performed by other research groups
  - Tree of Life

Supertree Estimation

Challenges:

- Tree compatibility is NP-complete (therefore, even if subtrees are correct, supertree estimation is hard)
- Estimated subtrees have error

Advantages:

- Estimating individual gene trees can be computationally feasible (compared to the combined analysis of many genes)
- Can use different types of data for each gene

Many Supertree Methods

Matrix Representation with Parsimony
(Most commonly used and most accurate)

- MRP
- weighted MRP
- MRF
- MRD
- Robinson-Foulds Supertrees
- Min-Cut
- Modified Min-Cut
- Semi-strict Supertree
- QMC
- Q-imputation
- SDM
- PhySIC
- Majority-Rule Supertrees
- Maximum Likelihood Supertrees
- and many more ...

MRP and MRL

- MRP (Matrix Representation with Parsimony)
- MRL (Matrix Representation with Likelihood)
  - The input set of source trees is represented by a binary matrix with ?s for missing data
  - Each edge in each source tree gives you one column in the matrix (based on the bipartition it defines on its leafset)
  - Then you analyze the matrix using parsimony or CFN maximum likelihood

Single gene vs. multi-gene analyses

- Most methods analyze *single* genes (or other genomic region). These produce estimated "gene trees".

- But species trees are estimated using *multiple* genes.

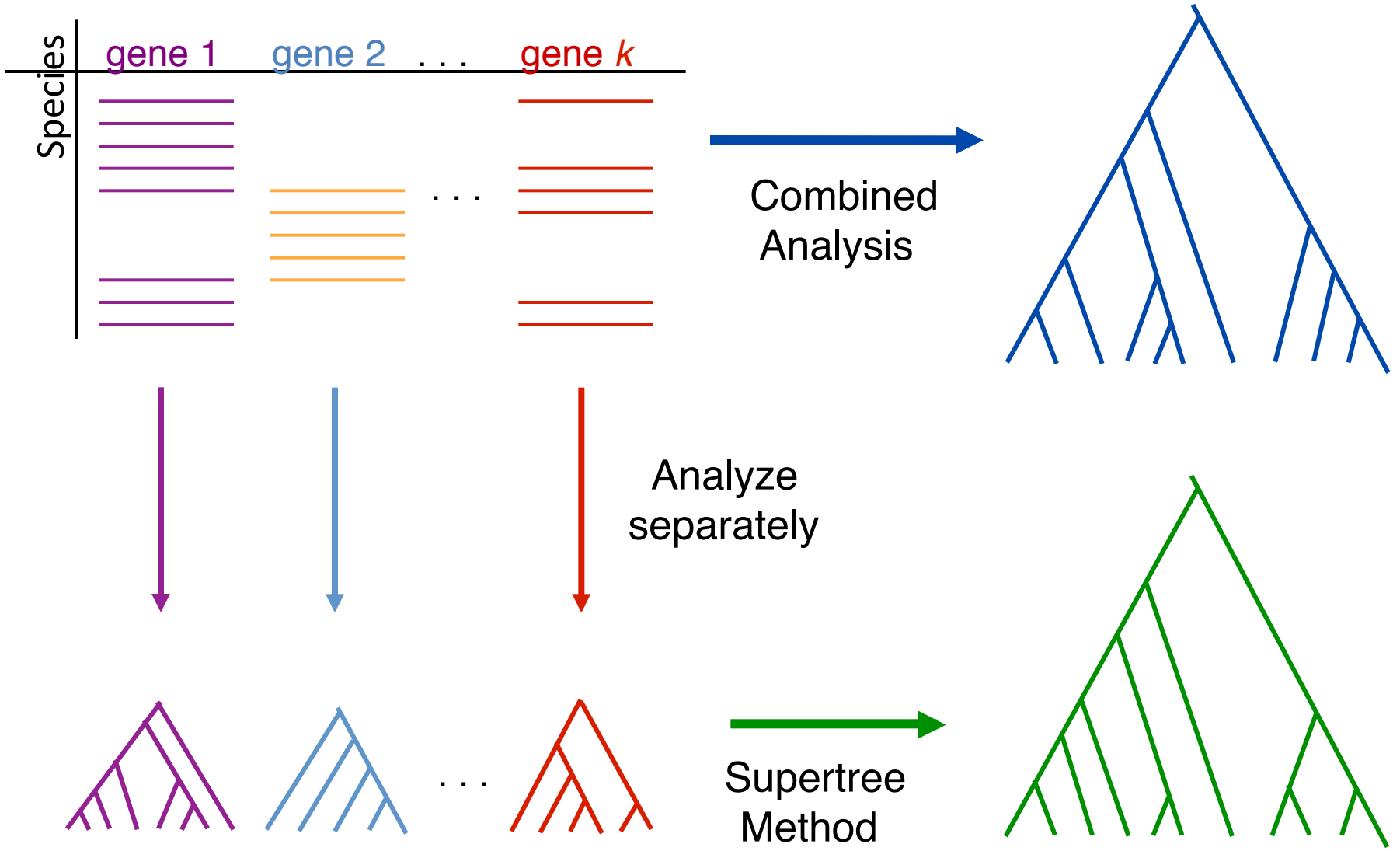
Not all genes present in all species

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

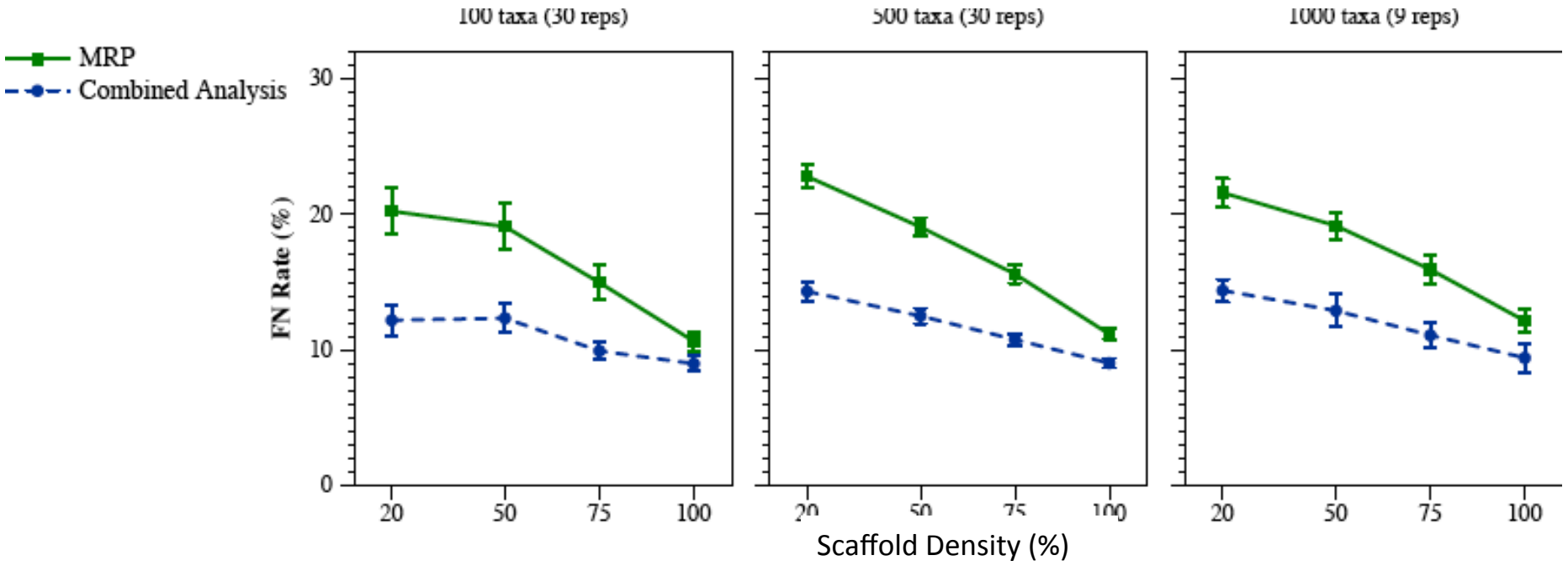
	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

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

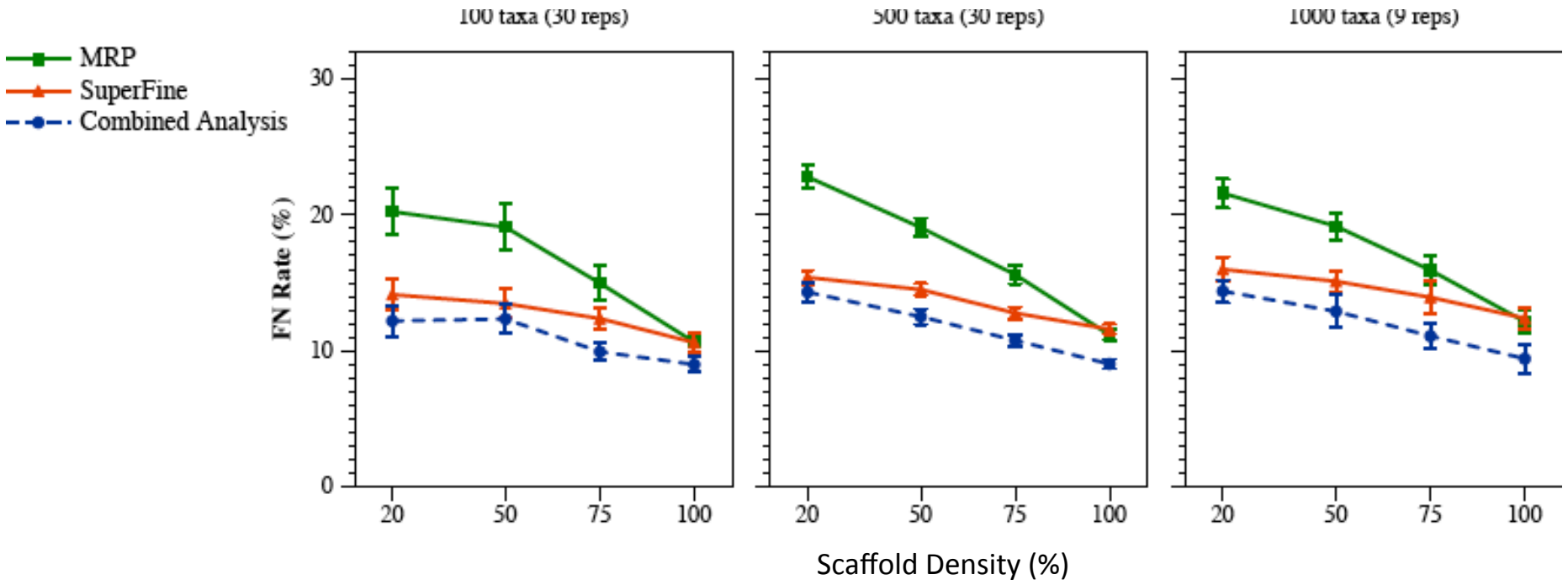
Two competing approaches



FN rate of MRP vs. combined analysis



SuperFine-boosting: improves accuracy of MRP



(Swenson et al., Syst. Biol. 2012)

SuperFine

- First, construct a supertree with low false positives
- Then, refine the tree to reduce false negatives by resolving each polytomy using a "base" supertree method (e.g., MRP)

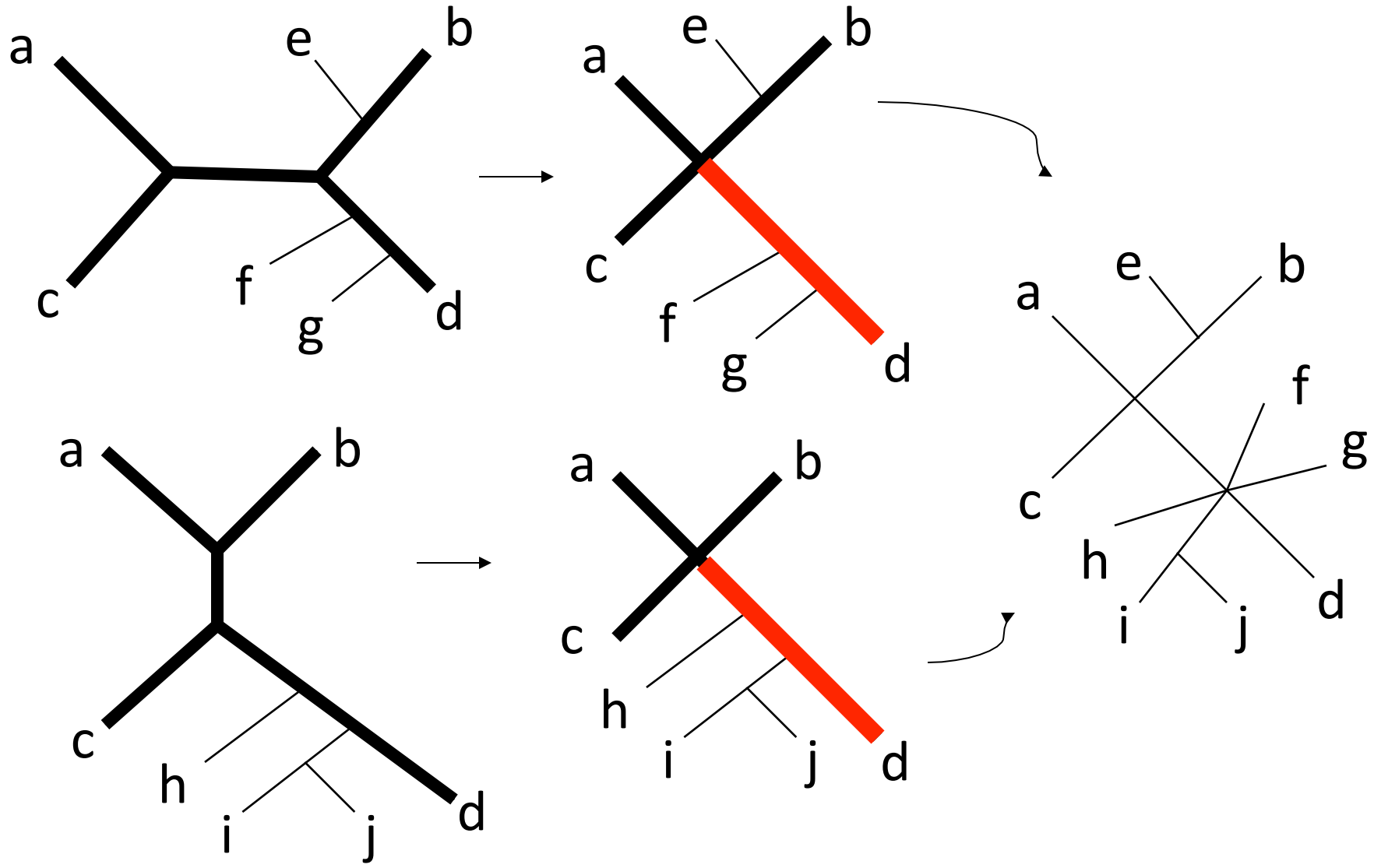
Obtaining a supertree with low FP

The Strict Consensus Merger (SCM)

SCM of two trees

- Computes the strict consensus on the common leaf set
- Then superimposes the two trees, contracting more edges in the presence of "collisions"

Strict Consensus Merger (SCM)



Theoretical results for SCM

- SCM can be computed in polynomial time
- For certain types of inputs, the SCM method solves the NP-hard "Tree Compatibility" problem
- All splits in the SCM "appear" in at least one source tree (and are not contradicted by any source tree)

Performance of SCM

- Low false positive (FP) rate

  (Estimated supertree has few false edges)

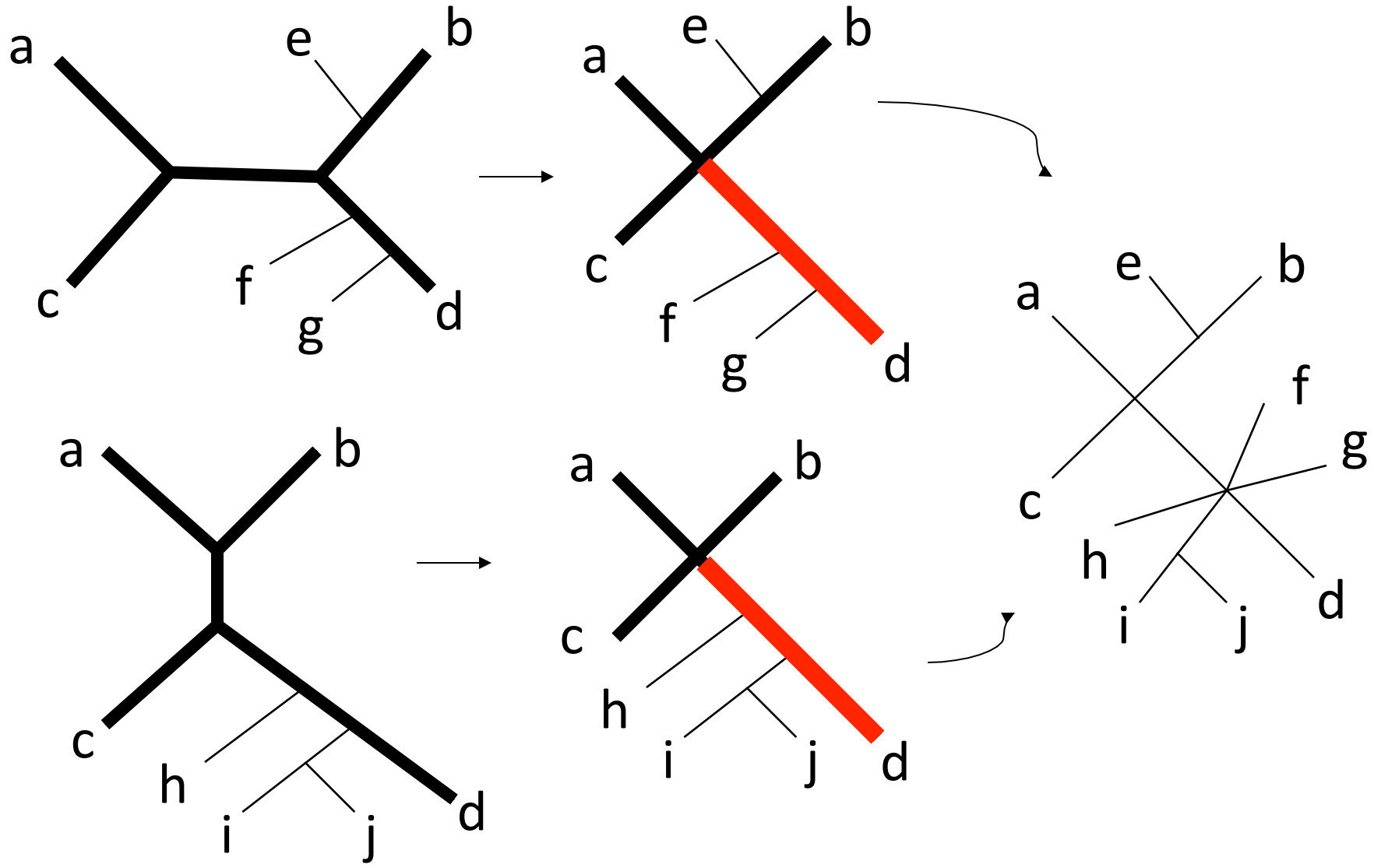
- High false negative (FN) rate

  (Estimated supertree is missing many true edges)

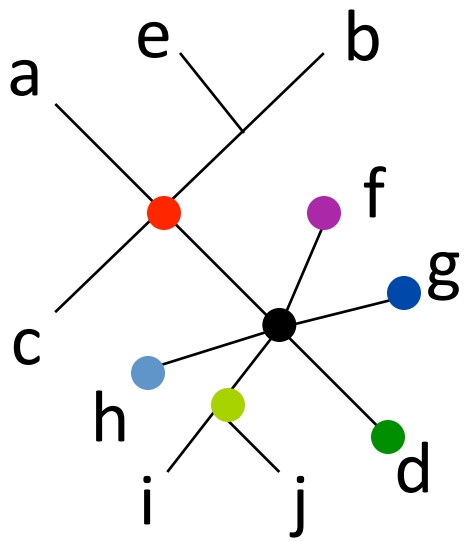
Part II of SuperFine

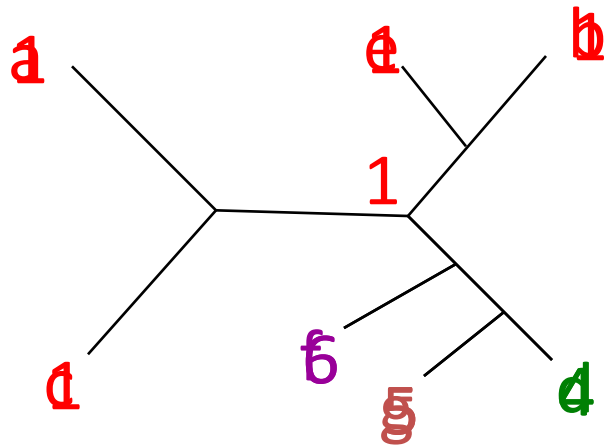
- Refine the tree to reduce false negatives by resolving each polytomy using a base supertree method (e.g., MRP)

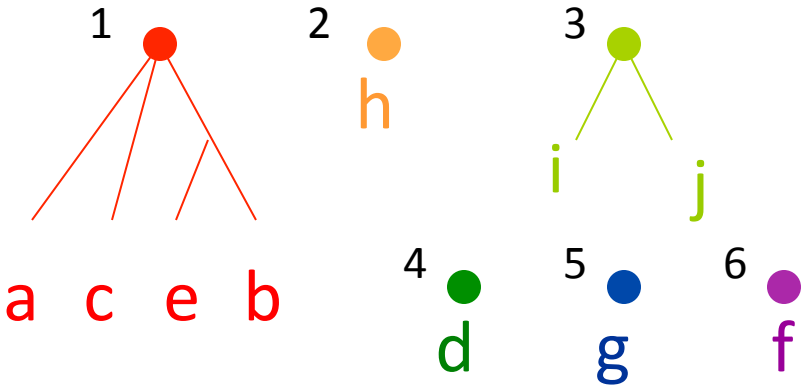
Part 1 of SuperFine

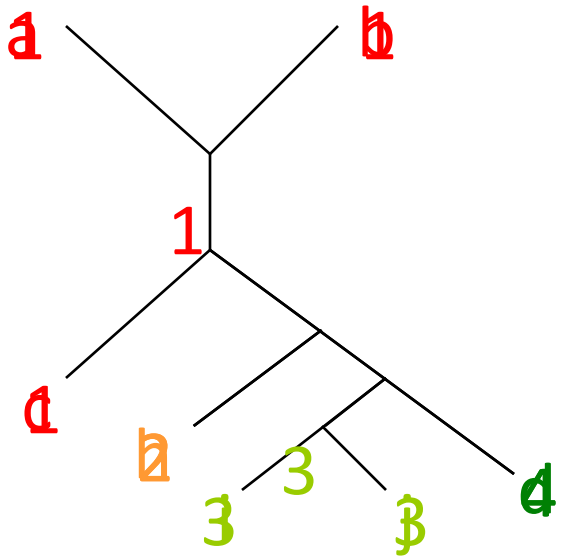


Part 2 of SuperFine









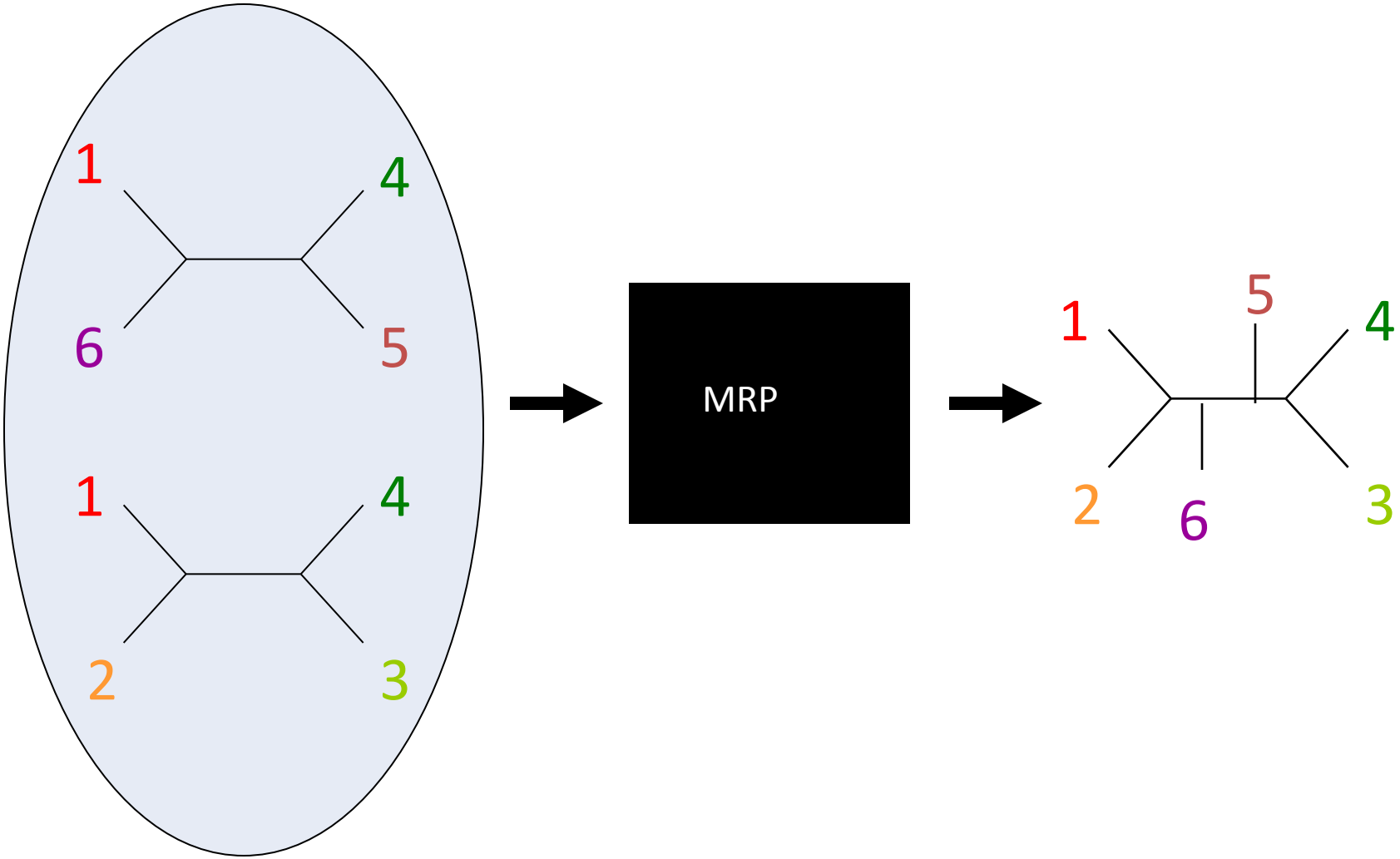
Theorem

Given

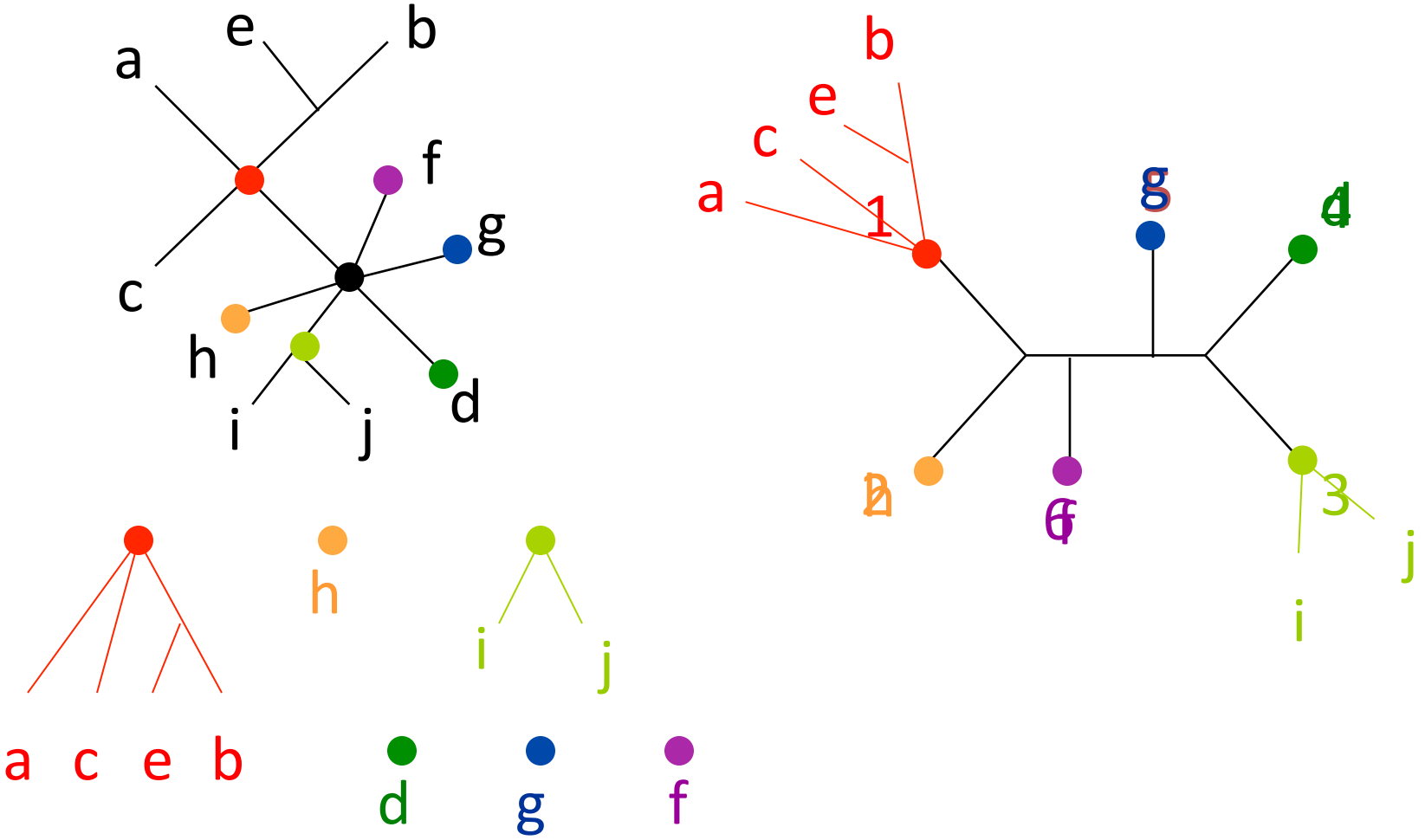
- a set of source trees,
- SCM tree T,
- and a polytomy in T,

after relabelling and reducing, each source tree has *at most one leaf with each label*.

Step 2: Apply MRP to the collection of reduced source trees



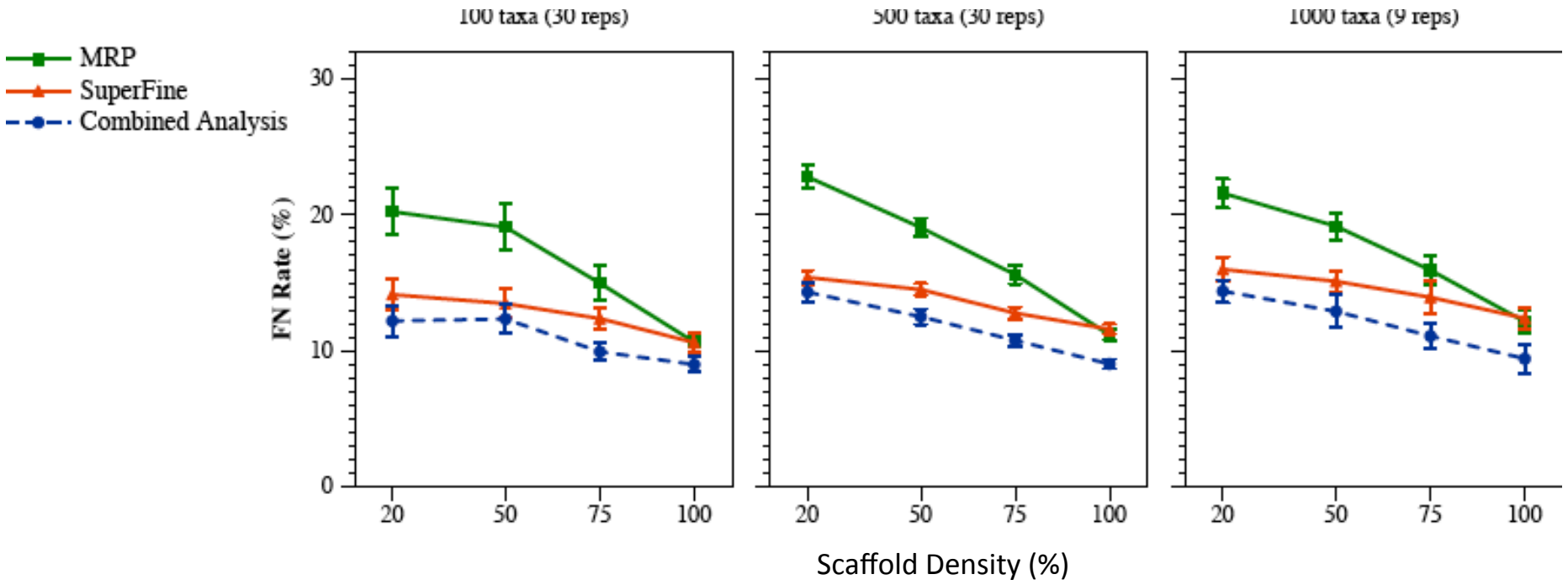
Replace polytomy using tree from MRP



Resolving a single polytomy, *v, using MRP*

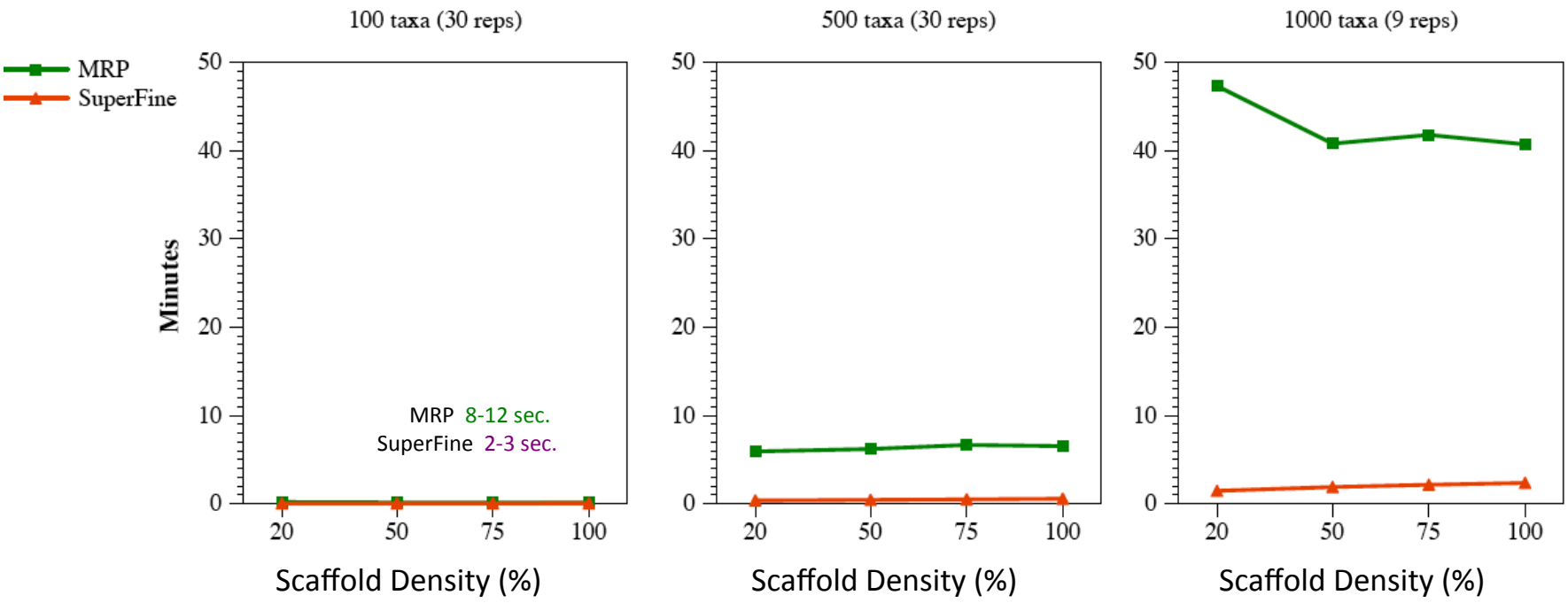
- Step 1: Reduce each source tree to a tree on leafset, {1,2,...,*d*} where *d*=degree(*v*)
- Step 2: Apply MRP to the collection of reduced source trees, to produce a tree t on {1,2,...,*d*}
- Step 3: Replace the star tree at *v* by tree *t*

SuperFine-boosting: improves accuracy of MRP



(Swenson et al., Syst. Biol. 2012)

SuperFine is also much faster



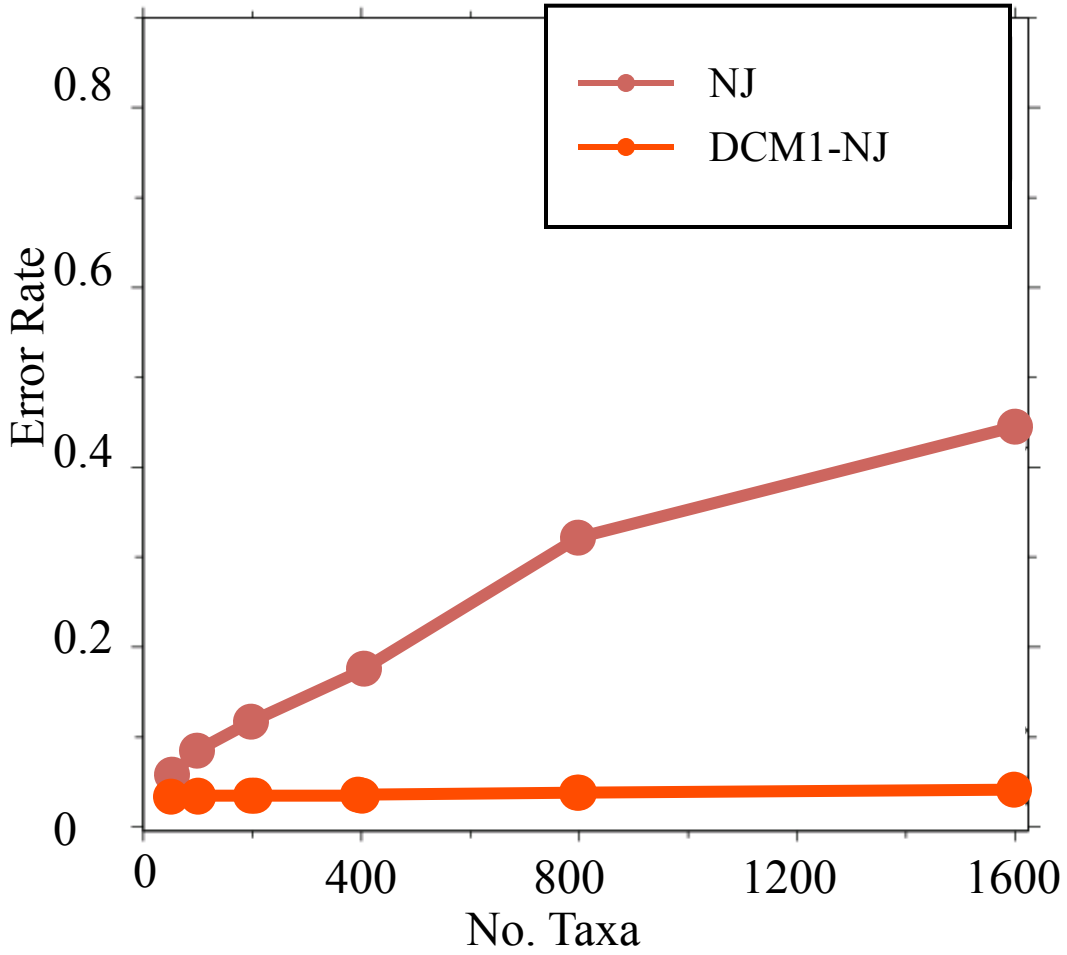
Uses of Supertree Methods

- DACTAL: divide-and-conquer trees almost without alignments (Nelesen et al, 2012)
- DCM1-boosting

In these methods, the dataset is divided into subsets (using chordal graph theory), and then trees on the subsets are combined (using supertree methods).

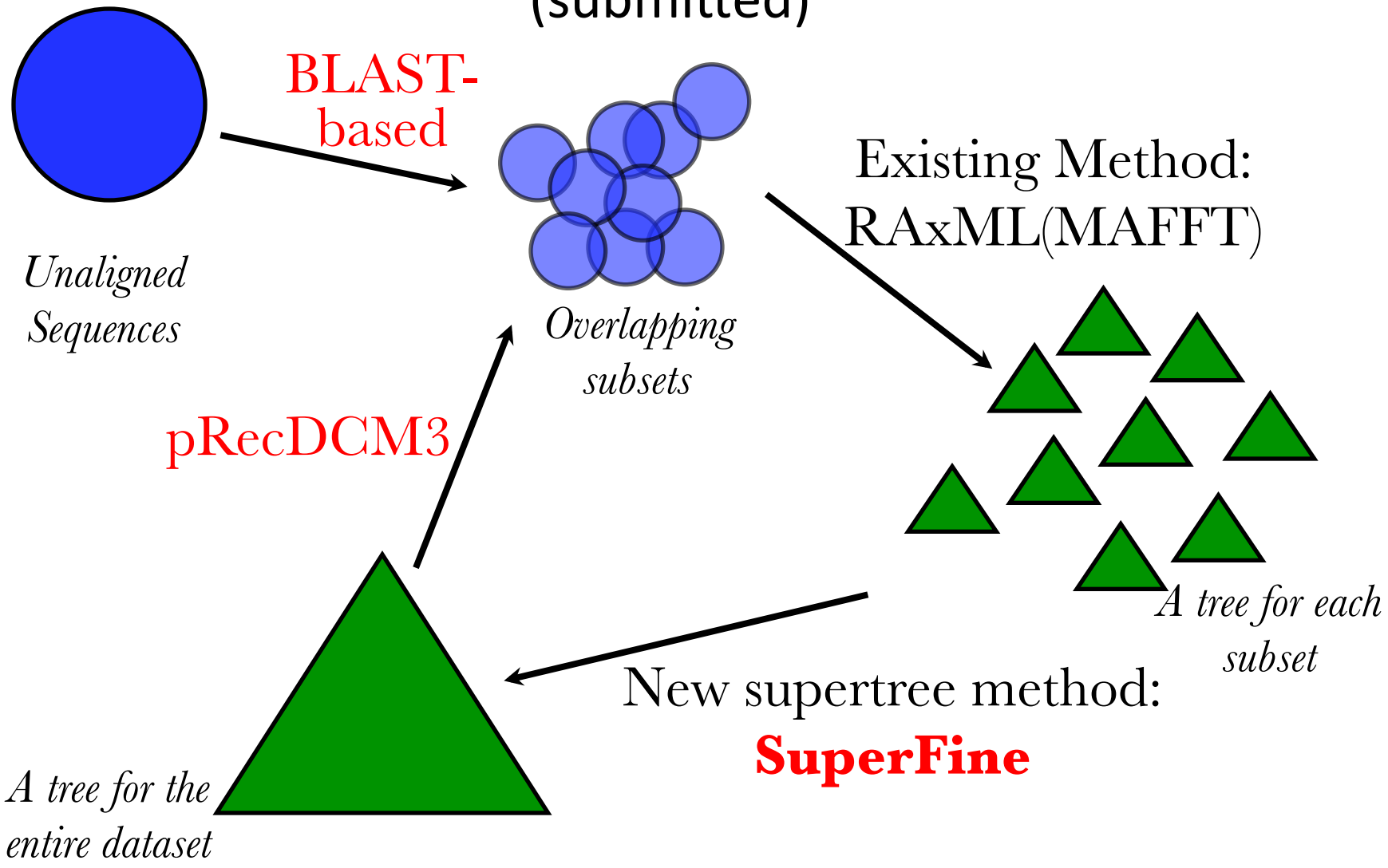
Proofs about the algorithm guarantees are established using chordal graph theory.

Chordal graph algorithms yield phylogeny estimation from *polynomial length* sequences



- Theorem (Warnow et al., SODA 2001): DCM1-NJ correct with high probability given sequences of length $O(\ln n\, e^{O(\ln n)})$
- Simulation study from Nakhleh et al. ISMB 2001

DACTAL: divide-and-conquer trees without alignment (submitted)



DACTAL more accurate than all standard methods, and much faster than **SATé**

Average results on 3 large RNA datasets (6K to 28K)

CRW: Comparative RNA database, structural alignments

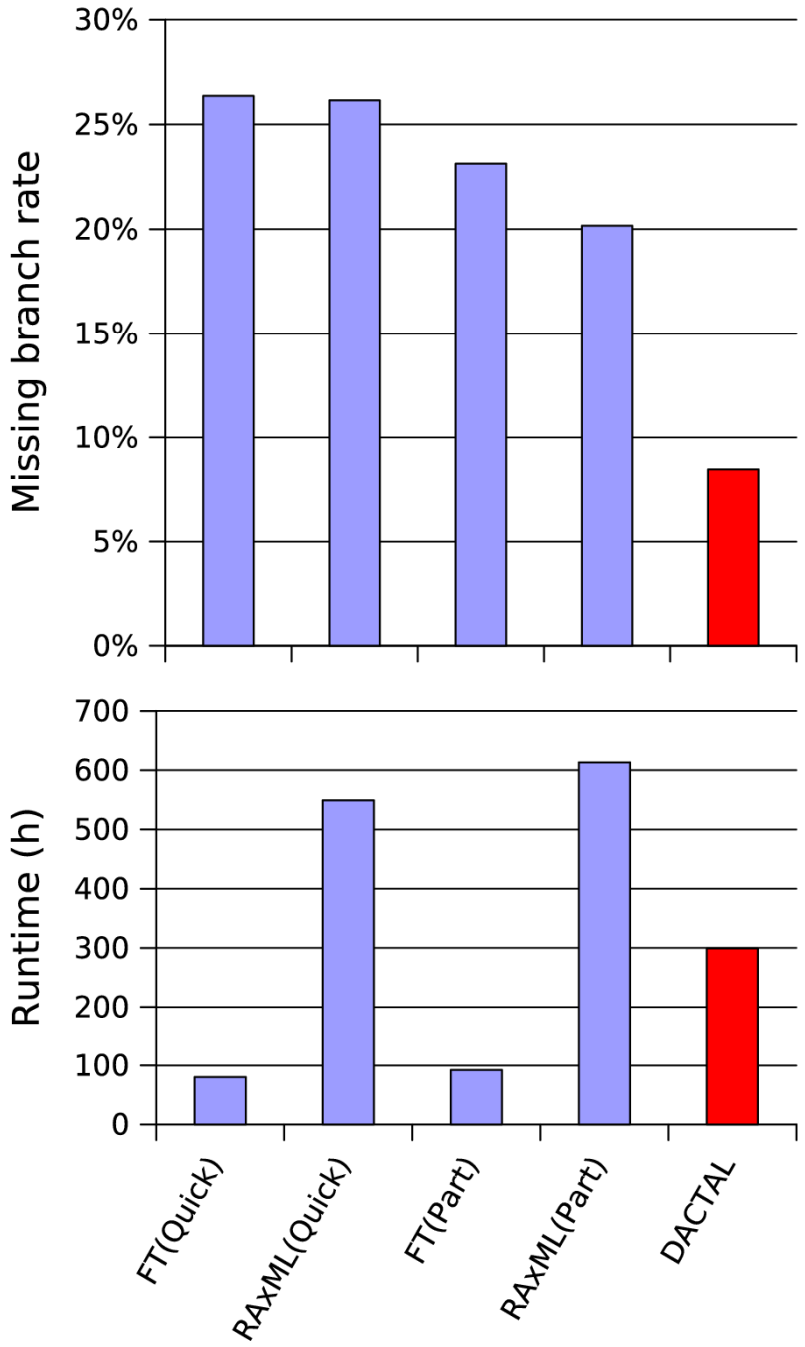
3 datasets with 6,323 to 27,643 sequences

Reference trees: 75% RAxML bootstrap trees

DACTAL (shown in red) run for 5 iterations starting from FT(Part)

SATé-1 fails on the largest dataset

SATé-2 runs but is not more accurate than DACTAL, and takes longer



More divide-and-conquer

- Recall that SATe and PASTA use divide-and-conquer (and also iteration) to improve alignment estimation.
- Alignments on different subsets are merged together using techniques like OPAL and Muscle.
- However, alignments can also be merged together using HMM-HMM alignment.
- You should think of your own algorithmic designs for improving scalability and accuracy, whether for MSA or for tree estimation!