High-performance phylogeny reconstruction

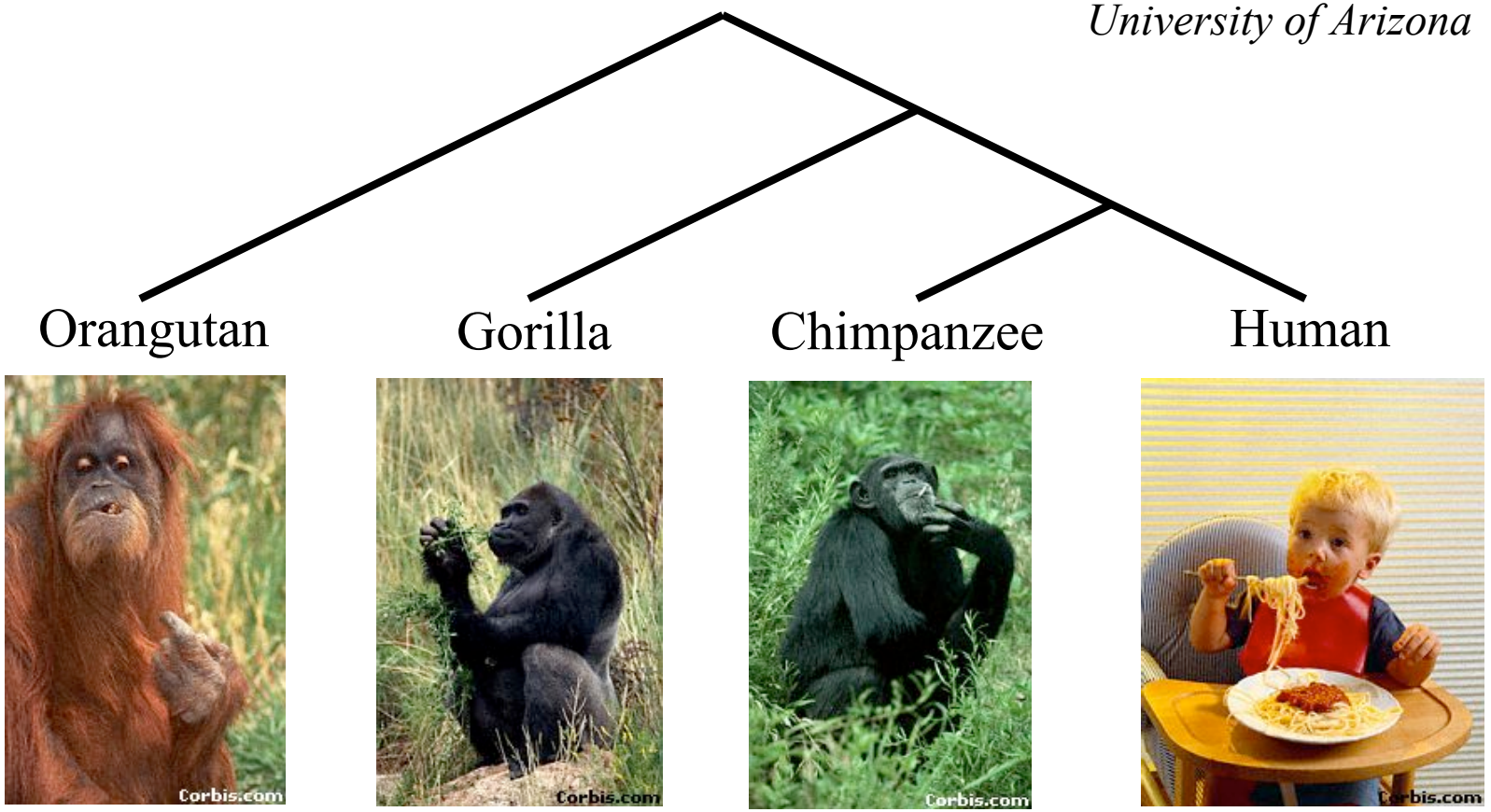
Tandy Warnow

Radcliffe Institute for Advanced Study

Center for Computational Biology and Bioinformatics, UT-Austin

Phylogeny

From the Tree of the Life Website,
University of Arizona



Evolution informs about everything in biology

- Big genome sequencing projects just produce data -- so what?
- Evolutionary history relates all organisms and genes, and helps us understand and predict
  - interactions between genes (genetic networks)
  - drug design
  - predicting functions of genes
  - influenza vaccine development
  - origins and spread of disease
  - origins and migrations of humans

CIPRES Project

- Cyber Infrastructure for Phylogenetic Research
- Funded by $11.6M ITR (Information Technology) Grant from NSF
- 5 lead institutions: UNM, UT-Austin, Florida State University, UC Berkeley, and UC San Diego, plus 8 other institutions
- Purpose: to create a national infrastructure of hardware, algorithms, database technology, etc., necessary to infer the Tree of Life
- 33 biologists, computer scientists, and mathematicians collaborating on the project

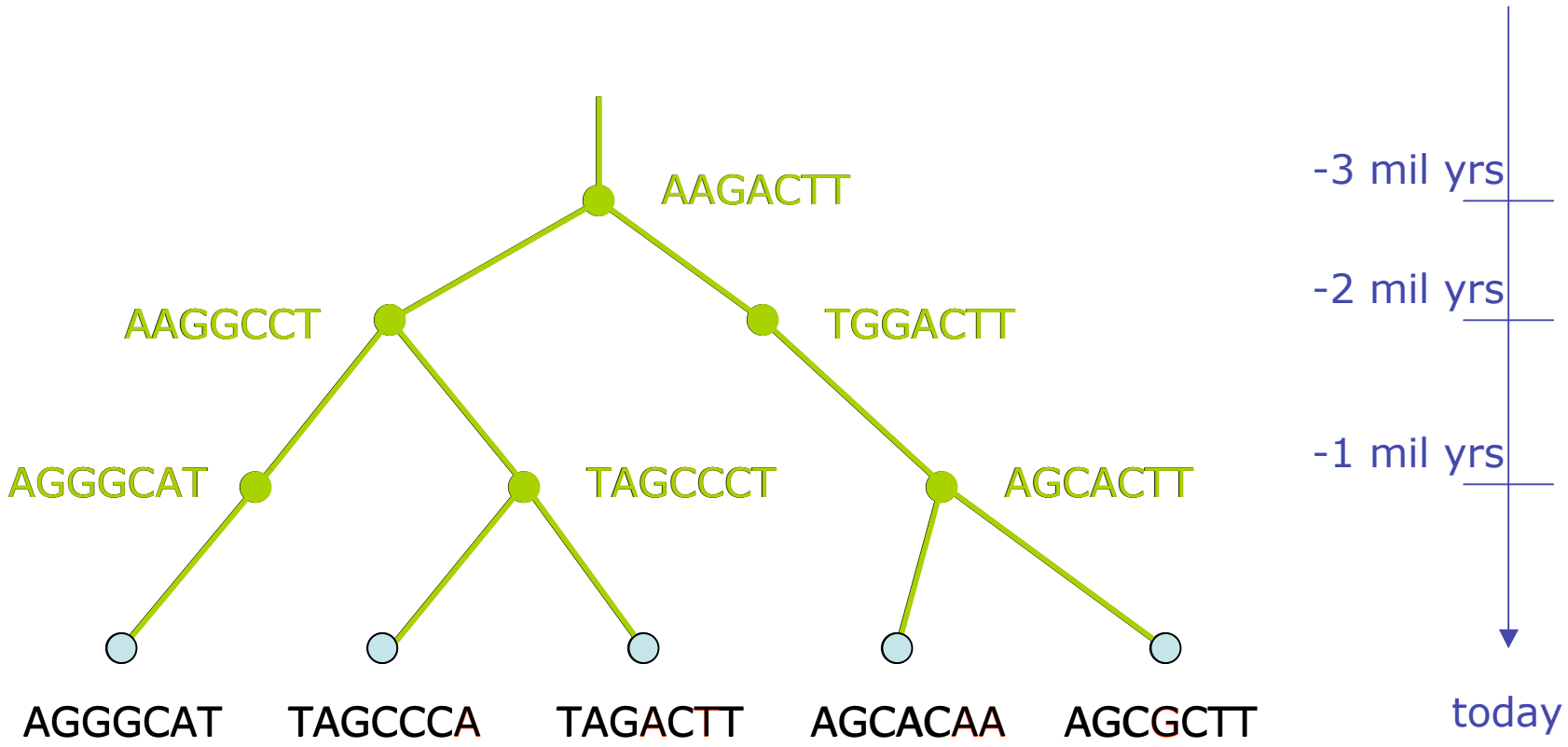
Projects

- Meta-methods for Maximum Parsimony and Maximum Likelihood: Bernard Moret, Usman Roshan, and Tiffani Williams
- Reticulate evolution: Randy Linder, Bernard Moret, Luay Nakhleh
- Gene Order Phylogeny: Bernard Moret, Jijun Tang, Li-San Wang, Bob Jansen, and Linda Raubeson
- Fast Converging Methods: Bernard Moret, Usman Roshan, Luay Nakhleh, and Katherine St. John
- Visualizing large trees: Nina Amenta, Katherine St John, Randy Linder, Bob Jansen, and David Hillis

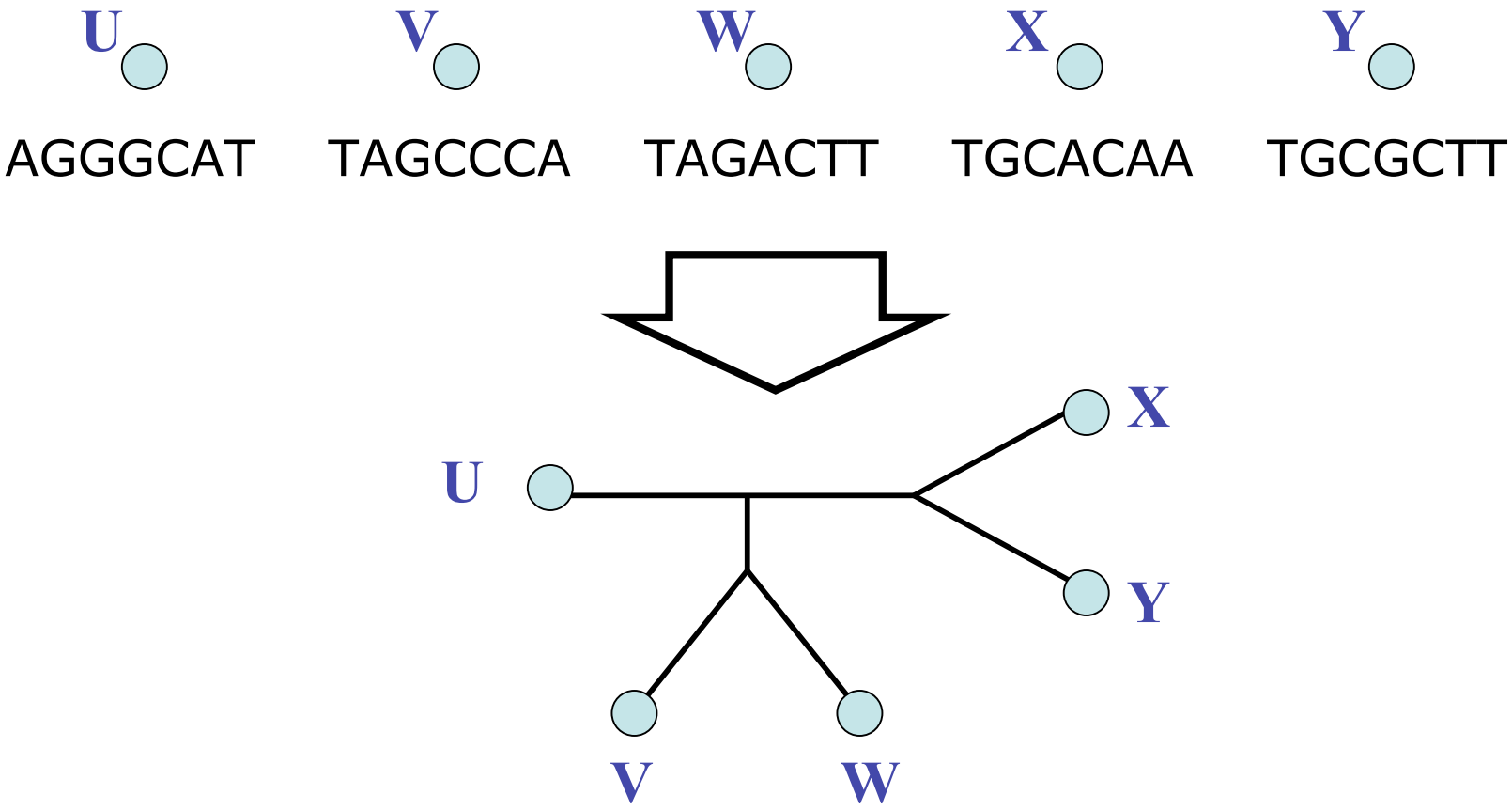
This talk

- IDCM3: a meta-method for speeding up base methods for maximum parsimony or maximum likelihood
- SpNet: a new approach for inferring reticulate evolutionary histories
- GRAPPA: software for gene order phylogeny reconstruction

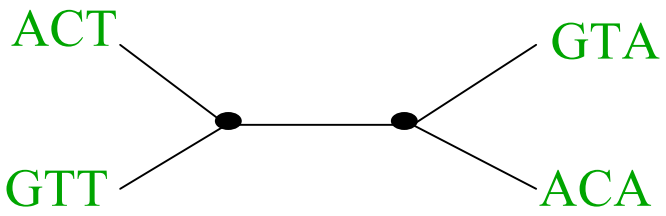
DNA Sequence Evolution

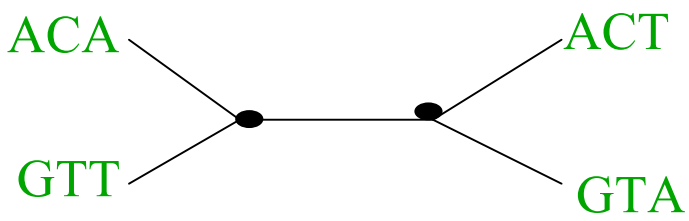


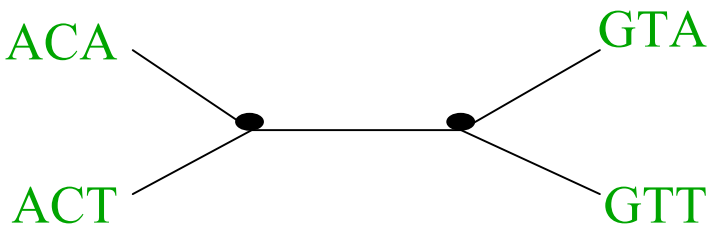
Molecular Systematics



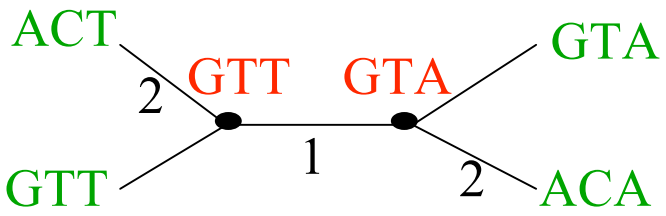
Maximum Parsimony



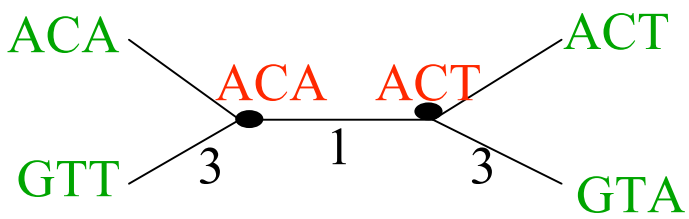




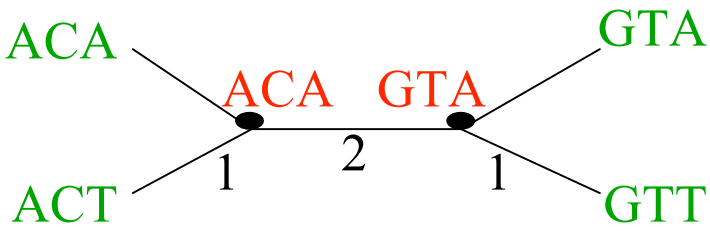
Maximum Parsimony



MP score = 5



MP score = 7

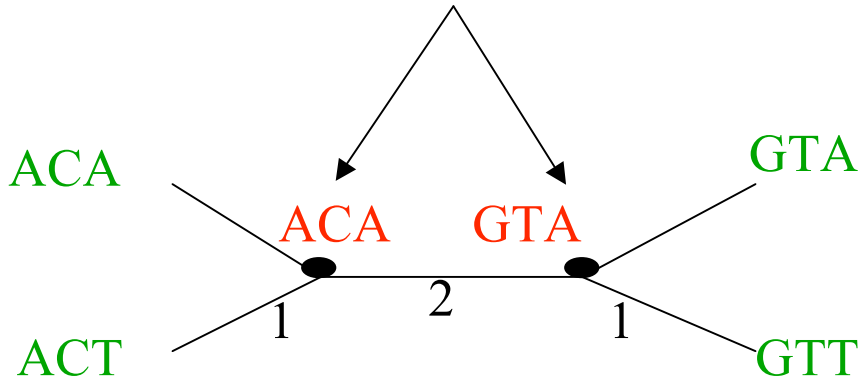


MP score = 4

Optimal MP tree

Maximum Parsimony: computational complexity

Optimal labeling can be computed in linear time O(nk)

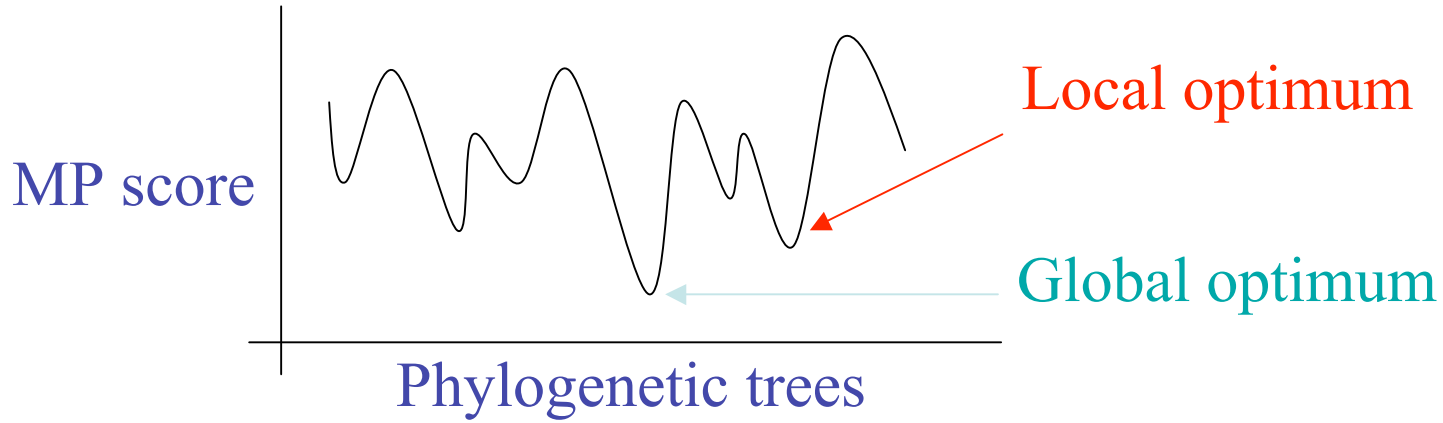


MP score = 4

Finding the optimal MP tree is **NP-hard**

Solving MP (maximum parsimony) and ML (maximum likelihood)

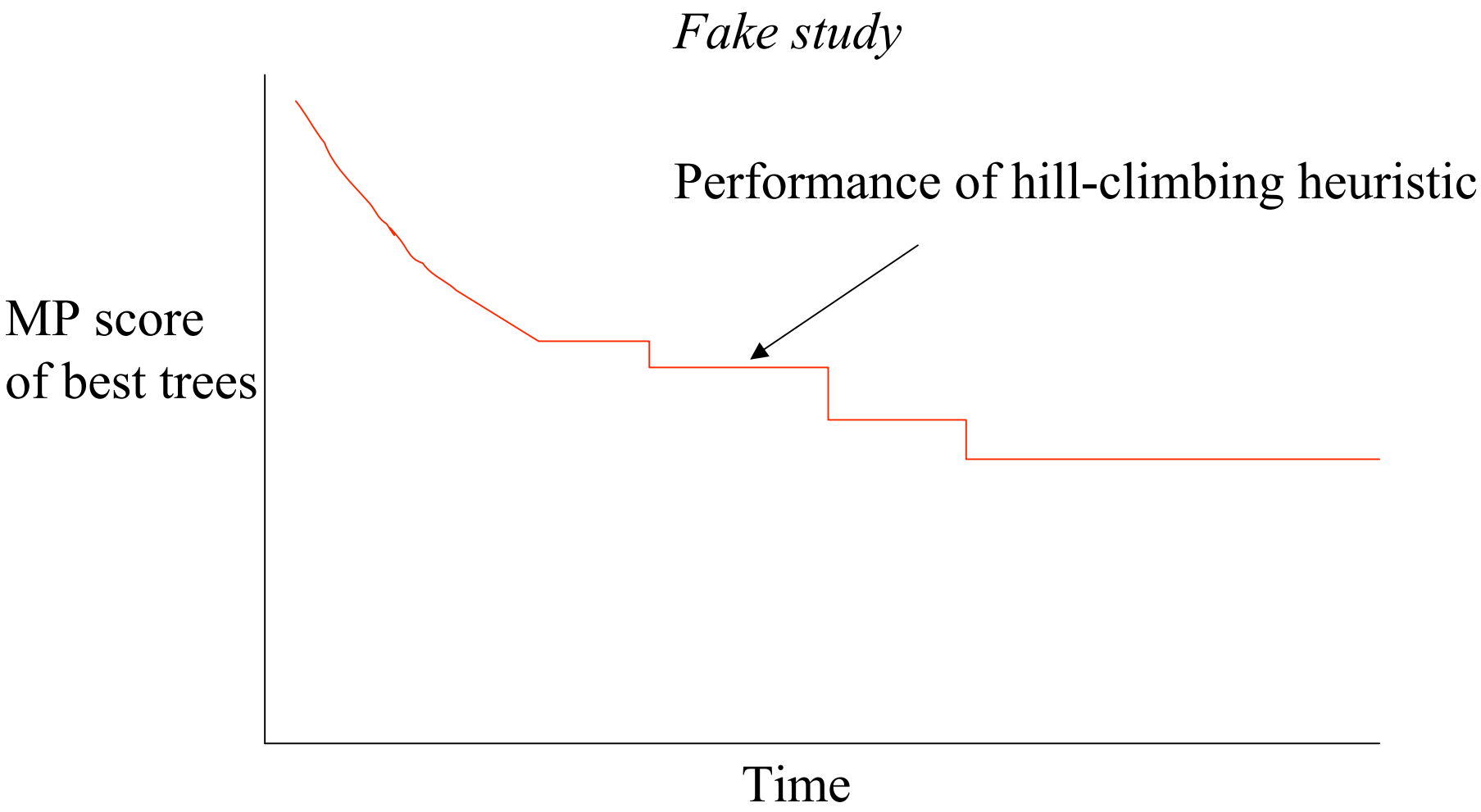
- Why are MP and ML hard? The search space is huge -- there are (2n-5)!! trees, it is easy to get stuck in local optima, and there can be many optimal trees.
- Why try to solve MP or ML? Our experimental studies show that polynomial time algorithms don't do as well as MP or ML when trees are big and have high rates of evolution.
- Why solve MP and ML well? Because trees can change in biologically significant ways with small changes in objective criterion. (Open problem!)



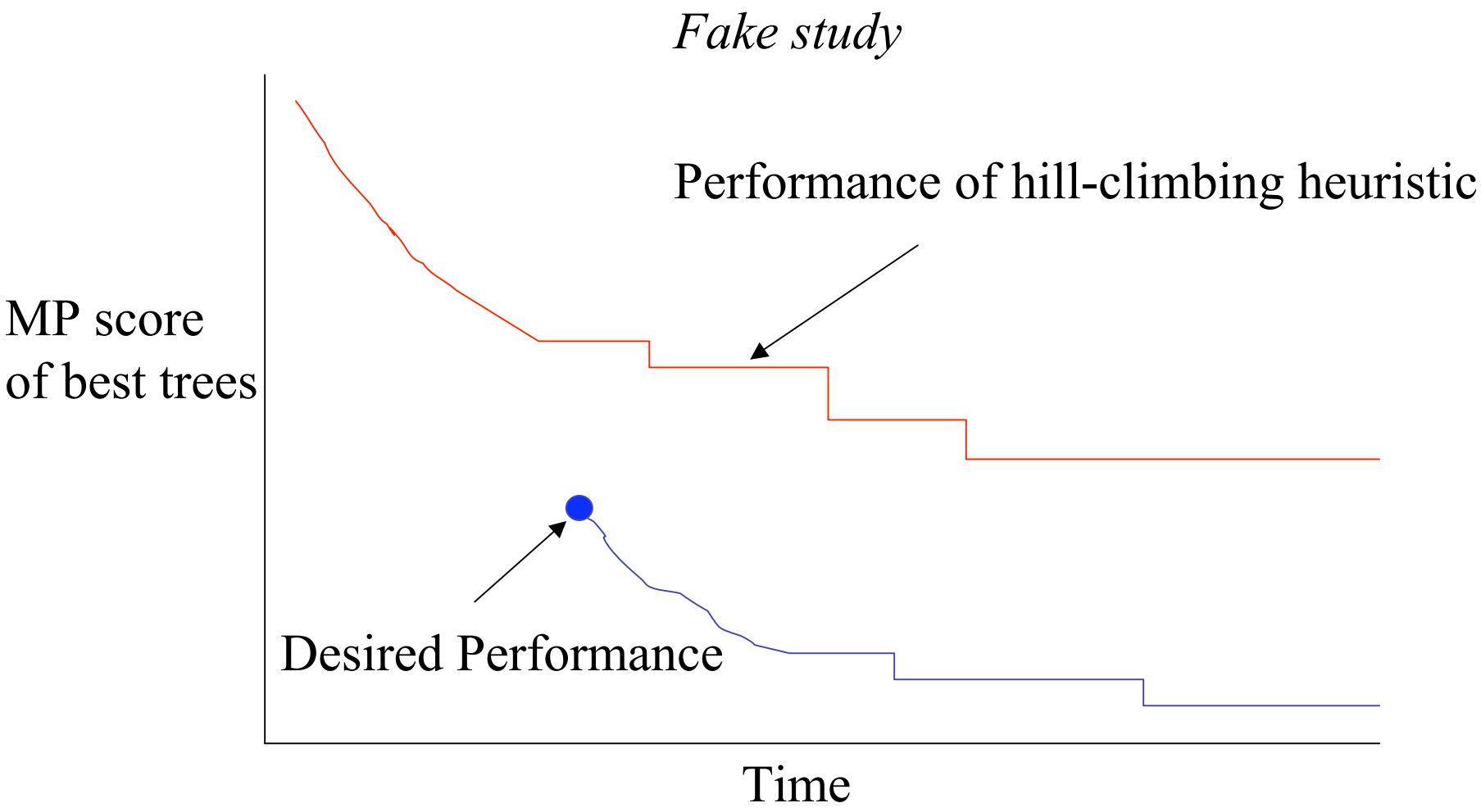
Current software for solving MP

- PAUP*4.0: Popular phylogeny software package which implements local search
- TNT: More recent software package which implements very fast tree searching routines
- And many more…

MP/ML heuristics



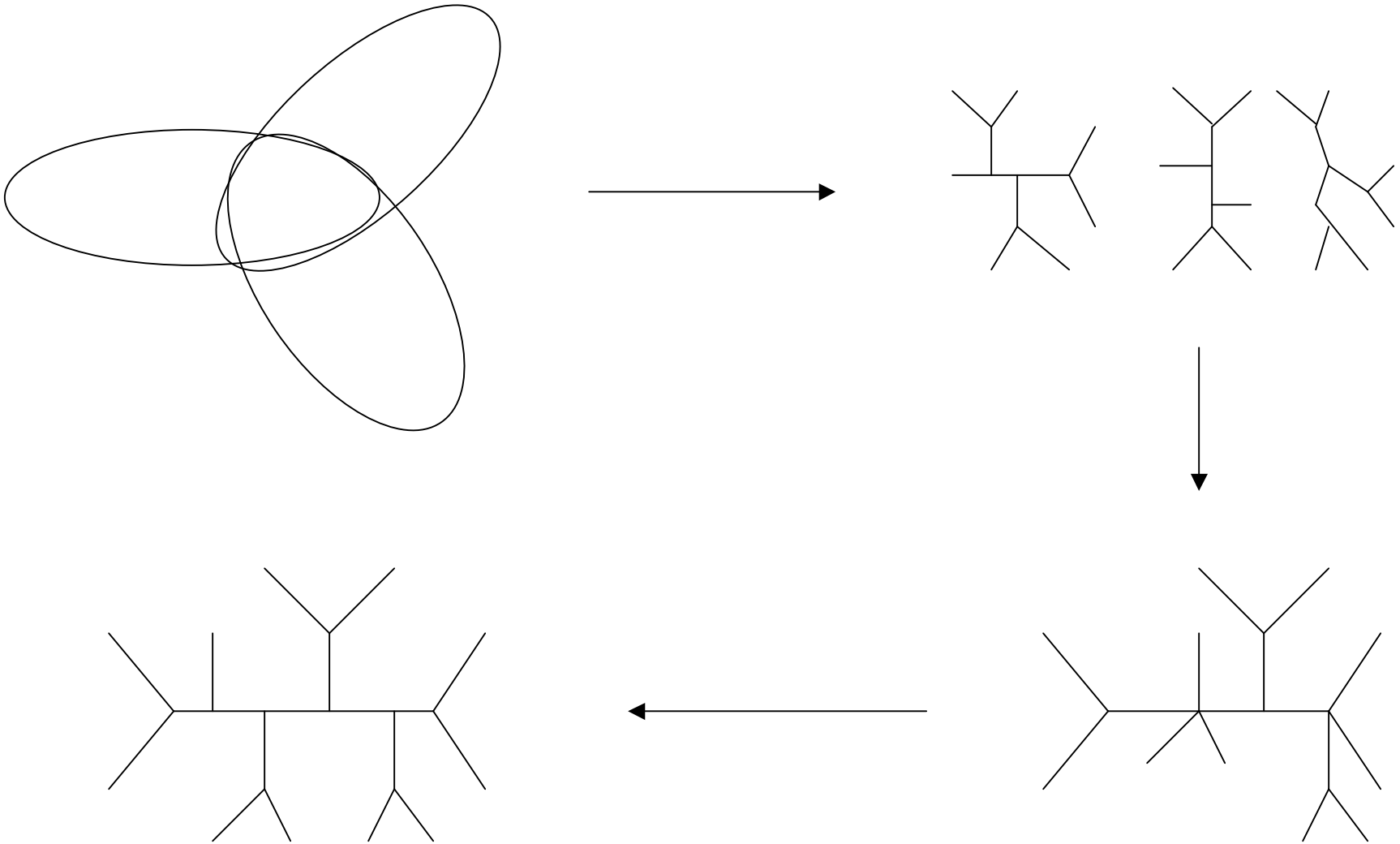
Speeding up MP/ML heuristics



Using divide-and-conquer for MP and ML

- Conjecture: better (more accurate) solutions will be found in less time, if we analyze a small number of smaller subsets and then combine solutions
- Need:
  - 1. techniques for decomposing datasets,
  - 2. base methods for subproblems, and
  - 3. techniques for combining subtrees

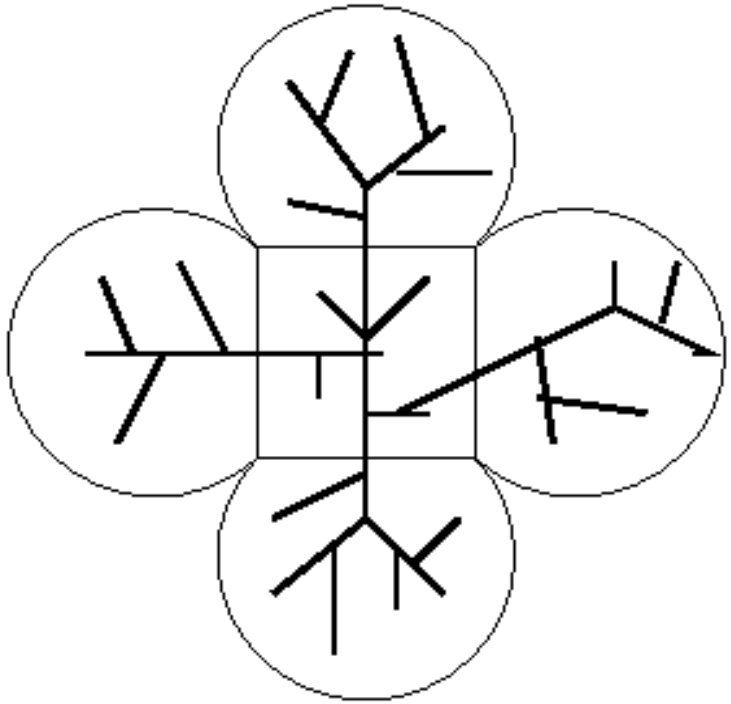
The DCM3 technique for speeding up MP/ML searches



DCM3 Decompositions

Input: Set *S* of sequences, and guide-tree T

1. Compute "short subtree" graph G(S,T), based upon T
2. Find clique separator in the graph G(S,T), and form subproblems

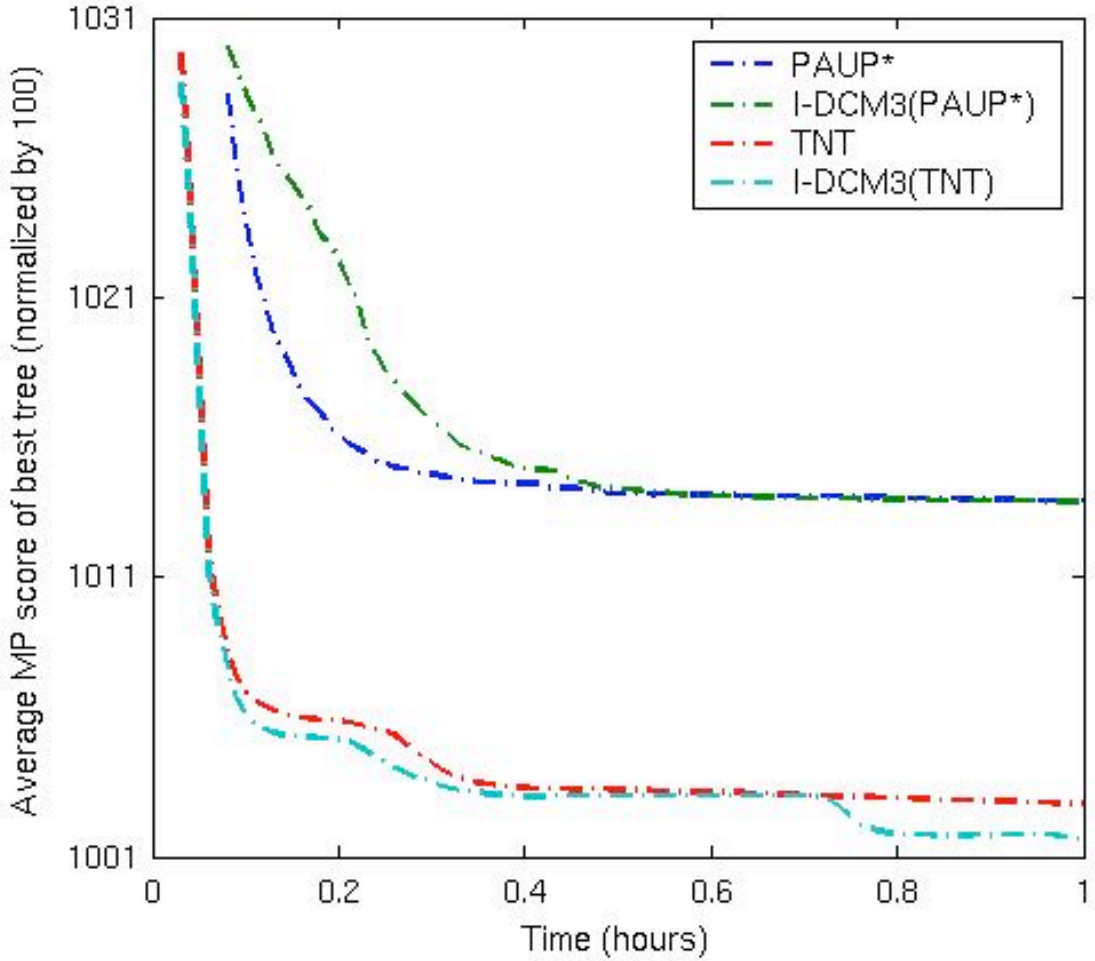


IDCM3 boosting of current techniques

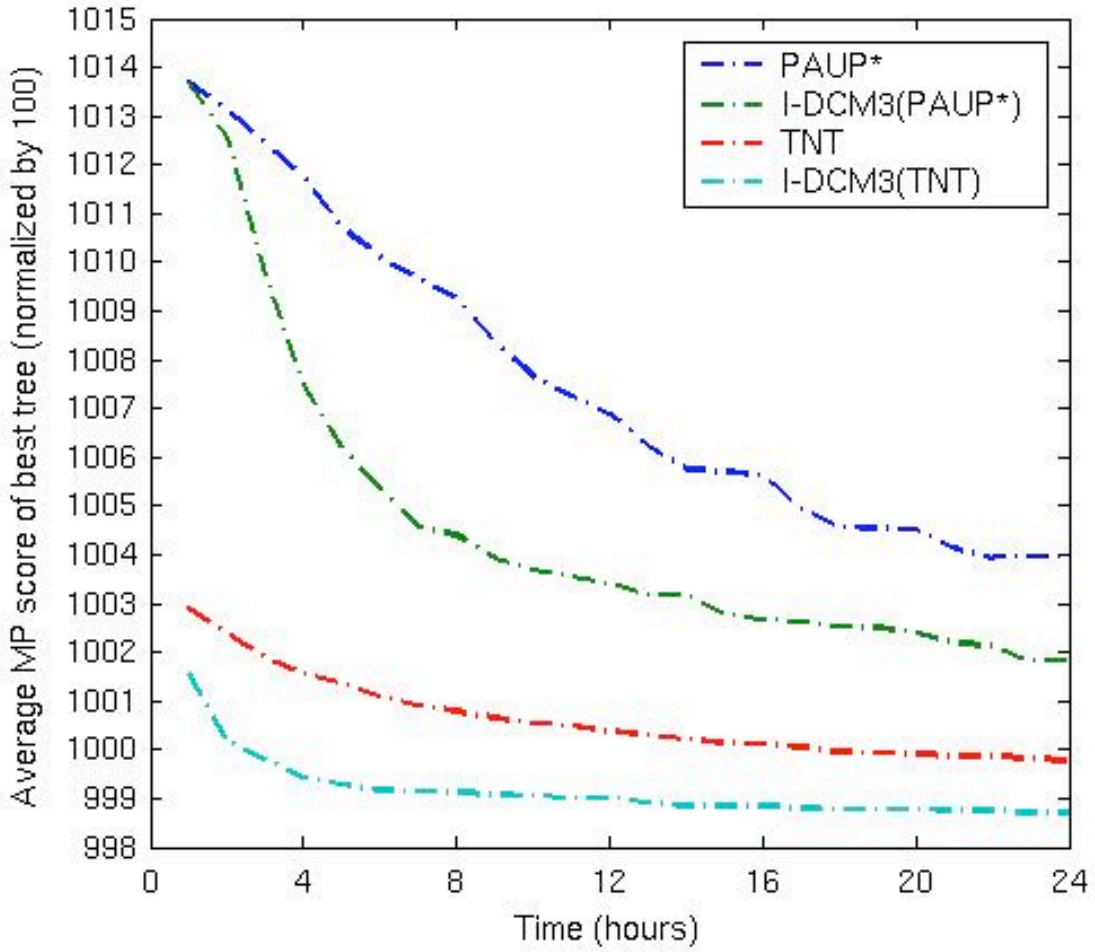
Datasets

- 429 Eukaryotes rDNA (Lipscomb et. al.)
- 576 Metazoa DNA (Goloboff)
- 500 rbcL DNA (Rice et. al.)
- 567 rbcL, atpb, and 18s DNA (Soltis et. al.)
- 854 rbcL DNA (Goloboff)
- 921 Avian Cytochrome DNA (Johnson)
- 2000 Eukaryotes sRNA (Gutell et. al.)
- 2594 rbcL DNA (Kallersjo et. al.)
- 7180 RNA (Gutell et. al.)
- 8506 RNA (Gutell et. al.)

Dataset of 8,506 RNA sequences (first hour)



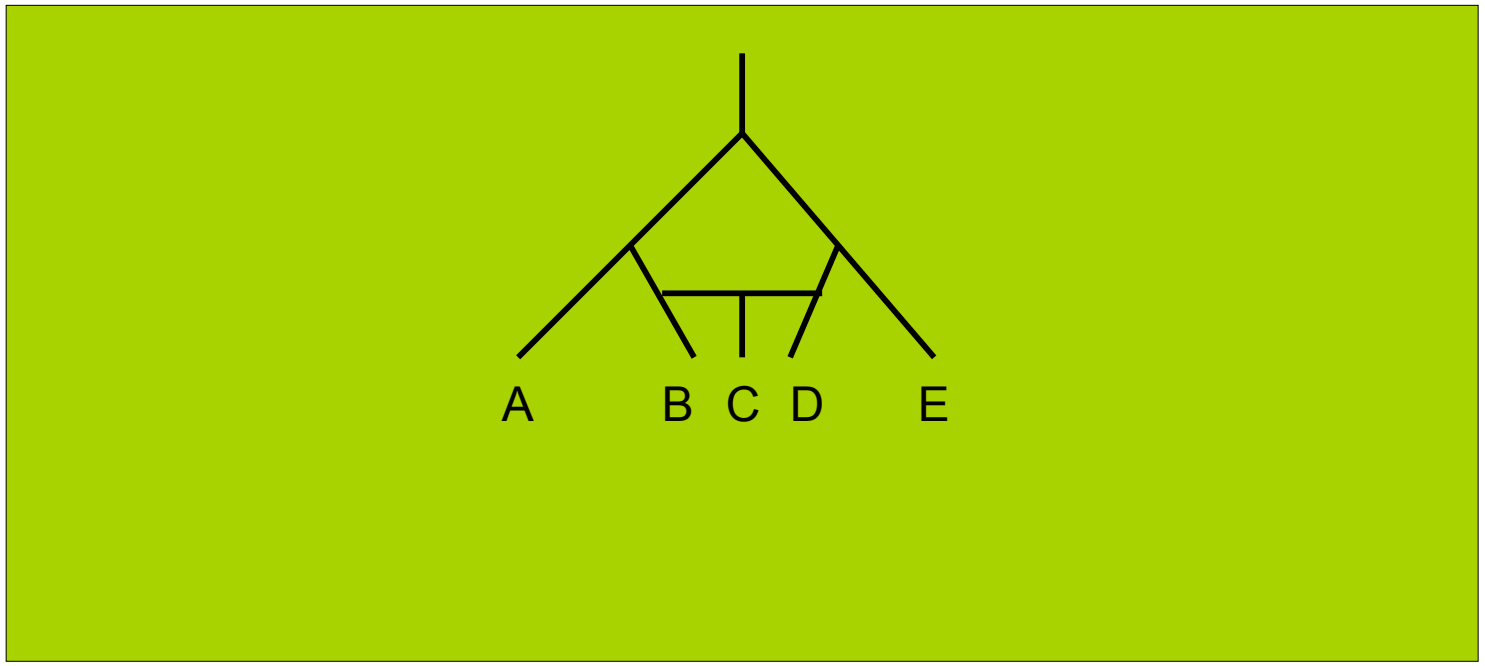
Dataset of 8,506 RNA sequences (first 24 hours)



Summary of IDM3-boosting

- For MP: IDCM3 (so far) has improved upon all base methods for almost all datasets. The harder the problem, or the harder the dataset, the bigger the improvement. (Improving a slow method is easiest -- improving TNT searches turns out to be harder.)
- For ML: our limited study (based upon PAUP* searches) show dramatic improvements
- Much still needs to be done!

Reticulate evolution: hybridizing speciation



Why Networks?

- Lateral gene transfer (LGT)
  - Ochman estimated that 755 of 4,288 ORF's in E.coli were from at least 234 LGT events
- Hybridization
  - Estimates that as many as 30% of all plant lineages are the products of hybridization
  - Fish
  - Some frogs

Reconstructing Phylogenetic Networks

Main question: to combine, or not to combine?

Separate analysis:

- Analyze individual genes separately
- Reconcile the resulting phylogenies

Combined analysis:

- Combine (via concatenation) the datasets, and attempt to infer the evolutionary history

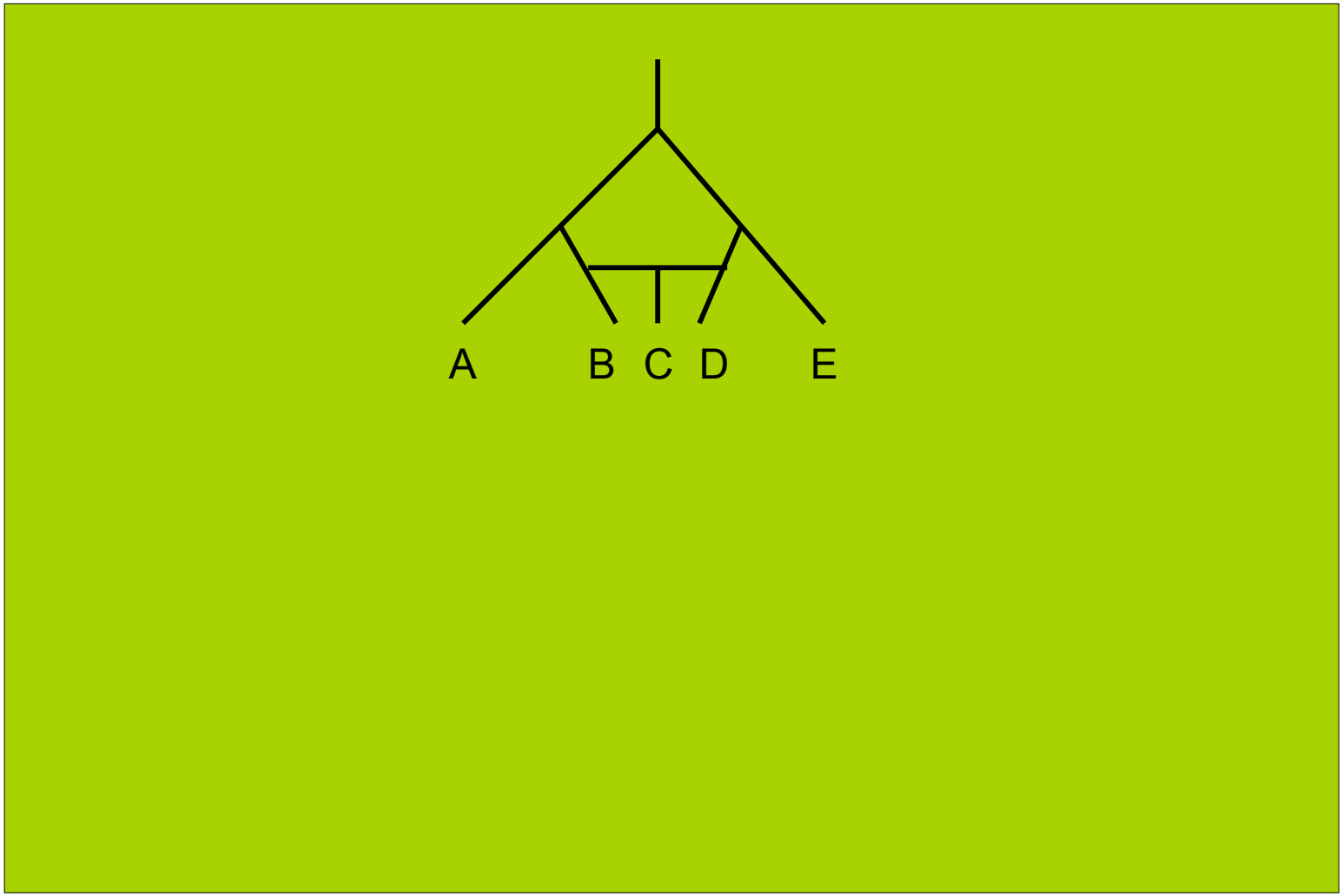
Wayne Maddison's Observation

Syst. Biol., 46(3):523-536, 1997

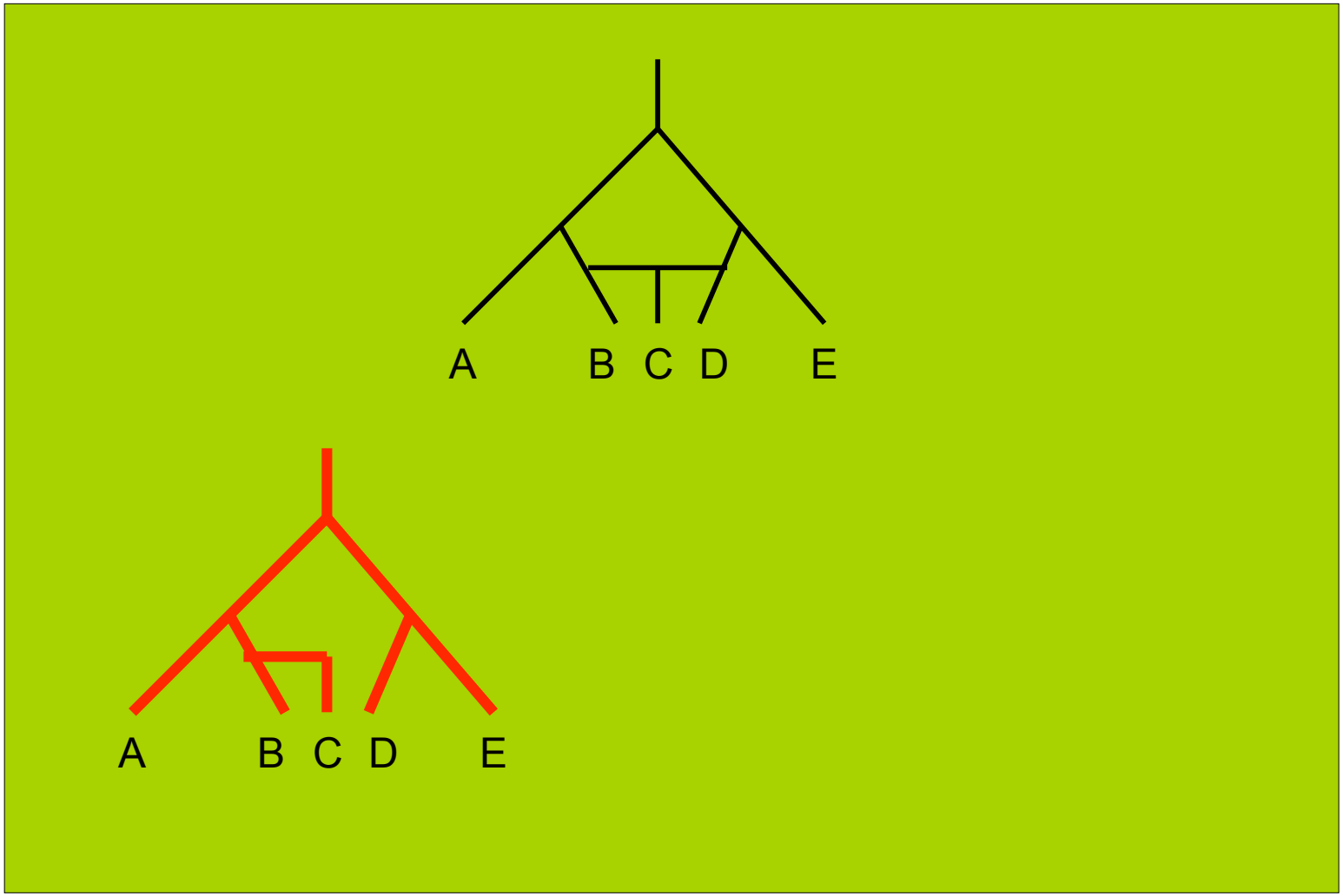
To paraphrase Wayne Maddison:

Genes evolve down trees contained within the network describing the evolutionary history, and so reconstructing phylogenetic networks can be done by combining individual gene trees.

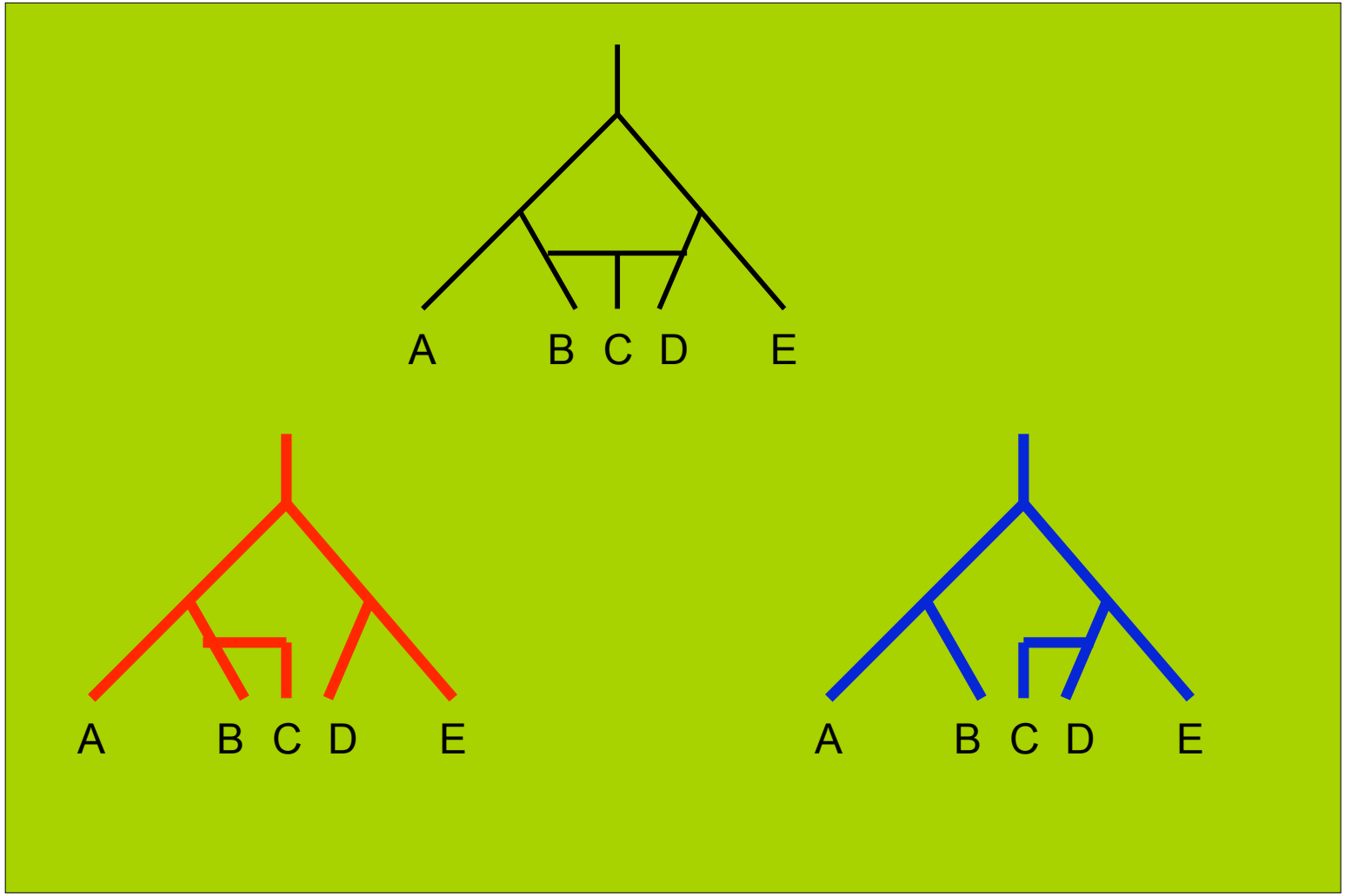
Species Networks



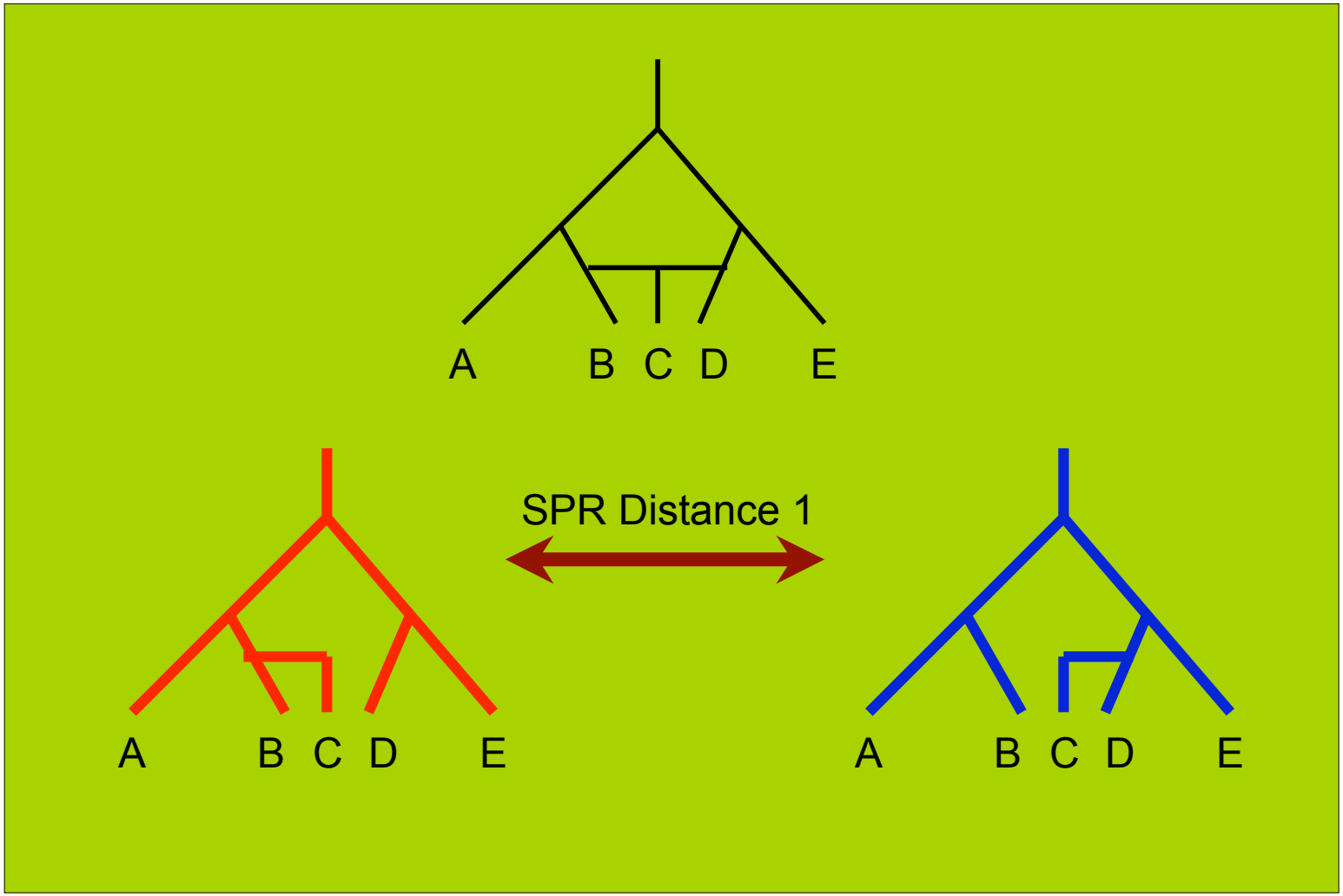
Gene Tree I in Species Networks



Gene Tree II in Species Networks



SPR Distances Among Gene Trees



Maddison's Method

Given two gene datasets

- Construct two gene trees T1 and T2
- If SPR(T1,T2)=0
  - Return a tree
- If SPR(T1,T2)=1
  - Return a network with one reticulation event

Open problem: extend to reconstructing a network with m reticulation events

Challenges

(1) Computational

– Computing SPR distances is of unknown computational complexity (probably hard)

Solving the Computational Challenge

- Galled-tree (GT) networks: reticulation events are independent
- Given two gene trees T1 and T2 on n leaves from a GT-network M with m reticulations, we can find the network N in O(mn) time (and it is unique)

Challenges

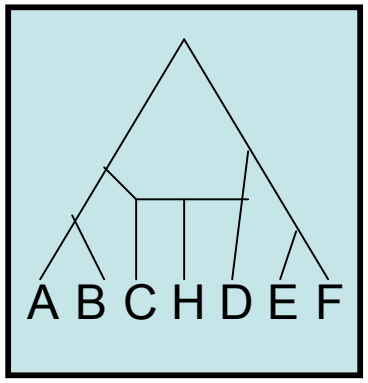
(2) Systematic

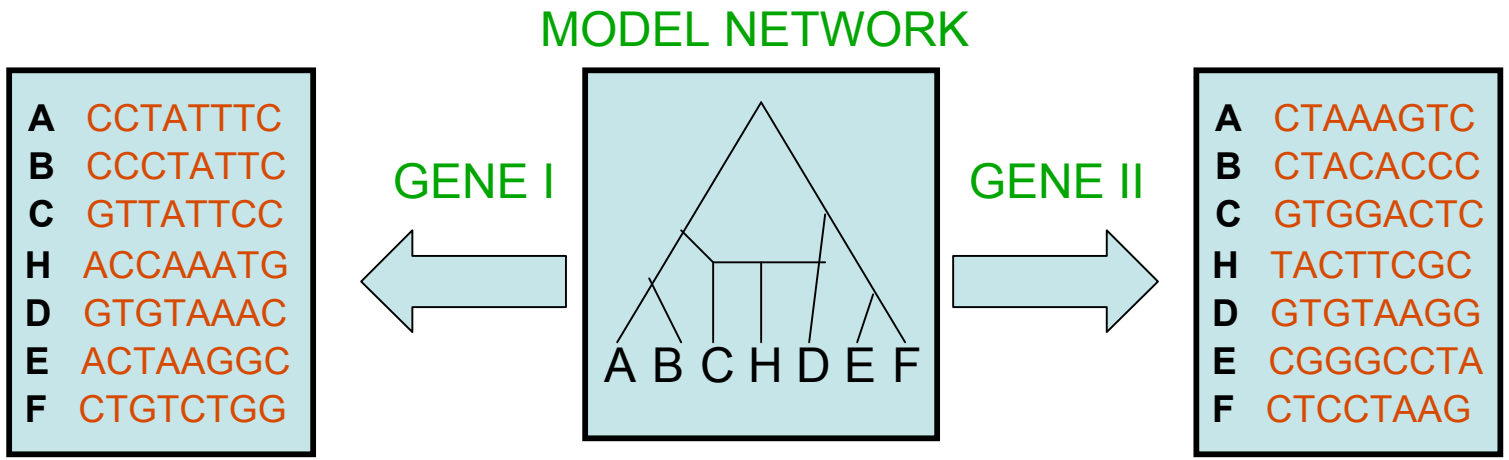
- Obtaining the correct gene trees in practice is very hard (due to missing data, inaccuracy of tree reconstruction methods, wrong assumptions, etc.)

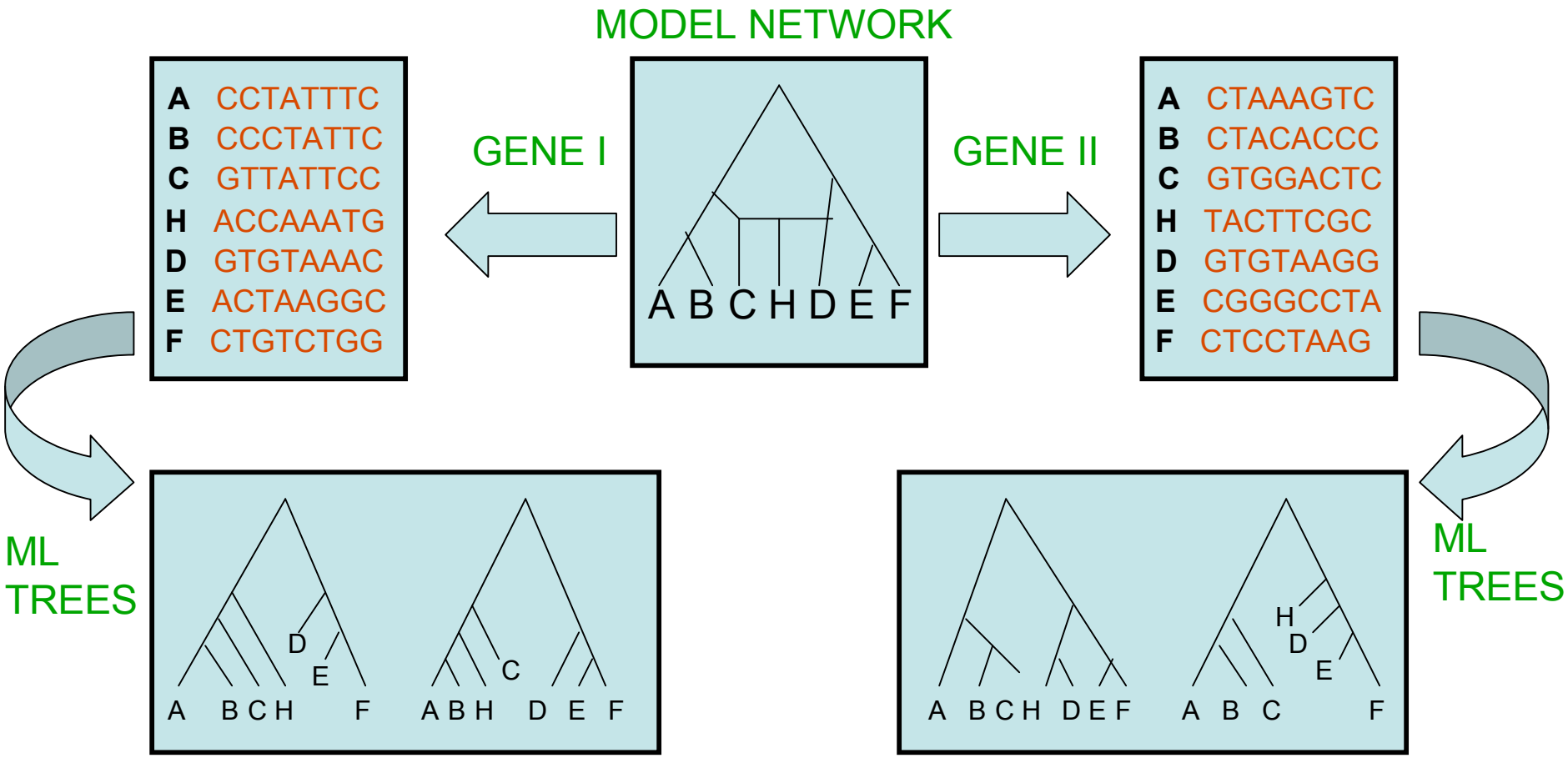
OUR METHOD

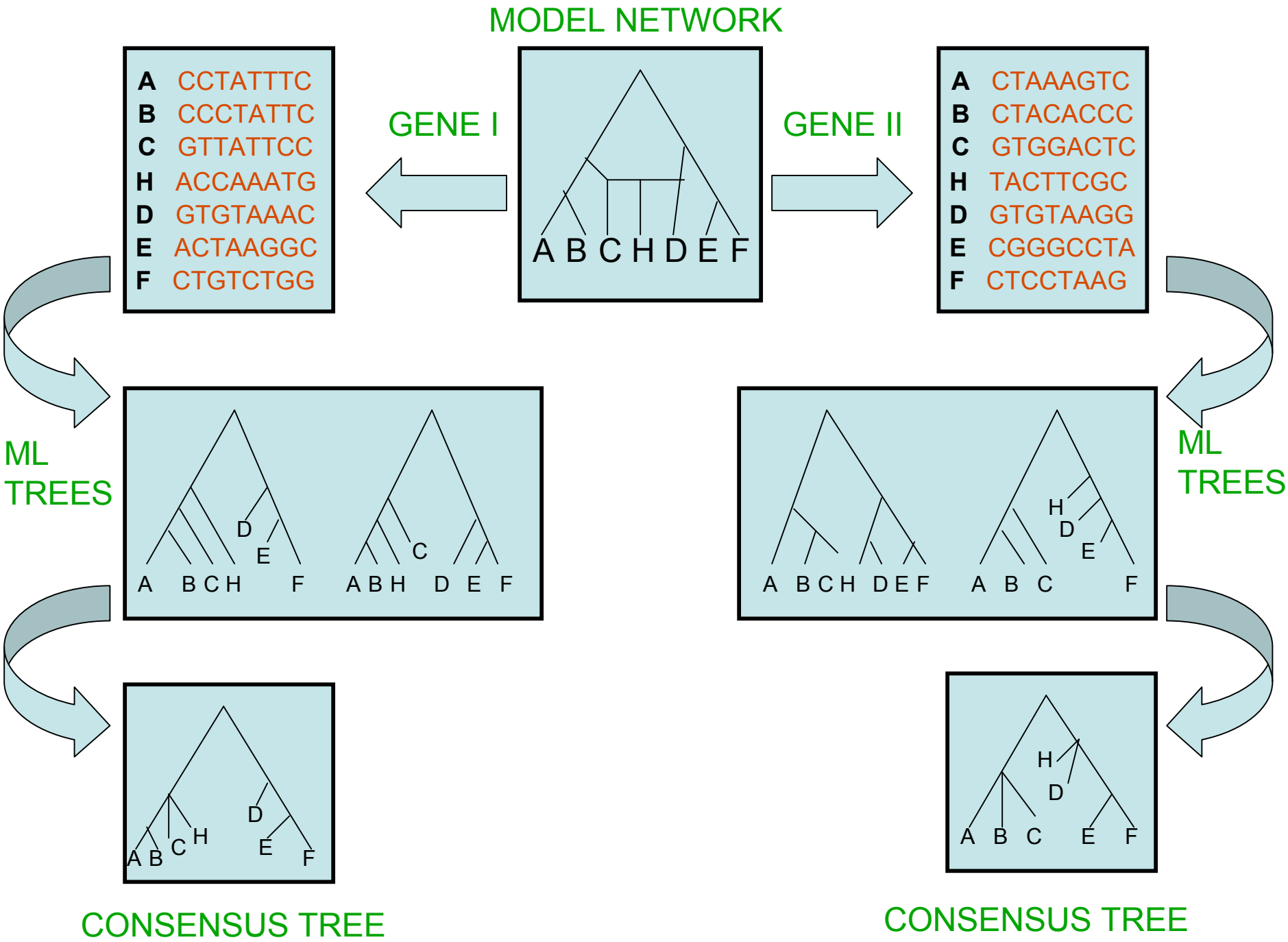
SpNet

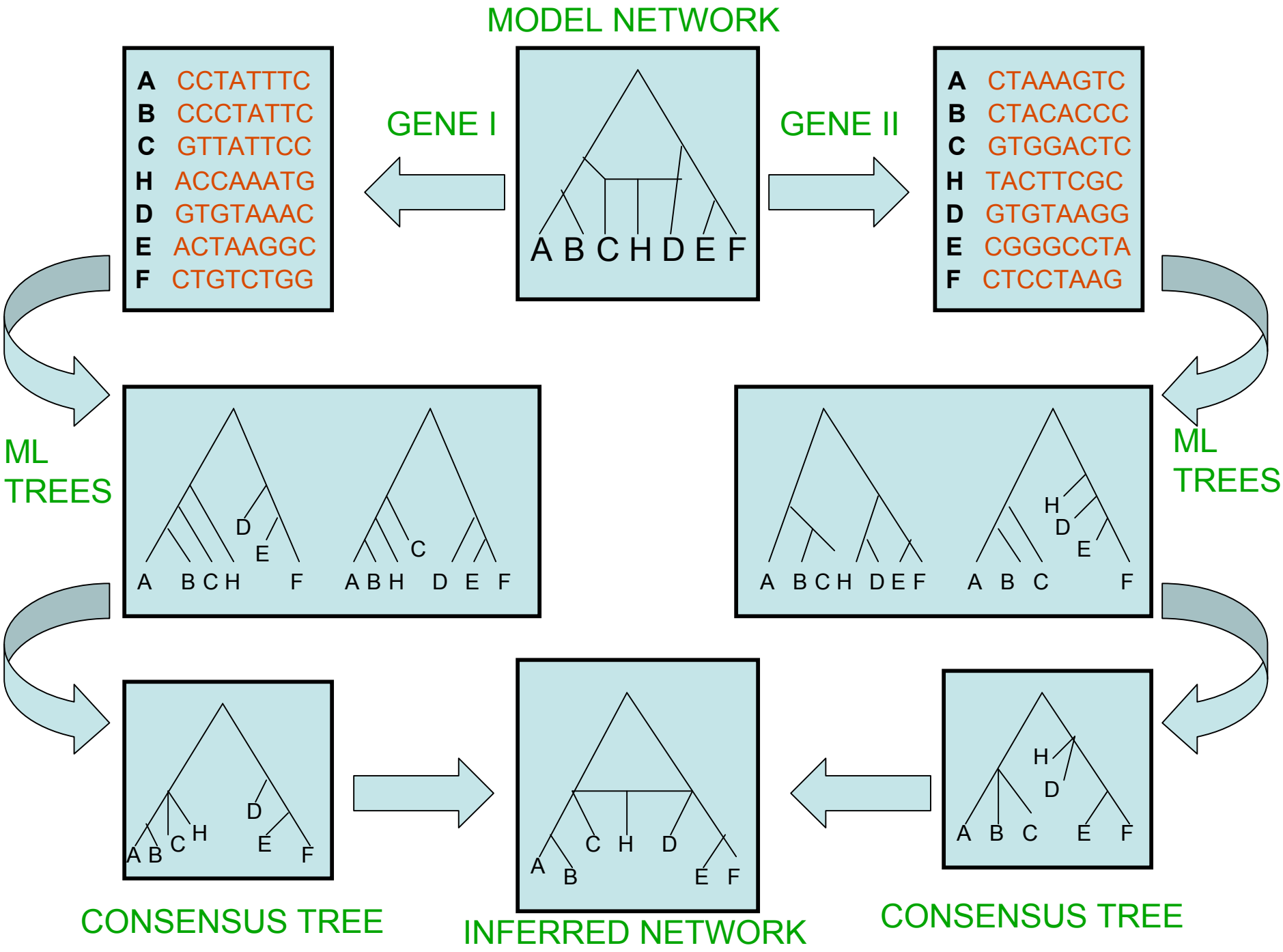
MODEL NETWORK




MODEL NETWORK
A CCTATTTC
B CCCTATTC
C GTTATTCC
H ACCAAATG
D GTGTAAAC
E ACTAAGGC
F CTGTCTGG
GENE I
A B C H D E F
GENE II
A CTAAAGTC
B CTACACCC
C GTGGACTC
H TACTTCGC
D GTGTAAGG
E CGGGCCTA
F CTCCTAAG


MODEL NETWORK
A CCTATTTC
B CCCTATTC
C GTTATTCC
H ACCAAATG
D GTGTAAAC
E ACTAAGGC
F CTGTCTGG
GENE I
A B C H D E F
GENE II
A CTAAAGTC
B CTACACCC
C GTGGACTC
H TACTTCGC
D GTGTAAGG
E CGGGCCTA
F CTCCTAAG
ML TREES
A B C H D E F
A B H C D E F
A B C H D E F
A B C H D E F
ML TREES


MODEL NETWORK
A CCTATTTC
B CCCTATTC
C GTTATTCC
H ACCAAATG
D GTGTAAAC
E ACTAAGGC
F CTGTCTGG
GENE I
A B C H D E F
GENE II
A CTAAAGTC
B CTACACCC
C GTGGACTC
H TACTTCGC
D GTGTAAGG
E CGGGCCTA
F CTCCTAAG
ML TREES
A B C H D E F
A B H C D E F
A B C H D E F
A B C H D E F
ML TREES
A B C H D
A B C D E F
CONSENSUS TREE
A B C H D E F
CONSENSUS TREE


MODEL NETWORK
A CCTATTTC
B CCCTATTC
C GTTATTCC
H ACCAAATG
D GTGTAAAC
E ACTAAGGC
F CTGTCTGG
GENE I
A B C H D E F
GENE II
A CTAAAGTC
B CTACACCC
C GTGGACTC
H TACTTCGC
D GTGTAAGG
E CGGGCCTA
F CTCCTAAG
ML TREES
A B C H D E F
A B H C D E F
ML TREES
A B C H D E F
A B C H D E F
CONSENSUS TREE
A B C H D E F
INFERRED NETWORK
A B C H D E F
CONSENSUS TREE
A B C H D E F

SpNet: Running Time

- We have a linear-time algorithm for the single hybrid case (implementation and experimental results are available as well)
- We are working on the general case of arbitrary number of reticulation events

Experimental Study

- Generated random networks on 10 and 20 taxa, with 0, 1, and 2 hybrids
- Evolved sequences under the GTR+Gamma model of evolution with invariant sites
- We studies the topological accuracy based on the splits defined by the model and inferred network

Evaluation Criteria

What is the topological accuracy of the inferred phylogeny?

- False positives (splits returned that aren't in the model phylogeny)
- False negatives (splits in the model phylogeny that are missing in the inferred phylogeny)

Methods

- SpNet: Do ML analysis of each dataset, and compute the strict consensus of the best two trees. Compare these consensus trees.

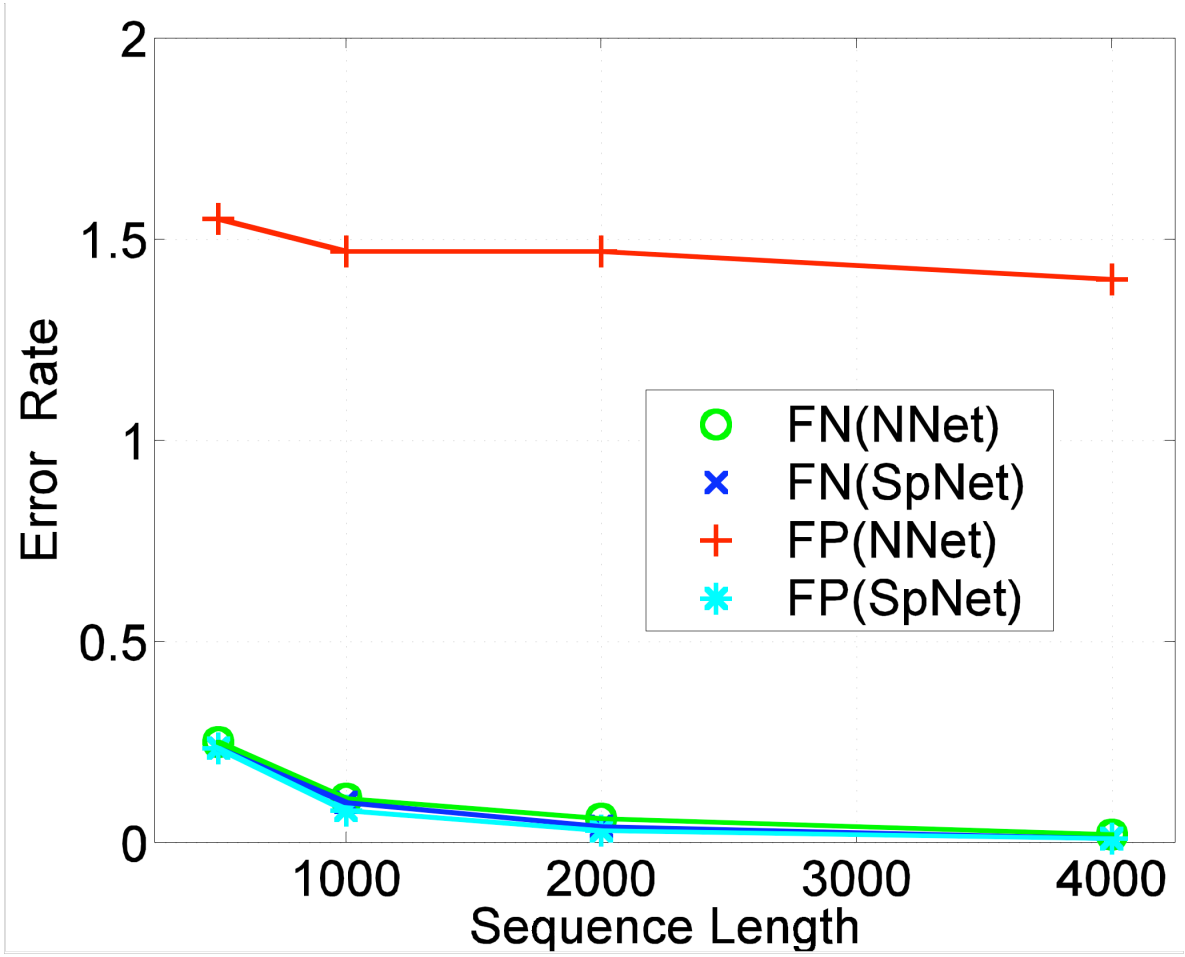
- Maddison: Do ML analysis, and compare best two trees.

- NNet: The method of Bryant and Moulton (combines agglomerative clustering technique from NJ with splits-graph representations of Bandelt and Dress).

Observations

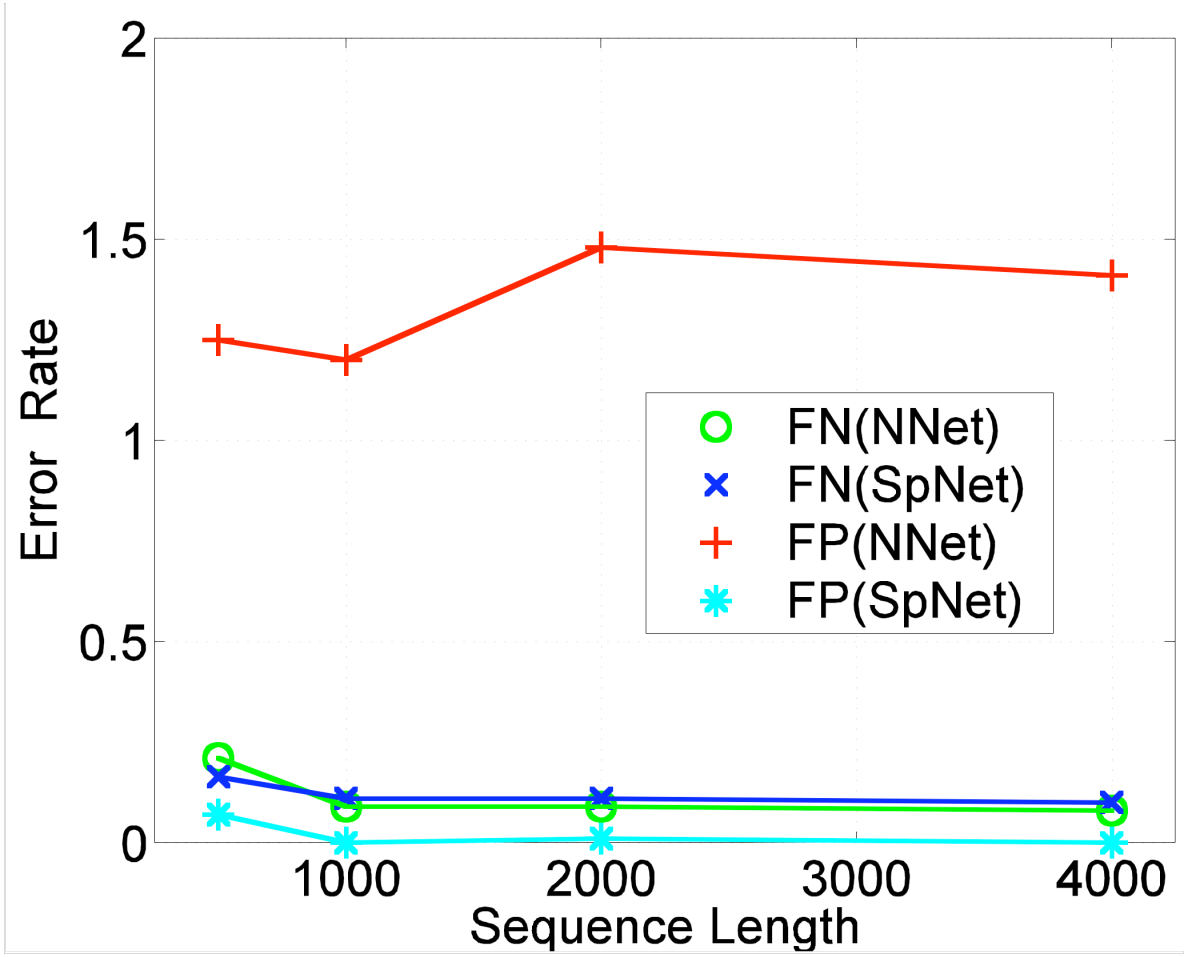
- Initial experiments established that Maddison's approach is usually inaccurate -- because individual gene trees cannot have any error
- Using ML instead of MP in SpNet seems to result in better estimates of gene trees
- Neighbor Net (NNet) produces many extra edges -- needs a postprocessing step!

Reconstruction Quality
Model Phylogeny: 20-taxon tree



Reconstruction Quality

Model Phylogeny: 20-taxon 1-hybrid network



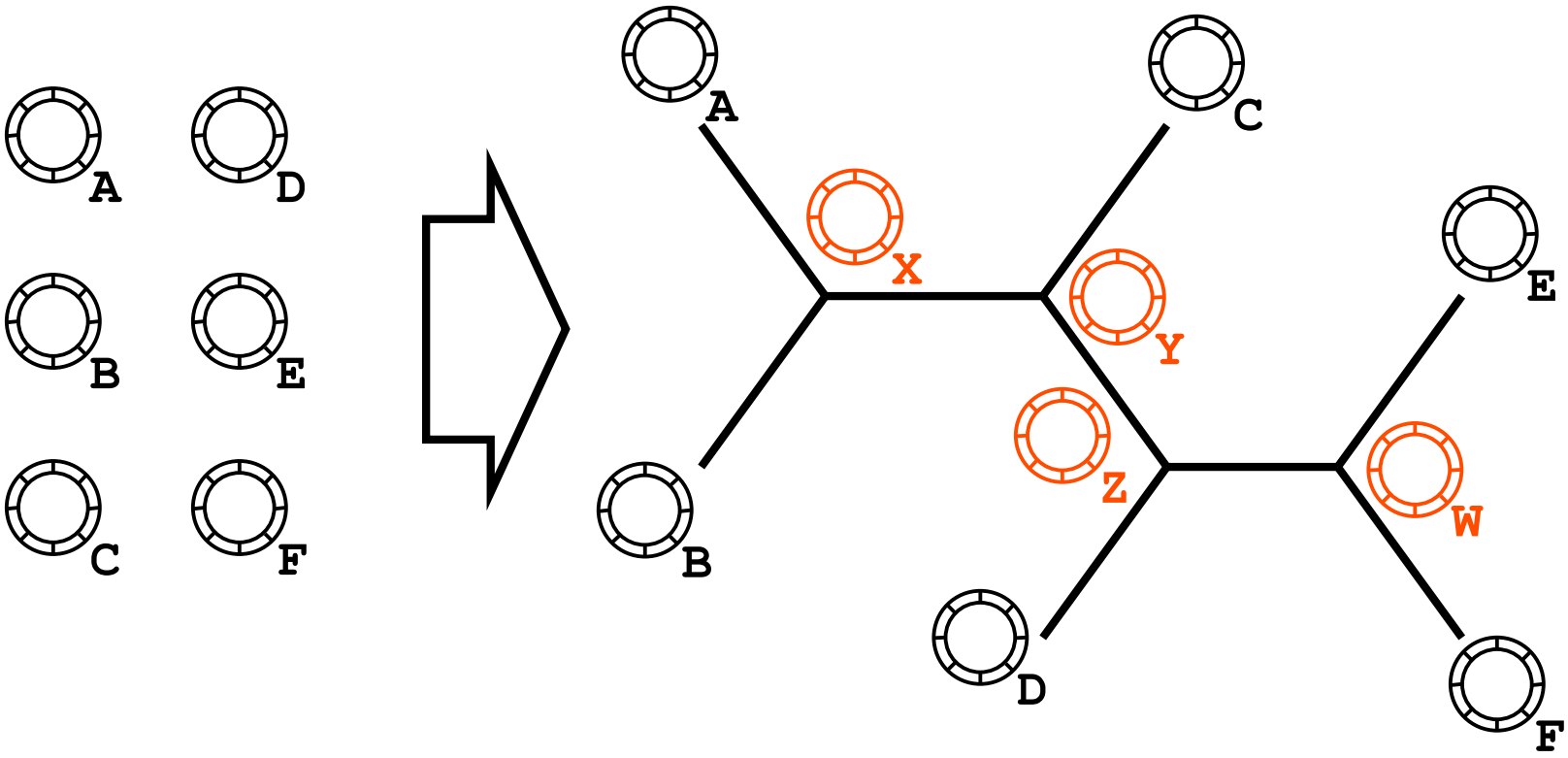
Conclusions

- Considering a set of “good” trees rather than a single optimal tree is advantageous in network reconstruction
- Separate analysis approaches outperform combined analysis approaches

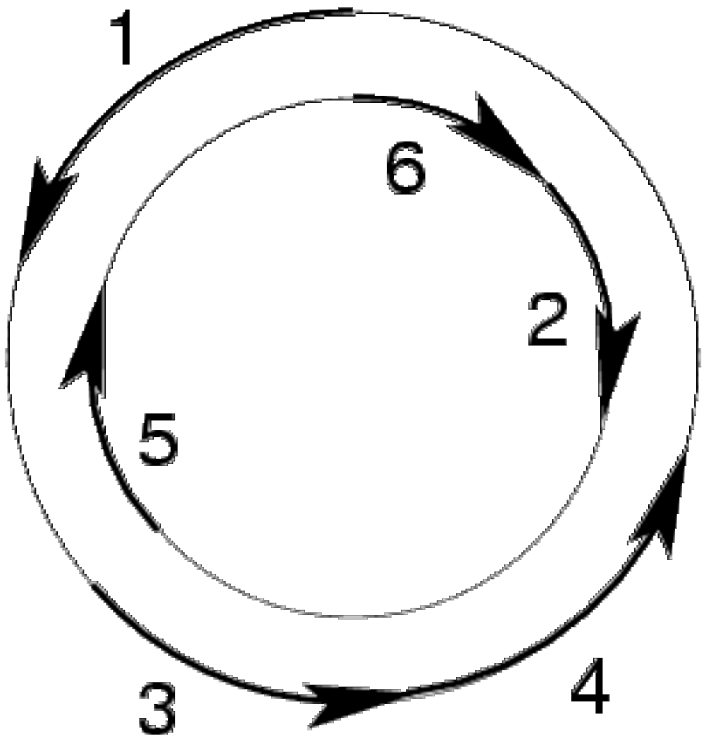
Ongoing research

- Using other techniques for obtaining unresolved trees (e.g., Bayesian analyses, bootstrapping, etc.)
- Detection vs. reconstruction – visualization and clustering techniques may also be useful (collaboration with St John)
- Extensions to multiple reticulations!

Whole-Genome Phylogenetics



Genomes As Signed Permutations



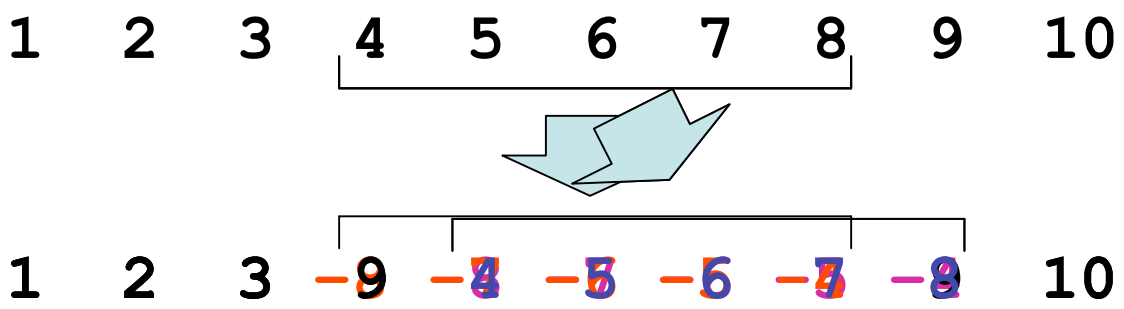
1 –5 3 4 -2 -6

or

6 2 -4 –3 5 –1

etc.

Genomes Evolve by Rearrangements



- Inversion (Reversal)
- Transposition
- Inverted Transposition

Genome Rearrangement Has A Huge State Space

- DNA sequences : 4 states per site
- Signed circular genomes with n genes:

$$2^{n-1}(n-1)!$$ states, 1 site

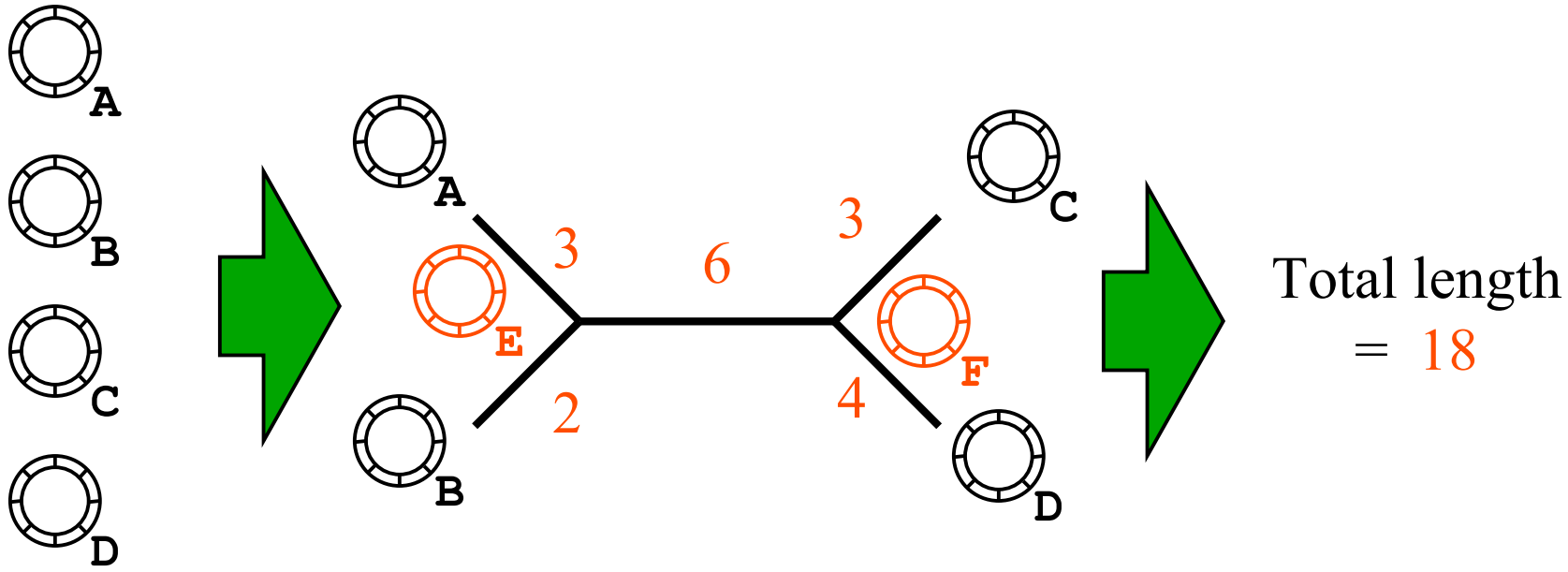
- Circular genomes (1 site)
  - with 37 genes: 2.56×10^{52} states
  - with 120 genes: 3.70×10^{232} states

Why use gene orders?

- “Rare genomic changes”: huge state space and relative infrequency of events (compared to site substitutions) could make the inference of deep evolution easier, or more accurate.
- Our research shows this is true, but accurate analysis of gene order data is computationally very intensive!

Maximum Parsimony on Rearranged Genomes (MPRG)

- The leaves are rearranged genomes.
- Find the tree that minimizes the total number of rearrangement events

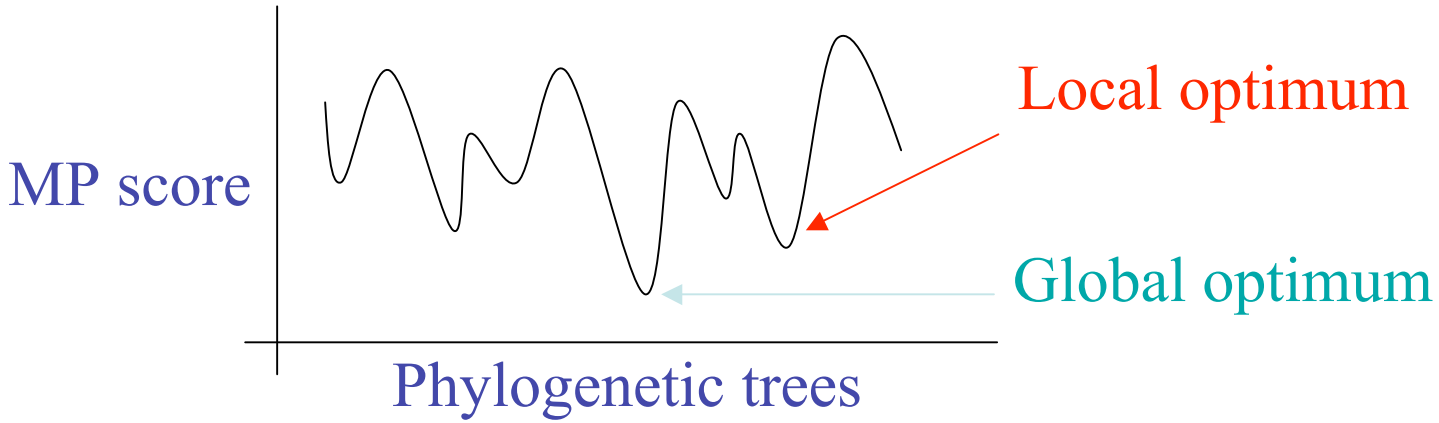


Optimization problems for gene order phylogeny

- Breakpoint phylogeny: find the phylogeny which minimizes the total number of breakpoints (NP-hard, even to find the median of three genomes)
- Inversion phylogeny: find the phylogeny which minimizes the sum of inversion distances on the edges (NP-hard, even to find the median of three genomes)

Inversion and Breakpoint phylogenies

- When the data are close to saturated, even Weighbor(EDE) analyses are insufficiently accurate. In these cases, our initial investigations suggest that the inversion and breakpoint phylogeny approaches may be superior.
- Problem: finding the best trees is enormously hard, since even the "point estimation" problem is hard (worse than estimating branch lengths in ML).



GRAPPA (Genome Rearrangement Analysis under Parsimony and other Phylogenetic Algorithms)

http://www.cs.unm.edu/~moret/GRAPPA/

- Heuristics for NP-hard optimization problems
- Fast polynomial time distance-based methods
- Contributors: U. New Mexico,U. Texas at Austin, Universitá di Bologna, Italy
- Freely available in source code at this site.
- Project leader: Bernard Moret (UNM) (moret@cs.unm.edu)

Benchmark gene order dataset: *Campanulaceae*

- 12 genomes + 1 outgroup (*Tobacco*), 105 gene segments
- NP-hard optimization problems: breakpoint and inversion phylogenies (techniques score every tree)

1997: **BPAnalysis** (Blanchette and Sankoff): 200 years (est.)

Benchmark gene order dataset: *Campanulaceae*

- 12 genomes + 1 outgroup (*Tobacco*), 105 gene segments
- NP-hard optimization problems: breakpoint and inversion phylogenies (techniques score every tree)

1997: **BPAnalysis** (Blanchette and Sankoff): 200 years (est.)

2000: Using **GRAPPA** v1.1 on the 512-processor Los Lobos Supercluster machine: 2 minutes (200,000-fold speedup per processor)

Benchmark gene order dataset: *Campanulaceae*

- 12 genomes + 1 outgroup (*Tobacco*), 105 gene segments
- NP-hard optimization problems: breakpoint and inversion phylogenies (techniques score every tree)

1997: **BPAnalysis** (Blanchette and Sankoff): 200 years (est.)

2000: Using **GRAPPA** v1.1 on the 512-processor Los Lobos Supercluster machine: 2 minutes (200,000-fold speedup per processor)

2003: Using latest version of **GRAPPA**: 2 minutes on a single processor (1-billion-fold speedup per processor)

Limitations and ongoing research

- Current methods limited to single chromosomes with equal gene content (or very small amounts of deletions and duplications) -- we are working on developing reliable techniques for genomes with unequal gene content

Acknowledgements

GRAPPA: Bernard Moret, David Bader, Li-San Wang, Jijun Tang, Bob Jansen, and Linda Raubeson

SpNet: Luay Nakhleh, Randy Linder, and Bernard Moret

- IDCM3: Usman Roshan, Tiffani Williams, Bernard Moret
- Funding: NSF, the David and Lucile Packard Foundation, the Institute for Cellular and Molecular Biology, and the Radcliffe Institute for Advanced Studies

Phylolab, U. Texas

Please visit us at
http://www.cs.utexas.edu/users/phylo/