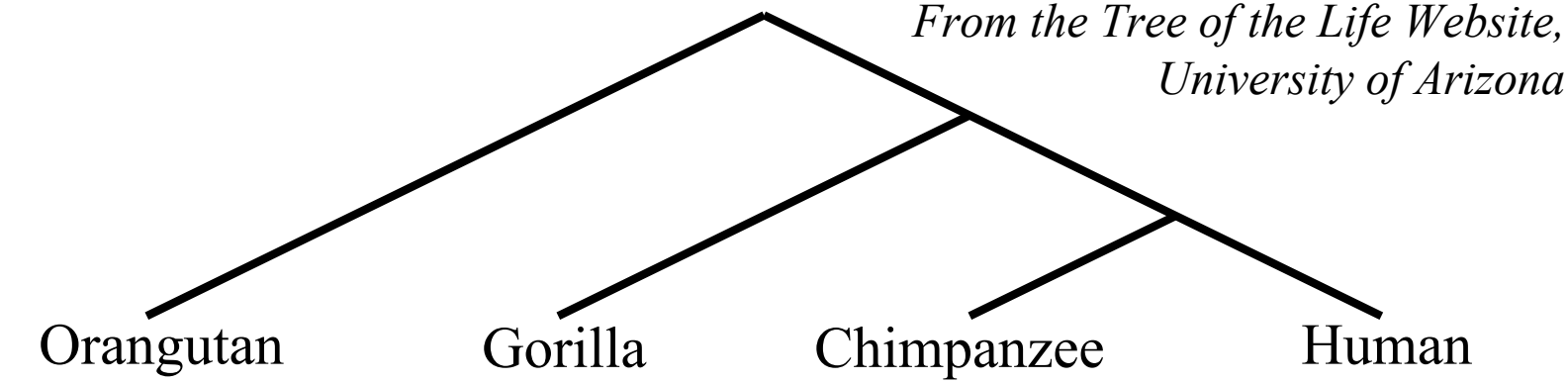
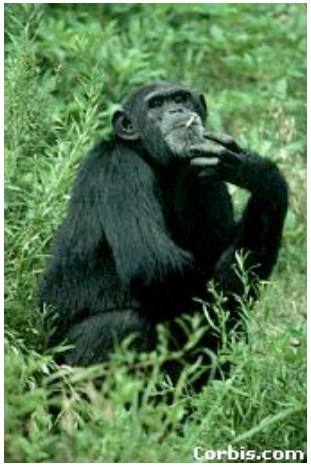
Combinatorial and graph-theoretic problems in evolutionary tree reconstruction

Tandy Warnow
Department of Computer Sciences
University of Texas at Austin

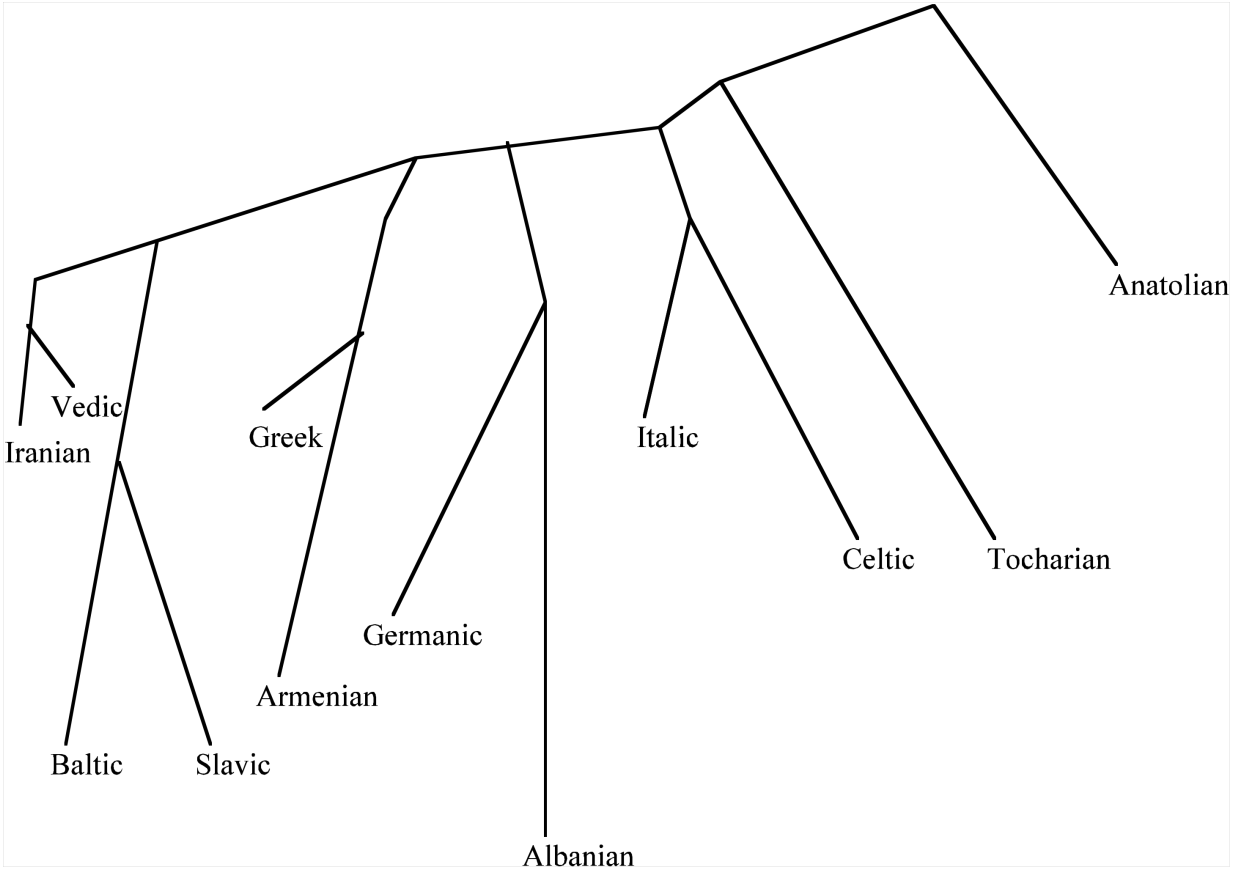
Phylogeny

From the Tree of the Life Website,
University of Arizona

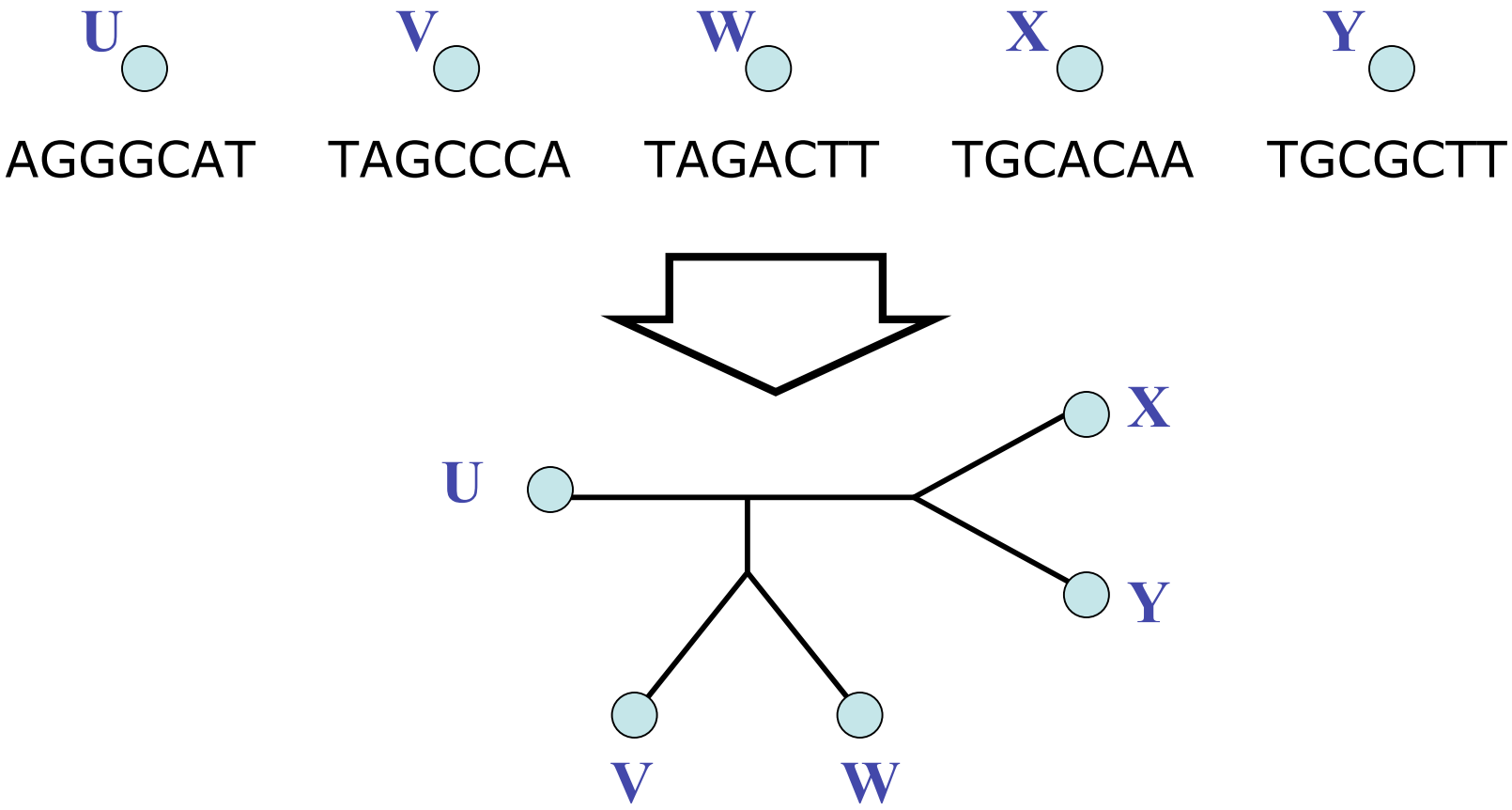




Possible Indo-European tree (Ringe, Warnow and Taylor 2000)

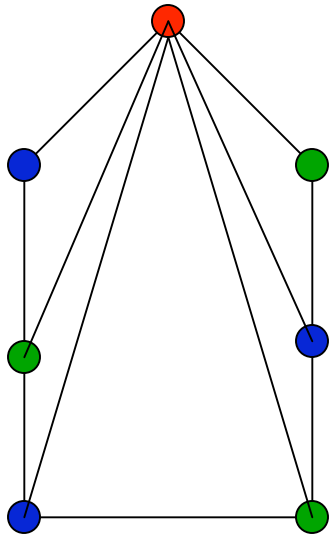


Phylogeny Problem



Triangulated Graphs

- Definition: A graph is triangulated if it has no simple cycles of size four or more.

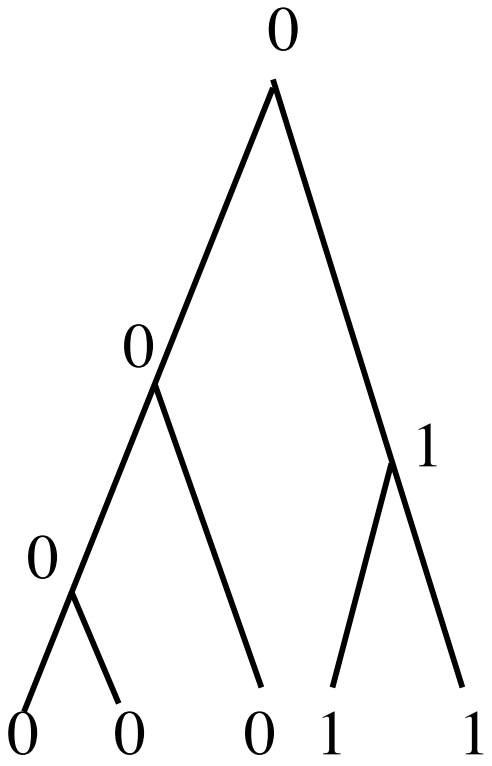


This talk: Triangulated graphs and phylogeny estimation

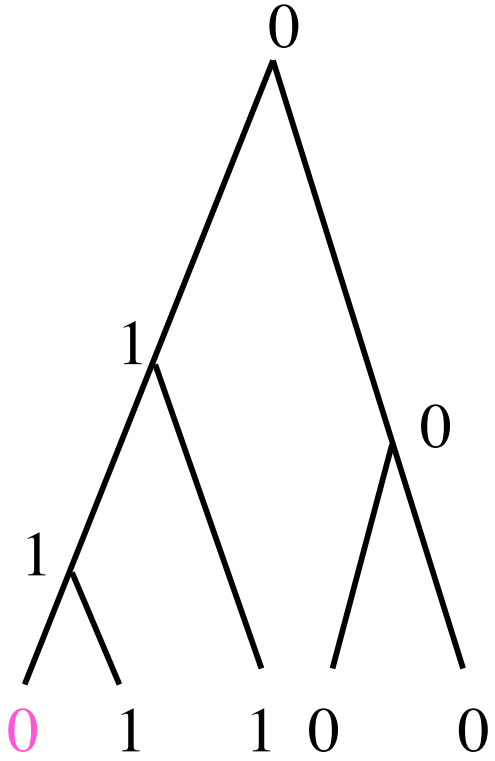
- The "Triangulating Colored Graphs" problem and an application to historical linguistics
- Using triangulated graphs to improve the accuracy and sequence length requirements phylogeny estimation in biology
- Using triangulated graphs to speed-up heuristics for NP-hard phylogenetic estimation problems

Part 1: Using triangulated graphs for historical linguistics

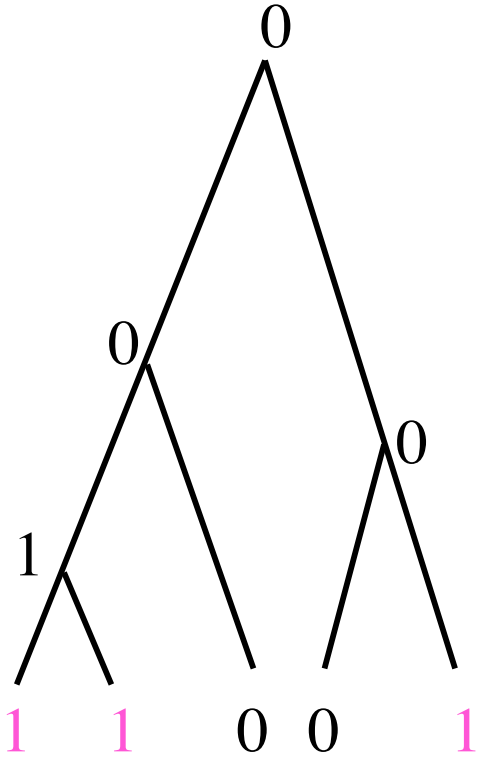
Some useful terminology: homoplasy



no homoplasy



back-mutation



parallel evolution

Perfect Phylogeny

- A phylogeny T for a set S of taxa is a perfect phylogeny if each state of each character occupies a subtree (no character has back-mutations or parallel evolution)

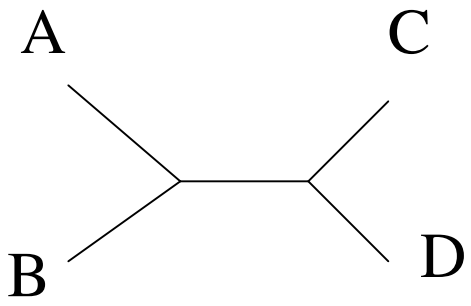
Perfect phylogenies, cont.

- A=(0,0), B=(0,1), C=(1,3), D=(1,2) has a perfect phylogeny!

- A=(0,0), B=(0,1), C=(1,0), D=(1,1) does not have a perfect phylogeny!

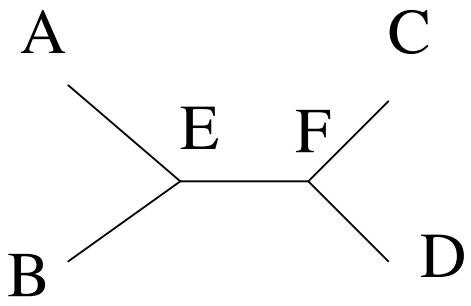
A perfect phylogeny

- A = 0 0
- B = 0 1
- C = 1 3
- D = 1 2



A perfect phylogeny

- A = 0 0
- B = 0 1
- C = 1 3
- D = 1 2
- E = 0 3
- F = 1 3



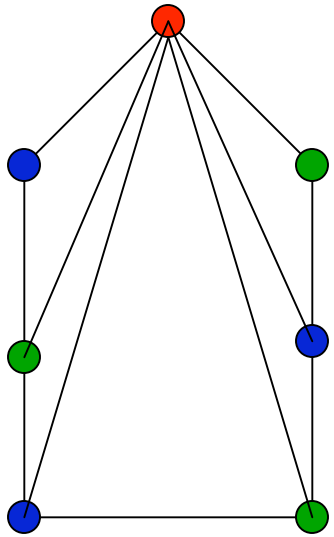
The Perfect Phylogeny Problem

- Given a set S of taxa (species, languages, etc.) determine if a perfect phylogeny T exists for S.

- The problem of determining whether a perfect phylogeny exists is NP-hard (McMorris *et al*. 1994, Steel 1991).

Triangulated Graphs

- Definition: A graph is triangulated if it has no simple cycles of size four or more.

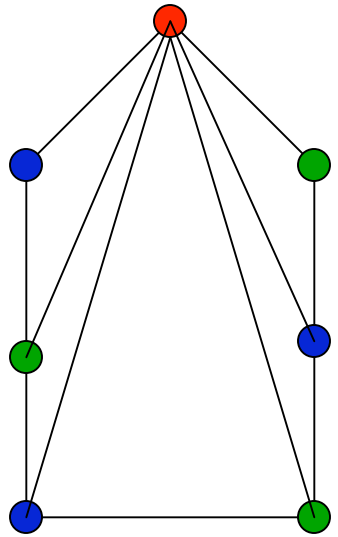


Triangulated graphs and trees

- A graph G=(V,E) is triangulated if and only if there exists a tree T so that G is the intersection graph of a set of subtrees of T.
  - vertices of G correspond to subtrees (f(v) is a subtree of T)
  - (v,w) is an edge in G if and only if f(v) and f(w) have a non-empty intersection

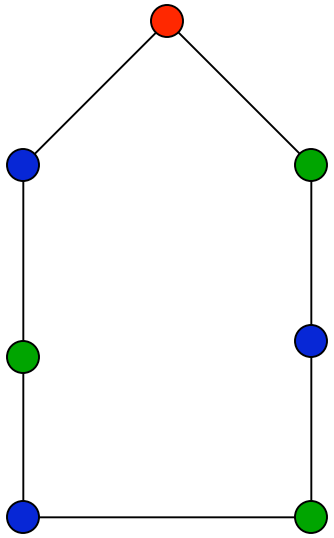
c-Triangulated Graphs

- A vertex-colored graph is c-triangulated if it is triangulated, but also properly colored!



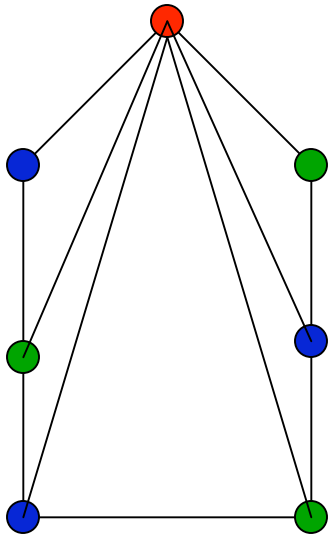
Triangulating Colored Graphs: An Example

A graph that can be c-triangulated



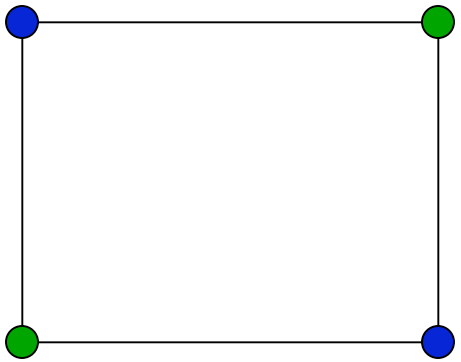
Triangulating Colored Graphs: An Example

A graph that can be c-triangulated



Triangulating Colored Graphs: An Example

A graph that cannot be c-triangulated



Triangulating Colored Graphs (TCG)

Triangulating Colored Graphs: given a vertex-colored graph G, determine if G can be c-triangulated.

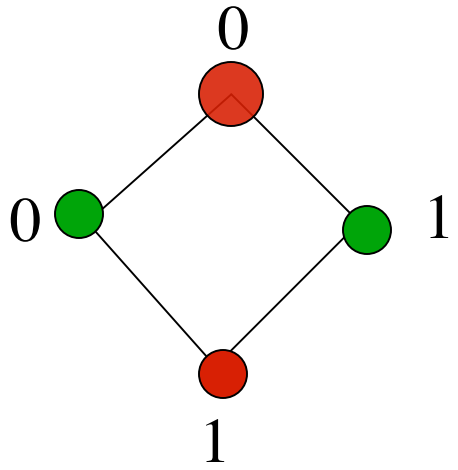
The PP and TCG Problems

- **Buneman's Theorem:**
  A perfect phylogeny exists for a set S *if and only if* the associated character state intersection graph can be *c*-triangulated.

- The PP and TCG problems are polynomially equivalent and NP-hard.

A no-instance of Perfect Phylogeny

- A = 0 0
- B = 0 1
- C = 1 0
- D = 1 1

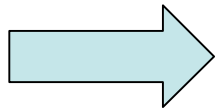


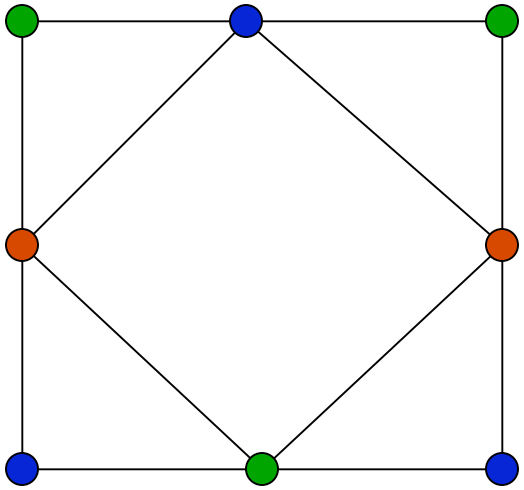
An input to perfect phylogeny (left) of four sequences described by two characters, and its character state intersection graph. Note that the character state intersection graph is 2-colored.

Solving the PP Problem Using Buneman's Theorem

"Yes" Instance of PP:

	c1	c2	c3
s1	3	2	1
s2	1	2	2
s3	1	1	3
s4	2	1	1

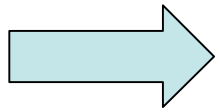


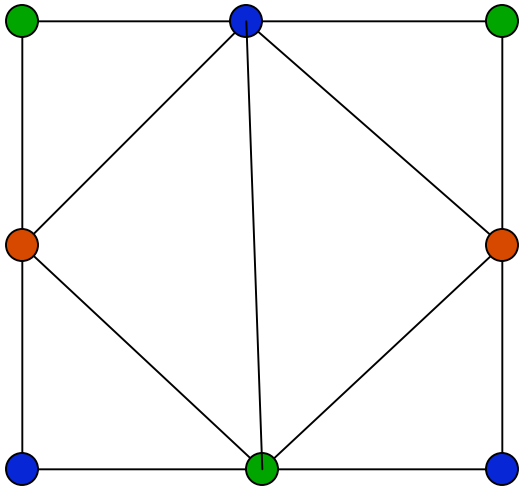


Solving the PP Problem Using Buneman's Theorem

"Yes" Instance of PP:

	c1	c2	c3
s1	3	2	1
s2	1	2	2
s3	1	1	3
s4	2	1	1





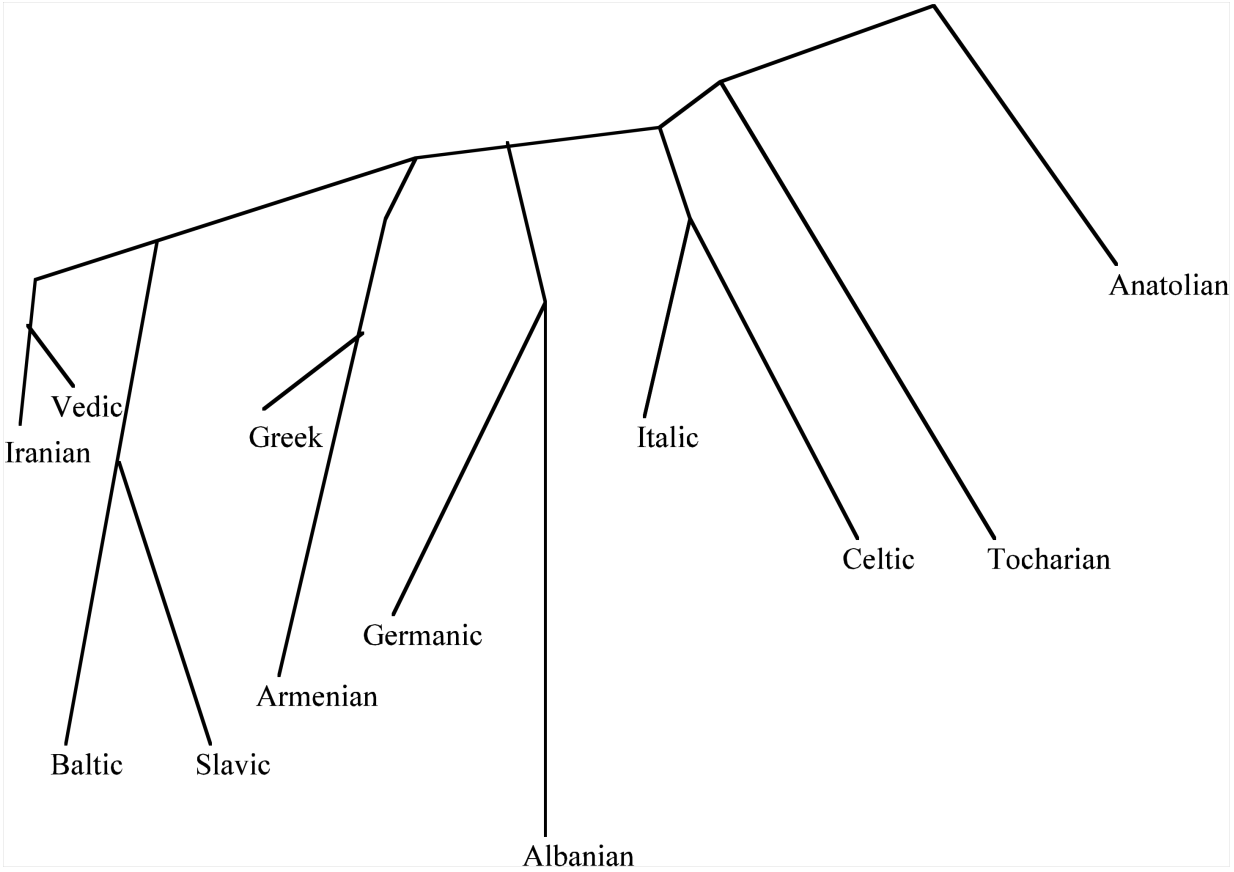
Some special cases are easy

- Binary character perfect phylogeny solvable in linear time
- r-state characters solvable in polynomial time for each r (combinatorial algorithm)
- Two character perfect phylogeny solvable in polynomial time (produces 2-colored graph)
- k-character perfect phylogeny solvable in polynomial time for each k (produces k-colored graphs -- connections to Robertson-Seymour graph minor theory)

Phylogenies of Languages

- Languages evolve over time, just as biological species do (geographic and other separations induce changes that over time make different dialects incomprehensible -- and new languages appear)
- The result can be modelled as a rooted tree
- The interesting thing is that many characteristics of languages evolve without back mutation or parallel evolution -- so a “perfect phylogeny” is possible!

Possible Indo-European tree (Ringe, Warnow and Taylor 2000)

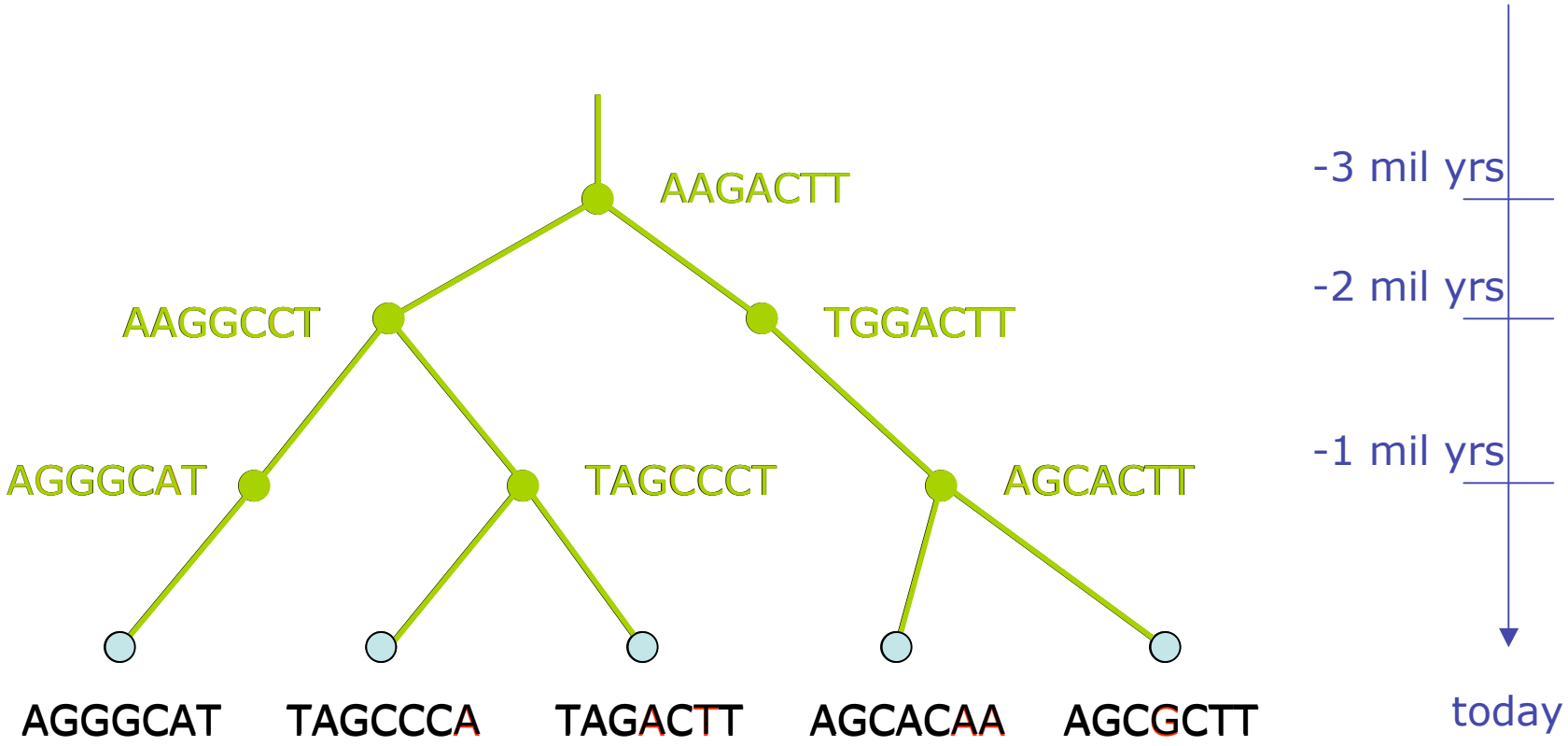


Part 2: Phylogeny estimation in biology

- Using triangulated graphs to improve the topological accuracy of distance-based methods

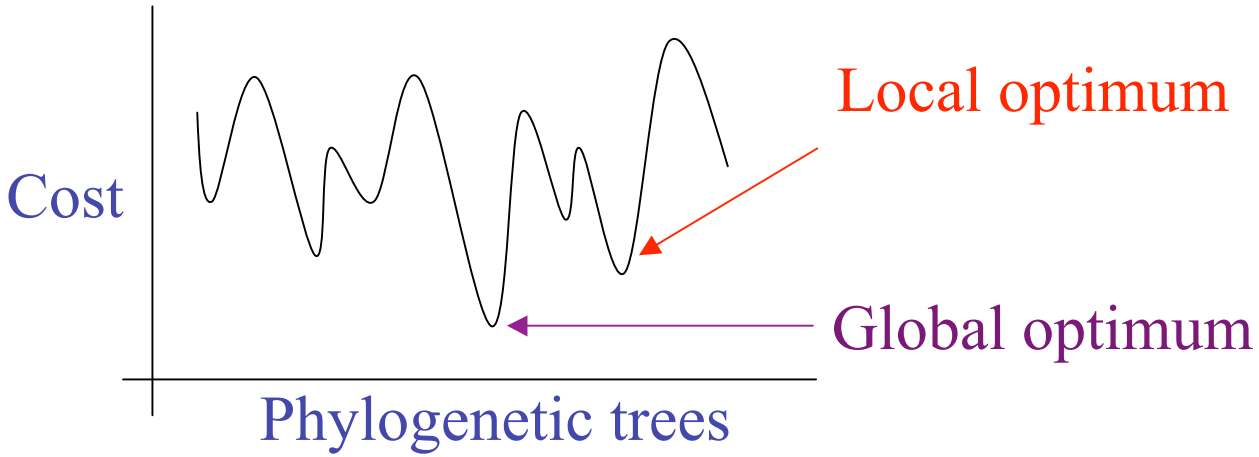
- Using triangulated graphs to speed up heuristics for NP-hard optimization problems

DNA Sequence Evolution



Phylogenetic reconstruction methods

1. Heuristics for NP-hard optimization criteria (Maximum Parsimony and Maximum Likelihood)



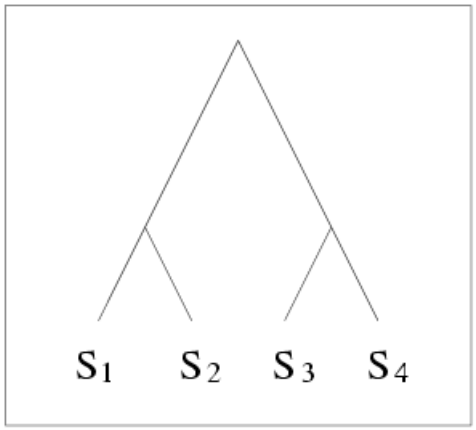
2. Polynomial time distance-based methods: Neighbor Joining, FastME, etc.
3. Bayesian MCMC methods.

Evaluating phylogeny reconstruction methods

- In simulation: how “topologically” accurate are trees reconstructed by the method?

- On real data: how good are the “scores” (typically either maximum parsimony or maximum likelihood) obtained by the method, as a function of time?

Distance-based Phylogenetic Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

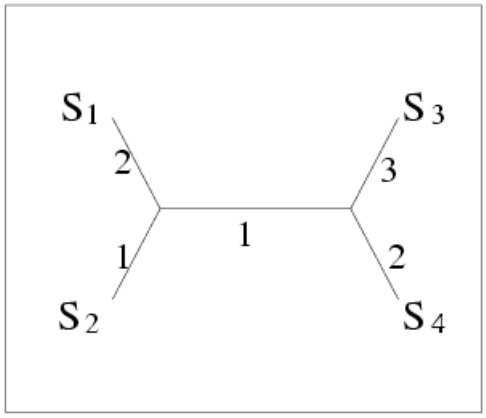
DNA SEQUENCES

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

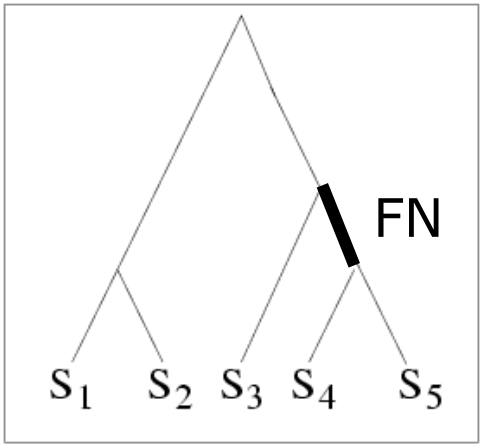
DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING

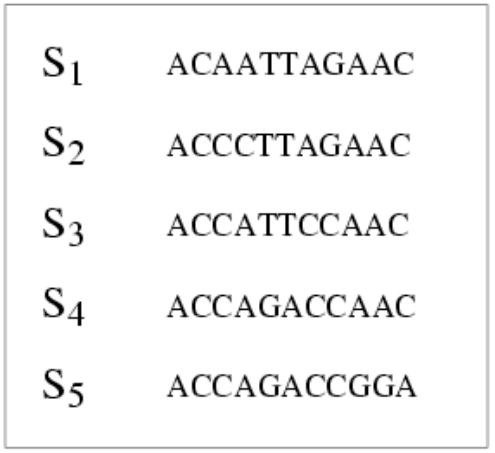


INFERRED TREE

Quantifying Error



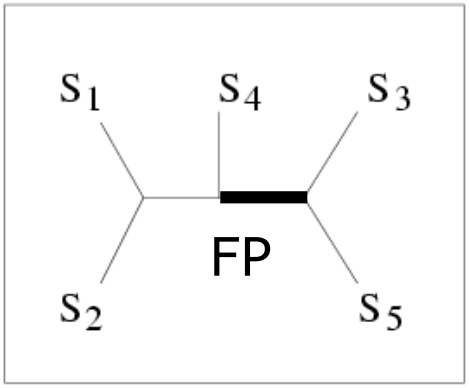
TRUE TREE



DNA SEQUENCES

FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

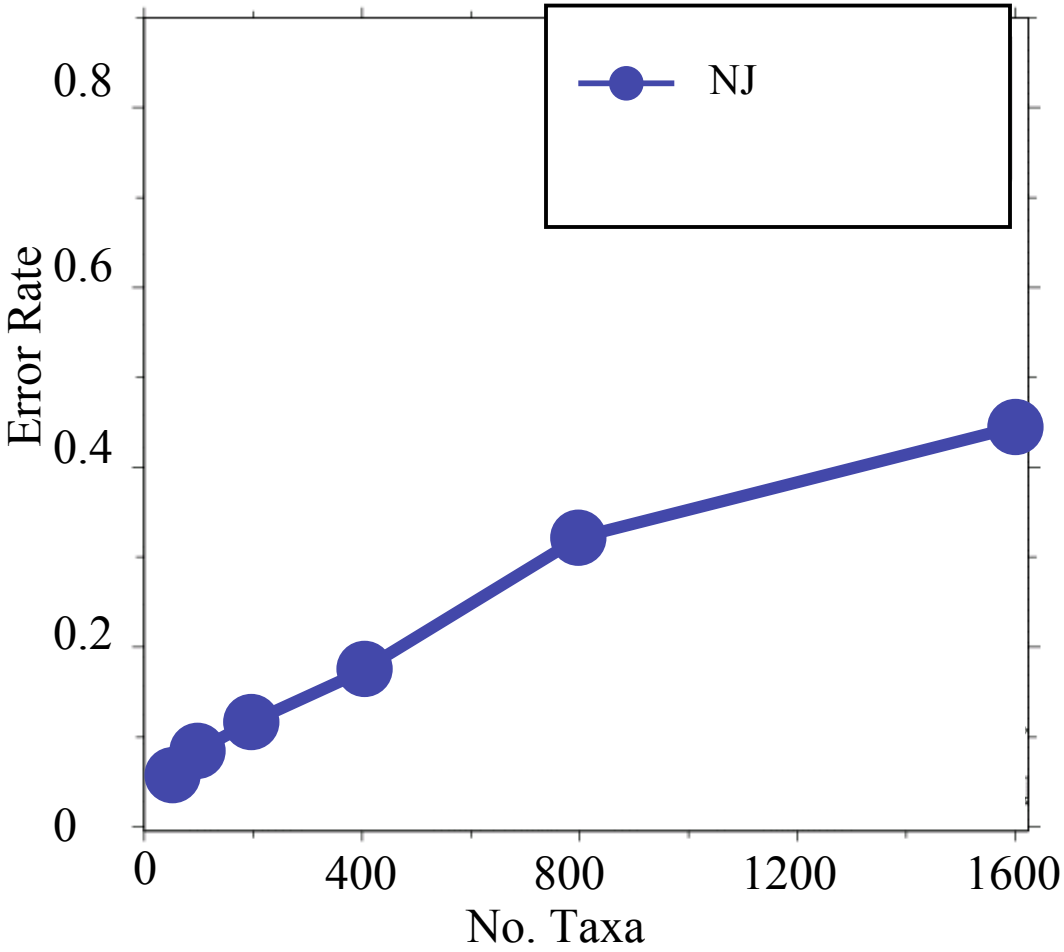
50% error rate



INFERRED TREE

Neighbor joining has poor accuracy on large diameter model trees

[Nakhleh et al. ISMB 2001]



Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

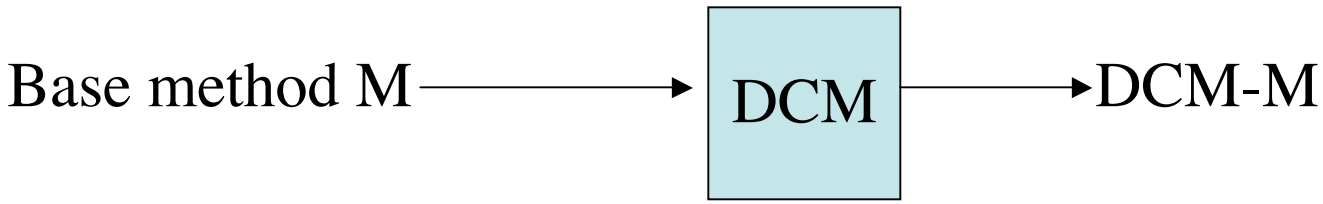
Error rates reflect proportion of incorrect edges in inferred trees.

Neighbor Joining's sequence length requirement is exponential!

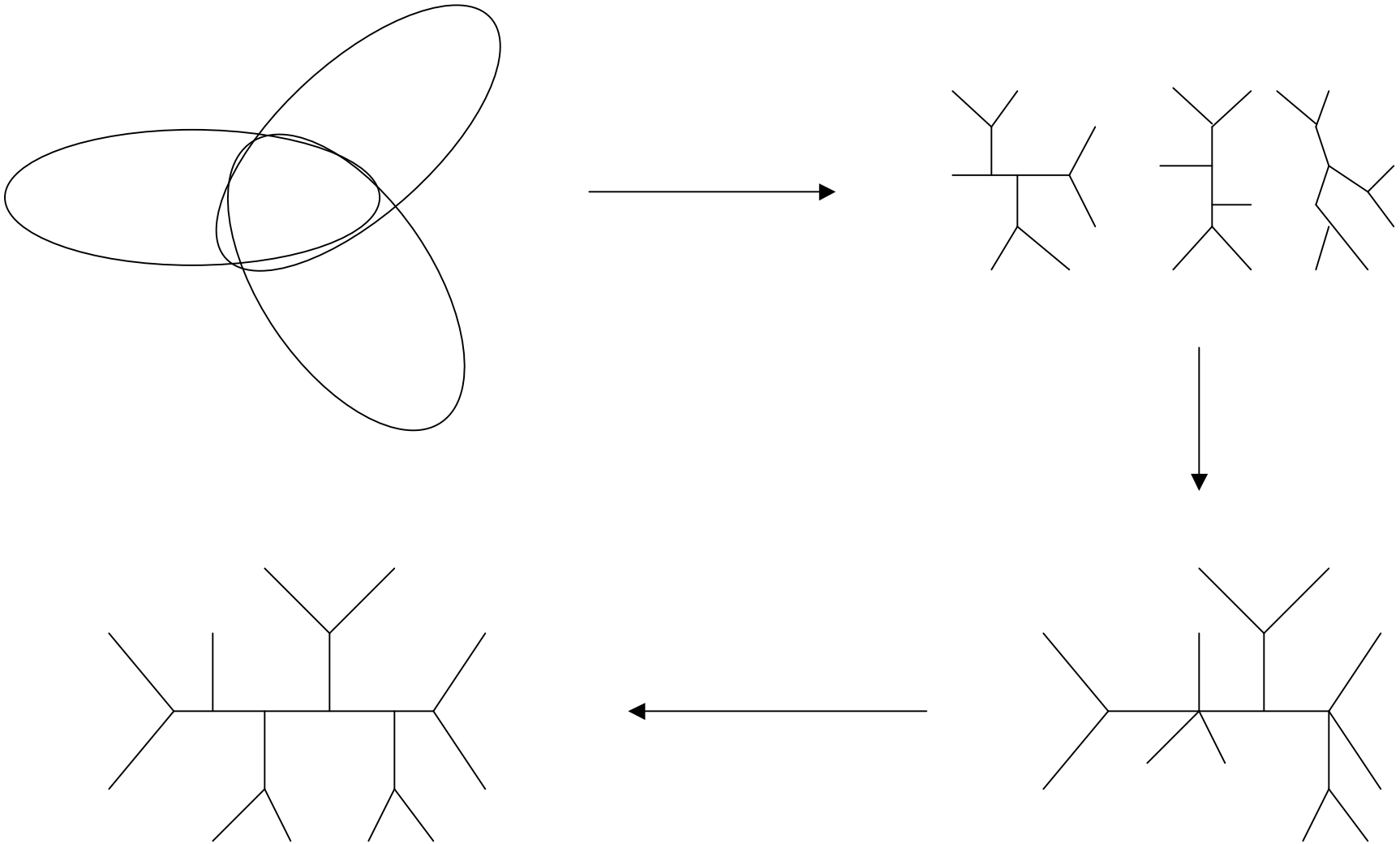
- Atteson: Let T be a General Markov model tree defining additive matrix D. Then Neighbor Joining will reconstruct the true tree with high probability from sequences that are of length at least **$O(\lg n \, e^{\max Dij})$.**

“Boosting” phylogeny reconstruction methods

- DCMs “boost” the performance of phylogeny reconstruction methods.



Divide-and-conquer for phylogeny estimation



Graph-theoretic divide-and-conquer (DCM's)

- Define a **triangulated** graph so that its vertices correspond to the input taxa
- Compute a decomposition of the graph into overlapping subgraphs, thus defining a decomposition of the taxa into overlapping subsets.
- Apply the "base method" to each subset of taxa, to construct a subtree
- Merge the subtrees into a single tree on the full set of taxa.

DCM1 Decompositions

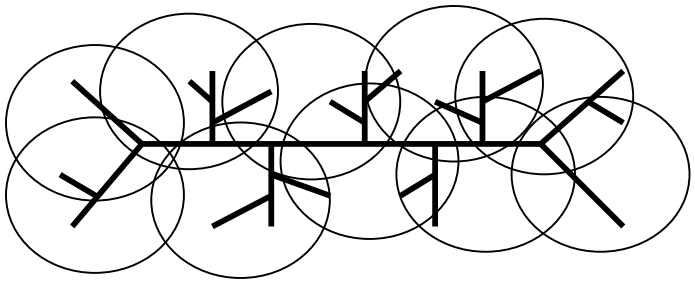
Input: Set S of sequences, distance matrix d, threshold value $q \in \{d_{ij}\}$

1. Compute threshold graph

$$G_q = (V, E), V = S, E = \{(i, j) : d(i, j) \leq q\}$$

2. Perform minimum weight triangulation (note: if d is an additive matrix, then the threshold graph is provably **triangulated**).

DCM1 decomposition : Compute maximal cliques

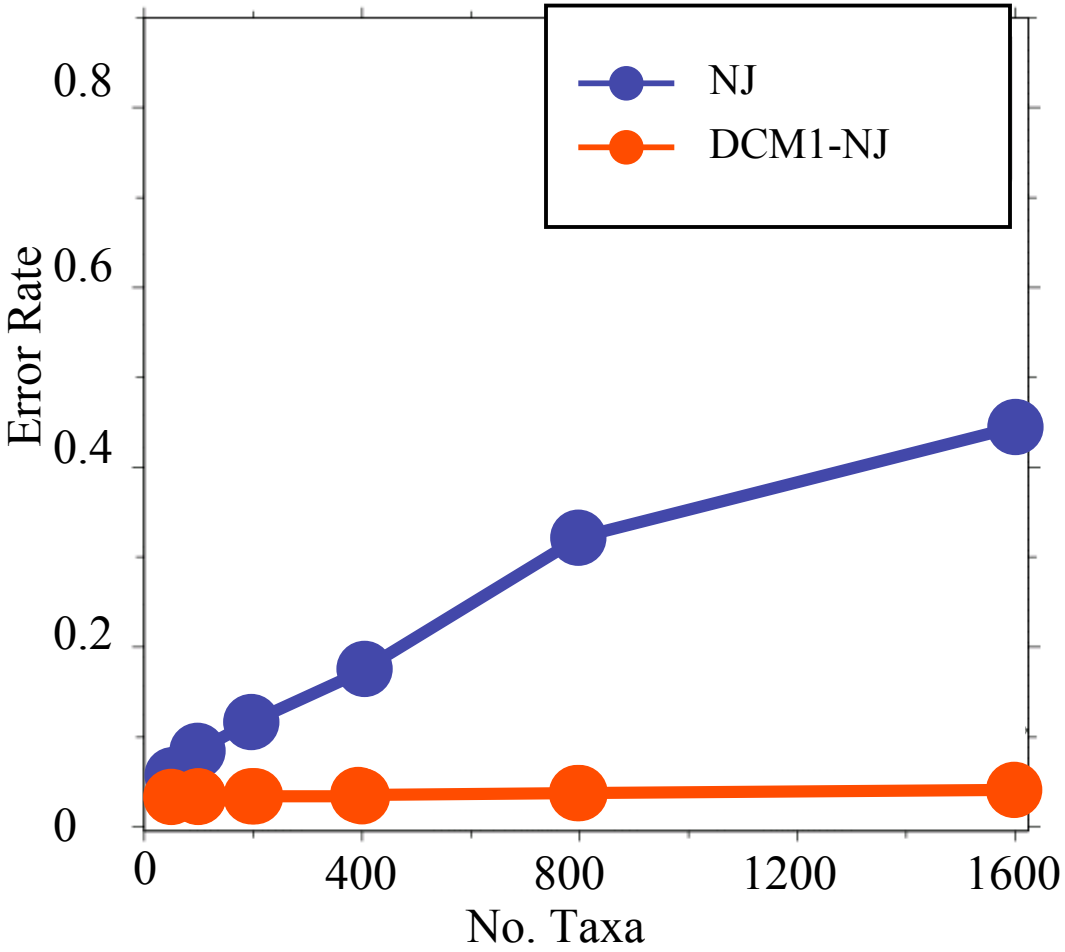


Improving upon NJ

- Construct trees on a number of smaller diameter subproblems, and merge the subtrees into a tree on the full dataset.
- Our approach:
  - Phase I: produce $O(n^2)$ trees (one for each diameter)
  - Phase II: pick the "best" tree from the set.

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001 and Warnow et al. SODA 2001]



Theorem: DCM1-NJ converges to the true tree from polynomial length sequences

What about solving MP and ML?

- Maximum Parsimony (MP) and maximum likelihood (ML) are the major phylogeny estimation methods used by systematists.

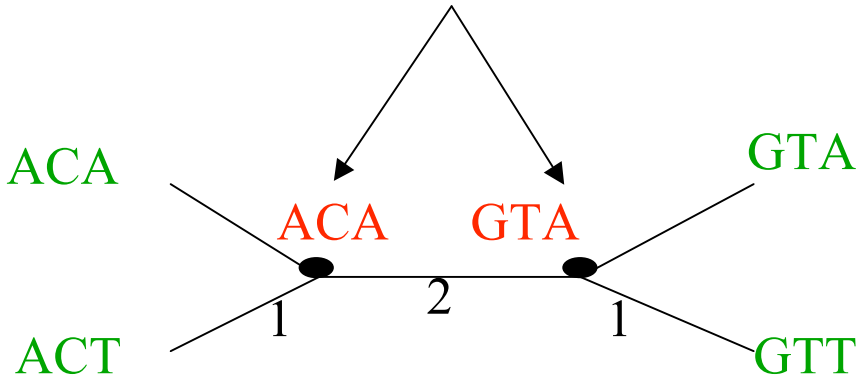
Maximum Parsimony

- **Input**: Set S of n aligned sequences of length k
- **Output**: A phylogenetic tree T
  - leaf-labeled by sequences in S
  - additional sequences of length k labeling the internal nodes of T

such that $\sum_{(i,j)\in E(T)} H(i,j)$ is minimized.

Maximum Parsimony: computational complexity

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



MP score = 4

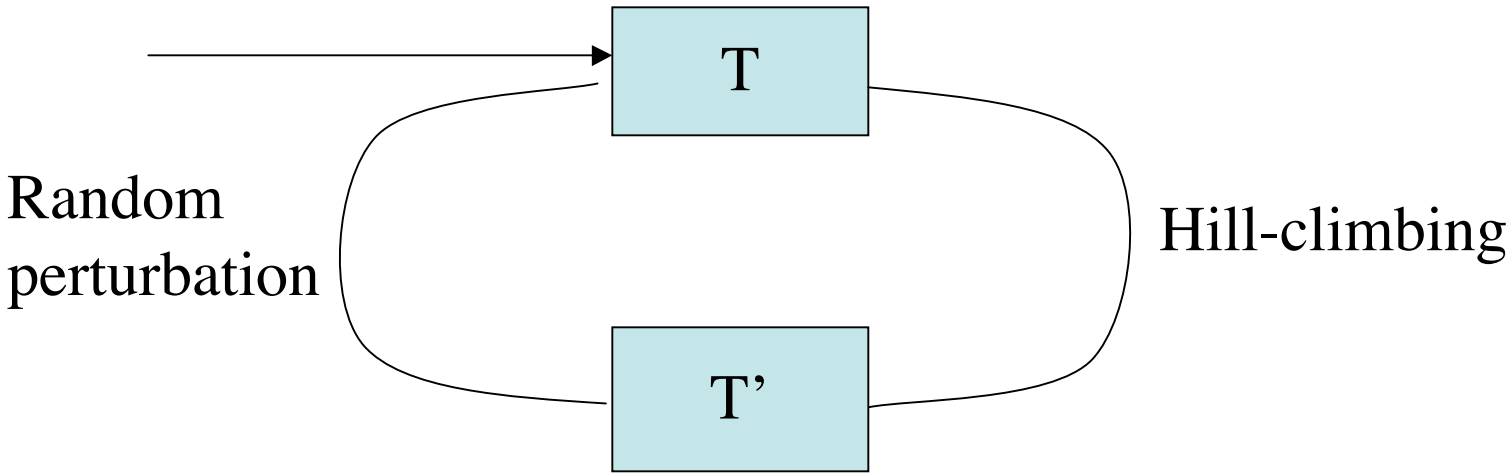
Finding the optimal MP tree is **NP-hard**

Solving NP-hard problems exactly is … unlikely

- Number of (unrooted) binary trees on *n* leaves is (2n-5)!!
- If each tree on **1000** taxa could be analyzed in **0.001** seconds, we would find the best tree in **2890 millennia**

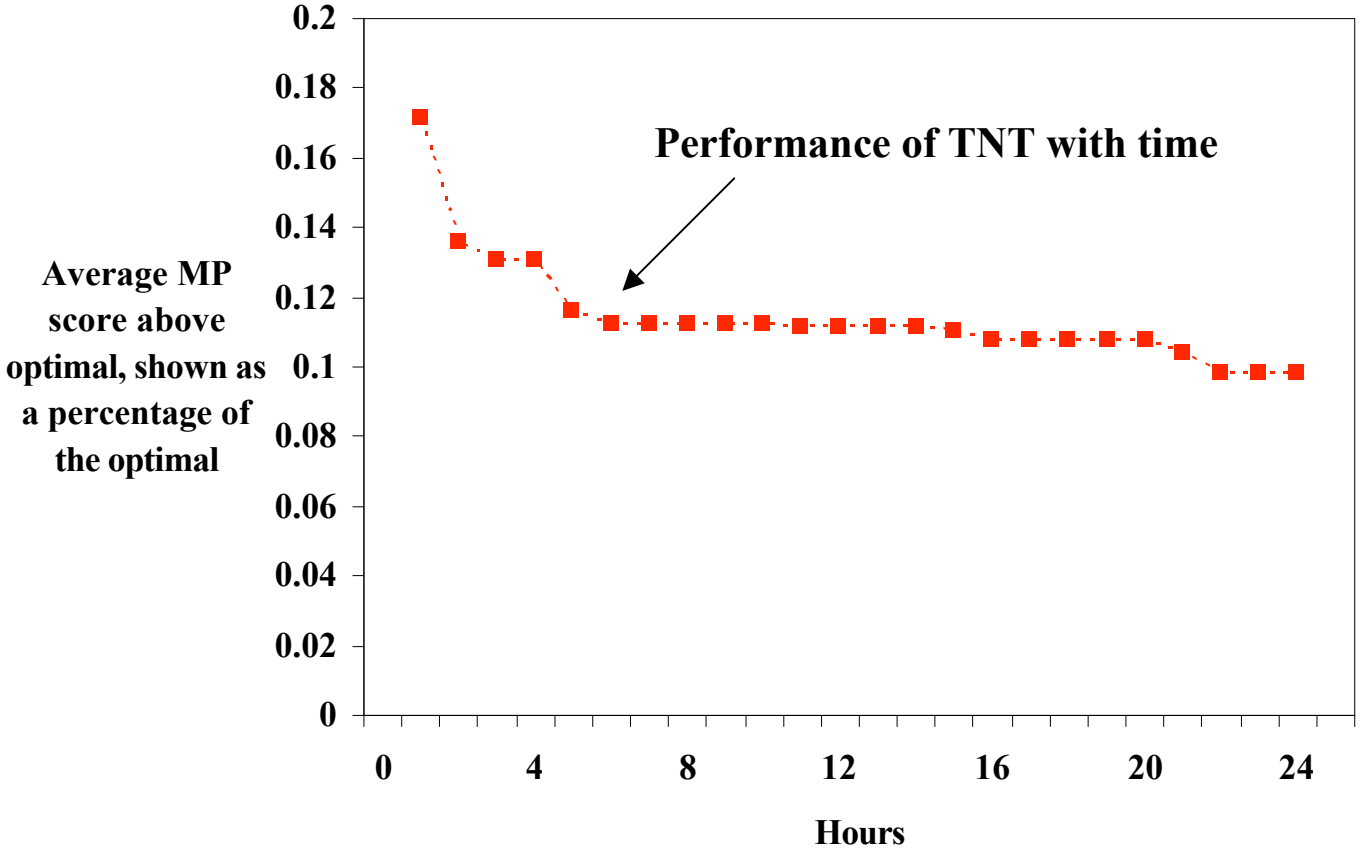
#leaves	#trees
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
20	2.2×10^{20}
100	4.5×10^{190}
1000	2.7×10^{2900}

Standard heuristic search



Problems with current techniques for MP

Shown here is the performance of the TNT software for maximum parsimony on a real dataset of almost 14,000 sequences. The required level of accuracy with respect to MP score is no more than **0.01% error** (otherwise high topological error results). ("Optimal" here means best score to date, using any method for any amount of time.)

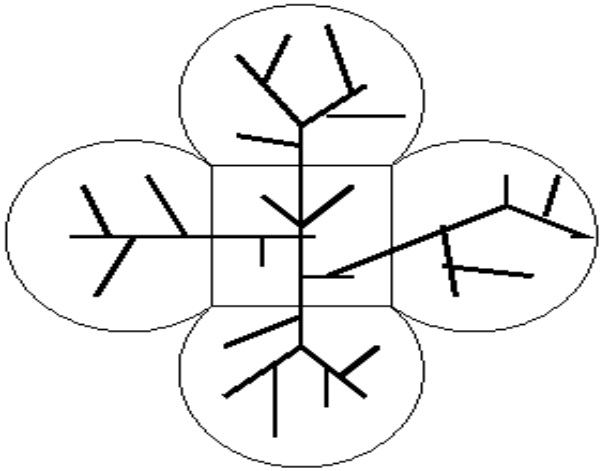


New DCM3 decomposition

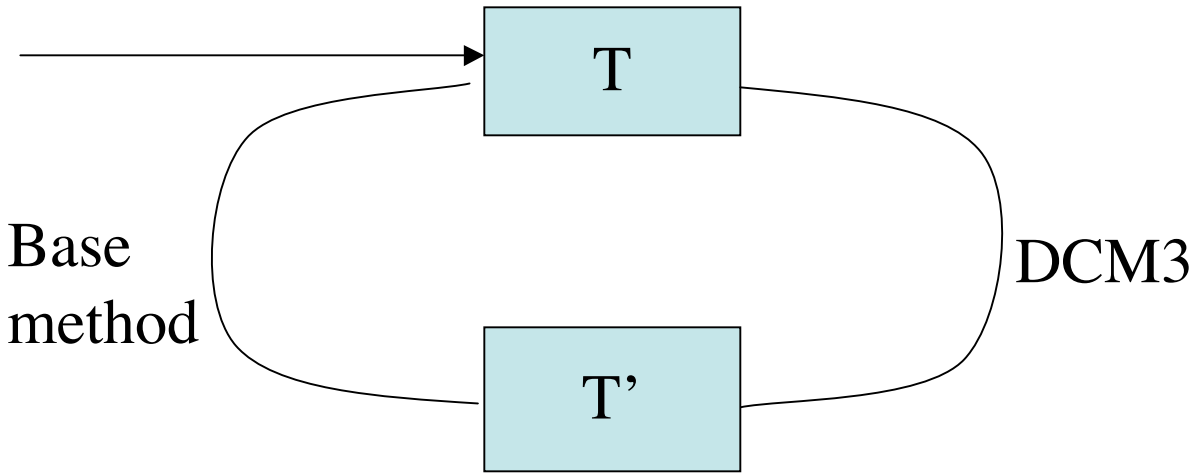
Input: Set *S* of sequences, and guide-tree *T*

1. We use a new graph ("*short subtree* graph") *G(S,T))*
   Note: G(S,T) is ***triangulated****!*
2. Find clique separator in *G(S,T)* and form subproblems

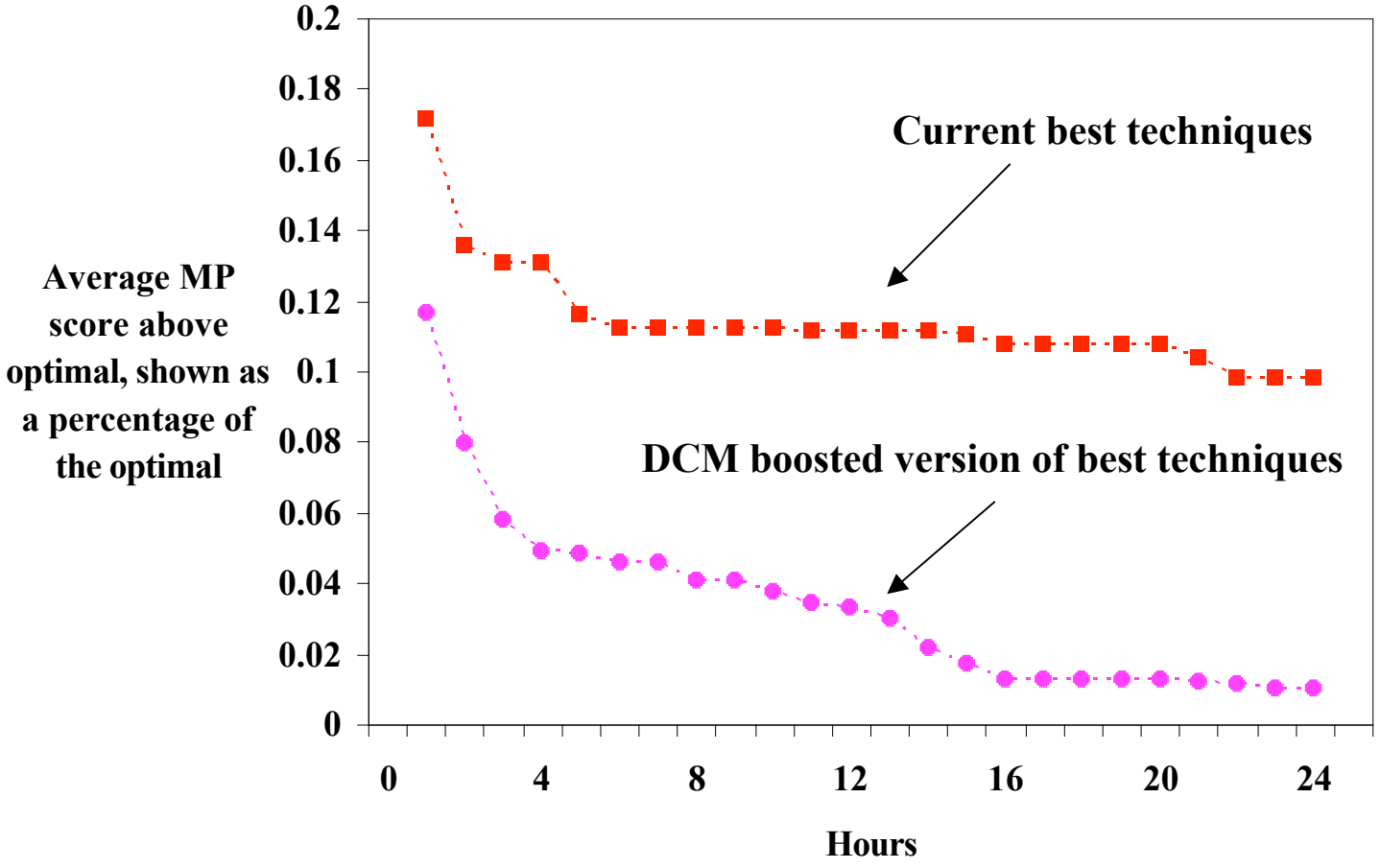
DCM3 decompositions
(1) can be obtained in **O(n)** time
(2) yield small subproblems
(3) can be used iteratively



Iterative-DCM3



Rec-I-DCM3 significantly improves performance



Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset

Summary

- NP-hard optimization problems abound in phylogeny reconstruction, and in computational biology in general, and need very accurate solutions.
- Many real problems have beautiful and natural combinatorial and graph-theoretic formulations.

Acknowledgments

- The CIPRES project www.phylo.org (and the US National Science Foundation more generally)
- The David and Lucile Packard Foundation
- The Program for Evolutionary Dynamics at Harvard, The Radcliffe Institute for Advanced Research, and the Institute for Cellular and Molecular Biology at UT-Austin
- Collaborators: Bernard Moret, Usman Roshan, Tiffani Williams, Daniel Huson, and Donald Ringe.