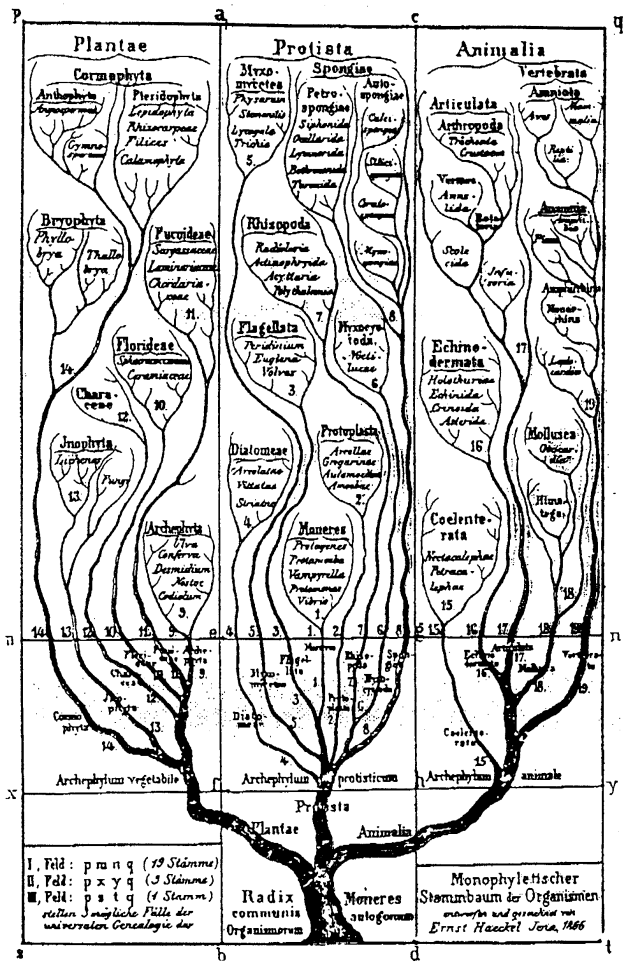
# The Disk-Covering Method for Tree Reconstruction

Daniel Huson
PACM, Princeton University

Tandy Warnow
CIS, University of Pennsylvania

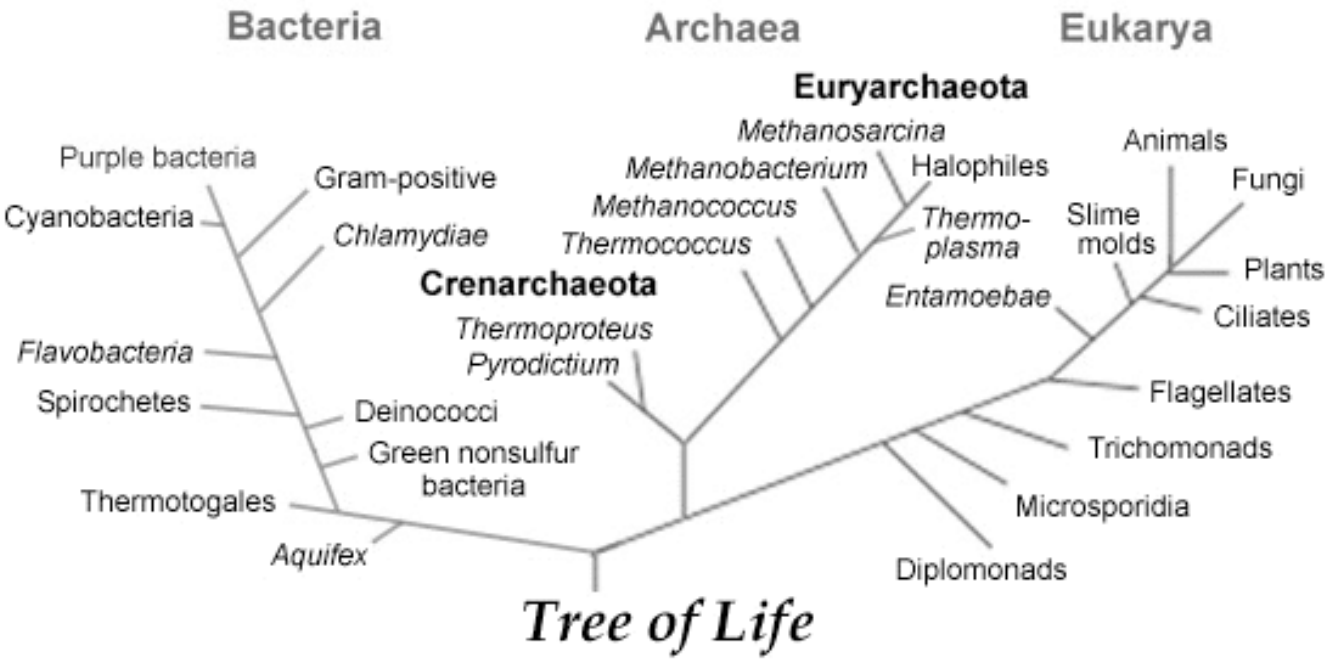
Kaikoura, 1999

# The Tree of Life




The first phylogenetic tree of life, Ernst Haeckel (1866)

# The Tree of Life




A modern view of the tree of life

# Big Trees are Hard to Infer

# Big Trees are Hard to Infer

- Sequence-based methods (e.g. Maximum Parsimony and maximum likelihood) are **computationally expensive**.

# Big Trees are Hard to Infer

- Sequence-based methods (e.g. Maximum Parsimony and maximum likelihood) are **computationally expensive**.

- Distance-based methods (e.g. Neighbor-Joining, Buneman tree, split decomposition) are fast, but **degrade in accuracy** with high evolutionary divergence.

# Big Trees are Hard to Infer

- Sequence-based methods (e.g. Maximum Parsimony and maximum likelihood) are **computationally expensive**.

- Distance-based methods (e.g. Neighbor-Joining, Buneman tree, split decomposition) are fast, but **degrade in accuracy** with high evolutionary divergence.

- Parsimony does best if all branches are short, so that large numbers of taxa may be needed for accurate tree reconstruction using Parsimony.

# Big Trees are Hard to Infer

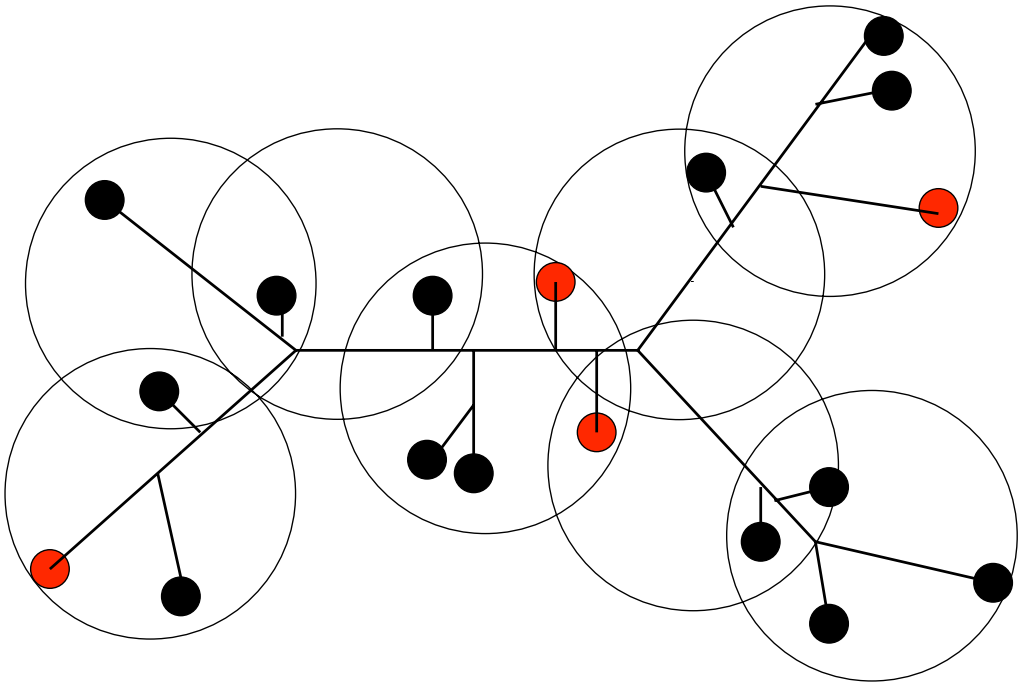
- Sequence-based methods (e.g. Maximum Parsimony and maximum likelihood) are **computationally expensive**.

- Distance-based methods (e.g. Neighbor-Joining, Buneman tree, split decomposition) are fast, but **degrade in accuracy** with high evolutionary divergence.

- Parsimony does best if all branches are short, so that large numbers of taxa may be needed for accurate tree reconstruction using Parsimony.

- Year-long Parsimony analyses (Rice *et al.*) of large divergent datasets are infeasible for most researchers.

# The Disk-Covering Method (DCM)



A divide-and-conquer approach based on the idea of covering given sequence data with small overlapping disks

- Each disk contains a small number of taxa.
- Taxa within a disk are very similar.
- Apply given *base-method* to subproblems.
- Use overlap to merge subtrees to obtain final tree.

# The DCM Algorithm

# The DCM Algorithm

- Input: distances and sequences
- Choose base-method (e.g. Parsimony, MLE, NJ, or d-splits)

# The DCM Algorithm

- Input: distances and sequences
- Choose base-method (e.g. Parsimony, MLE, NJ, or d-splits)

  - **For a given threshold $w$:**

    - Compute threshold graph $G$
    - Compute triangulation $G^*$ of threshold graph
    - Apply base-method to all maximal cliques in $G^*$
    - Merge trees guided by perfect elimination scheme

# The DCM Algorithm

- Input: distances and sequences
- Choose base-method (e.g. Parsimony, MLE, NJ, or d-splits)

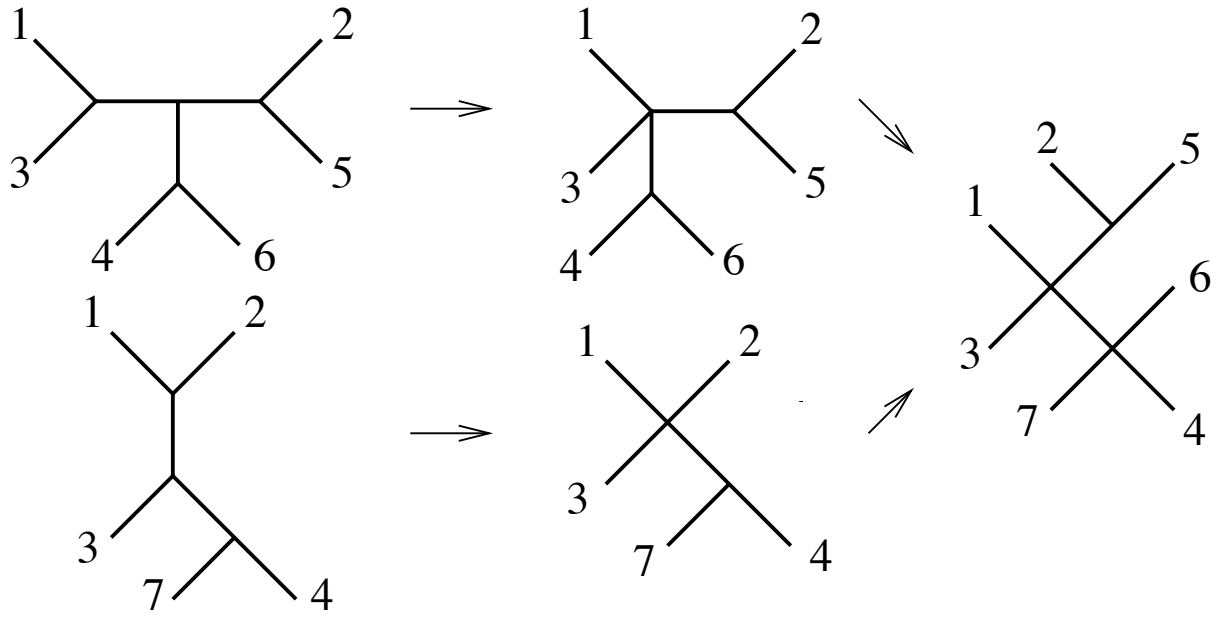
  - **For a given threshold $w$:**

    - Compute threshold graph $G$
    - Compute triangulation $G^*$ of threshold graph
    - Apply base-method to all maximal cliques in $G^*$
    - Merge trees guided by perfect elimination scheme

- Infer consensus of $\{T_w\}$.

# Merging Two Trees

Given trees on two overlapping sets of taxa, e.g. $\{1,2,3,4,5,6\}$ and $\{1,2,3,4,7\}$.




To merge the two trees together, first transform them (through edge contractions) so that they induce the same subtrees on their shared leaves and then combine them.

# Experimental Simulation Studies

- Choose model tree $T$ (e.g. inspired by biology)

# Experimental Simulation Studies

- Choose model tree $T$ (e.g. inspired by biology)

- Choose model of evolution (e.g. Jukes-Cantor model)

# Experimental Simulation Studies

- Choose model tree $T$ (e.g. inspired by biology)
- Choose model of evolution (e.g. Jukes-Cantor model)
- Evolve sequences along the model tree

# Experimental Simulation Studies

- Choose model tree $T$ (e.g. inspired by biology)
- Choose model of evolution (e.g. Jukes-Cantor model)
- Evolve sequences along the model tree
- Apply tree reconstruction method $M$ to evolved sequences $S$

# Experimental Simulation Studies

- Choose model tree $T$ (e.g. inspired by biology)
- Choose model of evolution (e.g. Jukes-Cantor model)
- Evolve sequences along the model tree
- Apply tree reconstruction method $M$ to evolved sequences $S$
- Compare estimation $M(S)$ with model tree $T$ for topological accuracy

# Experimental Simulation Studies

- Choose model tree $T$ (e.g. inspired by biology)
- Choose model of evolution (e.g. Jukes-Cantor model)
- Evolve sequences along the model tree
- Apply tree reconstruction method $M$ to evolved sequences $S$
- Compare estimation $M(S)$ with model tree $T$ for topological accuracy

Software: ecat, seqgen, PAUP, Phylip. Our programs in C++, LEDA.

# Objective: Topological Accuracy

One main goal in biology is to correctly infer the *order* of speciation events, hence the objective is to minimize:

# Objective: Topological Accuracy

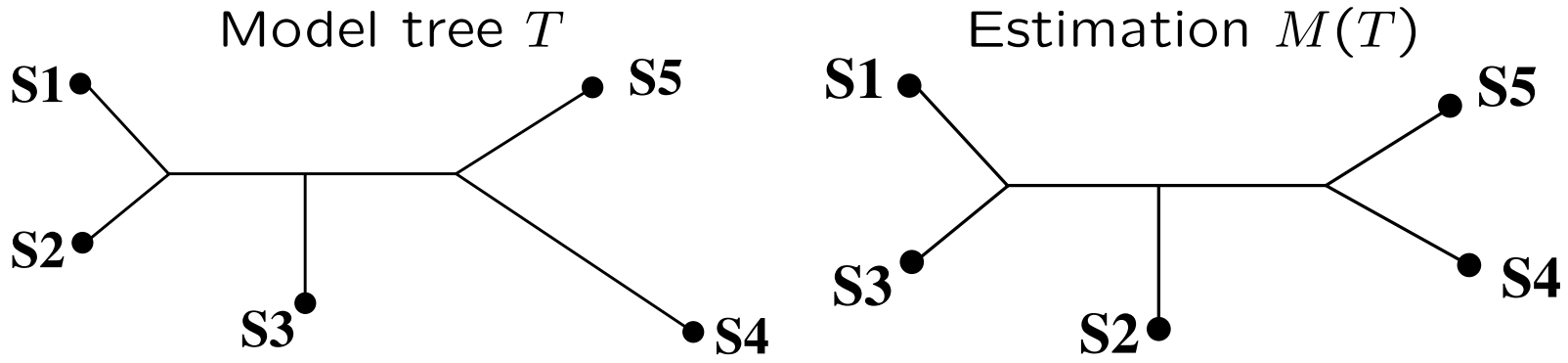
One main goal in biology is to correctly infer the *order* of speciation events, hence the objective is to minimize:

- *False positives*: wrongly inferred edges
- *False negatives*: missing edges

# Objective: Topological Accuracy

One main goal in biology is to correctly infer the *order* of speciation events, hence the objective is to minimize:

- *False positives*: wrongly inferred edges
- *False negatives*: missing edges

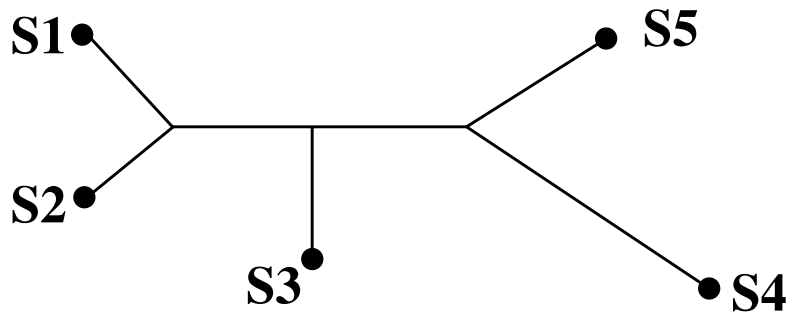


# Objective: Topological Accuracy

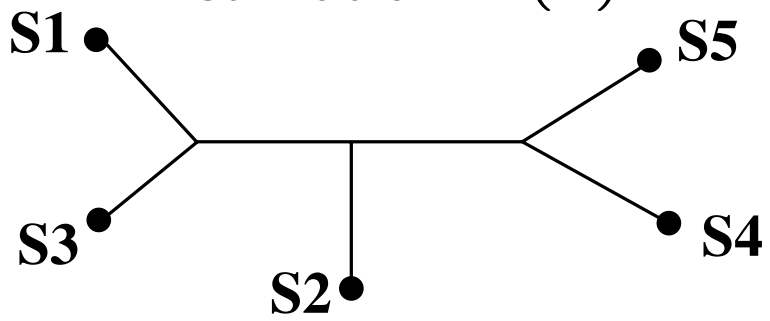
One main goal in biology is to correctly infer the *order* of speciation events, hence the objective is to minimize:

- *False positives*: wrongly inferred edges
- *False negatives*: missing edges

Model tree $T$

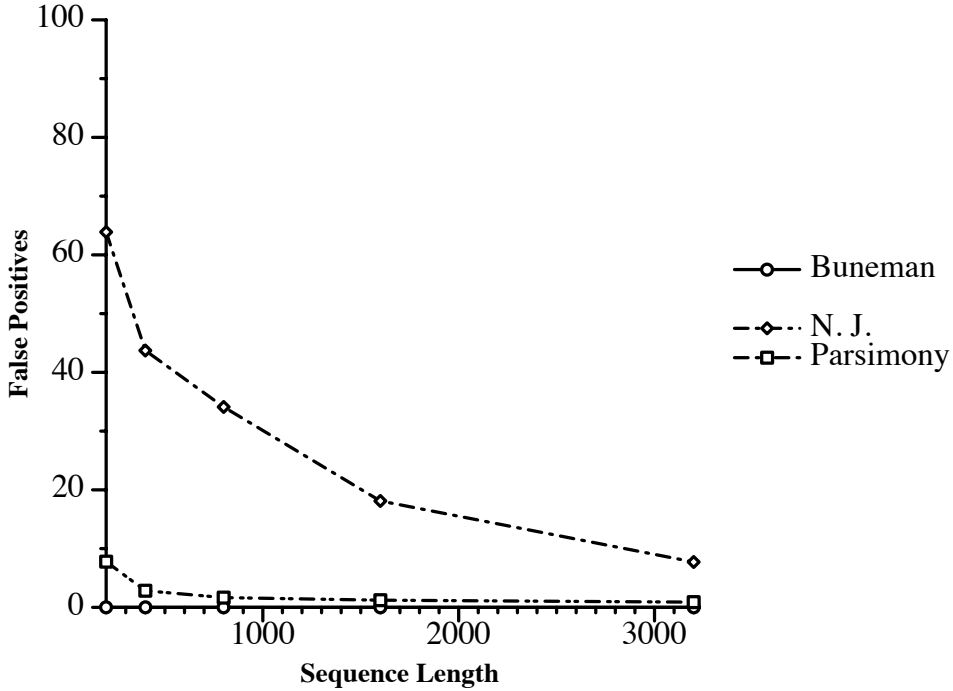



Estimation $M(T)$



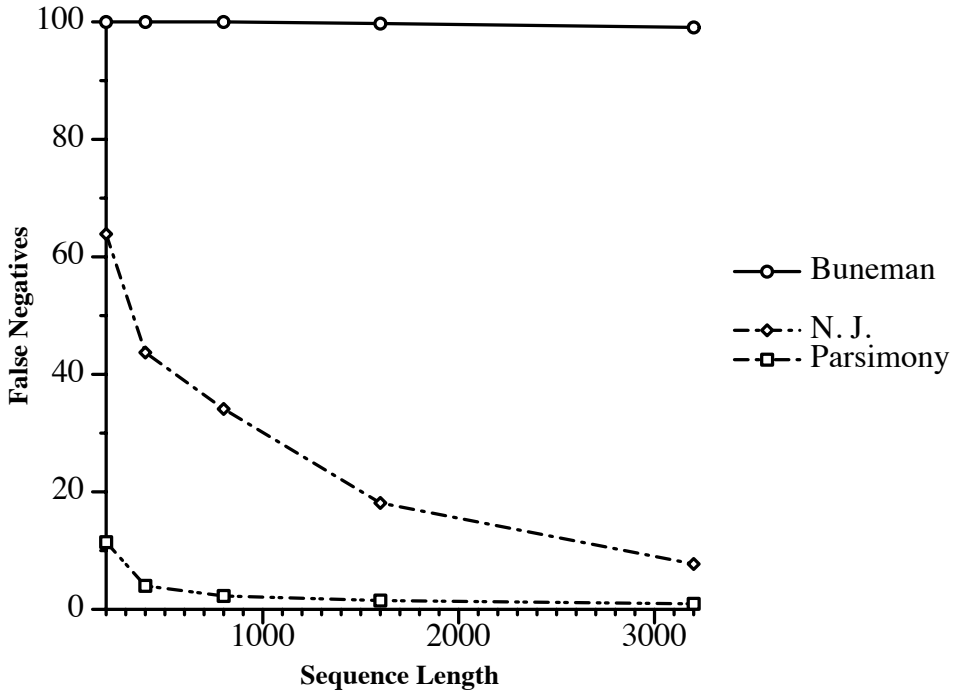

- One false positive: $\{S_1, S_3\}$ vs $\{S_2, S_4, S_5\}$
- One false negative: $\{S_1, S_2\}$ vs $\{S_3, S_4, S_5\}$

# Comparison of False Positive Rates



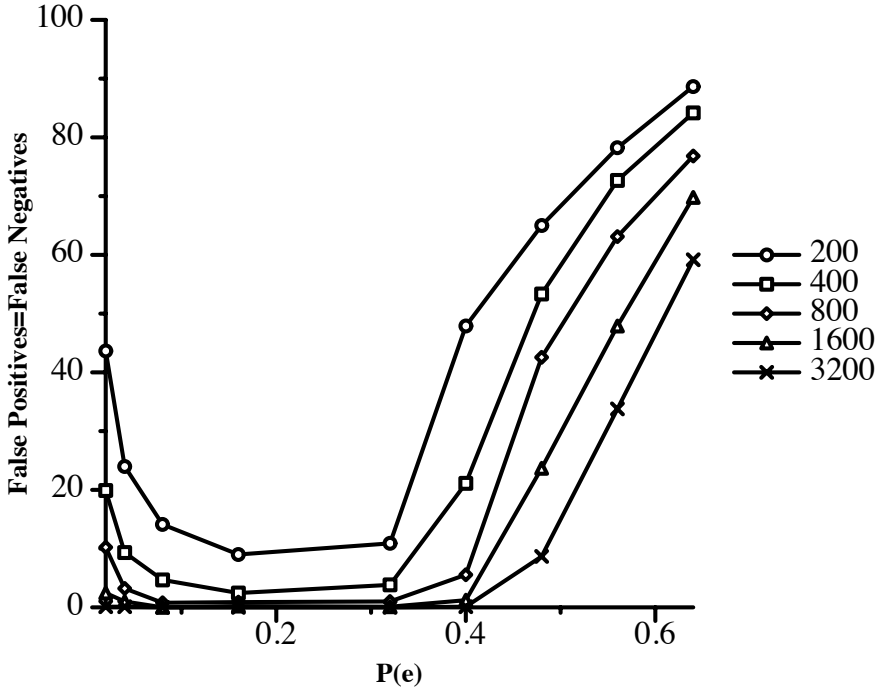

- 93 taxon tree (from 500 taxon *rbc*L dataset)
- maximum substitution probability $p(e)$ is 0.48
- 20 experiments per point

# Comparison of False Negative Rates




- 93 taxon tree (from 500 taxon *rbc*L dataset)
- maximum substitution probability $p(e)$ is 0.48
- 20 experiments per point

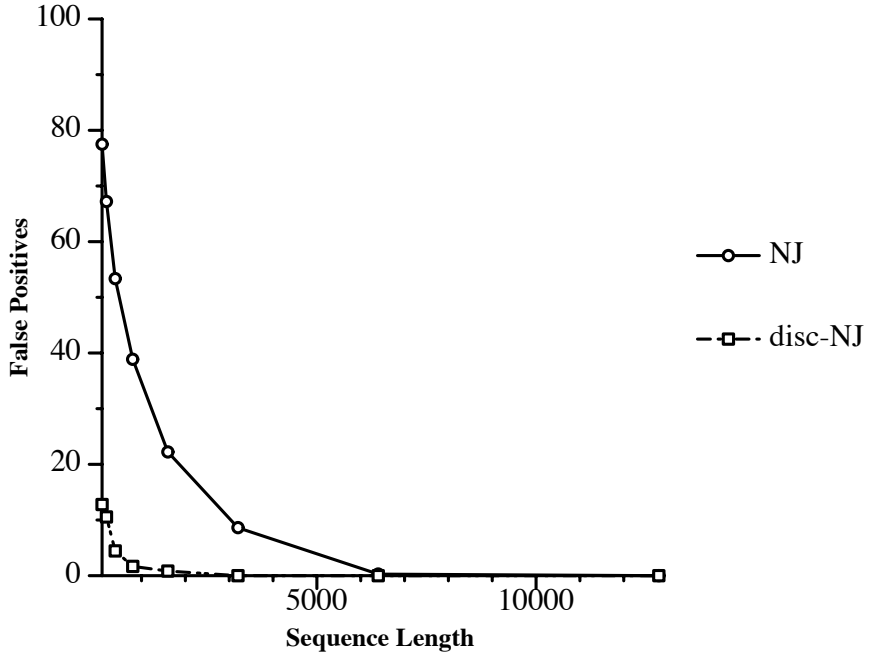
# Performance of Neighbor-Joining




P(e) vs. FP by Length

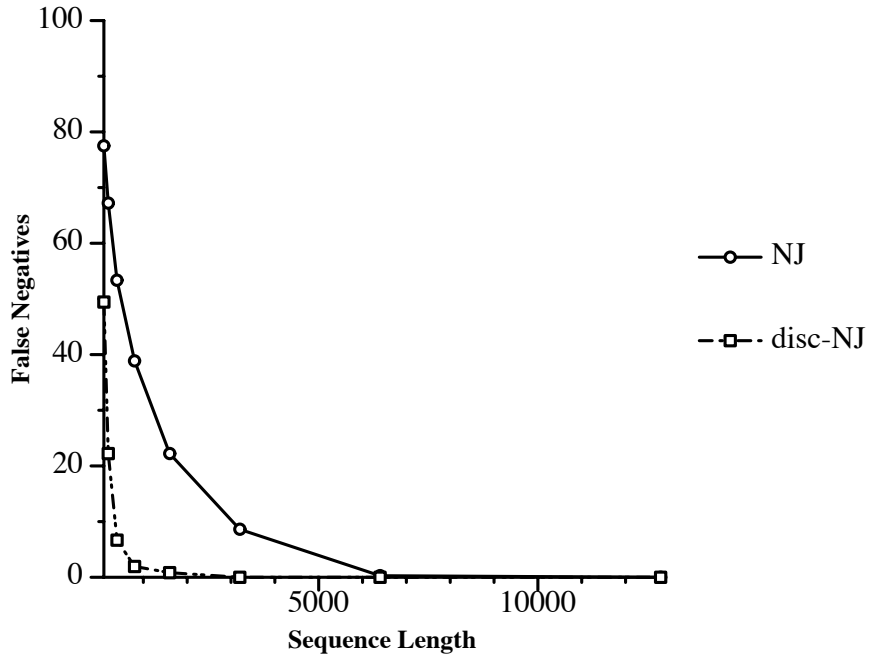
- 93 taxon tree (from 500 taxon *rbc*L dataset)
- maximum mutation probability $p(e)$ vs. FP (=FN) rate
- 20 experiments per point

# Neighbor-Joining vs DCM-NJ, **false positives**




- 93 taxon tree, maximum mutation probability $p(e) = 0.48$
- 10 experiments per point
- *Greedy asymmetric median tree*, i.e. consensus over all trees $\{T_w\}$.

# Neighbor-Joining vs DCM-NJ, **false negatives**




- 93 taxon tree, maximum mutation probability $p(e) = 0.48$
- *Greedy asymmetric median tree*, i.e. consensus over all trees $\{T_w\}$.
- 10 experiments per point

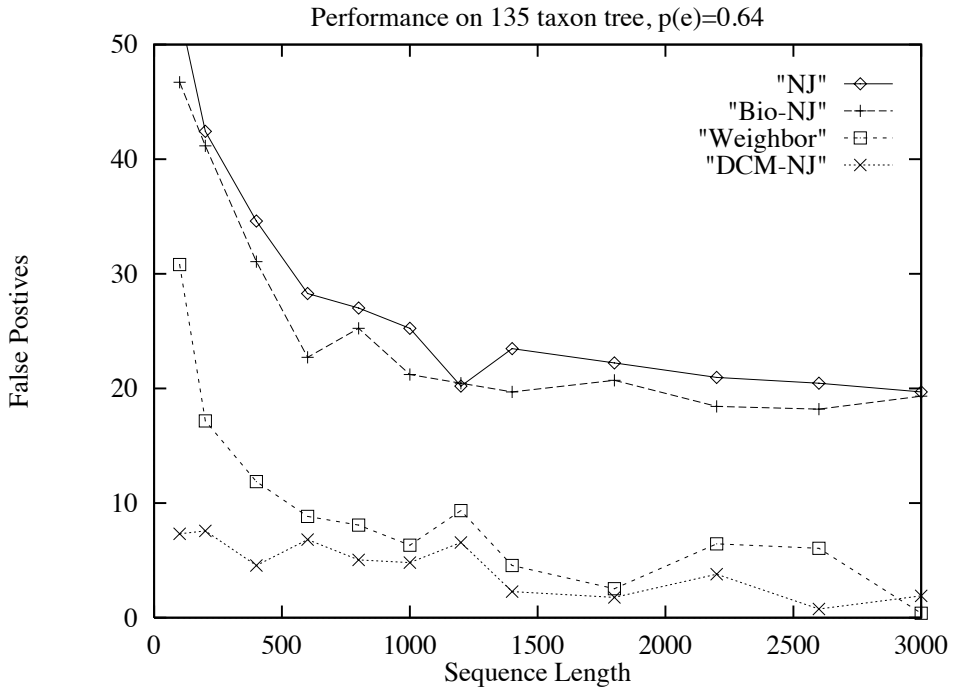
# New Variants of Neighbor-Joining

- Bio-NJ
  - Improved version of the NJ algorithm based on a simple model of sequence data
  - Oliver Gascuel, 1997.

# New Variants of Neighbor-Joining

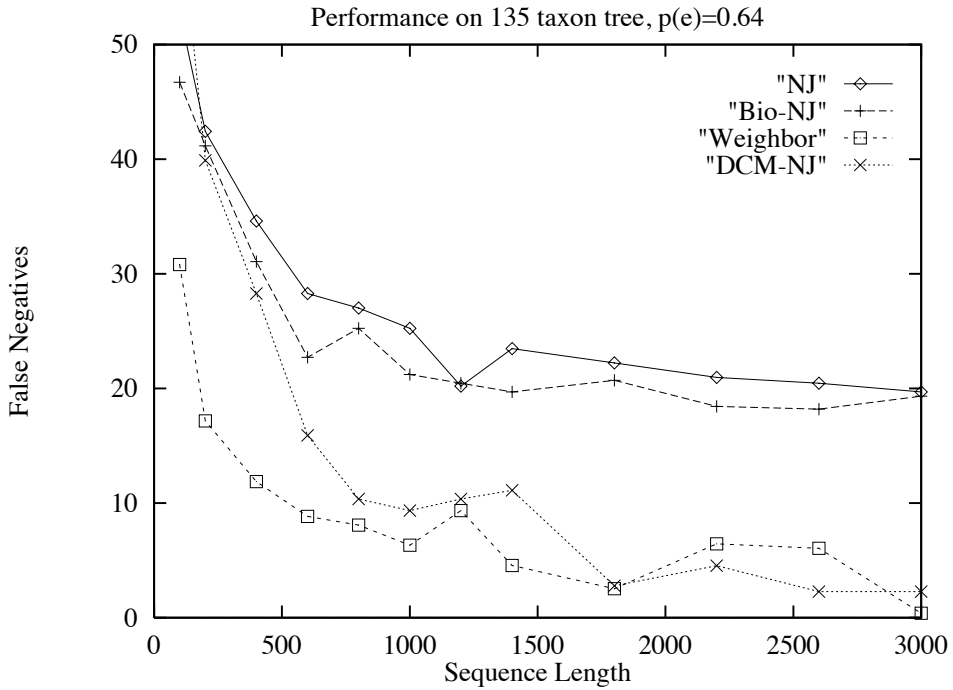
- Bio-NJ
  - Improved version of the NJ algorithm based on a simple model of sequence data
  - Oliver Gascuel, 1997.
- Weighbor
  - Weighted Neighbor-Joining
  - W. Bruno, A. Halpern & N. D. Socci, 1998

# NJ, Bio-NJ, Weighbor & DCM-NJ, FP



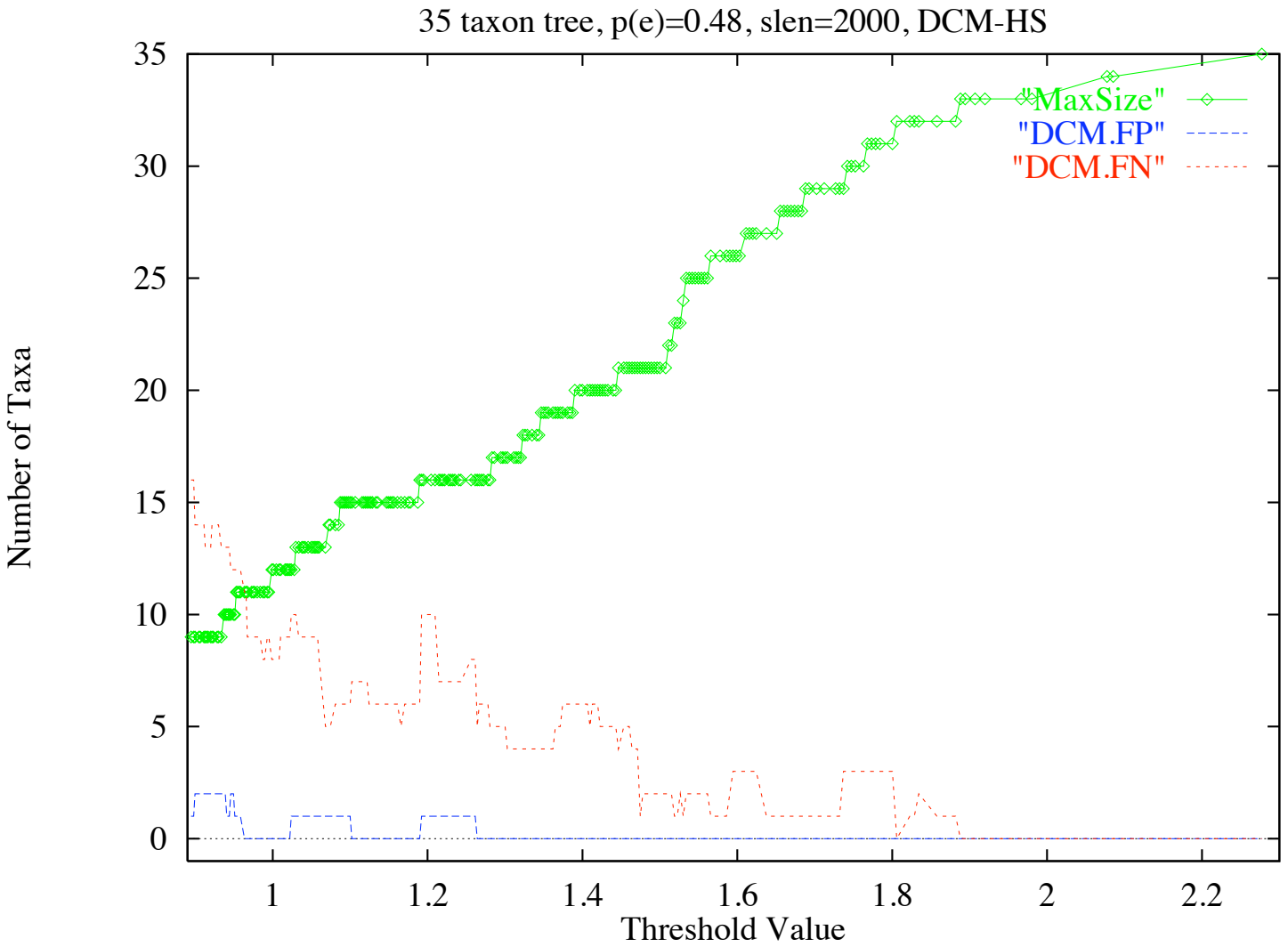

- 135 taxon tree, maximum mutation probability $p(e) = 0.64$
- *Greedy asymmetric median tree* of a small subset of $\{T_w\}$.
- 3-5 experiments per point, work in progress. . .

# NJ, Bio-NJ, Weighbor & DCM-NJ, FN

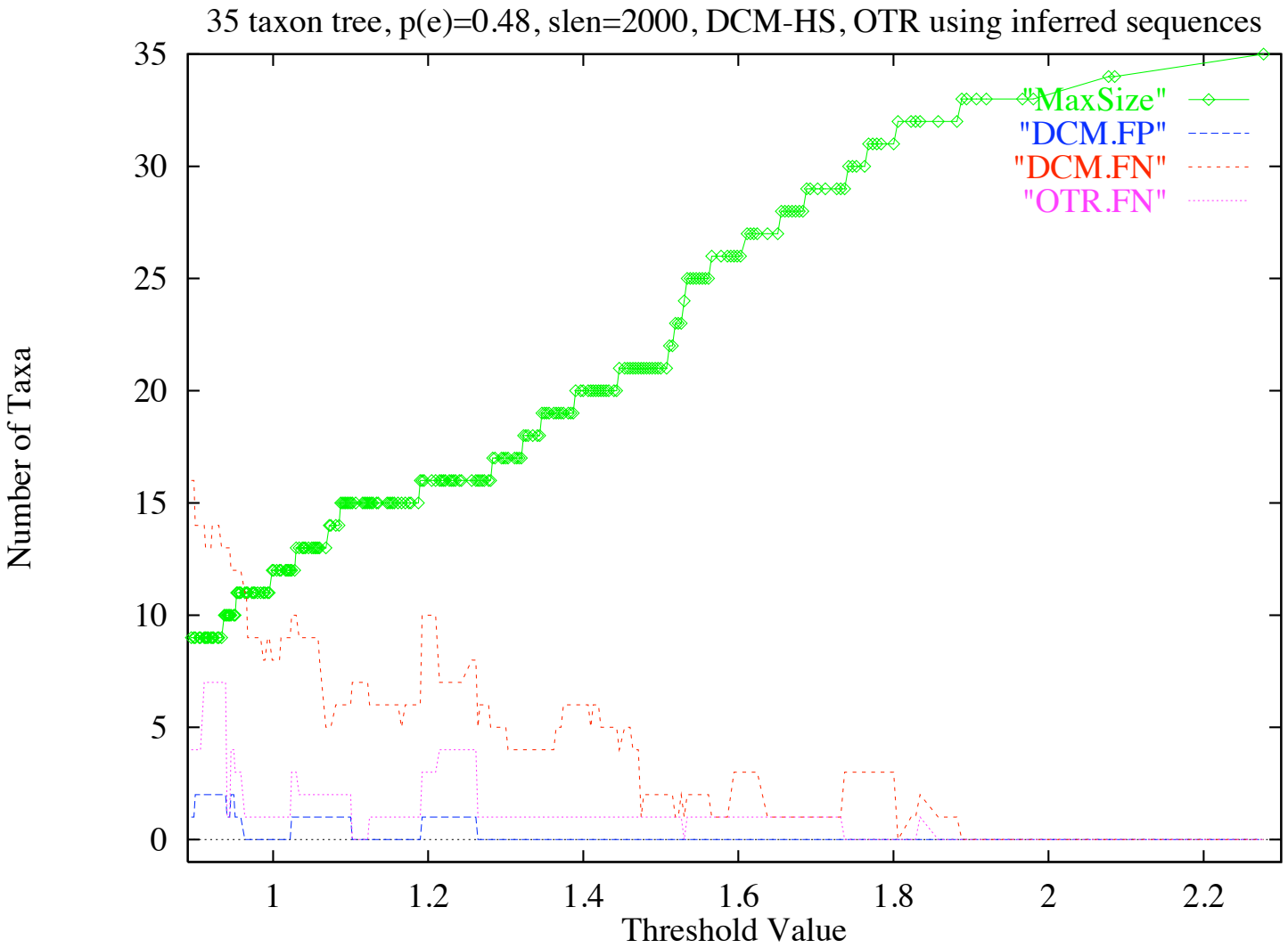



- 135 taxon tree, maximum mutation probability $p(e) = 0.64$
- *Greedy asymmetric median tree* of a small subset of $\{T_w\}$.
- 3-5 experiments per point, work in progress. . .

# Speeding up Parsimony and MLE

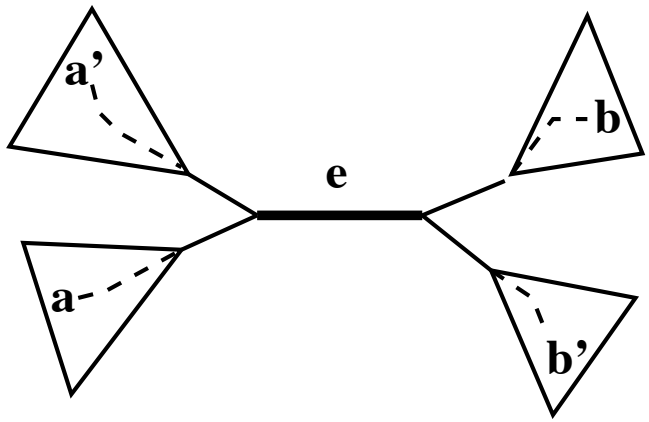


# Speeding up Parsimony and MLE



# Threshold and Merge Step

Let $T$ be a model tree and $d$ an estimated distance matrix. A *short quartet* around an internal edge $e$ is a set of four taxa $a, a', b, b'$ that lie in the four subtrees induced by $e$, of minimal width.




**Theorem** If the threshold $w$ is chosen large enough such that every short quartet induces a four-clique in $G^*$, then every merger is unique and a DCM method will recover the model tree $T$, if the base method is accurate on the base problems.

# Sequence Lengths Required for Accuracy

The length of biological sequences obtainable for phylogenetic analysis is bounded by a *few thousand* base pairs, so the question how sequence length affects performance is critical.

The sequence lengths that suffice for accuracy of distance methods such as Neighbor-Joining or the Buneman Tree grow **exponentially** in the divergence of the model tree.
(Atteson 1997, Erdös et al. 1997)

For DCM-boosted distances methods we can show:

> For almost all trees, **polylogarithmic** length suffices for accuracy with high probability, and **polynomial** length suffices for all trees with high probability.

# Conclusion and Future Research

By reduction to small and closely related-datasets, the DCM-method can substantially improve the accuracy and/or time requirements of phylogenetic tree reconstruction methods for large and divergent datasets.

Future research will focus on:

- More advanced simulations studies
- Parsimony and MLE
- exploring the application to multiple sequence alignment
- investigating performance on real datasets
- **developing a public version of the software.**

# References


- D.H. Huson, S. Nettles, L. Parida, T.J. Warnow and S. Yooseph. The Disk-Covering Method for Tree Reconstruction. In: R. Battiti and A.A. Bertossi eds., Proceedings of *Algorithms and Experiments* (ALEX'98) (Trento, Italy, Feb. 9–11, 1998), 62-75, 1998.

- D.H. Huson and K. Rice and S. Nettles and T.J. Warnow and S. Yooseph. Hybrid Tree Construction Methods. *Workshop on Algorithms Engineering* (WAE'98). Submitted by invitation to special edition of *JEA*.

- D.H. Huson, S. Nettles and T.J. Warnow. Obtaining highly accurate topology and evolutionary estimates of evolutionary trees from very short sequences. *Recomb'99*.

- D.H. Huson, L. Vawter and T.J. Warnow. Solving Large Scale Phylogenetic Problems using DCM. Submitted to: *ISMB99*.


http://www.mathematik.uni-bielefeld.de/~huson