Phylogenetic Reconstruction: Handling Large Scale and Complex Data

Bernard M.E. Moret

Department of Computer Science
University of New Mexico

Acknowledgments

- Main collaborators:

  Tandy Warnow (UT Austin CS)
  Robert Jansen and Randy Linder (UT Austin Biology)
  David Bader (UNM Comp. Eng.)

- Support:

  US National Science Foundation
  US National Institutes of Health
  Alfred P. Sloan Foundation
  IBM Corporation

Overview

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Phylogenies

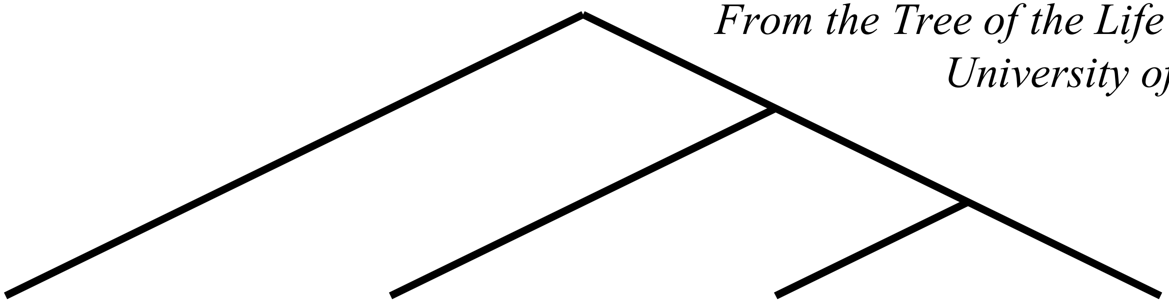
A phylogeny is a reconstruction of the evolutionary history of a collection of organisms.

It usually takes the form of a tree.

- Modern organisms are placed at the leaves.
- Edges denote evolutionary relationships.
- “Species” correspond to edge-disjoint paths.

The Great Apes

Phylogeny



From the Tree of the Life Website,
University of Arizona

Orangutan Gorilla Chimpanzee Human

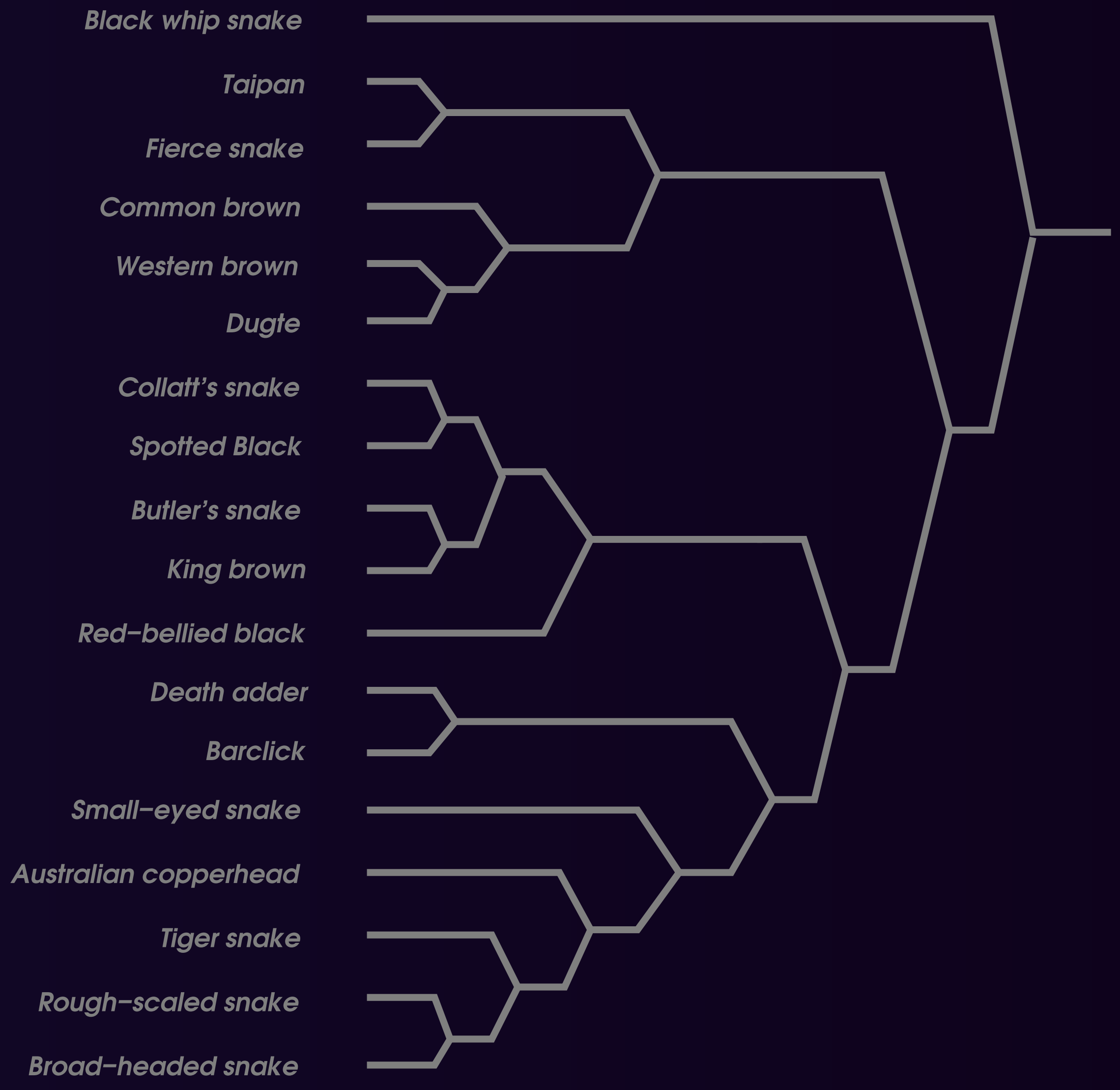
Phylogenies: Why?

Phylogenies provide the framework around which to organize all biological and biomedical knowledge.

They help us understand and predict:

- functions of and interactions between genes
- relationship between genotype and phenotype
- host/parasite co-evolution
- drug and vaccine development
- origins and spread of disease
- origins and migrations of humans

Example: Antivenins

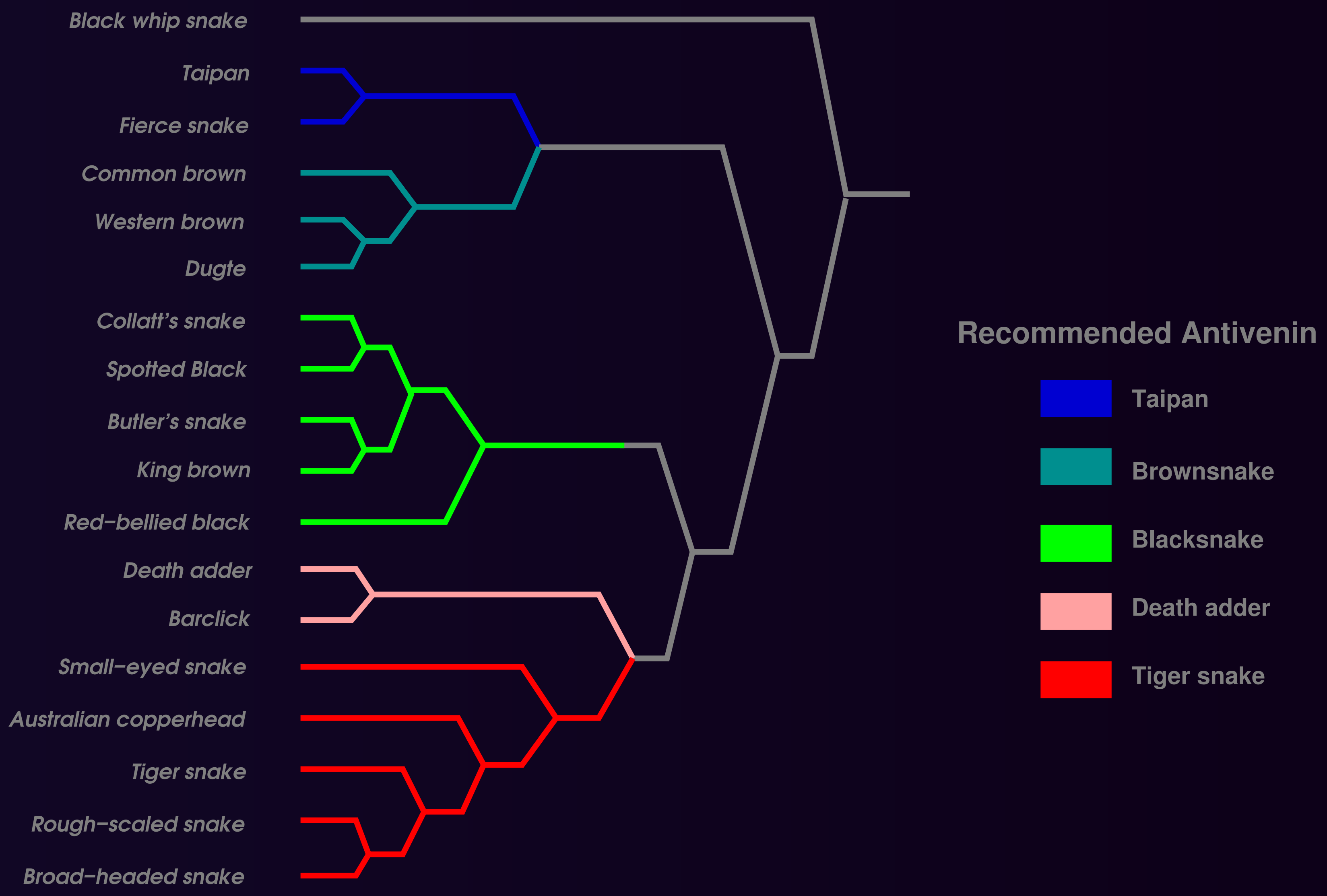


Venomous

Australian

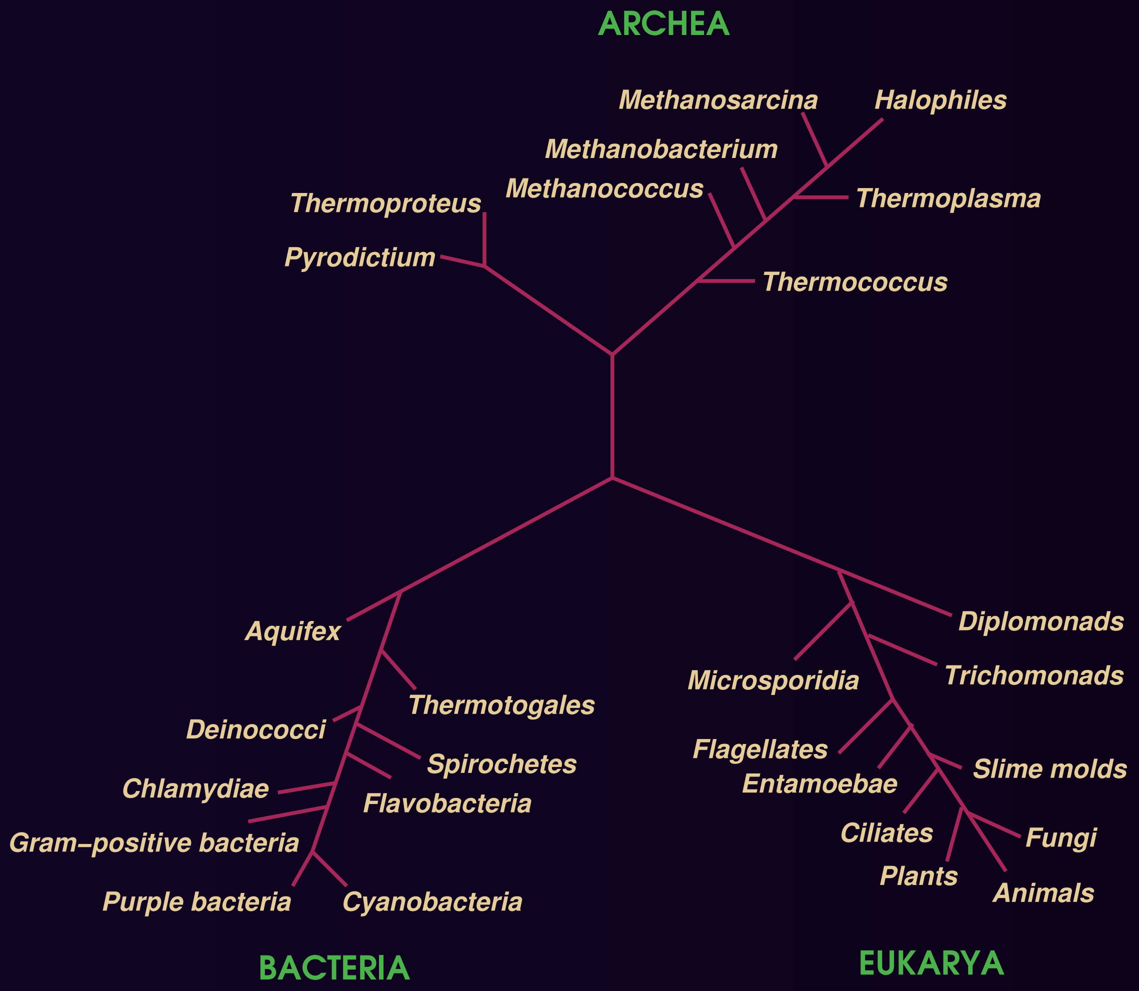
Snakes

Example: Antivenins



The Tree of Life

It is to Biology what the periodic table is to Chemistry



Scale of The Tree of Life

- 1,5 million described species.
- 10 million to 200 million existing species.
- Reconstruction tools can handle around 500 organisms.
- Reconstruction tools scale exponentially with the amount of data.

- Phylogenies: What and Why?
- Phylogenetic Reconstruction:
  a fast review from a CS standpoint
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Phylogenetic Reconstruction

Two categories of methods:

- Criterion-Based methods, such as Maximum Parsimony (MP) and Maximum Likelihood (ML)
- Ad hoc, usually *distance-based* and using clustering ideas, such as Neighbor-Joining (NJ)

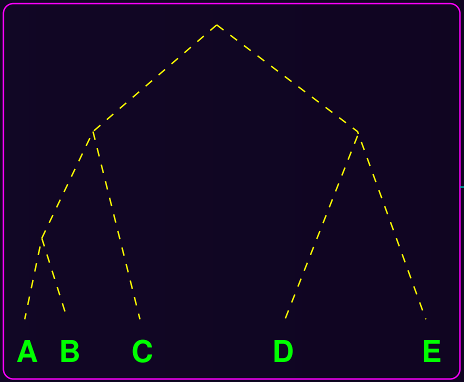
In addition:

- Meta-methods decompose the data into smaller subsets, construct trees on those subsets, and use the resulting trees to build a tree for the entire dataset (quartets, disk-covering)

Phylogenetic Distances

- **True evolutionary distance:**
  the *actual* number of evolutionary events that took place to transform one datum into the other.
- **Edit distance:**
  the *minimum* number of permitted evolutionary events that can transform one datum into the other.
- **Estimated evolutionary distance:**
  our best *estimate* of the true evolutionary distance, obtained heuristically or by correcting the edit distance according to a model of evolution.

Distance-Based Methods



(Unknown) True Tree

Extract data on extant taxa

A	acaattagaacta
B	acccttagaccta
C	acacttcgaccca
D	acacagagaacca
E	acccatagaacta

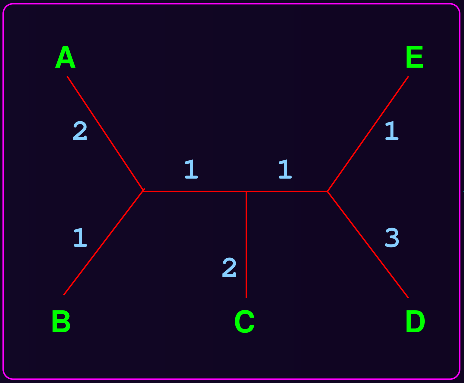
Molecular Data

Estimate pairwise distances

	B	C	D	E
A	3	5	6	3
B		4	6	3
C			5	6
D				4

Distance Matrix

Neighbor–joining



Inferred Tree

Parsimony-Based Methods

- Aim to minimize total *number of character changes*.
- Assume that characters are *independent*.
- Reconstruct *ancestral data*.
- Are known not to be statistically consistent with sequence data, but yield good results in most cases.
- Finding most parsimonious tree is NP-hard.
- Optimal solutions are limited to sizes around 30. Heuristic solutions are fairly good to sizes of 500.

Likelihood-Based Methods

- Aim to return tree with highest likelihood of having produced the observed data.
- Are based on a specific model of evolution and usually *estimate model parameters*.
- Produce *likelihood estimate* (prior or posterior conditional) for each tree.
- Are statistically consistent for most models.
- Even scoring a fixed tree is very expensive.
- Optimal solutions are limited to specific sets of 4 taxa. Heuristics run to completion on at most 10 taxa, but appear good to about 100 taxa (e.g., PhyML).

Meta-Methods

Decompose dataset into smaller, overlapping subsets, reconstruct trees for the subsets (with a *base* method), and combine results into a tree for the entire dataset.

- Quartet-based methods: use all possible smallest subsets (quartets); include Q* and Tree-Puzzle.

  Slow and inaccurate regardless of base method.

- Disk-Covering methods (DCMs): decompose the dataset into overlapping “disks" (tight subsets).

  High-powered machinery *succeeds*, especially when tree is imbalanced.

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Scaling Up: The Issues

- *Distance-based methods are (fairly) fast, but not accurate enough on large problems (large evolutionary diameter).*
- *Criterion-based methods take days for a few hundred taxa and scale exponentially.*
- *All methods perform better with longer sequences and larger state spaces, but biological sequences are bounded.*

Scaling Up: The Requirements

- *Distance-based methods are (fairly) fast, but not accurate enough on large problems.*

  Decompose large problems into smaller ones so as to reduce evolutionary diameter.

- *Criterion-based methods take days for a few hundred taxa and scale exponentially.*

  Use algorithmic techniques to bypass the exponential growth, such as divide-and-conquer.

- *All methods perform better with longer sequences and larger state spaces, but biological sequences are bounded.*

  Design methods that converge on short sequences, so-called *fast converging methods*.

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Scaling Up: Disk-Covering

Basic idea:

- decompose dataset into *overlapping* compact subsets—the *disks*
- reconstruct a tree for each subset
- assemble these trees into a single tree

Variations so far: DCM1, DCM2, DCM3, recursive versions, iterative versions

DCM Decompositions

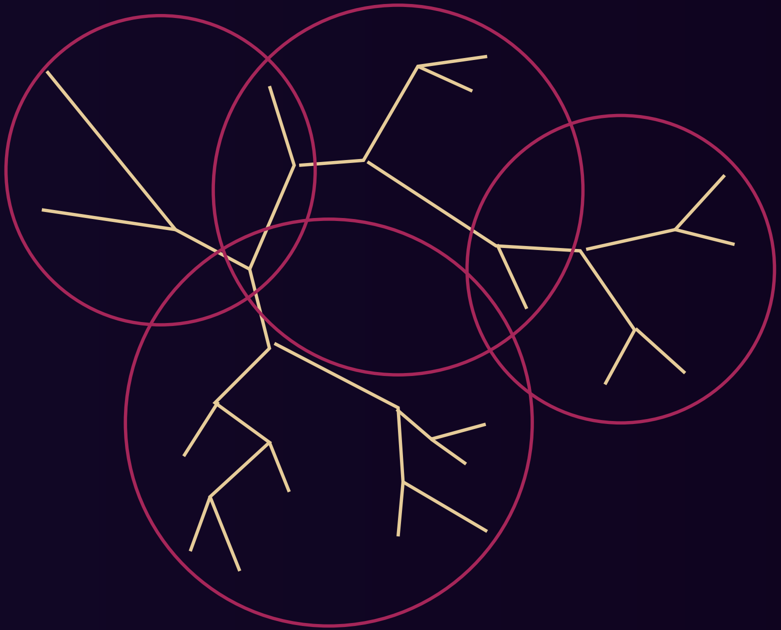
Given set S of items, distance matrix on S, $D = (d_{ij})$, and threshold T:

- *Construct threshold graph* $G = (S, \{(i, j) \mid d_{ij} \leq T\})$.
- *Compute min. triangulation of* G.
- DCM1: *find all maximal cliques, then each clique is a disk.*

  DCM2: *find graph separator* X, *let* $\{S_i\}$ *be connected components of* $G - X$, *then each* $X \cup S_i$ *is a disk.*

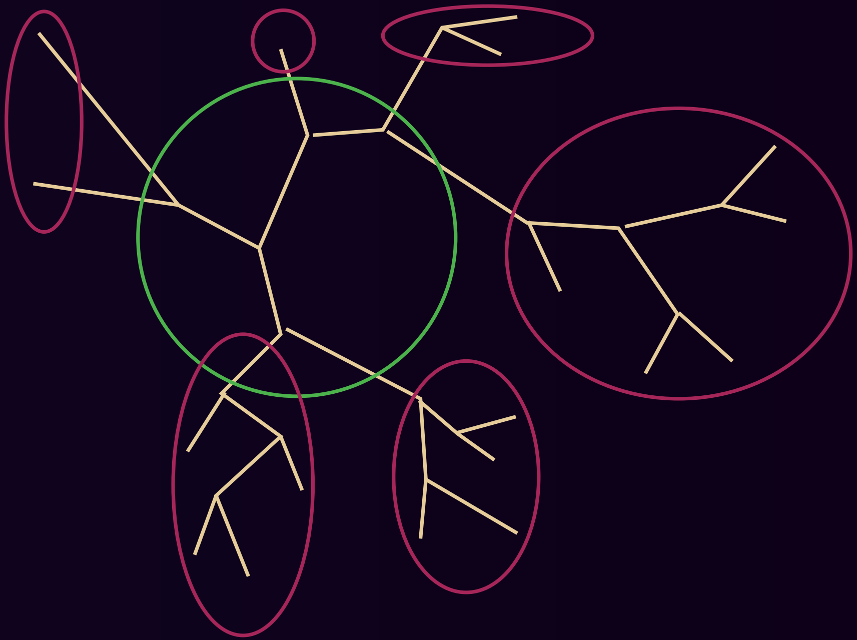
DCM1 and DCM2

DCM1



4 disks

DCM2



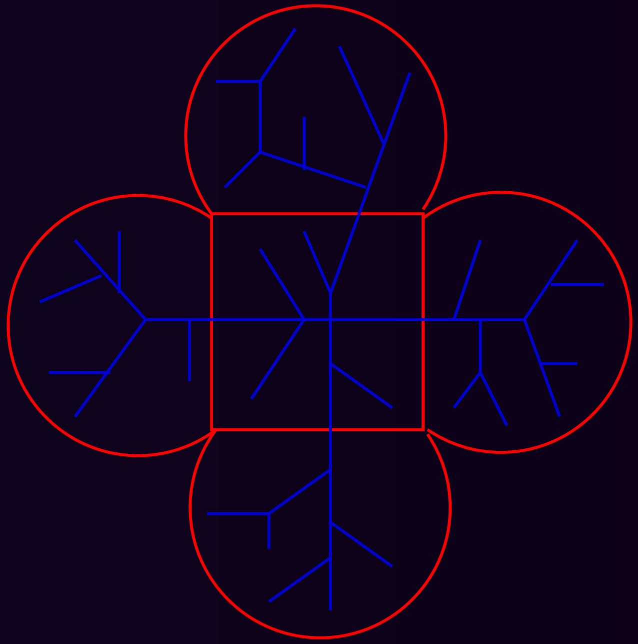
separator in green

Improvement: DCM3

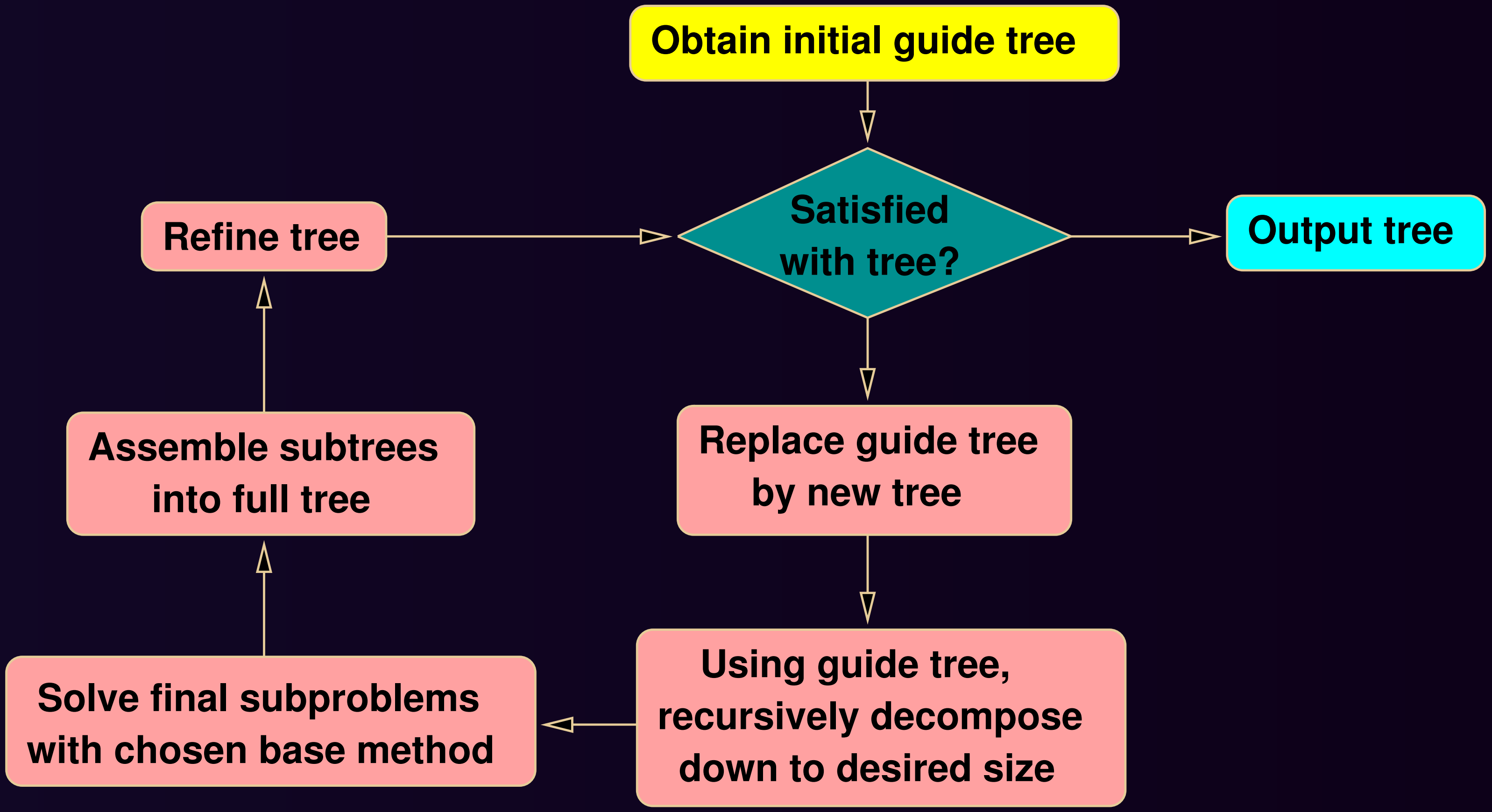
DCM1 and DCM2: decomposition based on distance matrix only

DCM3: use best tree so far to guide the decomposition

Given set S and tree T, compute short subtree graph $G(S,T)$ and find *clique separator* in G to form subproblems.



Using DCM3: Recurse & Iterate


Obtain initial guide tree

Satisfied with tree?

Output tree

Replace guide tree by new tree

Using guide tree, recursively decompose down to desired size

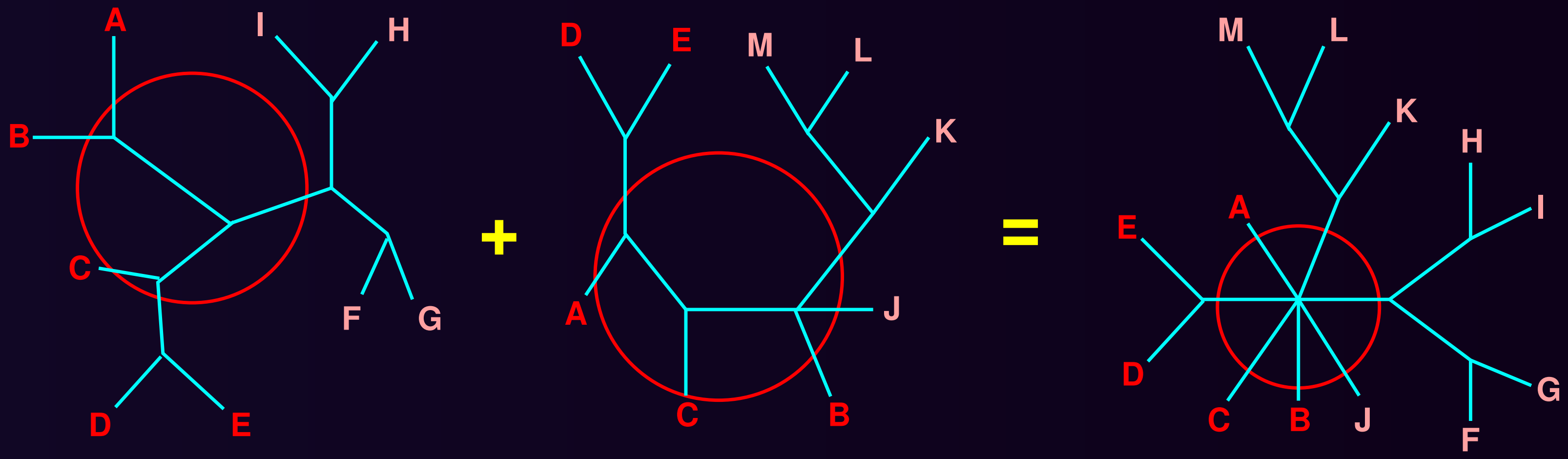
Solve final subproblems with chosen base method

Assemble subtrees into full tree

Refine tree

Merging Trees in DCM

We designed a specialized supertree method for DCMs: the *strict consensus merger (SCM)*

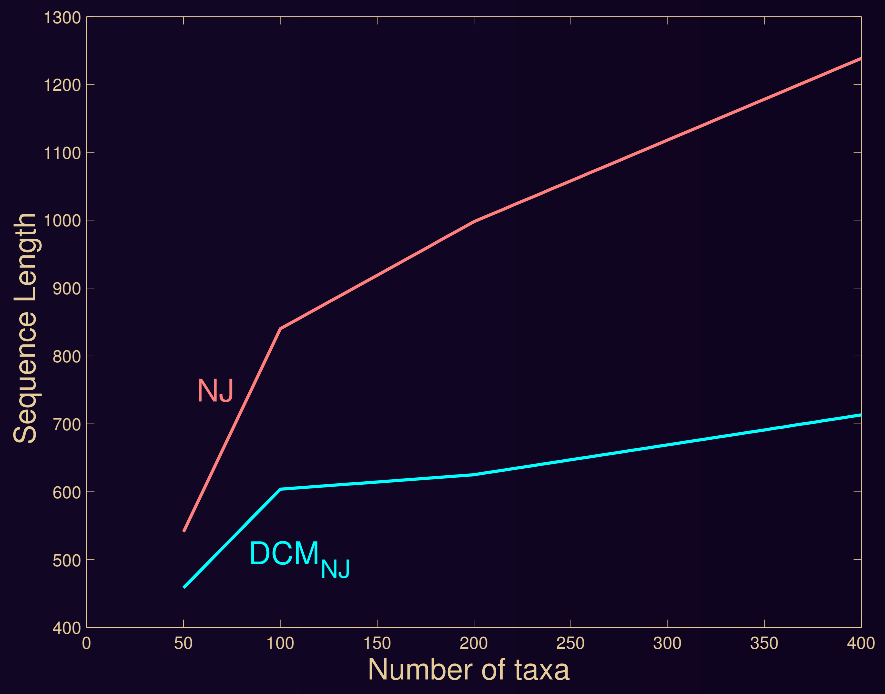


SCM tends to produce many polytomies

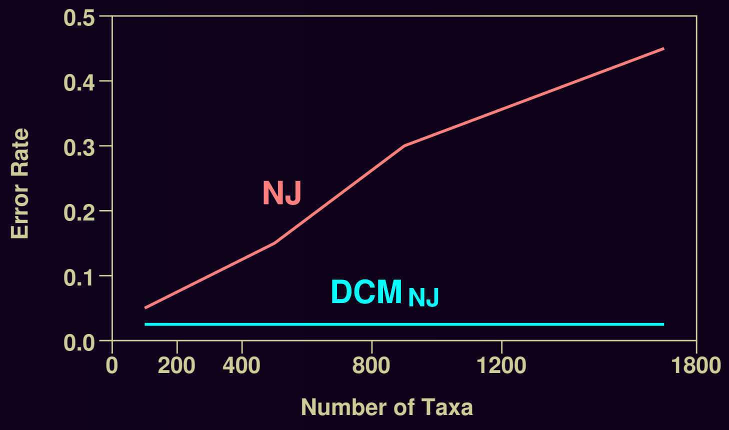
Results with DCM1 and NJ

using Kimura 2-parameter plus Γ model

reduced sequence length
(0.15 error rate)



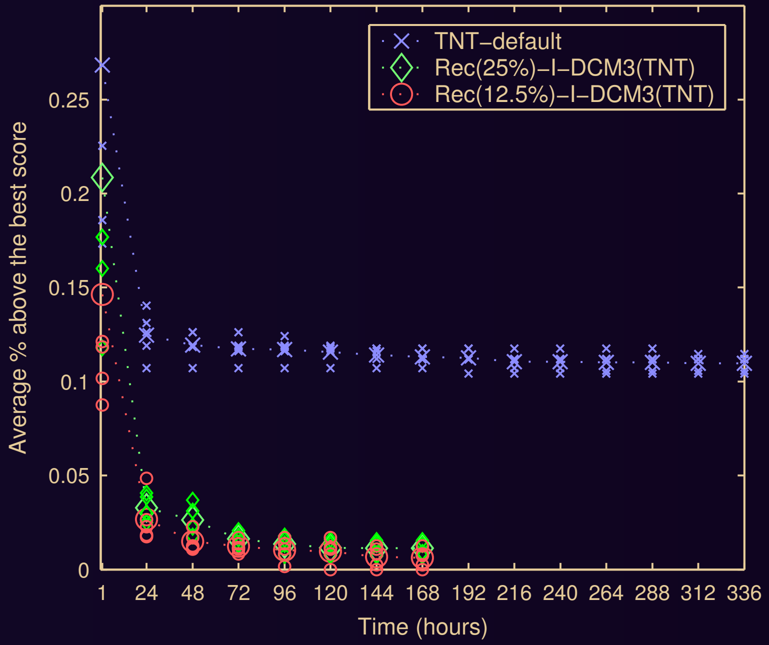
reduced error rate
(1,000 sequence length)



Results with Rec-I-DCM3 and MP

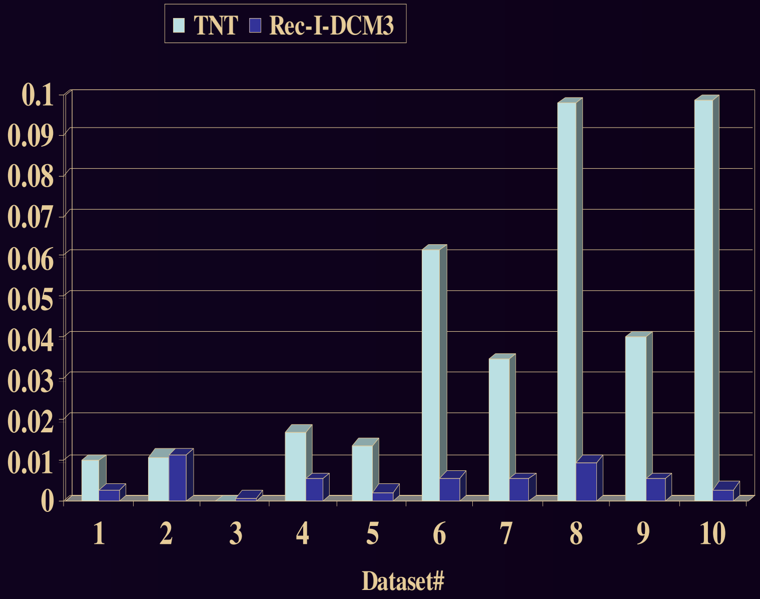
Rec-I-DCM3(TNT) vs. TNT

10,000 RNA sequences



10 datasets

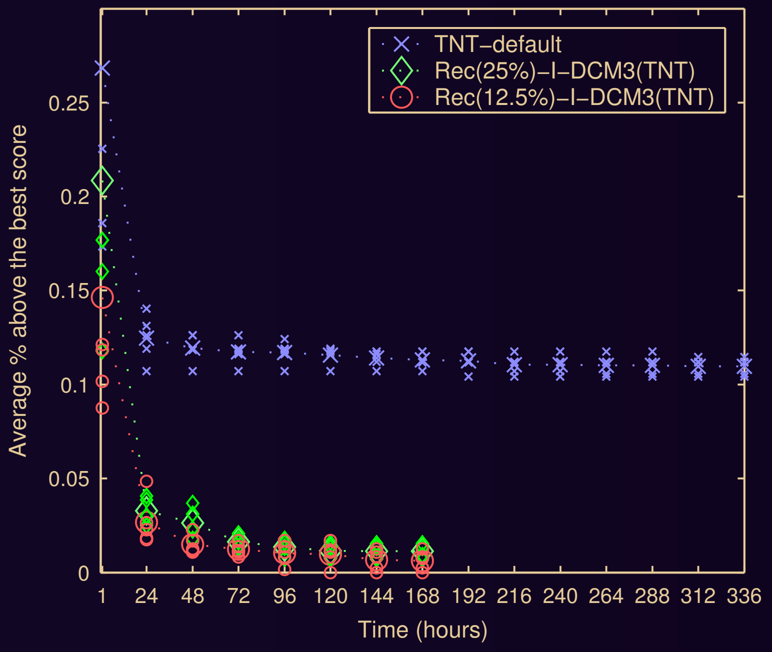
(from 4,000 to 15,000)



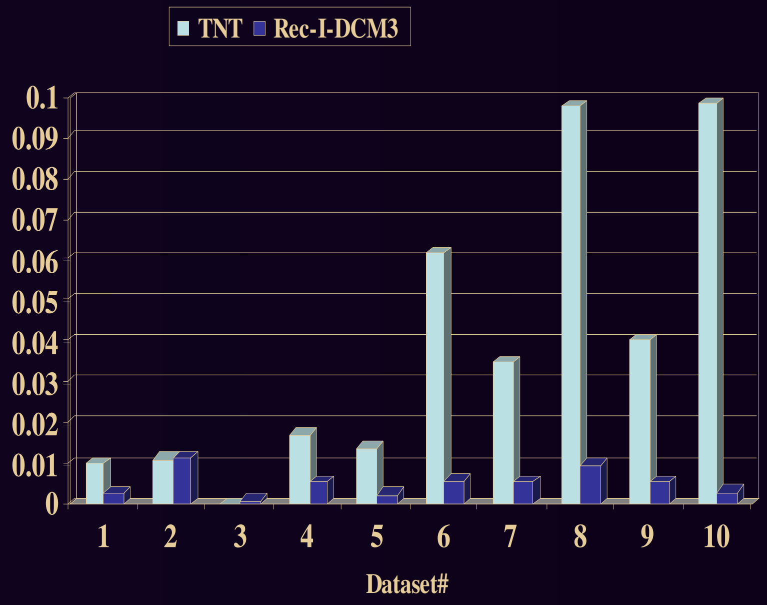
Results with Rec-I-DCM3 and MP

Rec-I-DCM3(TNT) vs. TNT

10,000 RNA sequences



10 datasets
(from 4,000 to 15,000)



Finding: 0.01% error is the maximum allowed!!

Scaling Up: Current and Future

- DCM4: combine DCM1 and guide tree to obtain smaller subsets.
- Develop a statistical framework to enable DCM approaches to Bayesian and ML reconstruction.
- Design new supertree algorithms to maximize resolution.
- Include direct database storage and retrieval within the algorithms (*lab notebook*).
- Test scaling to tens of millions of taxa using highly accurate simulations of sequence evolution.

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Phylogenetic Data

- All kinds of data have been used: behavioral, morphological, metabolic, etc.
- Current data of choice are molecular data.
- Two main kinds of molecular data:
    - **sequence data**
      (nucleotide/codon sequences from genes)
    - **gene-order data**
      (gene ordering on chromosomes)

Sequence Data: Attributes

- **Advantages:**
  - Large amounts of data.
  - Familiar data, many tools.
  - Accepted models of character evolution.
- **Problems:**
  - Few character states, so high risk of homoplasy.
  - Poor models of sequence evolution.
  - Multiple alignments poorly solved.
  - Gene evolution different from organism evolution; recombination problematic for lineage sorting.

Gene-Order Data

The ordered sequence of genes on one or more chromosomes.

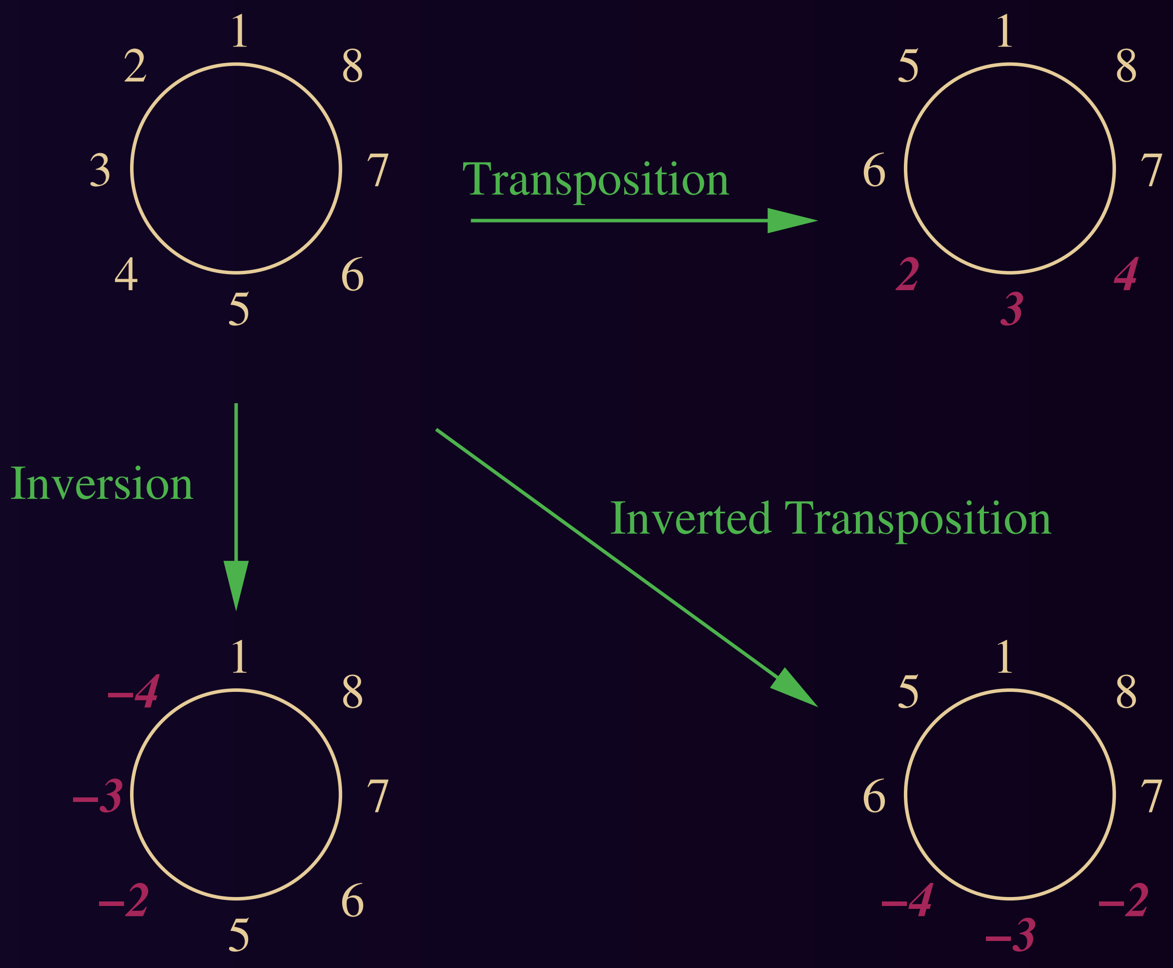
The entire gene order is a *single character*, which can assume a huge number of states.

Evolves through inversions, insertions (incl. duplications), and deletions; also transpositions (seen in mitochondria) and translocations (between chromosomes).

- Need to identify genes and gene families.
- Need to refine model for specific organisms to handle operons, exons, etc.

Genome Rearrangements

Model based on three types of rearrangements:



Gene-Order Data: Attributes

- **Advantages:**
  - Rare genomic events (*sensu* Rokas/Holland) and huge state space, so very low risk of homoplasy.
  - No need for alignments.
  - No gene tree/species tree problem.
- **Problems:**
  - Mathematics *much more complex* than for sequence data.
  - Models of evolution not well characterized.
  - Very limited data (mostly organelles and bacteria).

Gene-Order Data vs. Sequence Data

	Sequence	Gene-Order
evolution	*fast*	*slow*
data type	*a few genes*	*whole genome*
data amount	*abundant*	*sparse*
models	*good (sites)* *primitive (seqs.)*	*primitive*
computation	*easy*	*hard*

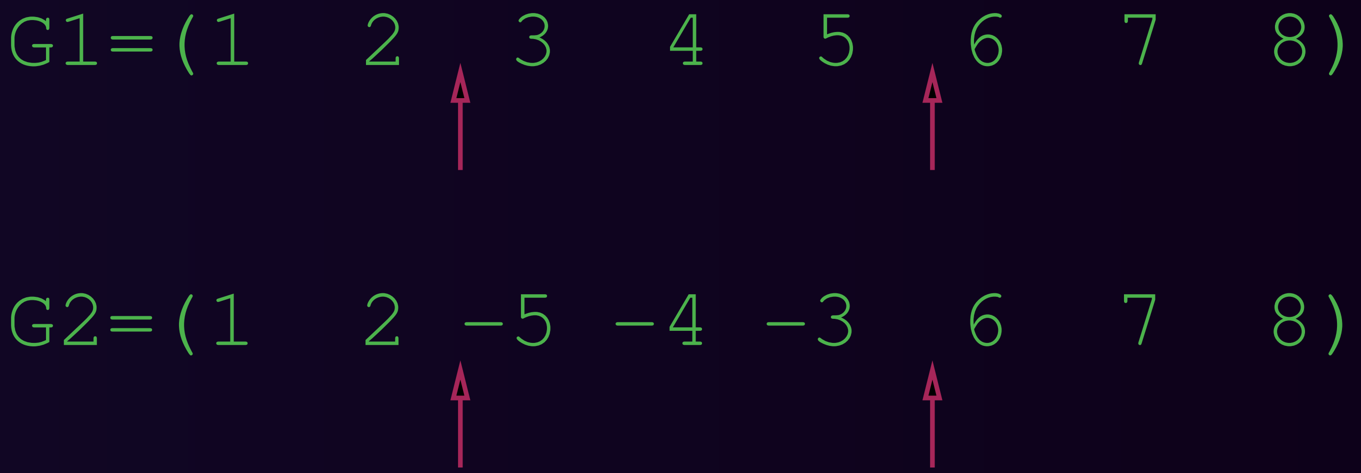
- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Distances for Gene-Order Data

- BP [Sankoff et al. 1998]:
  Distance counting the number of altered adjacencies (breakpoints) for identical gene content; linear time.
- INV [Bader/Moret/Yan 2001]:
  Edit distance (inversions) for identical gene content; linear time.
- EDE, IEBP [Moret et al. 2002, Wang/Warnow 2001]:
  Distance corrections to estimate true evolutionary distance; quadratic time.
- INV-DEL [El-Mabrouk 2000]:
  Edit distance (inversions and insertions/deletions, but no duplications); linear time [Liu/Moret 2003].
- ALL [Marron/Swenson/Moret 2003]:
  Estimated evolutionary distance (inversions, insertions/deletions, duplications).

Breakpoint Distance

The number of adjacencies present in one genome, but not the other.



```
G1=(1   2   3   4   5   6   7   8)

G2=(1   2  -5  -4  -3   6   7   8)
```

Inversion Distance

Given signed gene orders of equal content, compute the inversion-only edit distance.

- Hannenhalli/Pevzner 1995: cubic time
- Kaplan/Shamir/Tarjan 1997: quadratic time
- Bader/Moret/Yan 2001: linear time

Gene-Order Distances in General

Signed gene orders may include duplicates, need not have identical gene content.

Previous work (not useable for phylogeny):

- Exemplar heuristic for duplications by Sankoff (NP-hard).
- Exact inversions plus deletions, but no duplications allowed, by El-Mabrouk.
- Heuristic by Bourque, used only on very small sets.

Our work:

- Bounded approximation for unequal gene content.
- Direct estimate of evolutionary distance.

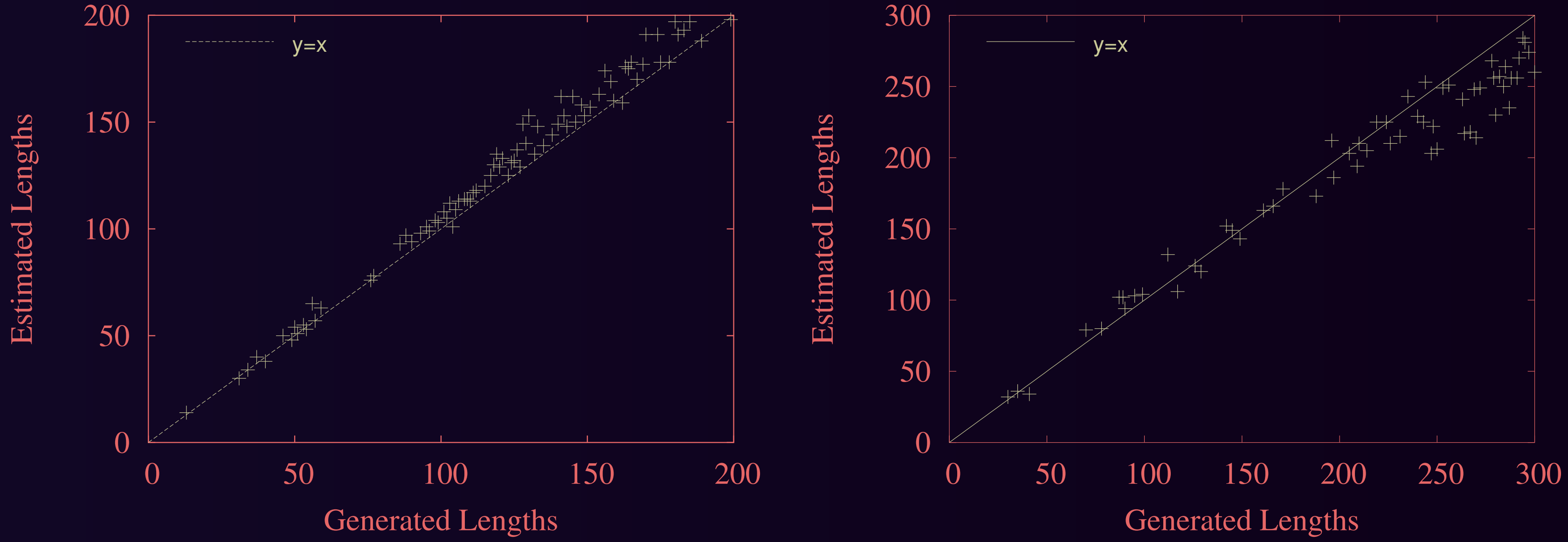
Direct Estimate of Distance

- Highly accurate even on large genomes with large distances.
- Accounts for insertions, duplications, deletions, and inversions.
- Key lemma: *there always exists a shortest sequence that first does all insertions, then all inversions, and finally all deletions*.
- Matches elements of gene families using optimal covering; treats unmatched elements as insertions/deletions.
- Tracks sequence of deletions and inversions backward to figure out how to parcel out insertions.

Direct Distance Estimate: Example

Simulated 800-gene genomes, 70% inversions (mean of 20, located uniformly), 16% deletions, 7% insertions, and 7% duplications (all mean 10).

left: expected pairwise distances from 40 to 160 events
right: expected pairwise distances from 80 to 320 events

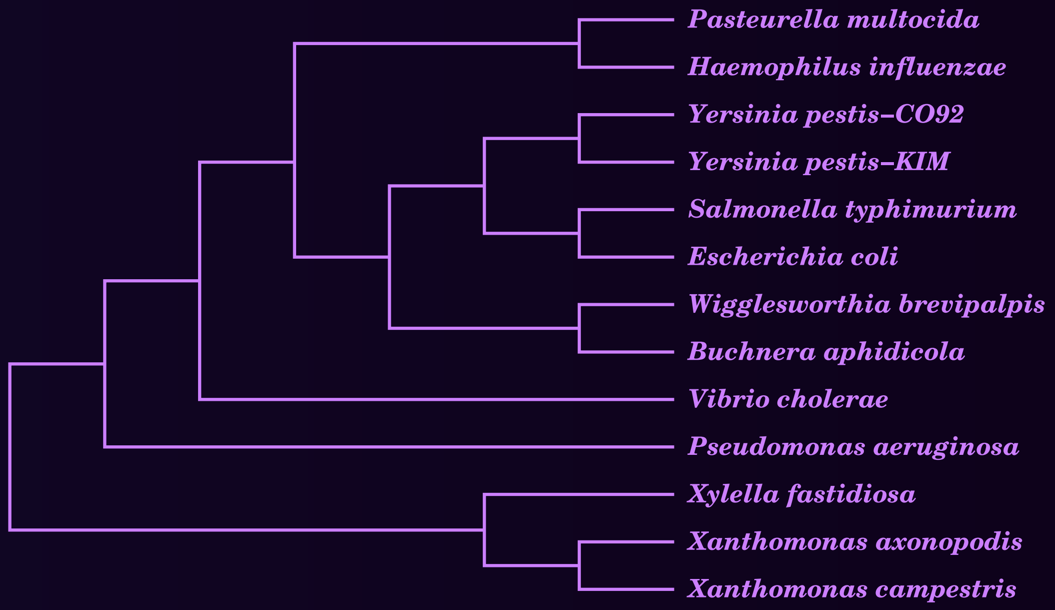


Using the Swenson et al. Estimate

(unpublished)

13 gamma proteobacteria (Lerat/Daubin/Moran 2003)

Over 3,400 genes, with 540–3,000 genes and 3%–30% duplications per genome; pairwise distances from 170 to 1700 events.



Reference phylogeny: 2 years of work, over 60 gene sequences.

*Using our distance estimates and naïve NJ:
1 hour to compute distances, 1 second to construct tree,
and* only one error *(long branch attraction, trivially fixed).*

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Reconstructing Ancestral Genomes

Goal: Reconstruct a signed gene order at each internal node in the tree to minimize sum of edge distances.

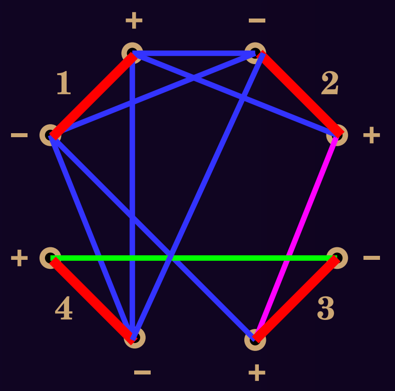
Problem is NP-hard even for just three leaves, no duplications, and simplest of distances (breakpoint, plain inversion)!

This is the median problem for signed genomes: given three genomes, produce a new genome that will minimize the sum of the distances from it to the other three.

Median Problem for Breakpoints

Sankoff showed to to convert MPB for identical gene content to the Travelling Salesperson Problem

+1 −2 +4 +3
+1 +2 −3 −4
+2 −3 −4 −1



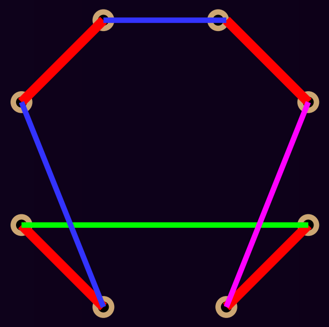
🟥 cost = − max

🟩 cost = 0

🟪 cost = 1

🟦 cost = 2

edges not shown have cost = 3



an optimal solution corresponding to genome +1 +2 −3 −4

Adjacency A B becomes an edge from A to −B

The cost of an edge A −B is the number of genomes that do NOT have the adjacency A B

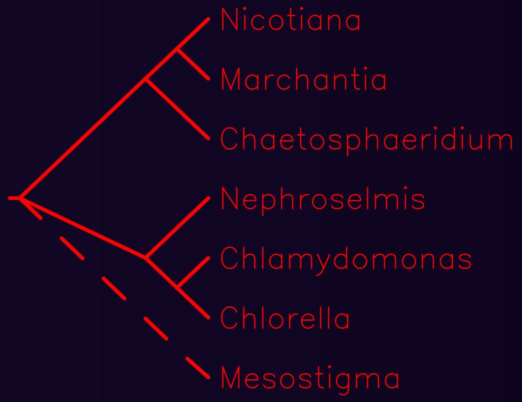
Median Problem for Inversions

No simple formulation in terms of a standard optimization problem.

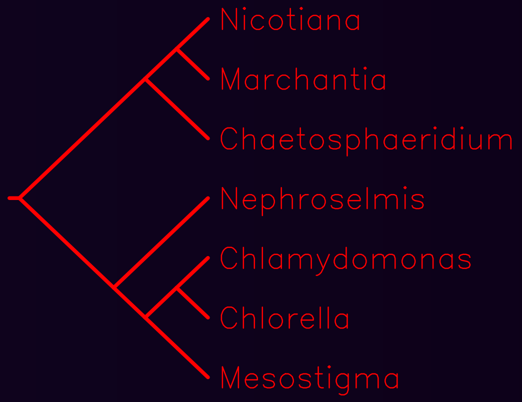
- Exact solutions given by Siepel/Moret and by Caprara for identical gene content; work well for distances to median of 0–15 inversions.
- Various heuristics proposed by Bourque and Pevzner and others.
- Extensions by Tang/Moret to handle distances up to 50-100 events.
- Inversion median shown preferable to breakpoint median (Siepel/Moret, Tang/Moret).

Medians with Unequal Gene Content

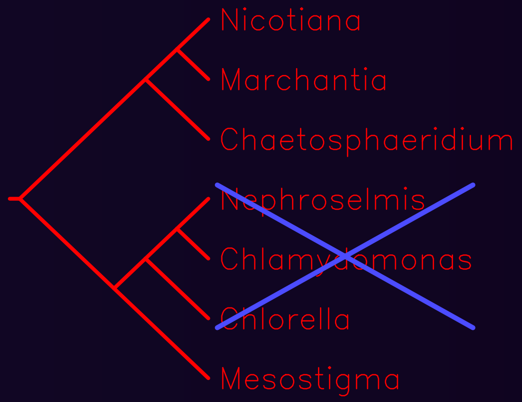
Tang/Moret/Cui/DePamphilis (2004): chloroplast data



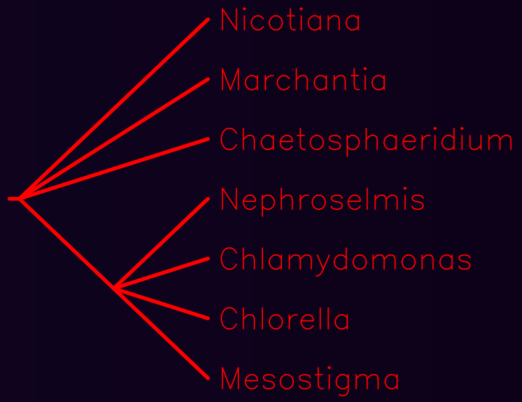
organismal



Tang/Moret GRAPPA



NJ (inv.)



breakpoint GRAPPA

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Reconstruction from Gene-Order Data

- **Distance methods**
  - NJ and Weighbor with corrected distances, with or without DCM
- **Parsimony-based methods**
  - Encoding approaches: MPBE, MPME
  - Direct approaches: BPAnalysis, GRAPPA, MGR, DCM-GRAPPA
- **Likelihood-based methods**

Direct Approaches: BPAnalysis

(Sankoff and Blanchette)

```
Initially label all internal nodes with gene orders
Repeat
    For each internal node $v$, with neighbors $A$, $B$, and $C$, do
        Solve the $MPB$ on $A, B, C$ to yield label $m$
        If relabelling $v$ with $m$ improves the tree score, then do it
until no internal node can be relabelled
```

GRAPPA

Genome **R**earrangements **A**nalysis under **P**arsimony & other **P**hylogenetic **A**lgorithms

GRAPPA

Genome Rearrangements Analysis under Parsimony & other Phylogenetic Algorithms

- Began as a reimplementation of BPAnalysis.
- Current version runs up to one billion times faster than BPAnalysis, thanks to *algorithmic engineering*. (Fast code, better bounding, caching results, ordering computations, etc.)
- Limit: every added taxon multiplies the running time by twice the number of taxa.
  So 13 taxa take 20 mins, 15 taxa two weeks, 16 taxa a year, 20 taxa over 2 million years, and . . .

GRAPPA: Speed-Ups

In order of increasing benefits:

- very fast generation of candidate trees
- hand-tuned code
  (unrolling loops, maintaining values in registers)
- parallelization
- minimizing memory usage and maximizing cache hits
- fast specialized TSP solver for breakpoint medians
- bounding trees to avoid scoring them
  (using a tour of the leaves)
- examining trees in increasing order by bound values

DCM-GRAPPA

Extension to GRAPPA to scale it
to large datasets (Tang and Moret 2003).

- *Scales gracefully to over 1,000 genomes (less than 2 days of computation).*
- *Retains accuracy of GRAPPA: error rates on 1,000-genome datasets are consistently below 3%.*
- *Uses DCM1 (early version), so can surely be improved.*

Very Tight Bounds from LP

(Tang/Moret CPM'05)

Use selected triangle inequalities on a tree as Linear Programming constraints to compute a lower bound.

- *With good selection, bound is very tight ($\geq$ 99%).*
- *Avoids scoring trees with `GRAPPA`: no median computation, so very fast.*
- *Allows `GRAPPA` to handle much larger genomes.*

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene-Order Data: What and Why?
- Computing with Gene-Order Data
- Ancestral Gene Orders
- Reconstruction from Gene-Order Data
- Some Open Problems

Some Open Problems

- Tree models
- Evolutionary models
- Extensions of Hannenhalli-Pevzner theory to handle
  - *transpositions and inversions*
  - *length-dependent rearrangements*
  - *position-dependent rearrangements*
  - *duplications*
- Good combinatorial formulation of the median problem for inversions and for more general cases.
- Tighter bounds on tree scores (our linear programming approach may be solving that).
- Extensions to phylogenetic networks.

Conclusions

- Disk-covering methods can extend the range of existing methods by several orders of magnitude—and we have just begun.

Conclusions

- Disk-covering methods can extend the range of existing methods by several orders of magnitude—and we have just begun.
- Gene-order data carry a strong phylogenetic signal and current algorithmic approaches scale to significant sizes.

Conclusions

- Disk-covering methods can extend the range of existing methods by several orders of magnitude—and we have just begun.
- Gene-order data carry a strong phylogenetic signal and current algorithmic approaches scale to significant sizes.
- Strong algorithmic design, good algorithm engineering, and high-performance computing are all crucial components of successful computational biology research.

Thank You!

Laboratory for
High-Performance Algorithm Engineering
and Computational Molecular Biology

compbio.unm.edu

CIPRES
Cyber Infrastructure
for Phylogenetic Research

www.phylo.org