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:

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

Overview

- Phylogenies: What and Why?
- The Tree of Life and CIPRES
- Phylogenetic Reconstruction
- Scaling Up
- Gene Content and Order Data
  - Gene-Order Data: What and Why?
  - Computing with Gene-Order Data
  - Ancestral Gene Orders
- Summary

Phylogenies

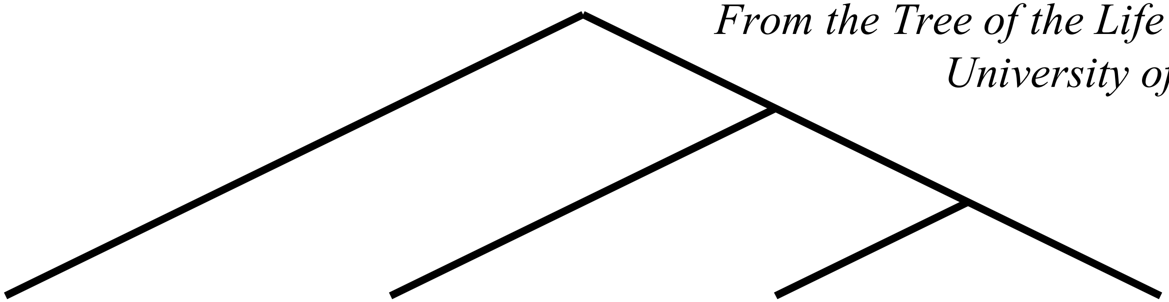
A phylogeny is a reconstruction of the evolutionary history of a collection of organisms.

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

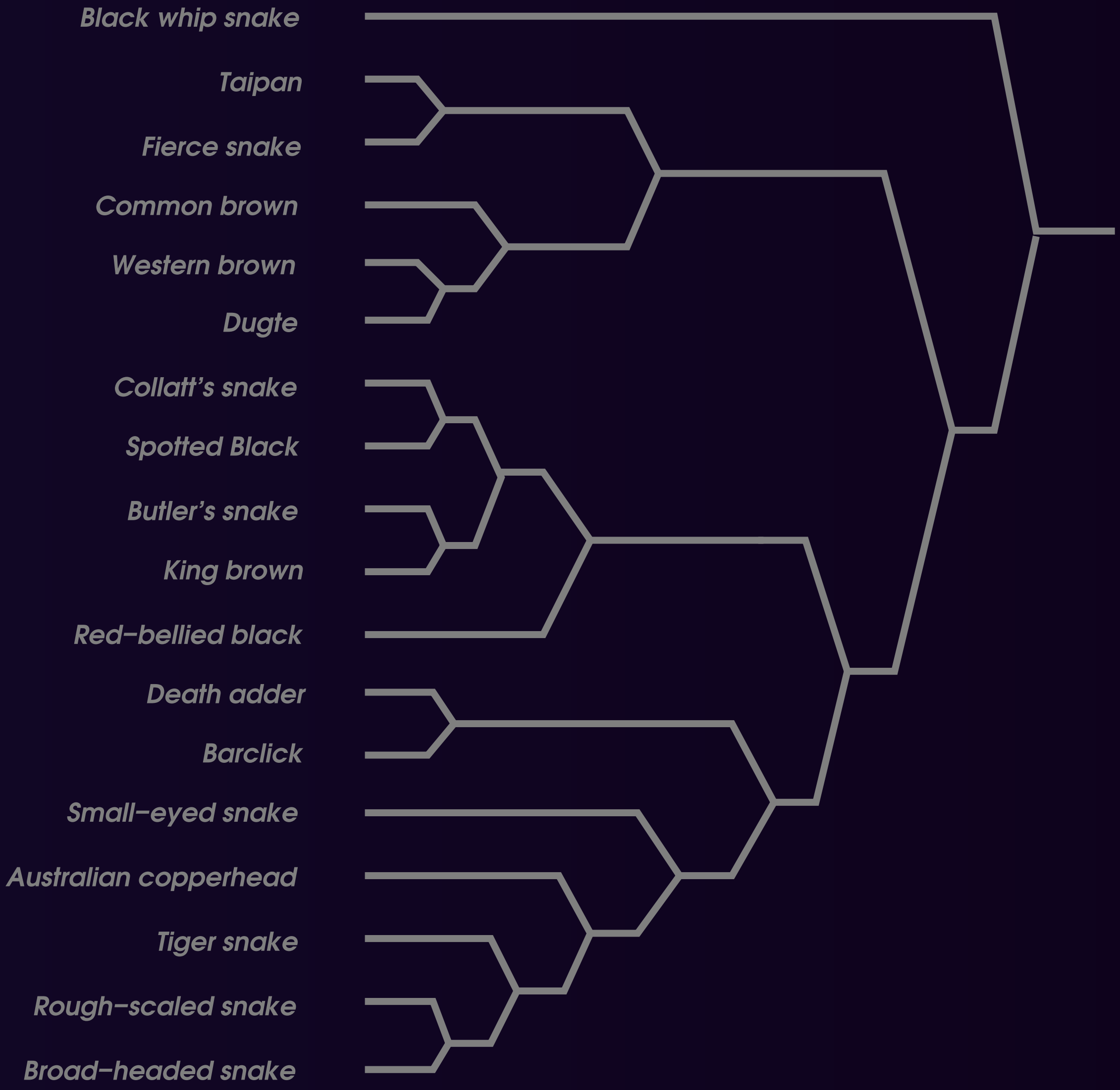
Phylogenies: Why?

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 targets
- origins and spread of disease
- origins and migrations of humans

Example: Antivenins



Black whip snake

Taipan

Fierce snake

Common brown

Western brown

Dugte

Collatt's snake

Spotted Black

Butler's snake

King brown

Red–bellied black

Death adder

Barclick

Small–eyed snake

Australian copperhead

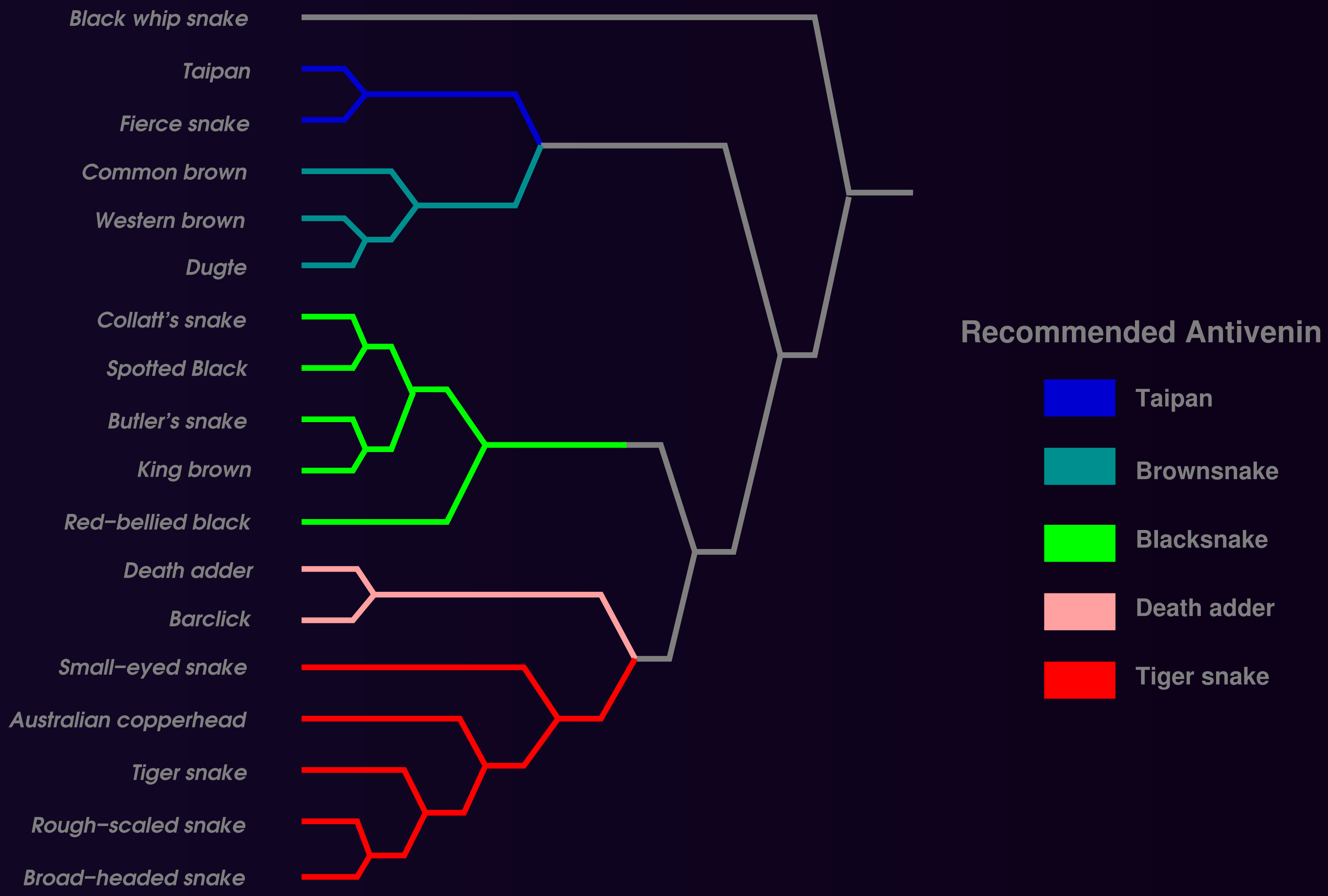
Tiger snake

Rough–scaled snake

Broad–headed snake

Venomous
Australian
Snakes

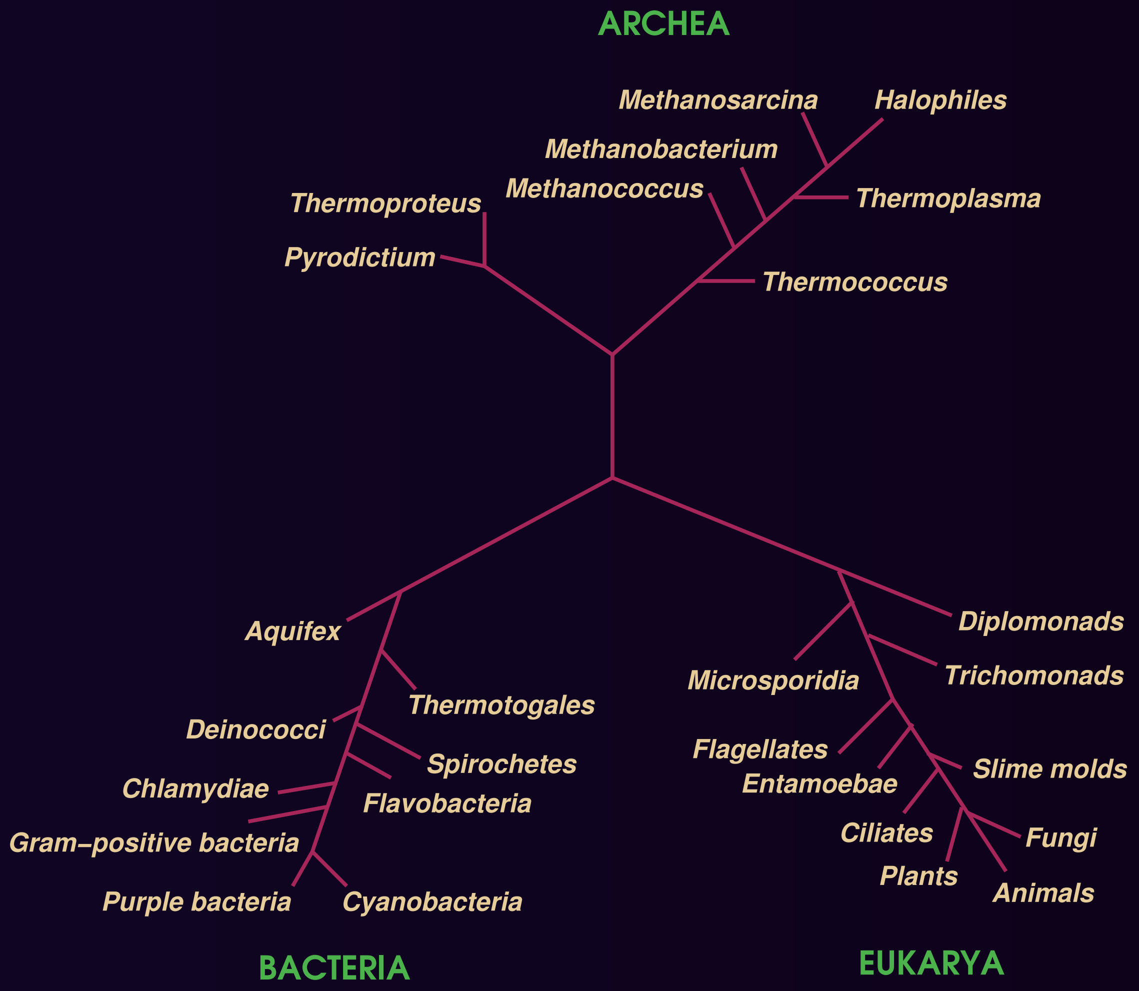
Example: Antivenins



- Phylogenies: What and Why?
- The Tree of Life and CIPRES
- Phylogenetic Reconstruction
- Scaling Up
- Gene Content and Order Data
    - Gene-Order Data: What and Why?
    - Computing with Gene-Order Data
    - Ancestral Gene Orders
- Summary

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.

The CIPRES Project

CIPRES

Cyber Infrastructure for Phylogenetic Research

www.phylo.org

A community project funded for 5 years by the NSF for $12M under the ITR program, with the aim to develop the infrastructure (hardware, software, and databases) necessary to support the reconstruction of the Tree of Life.

- *over 15 institutions, including three museums*
- *over 50 researchers, evenly split between CS and Biology*
- *research in algorithms, simulation and modelling, databases, software architecture, and high-performance computing*

CIPRES: Participants

U. New Mexico

UT Austin

Florida State U.

U. Connecticut

AMNH

UC Berkeley

Texas A&M

U. Arizona

Rice U.

Yale U.

UC San Diego

U. Pennsylvania

U. British Columbia

U. South Carolina

North Carolina State U.

- Phylogenies: What and Why?
- Phylogenetic Reconstruction:
  a fast review from a CS standpoint
- Scaling Up
- Gene Content and Order Data
  - Gene-Order Data: What and Why?
  - Computing with Gene-Order Data
  - Ancestral Gene Orders
- Summary

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)

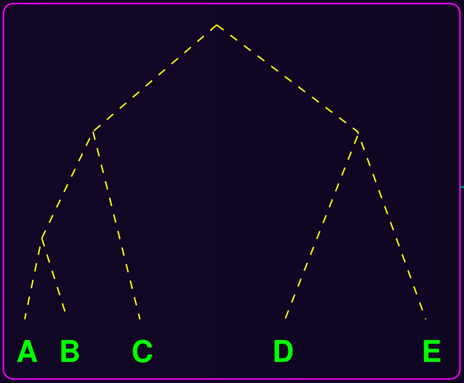
In addition:

- 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

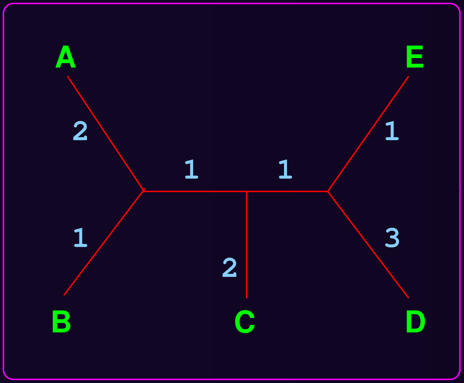
Molecular Data

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*.

Finding a most parsimonious tree is NP-hard.

Optimal solutions are limited to sizes around 30.

- Heuristic solutions appear good to about 500 taxa (e.g., TNT).

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.

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 15 taxa, but appear good to about 100 taxa (e.g., MrBayes, PhyML, RAxML).

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up
- Gene Content and Order Data
  - Gene-Order Data: What and Why?
  - Computing with Gene-Order Data
  - Ancestral Gene Orders
- Summary

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: Approaches

- *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*.

Scaling Up: A Solution

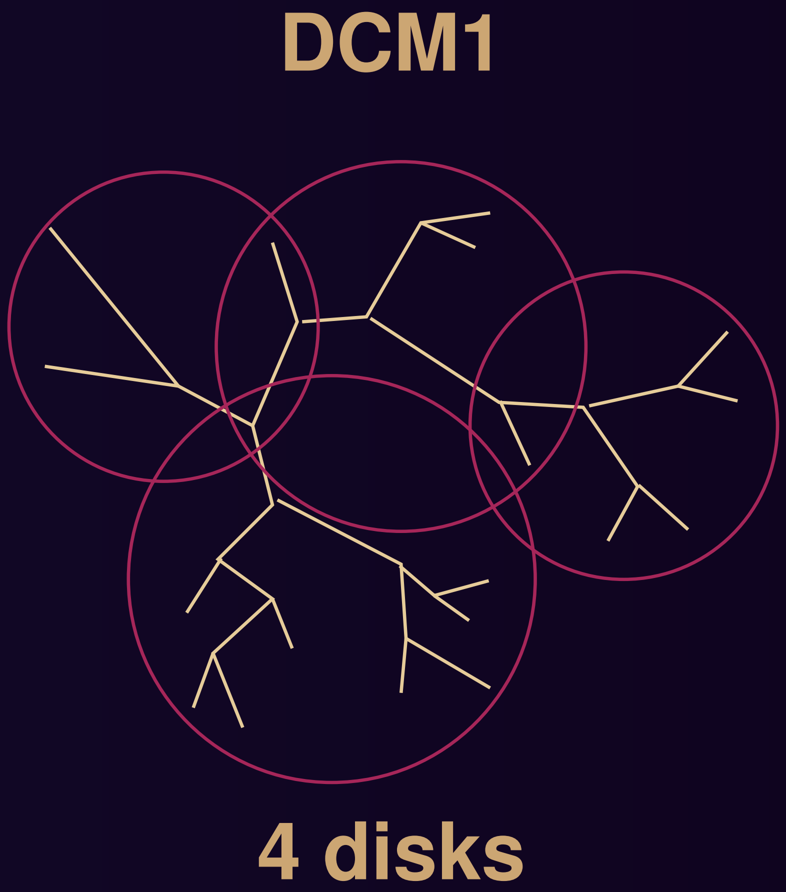
Disk-Covering Methods

Basic idea:

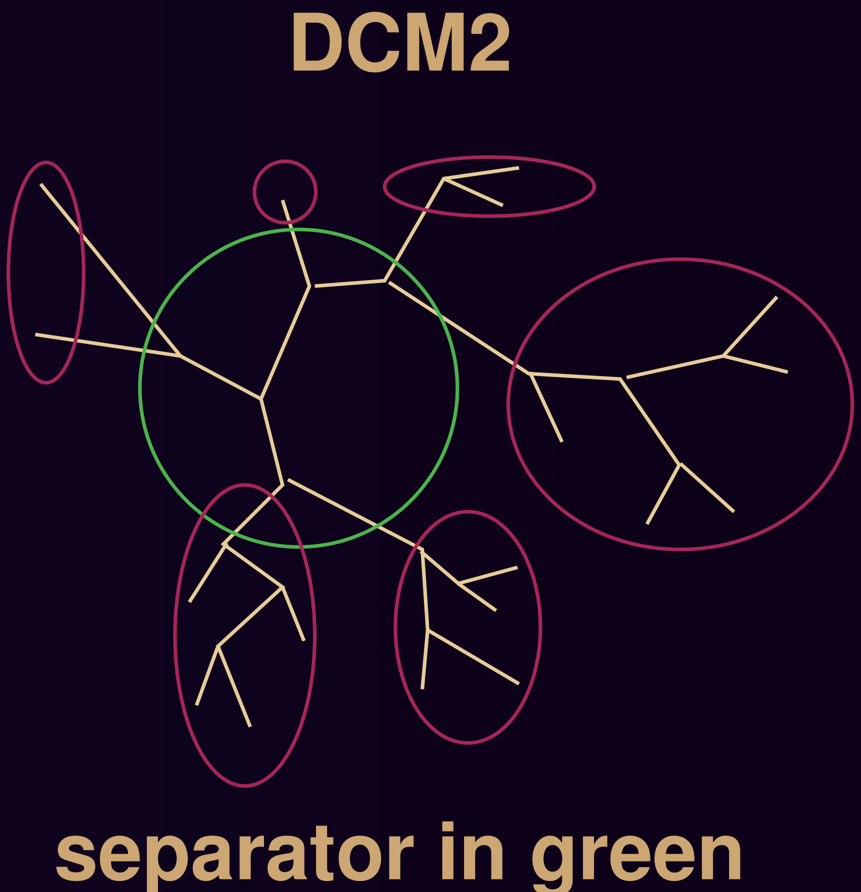
- decompose dataset into *overlapping* compact subsets—the *disks*
- reconstruct a tree for each subset
- assemble these trees into a single tree

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

DCM1 and DCM2


DCM1

4 disks


DCM2

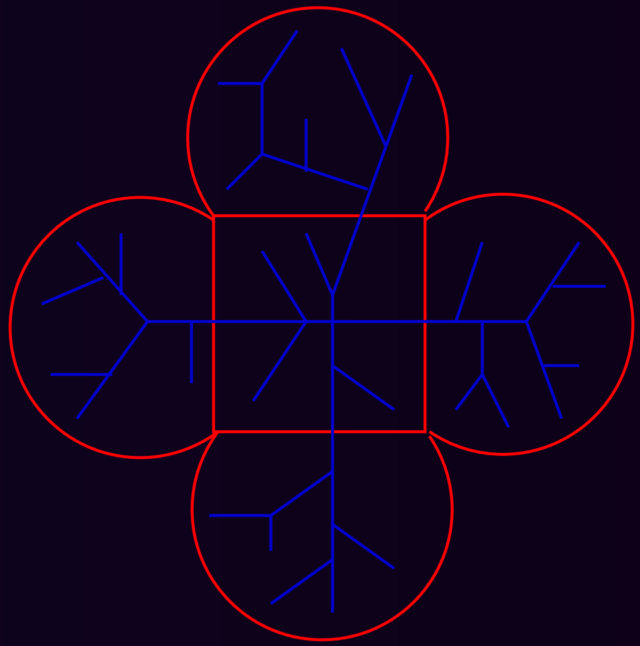
separator in green

Improvement: DCM3

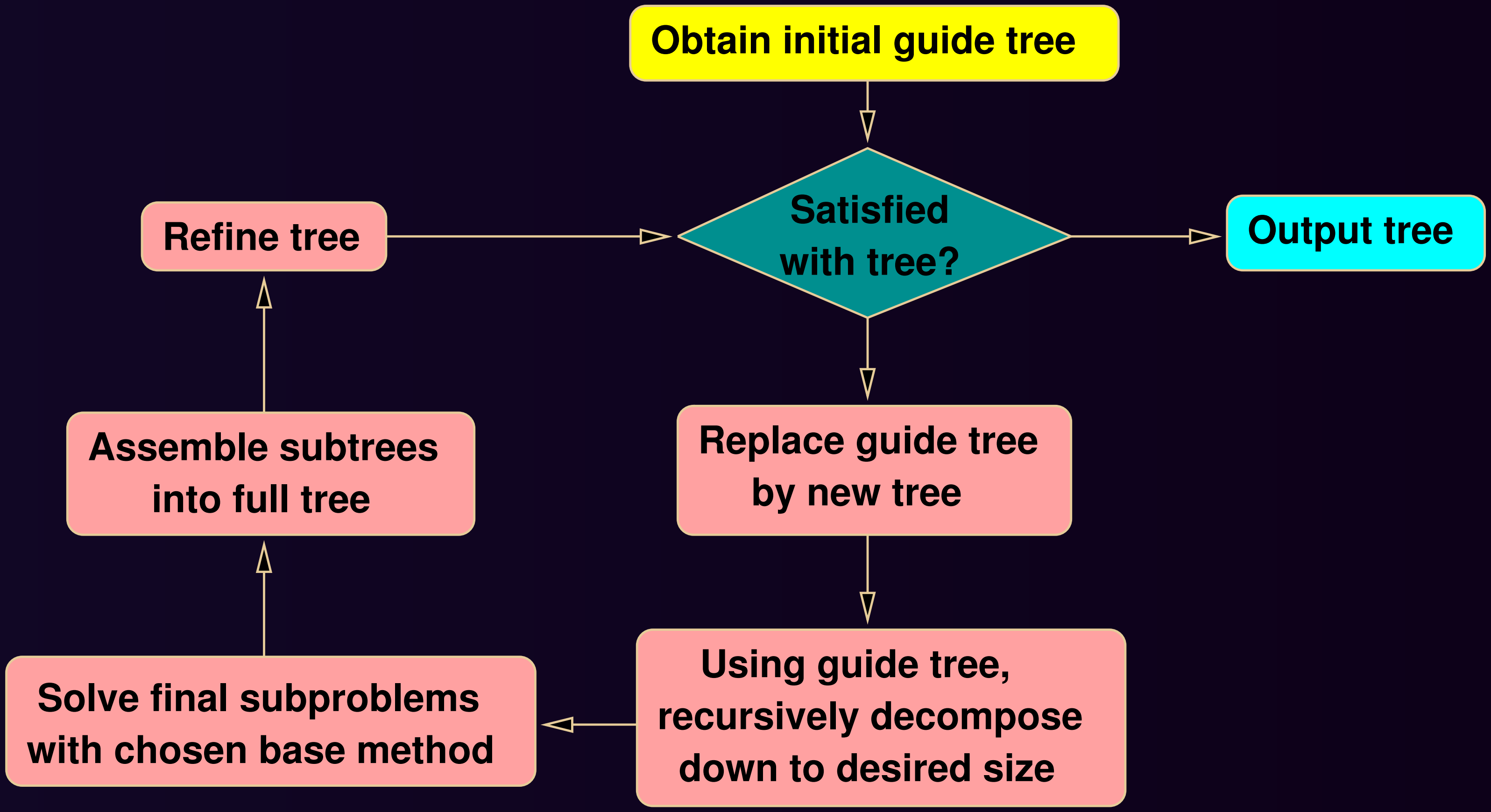
DCM1 and DCM2: decomposition based on distance matrix only

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

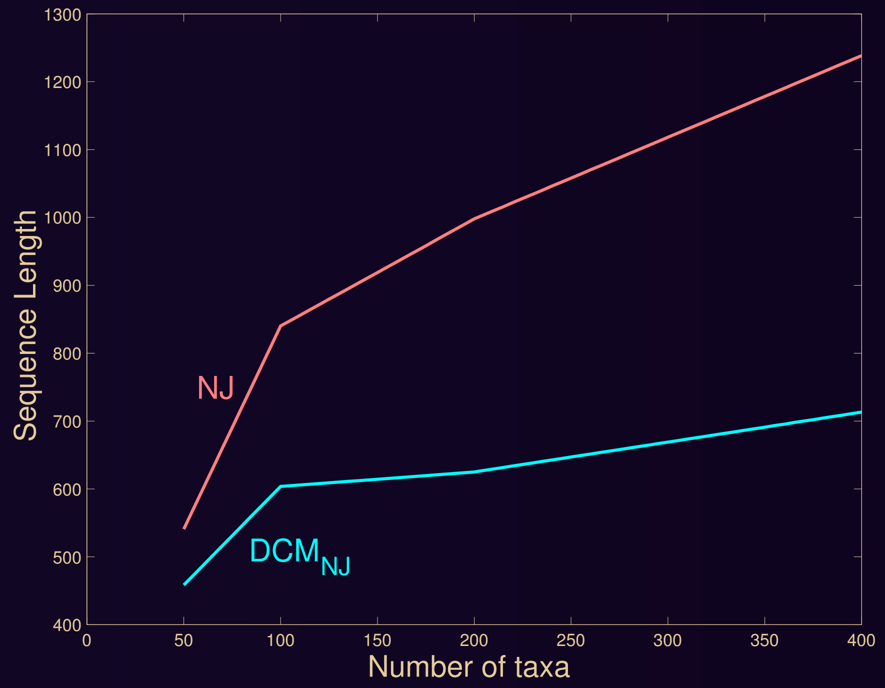
Assemble subtrees into full tree

Refine tree

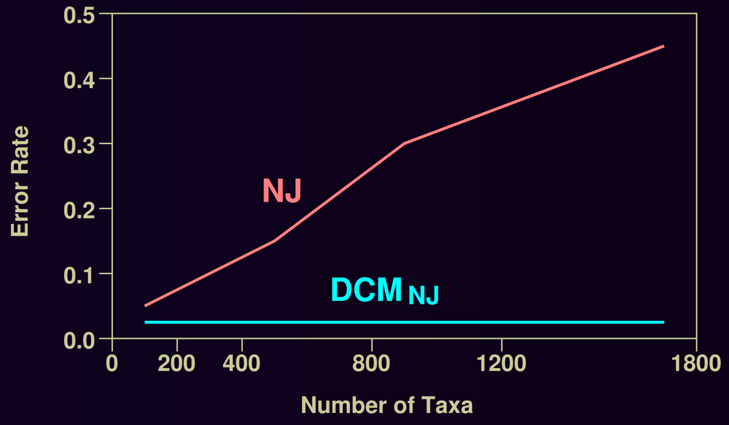
Results with DCM1 and NJ

using Kimura 2-parameter plus Γ model

reduced sequence length
(0.15 error rate)



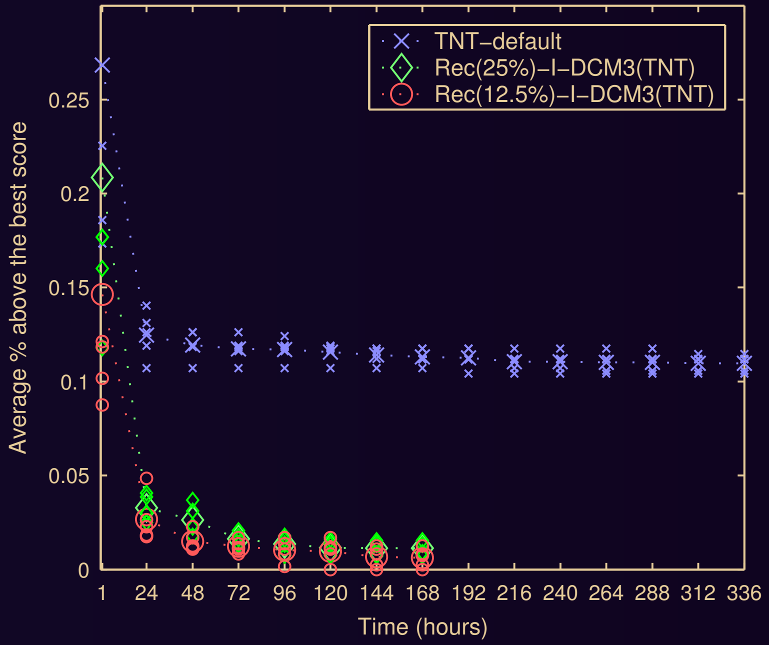
reduced error rate
(1,000 sequence length)



Results with Rec-I-DCM3 and MP

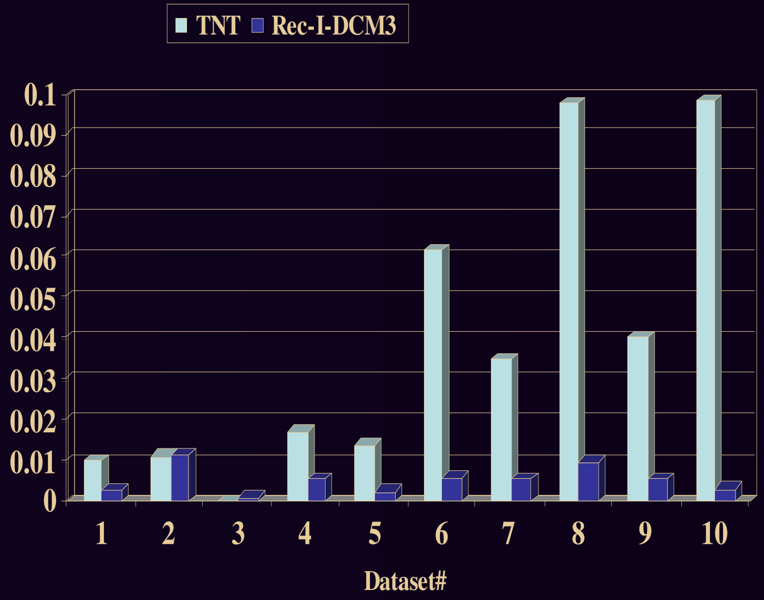
Rec-I-DCM3(TNT) vs. TNT

10,000 RNA sequences



10 datasets

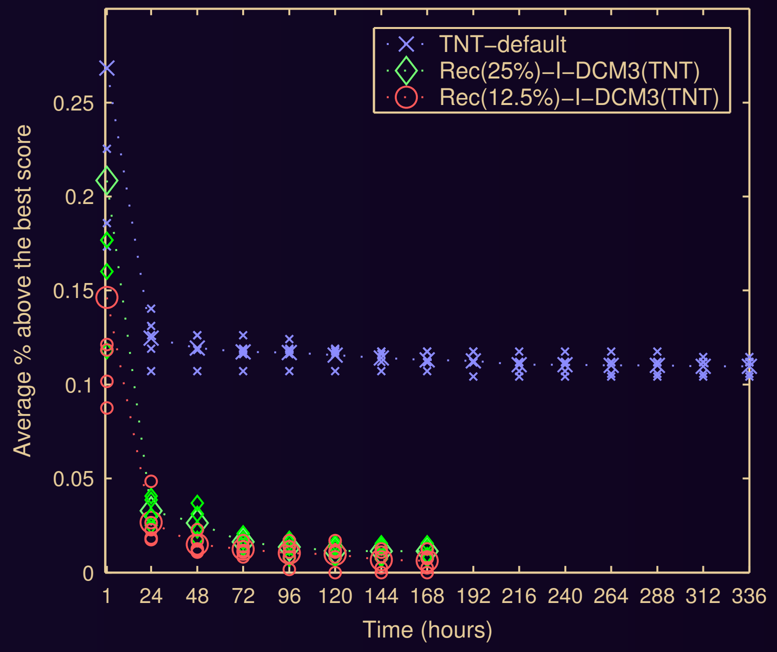
(from 4,000 to 15,000)



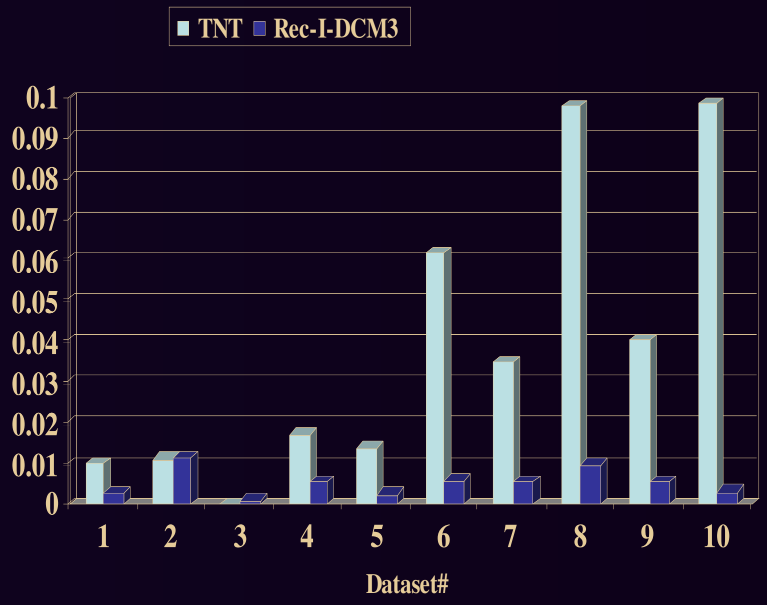
Results with Rec-I-DCM3 and MP

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!!

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up
- Gene Content and Order Data
    - Gene-Order Data: What and Why?
    - Computing with Gene-Order Data
    - Ancestral Gene Orders
- Summary

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

  (nucleotide/codon sequences from genes)

  gene content and order

  (gene ordering on chromosomes)

Gene-Order Data

The ordered sequence of genes on one or more chromosomes.

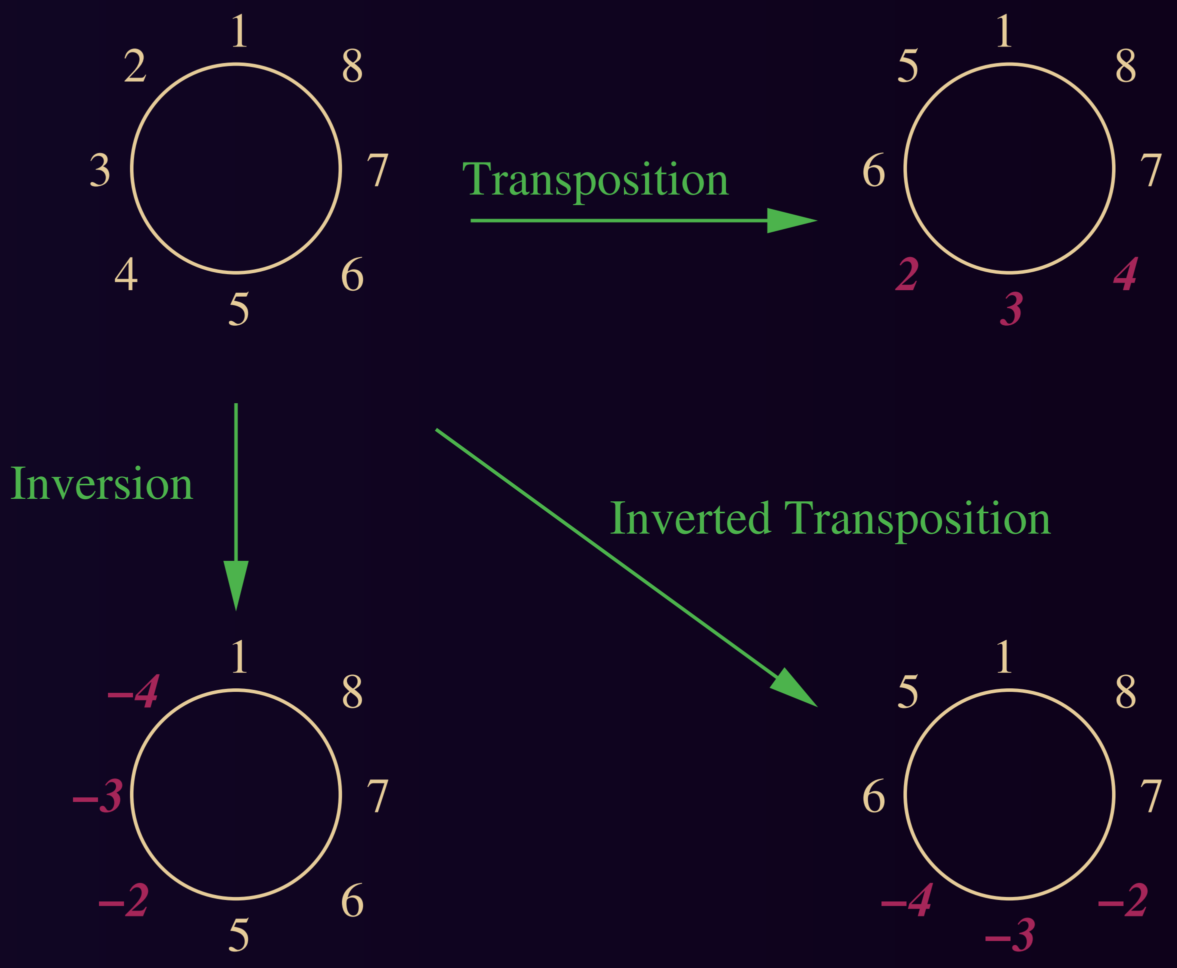
The entire gene order is a *single character*, which can assume a huge number of states.

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.

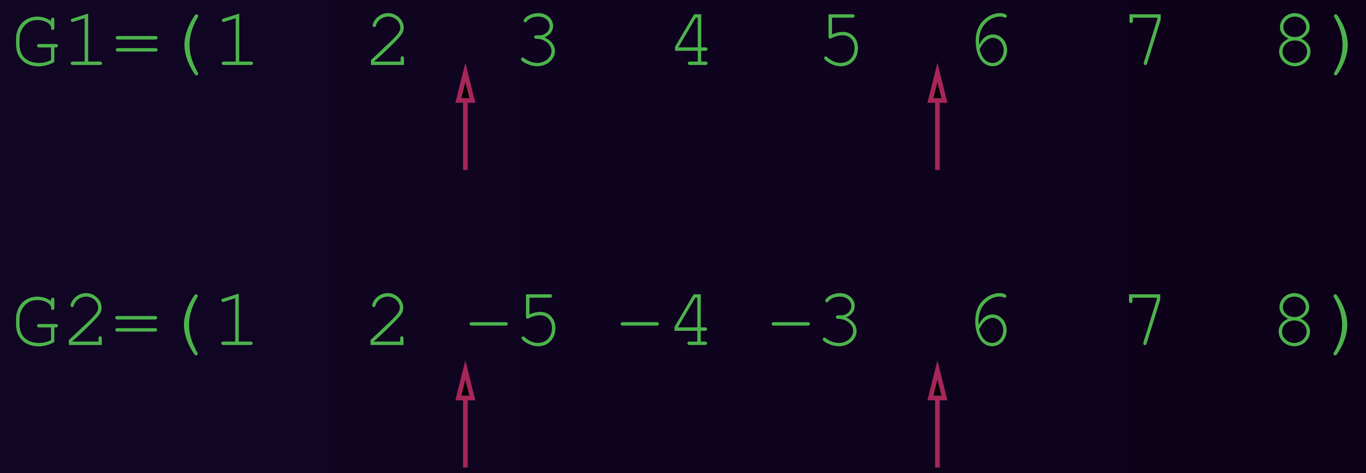
Problems:

- Mathematics *much more complex* than for sequence data.
- Models of evolution not well characterized.
- Very limited data (mostly organelles and bacteria).

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up: The Issues
- Scaling Up: A Solution
- Gene Content and Order Data
  - Gene-Order Data: What and Why?
  - Computing with Gene-Order Data
  - Ancestral Gene Orders
- Summary

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)
```

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.

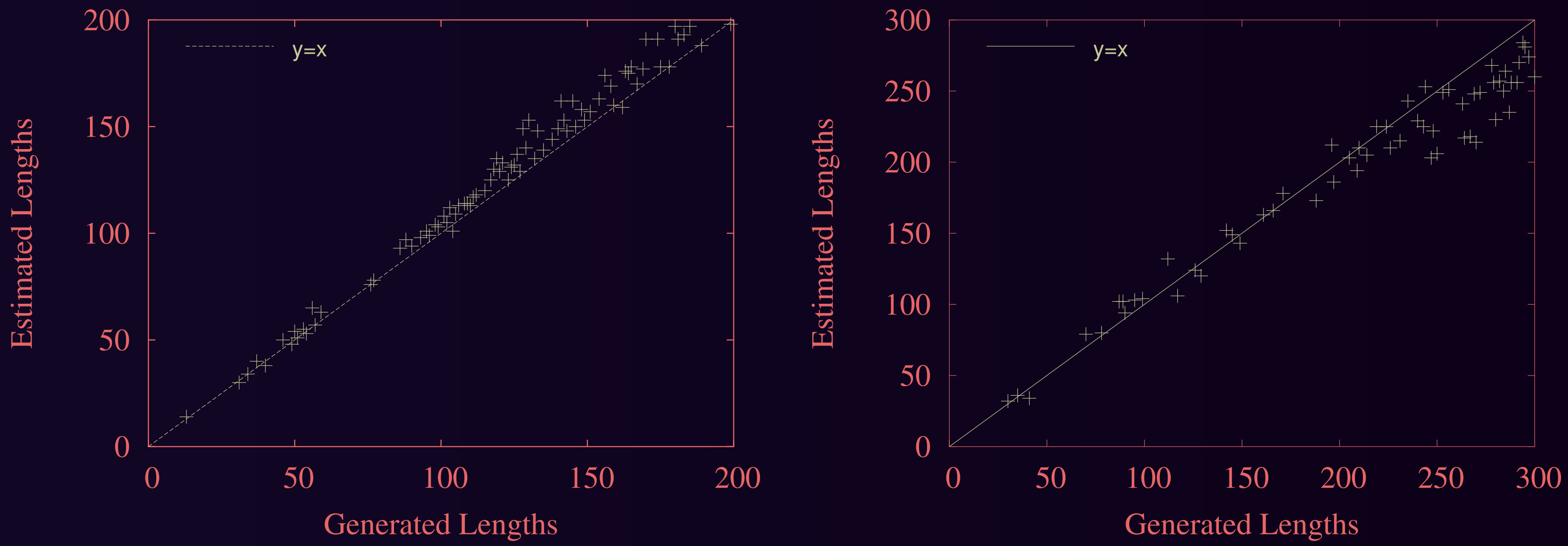
Our work:

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

Direct 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

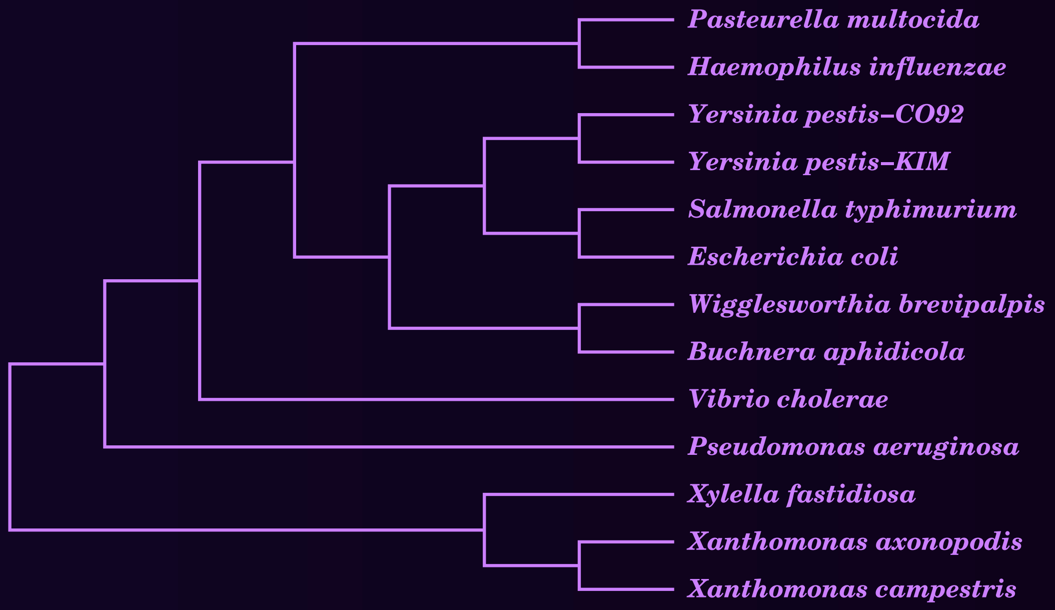


Using Our Distance Estimate

(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 Content and Order Data
    - Gene-Order Data: What and Why?
    - Computing with Gene-Order Data
    - Ancestral Gene Orders
- Summary

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)!

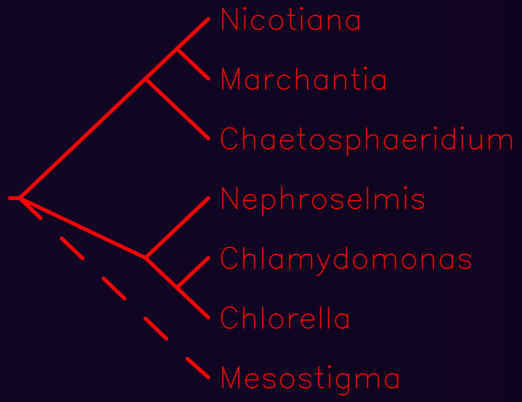
This is the median problem for signed genomes: given three genomes, produce a new genome that will minimize the sum of the distances from it to the other three.

Observations on Ancestral Genomes

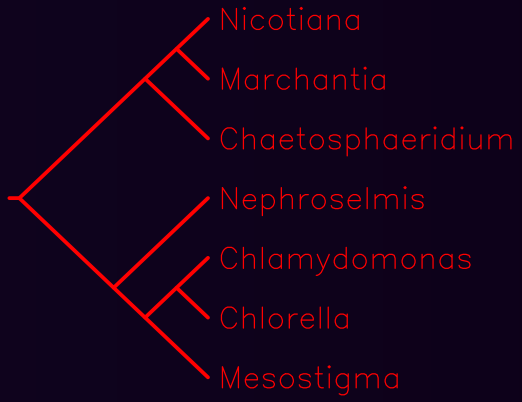
- *Successful for small, streamlined genomes (organelles).*
- *Must take into account gene content:*
  common simplification of reducing to just shared genes loses too much information.
- *Not doable for bacterial nuclear genomes without additional biological constraints: too many solutions.*
- *Possible additional constraints:*
  hot spots, lengths of affected segments, protected segments (centromere, origin of replication, etc.), nucleotide data around each gene, etc.

Medians with Unequal Gene Content

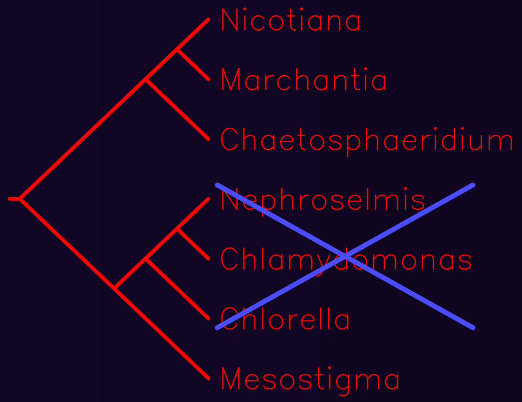
Tang/Moret/Cui/DePamphilis (2004): chloroplast data



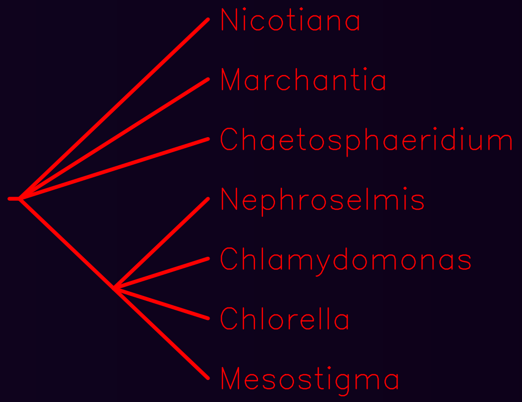
organismal



Tang/Moret GRAPPA



NJ (inv.)



breakpoint GRAPPA

- Phylogenies: What and Why?
- Phylogenetic Reconstruction
- Scaling Up
- Scaling Up
- Gene Content and Order Data
    - Gene-Order Data: What and Why?
    - Computing with Gene-Order Data
    - Ancestral Gene Orders
- Summary

Summary

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

Summary

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

Summary

- Scaling up: Disk-covering methods can extend the range of existing methods by several orders of magnitude—and we have just begun.
- Complex Data: Gene-order data carry a strong phylogenetic signal and current algorithmic approaches scale to significant sizes.
- Approach: Strong algorithmic design, constant iteration on evolutionary models, extensive testing on simulated data and biological data. Stimulate CS research (even if highly abstract) and biological 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