Acknowledgments

- **Constant Collaborators:**
  - Tandy Warnow (UT Austin)
  - David Bader (UNM)

- **Support:**
  - US National Science Foundation
  - US National Institutes of Health
  - Alfred P. Sloan Foundation
  - IBM Corporation
Overview

- Phylogenies: What and Why?
Overview

- Phylogenies: What and Why?
- Phylogenetic Reconstruction: How?
Overview

- Phylogenies: What and Why?
- Phylogenetic Reconstruction: How?
- Limitations and Challenges
Overview

- Phylogenies: What and Why?
- Phylogenetic Reconstruction: How?
- Limitations and Challenges
- The CIPRES Project
Overview

- Phylogenies: What and Why?
- Phylogenetic Reconstruction: How?
- Limitations and Challenges
- The CIPRES Project
- Research in my Lab
Overview

- Phylogenies: What and Why?
- Phylogenetic Reconstruction: How?
- Limitations and Challenges
- The CIPRES Project
- Research in my Lab
- Summary and Conclusions
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
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
- origins and spread of disease
- drug and vaccine development
- origins and migrations of humans
Herpes Viruses that Affect Humans
Epidemiology of West Nile Virus

- Romania 1996
- Israel 1952
- South Africa
- Egypt 1951
- Senegal 1979
- Italy 1998
- Romania 1996
- Kenya 1998
- New York 1999
- Israel 1998
- Central African Republic 1967
- Ivory Coast 1981
- Kunjin 1966–1991
- India 1955–1980
Drug Design: 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
Grand Challenge: The Tree of Life

BACTERIA
- Cyanobacteria
- Flavobacteria
- Spirochetes
- Deinococci
- Chlamydiae
- Gram-positive bacteria
- Purple bacteria

ARCHEA
- Thermococcus
- Thermoplasma
- Methanobacterium
- Methanosarcina
- Halophiles

EUKARYA
- Diplomonads
- Trichomonads
- Microsporidia
- Ciliates
- Entamoebae
- Plants
- Fungi
- Animals
- Slime molds
Scale of The Tree of Life

- 20 fully sequenced eukaryotic (plants, animals, protists) genomes
- 600 fully sequenced bacterial genomes
- Several sequenced genes for perhaps 50,000 species
- 1,5 million described species
- Estimates for existing species vary from 10 million to 200 million.
- Genome-based tools can handle 20–50 organisms.
- Gene-based tools can handle 200–500 organisms.
- Both sets of tools scale exponentially with the amount of data.
Phylogenetic Reconstruction

- **Data:**
  behavioral, morphological, metabolic, molecular, etc.
  Main data today are DNA sequence data.

- **Models:**
  models of speciation, of population evolution, of molecular character evolution, etc.

- **Algorithms:**
  clustering, optimization, estimation of distributions, and heuristics.
Molecular Data

Typically the DNA sequence of a few genes. Characters are individual positions in the string and can assume 4 states (nucleotides) or 20 states (codons). Evolve through point mutations, insertions (incl. duplications), and deletions.
Typically the DNA sequence of a few genes.
Characters are individual positions in the string and can assume 4 states (nucleotides) or 20 states (codons).
Evolve through point mutations, insertions (incl. duplications), and deletions.

- Find homologous genes across all organisms.
- Align gene sequences for the entire set (to identify gaps—insertions and deletions—and point mutations).
- Decide whether to use a single gene for each analysis or to combine the data.
- Lengths limited by size of genes (typically several hundred base pairs)
Sequence Data: Illustration

AGGCAT TAGCCCA TAGACTT AGCGCTTAGCACAATGAACTT

AAGACTT

AAGGCCT

AGGGCAT

AGGCAT

TAGCCCA

TAGACTT

TGAACCTT

AGCACAA

AGCGCCTT
Sequence Data: Attributes

- **Advantages:**
  - Large amounts of data available.
  - Accepted models of sequence evolution.
  - Models and objective functions provide a reasonable computational framework.

- **Problems:**
  - Fast evolution restricts use to a few million years.
  - Gene evolution need not be identical to organism evolution.
  - Multiple alignments are not well solved.
Gene-Order Data

The ordered sequence of genes on one or more chromosomes.
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 (in mitochondria) and translocations (between chromosomes).
Gene-Order Data

The ordered sequence of genes on one or more chromosomes.
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 (in mitochondria) and translocations (between chromosomes).

- Identify homologous genes, including duplications.
- Refine rearrangement model for given collection of organisms (e.g., handle bacterial operons or eukaryotic exons explicitly).
Gene-Order Data: Rearrangements

Inversion → Transposition

Inverted Transposition
Gene-Order Data: Attributes

- **Advantages:**
  - No need for multiple alignments.
  - No gene tree/species tree problem.
  - Rare evolutionary events and unlikely to cause “silent” changes—so can go back hundreds of millions years.

- **Problems:**
  - Mathematics *much more complex* than for sequence data.
  - Models of evolution not well characterized.
  - Very limited data (mostly organelles).
Other Data

- protein folds
  remarkably conserved, but give rise to very complex models

- metabolic pathways
  highly specific, but insufficient for large datasets

- morphological characters
  not as clearly inherited and inherently fuzzy

- etc.!
Good models emerge from collaborations among biologists, mathematicians, and computer scientists; they are:

- **biologically plausible**: they produce credible data and possess explanatory power.

- **mathematically sound**: it is possible to prove desirable properties (convergence, consistency, etc.).

- **computationally tractable**: producing data is easy and reversing the model is possible.
Speciation Models

Usually based on a *birth-death* process: in any time interval, there are given probabilities for extinction or speciation; also known as the *coalescent* or *Yule-Harding* model.

But need more data and refinements:

- *inheritance of tendency to speciate*
- *punctuated equilibrium*
- *connection to population genetics*
Molecular Evolution Models

From large amounts of data, models build transition matrices \((4 \times 4\) for nucleotides, \(20 \times 20\) for aminoacids).

- Widely used to estimate evolutionary rates and well supported by data.
- Still assume independence among sites (e.g., each nucleotide or codon evolves independently of the others).
- Remain unconnected to speciation model.
Two main categories of methods:

- **Distance-based methods** (UPGMA, neighbor-joining) work from a matrix of pairwise distances.
- **Criterion-based methods** (Minimum Evolution, Maximum Parsimony, and Maximum Likelihood) rely on an underlying model and attempt to infer or reconstruct additional data.

In addition:

- **Meta-methods** (quartet-based methods, disk-covering method) decompose the data into smaller subsets, construct trees on those subsets, and use the resulting trees to build a tree for the entire dataset.
Two main categories of methods:

- **Distance-based** methods (UPGMA, neighbor-joining) work from a matrix of pairwise distances.
- **Criterion-based** methods (Minimum Evolution, Maximum Parsimony, and Maximum Likelihood) rely on an underlying model and attempt to infer or reconstruct additional data.
Algorithms

Two main categories of methods:

- **Distance-based methods** (UPGMA, neighbor-joining) work from a matrix of pairwise distances.
- **Criterion-based methods** (Minimum Evolution, Maximum Parsimony, and Maximum Likelihood) rely on an underlying model and attempt to infer or reconstruct additional data.

In addition:

- **Meta-methods** (quartet-based methods, disk-covering method) decompose the data into smaller subsets, construct trees on those subsets, and use the resulting trees to build a tree for the entire dataset.
Evolutionary Distances

- **True evolutionary distance**: the actual number of permitted 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.

- **Expected true evolutionary distance**: obtained from the edit distance by correcting for the known (model or experiments) statistical relationship between true and edit distances.
Distance-Based Methods

- Use edit or expected true evolutionary distances.
- Usually run in *low polynomial time*.
- Reconstruct *only topologies*: no ancestral data.
- Prototype is Neighbor-Joining (NJ).
- NJ is optimal on *additive* distances (where the distance along a path in the true tree equals the pairwise distance in the matrix).
- NJ is *statistically consistent* (produces the true tree with probability 1 as the sequence length goes to infinity).
Parsimony-Based Methods

- Aim to minimize (weighted) total number of character changes.
- Assume that characters are independent.
- Reconstruct ancestral data.
- Known not to be statistically consistent with sequence data.
- Finding most parsimonious tree is NP-hard.
- Optimal solutions limited to about 30; heuristics appear fairly good to about 500.
Likelihood-Based Methods

- Based on a specific model of evolution and *estimate all model parameters*.
- Produce a *likelihood estimate* (prior or posterior conditional) for each tree.
- Statistically consistent.
- Reconstruct *only topologies*.
- Prone to numerical problems: likelihood of typical trees is infinitesimal.
- Presumably NP-hard; even scoring one tree is very expensive.
- Optimal solutions limited to 4; heuristic solutions appear fairly good to about 100.
Meta-Methods

**General Principle:**
decompose the dataset into smaller, overlapping subsets, reconstruct trees for the subsets (by some base method), and combine the results into a tree for the entire dataset.

- **Quartet-based methods:** use all possible smallest subsets (quartet: set of 4 genomes); best-known is Tree-Puzzle. Slow and inherently inaccurate for any base method.

- **Disk-covering method (DCM):** set up graph from distance matrix, nd overlapping triangulated subgraphs, use them for decomposition. High-powered machinery succeeds very well, especially when tree is imbalanced.

– p. 28
Meta-Methods

General Principle:
decompose the dataset into smaller, overlapping subsets, reconstruct trees for the subsets (by some base method), and combine the results into a tree for the entire dataset.

- **Quartet**-based methods:
  use all possible smallest subsets (quartet: set of 4 genomes); best-known is Tree-Puzzle.
  *Slow and inherently inaccurate* for any base method.

- **Disk-covering** method (**DCM**):
  set up graph from distance matrix, find overlapping triangulated subgraphs, use them for decomposition.
  High-powered machinery *succeeds* very well, especially when tree is imbalanced.
Limitations and Challenges

- **Accuracy**
  not a matter of optimization, but of *scientific truth*!
  how does it scale? how do we evaluate it?

- **Computational Demands**
  all criterion-based optimizations are NP-hard
  the more accurate the model, the worse the problem

- **Data Integration**
  a single type of data cannot answer all questions
  but integration is beyond our reach

- **Database Design**
  database “search” is often a linear search: complex objects give rise to difficult queries
Limitations on Accuracy

- True distances cannot be computed
- Insufficient sequence length
- Primitive or erroneous models
- Algorithmic idiosyncrasies
  (NJ suffers with high diameter, MP suffers from long branch attraction, ML cannot be optimized)
- Gene evolution is not species evolution
- Not a tree, but a directed acyclic graph
  (due to hybridization, lateral gene transfer, etc.)
Evaluating Accuracy

- There is only one instance!
- We want the truth, but it cannot be known or measured
- Optimization is done on surrogate criteria
- Simulation studies are only as good as models
- Parameter space is ridiculously large
- What matters: tree structure? edge lengths? data at internal nodes?
A simple query such as

what is the percentage of trees in the DB in which
organisms $x_1, \ldots, x_m$ and organisms $y_1, \ldots, y_n$
occur in distinct subtrees?

requires a linear search through the DB!

The famous BLAST algorithm was designed to speed up
a similar linear search.

How can we preprocess and store the data
so as to avoid linear searches?
Research in my Laboratory

- Scaling up methods through algorithm design, algorithm engineering, and high-performance computing.
- Whole-genome rearrangements in phylogenetic analysis and comparative genomics.
- Reticulate (non-tree) evolution and its reconstruction.
- Computing directly from databases (rather than in-core).

compbio.unm.edu
DCM Methods

Three varieties so far:
DCM Methods

Three varieties so far:

- **DCM-1**: Disks are cliques, to minimize diameter.
Three varieties so far:

- **DCM-1**: Disks are cliques, to minimize diameter.

- **DCM-2**: Disks are made of a graph separator plus a component, so all disks share same subset.
Three varieties so far:

- **DCM-1**: Disks are cliques, to minimize diameter.

- **DCM-2**: Disks are made of a graph separator plus a component, so all disks share same subset.

- **Rec-I-DCM-3**: Uses recursion and iteration, the latter controlled through a *guide tree*.
DCM-Boosted Methods

- DCM-1-NJ (with an MP last step) beats NJ and greedy MP on sequence data and is robust against size, rate, and other model variations.
- DCM-1-GRAPPA scales gracefully from the limit of 15 genomes for GRAPPA to at least 1,000 genomes.
- Rec-I-DCM-3 with MP does better than any other method on large real datasets and scales to at least 15,000 taxa.
Direct Approaches: BPAnalysis

(due to 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 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

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 ...
DCM-GRAPPA

Our extension to GRAPPA to scale it to large datasets (Tang and Moret).

- Scales gracefully to at least 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 the DCM-1 approach—may do even better with forthcoming DCM-3.
DCM-GRAPPA: Details

- Compute pairwise distances
- Check all possible threshold values
- For each threshold value
  - Discard values above threshold
  - Create graph from reduced distance matrix
  - Triangulate the graph
  - Find maximum cliques (disks) in the graph
  - Run GRAPPA (or recursive DCM-GRAPPA) on the disks
  - Merge the resulting trees
How to choose test sets?

- **Biological datasets** test performance where it matters, but can be used only for ranking, are too few to permit quantitative evaluations, and are often hard to obtain. *Good for anecdotal reports and “reality checks.”*

- **Simulated datasets** enable absolute evaluations of solution quality and can be generated in arbitrarily large numbers. *Only way to obtain valid characterizations.*
Results: Distance Methods

inversion/transposition/inverted transposition: 1:1:1 ratio
120 genes per genome, 10-20-40-80-160 genomes

![Graph showing normalized maximum pairwise inversion distance vs. false negative rate for different methods.]
Results: DCM-GRAPPA

inversion-only evolution, expected edge length 4
100 genes per genome, 20-40-80-160-320-640 genomes

Shown is total number of edges in error (log/log scale)
Fresh off the Press: Bacterial Distances

13 gamma-proteobacteria:
from 500 genes for *B. aphidicola* (obligate endosymbiont) to 5,000 genes for *P. aeruginosa*
3,430 gene families after dropping singletons. We concatenated the 2 chromosomes of *V. cholerae*.

<table>
<thead>
<tr>
<th></th>
<th>Baphi</th>
<th>Ecoli</th>
<th>Hinfl</th>
<th>Paer</th>
<th>Pmult</th>
<th>Styphi</th>
<th>Vchol</th>
<th>Wigg</th>
<th>Xaxo</th>
<th>Xcamp</th>
<th>Xfast</th>
<th>Y_CO92</th>
<th>Y_KIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>302</td>
<td>339</td>
<td>343</td>
<td>330</td>
<td>299</td>
<td>317</td>
<td>277</td>
<td>343</td>
<td>343</td>
<td>341</td>
<td>297</td>
</tr>
<tr>
<td>367</td>
<td>461</td>
<td>0</td>
<td>1141</td>
<td>1710</td>
<td>1175</td>
<td>465</td>
<td>1332</td>
<td>530</td>
<td>1484</td>
<td>1484</td>
<td>1040</td>
<td>981</td>
<td>991</td>
</tr>
<tr>
<td>484</td>
<td>484</td>
<td>809</td>
<td>0</td>
<td>889</td>
<td>574</td>
<td>815</td>
<td>827</td>
<td>520</td>
<td>829</td>
<td>825</td>
<td>745</td>
<td>824</td>
<td>823</td>
</tr>
<tr>
<td>343</td>
<td>492</td>
<td>1634</td>
<td>1050</td>
<td>0</td>
<td>1126</td>
<td>1624</td>
<td>1363</td>
<td>530</td>
<td>1574</td>
<td>1586</td>
<td>1006</td>
<td>1546</td>
<td>1577</td>
</tr>
<tr>
<td>497</td>
<td>497</td>
<td>871</td>
<td>643</td>
<td>987</td>
<td>0</td>
<td>873</td>
<td>895</td>
<td>519</td>
<td>894</td>
<td>891</td>
<td>784</td>
<td>880</td>
<td>890</td>
</tr>
<tr>
<td>343</td>
<td>451</td>
<td>458</td>
<td>1152</td>
<td>1725</td>
<td>1190</td>
<td>0</td>
<td>1350</td>
<td>520</td>
<td>1501</td>
<td>1505</td>
<td>1048</td>
<td>991</td>
<td>1035</td>
</tr>
<tr>
<td>457</td>
<td>467</td>
<td>1115</td>
<td>1017</td>
<td>1245</td>
<td>1063</td>
<td>1120</td>
<td>0</td>
<td>511</td>
<td>1202</td>
<td>1184</td>
<td>929</td>
<td>1082</td>
<td>1105</td>
</tr>
<tr>
<td>467</td>
<td>294</td>
<td>338</td>
<td>378</td>
<td>377</td>
<td>363</td>
<td>337</td>
<td>354</td>
<td>0</td>
<td>401</td>
<td>400</td>
<td>391</td>
<td>334</td>
<td>341</td>
</tr>
<tr>
<td>460</td>
<td>454</td>
<td>916</td>
<td>780</td>
<td>902</td>
<td>806</td>
<td>915</td>
<td>865</td>
<td>503</td>
<td>553</td>
<td>555</td>
<td>0</td>
<td>888</td>
<td>895</td>
</tr>
<tr>
<td>459</td>
<td>460</td>
<td>1385</td>
<td>934</td>
<td>1472</td>
<td>984</td>
<td>1403</td>
<td>1206</td>
<td>530</td>
<td>176</td>
<td>0</td>
<td>736</td>
<td>1275</td>
<td>1279</td>
</tr>
<tr>
<td>468</td>
<td>459</td>
<td>913</td>
<td>1110</td>
<td>1575</td>
<td>1147</td>
<td>923</td>
<td>1210</td>
<td>505</td>
<td>1340</td>
<td>1332</td>
<td>991</td>
<td>0</td>
<td>163</td>
</tr>
<tr>
<td>468</td>
<td>468</td>
<td>940</td>
<td>1114</td>
<td>1609</td>
<td>1170</td>
<td>967</td>
<td>1250</td>
<td>514</td>
<td>1346</td>
<td>1335</td>
<td>996</td>
<td>166</td>
<td>0</td>
</tr>
</tbody>
</table>
Results: Unequal Gene Content (3)

We used the reference phylogeny to compute gene content and to assign loss vs. gain—hence the asymmetry.

Some pairwise distances are enormous: up to $1.600$ events! On some paths, each edge has at least $400$ events on it.

NJ Reconstruction

Only one edge is in error: the edge leading to *V. cholerae*, due to our handling of its two chromosomes.
Notes:

- We used the reference phylogeny to compute gene content and to assign loss vs. gain—hence the asymmetry.
- Some pairwise distances are enormous: up to 1,600 events!
  On some paths, each edge has at least 400 events on it.
Notes:
- We used the reference phylogeny to compute gene content and to assign loss vs. gain—hence the asymmetry.
- Some pairwise distances are enormous: up to 1,600 events! On some paths, each edge has at least 400 events on it.

NJ Reconstruction
*Only one edge* is in error: the edge leading to *V. cholerae*, due to our handling of its two chromosomes.
Conclusions

Gene-order data carries very good phylogenetic information—much better than sequence data! Current algorithmic approaches scale to significant sizes (1,000 for DCM-GRAPPA)—comparable to the best achievable with sequence data and with better results. Current approaches remain unable to handle unequal gene content with duplications, but major progress has been made over the last 5 years. Data availability is increasing rapidly for organellar genomes, slowly for nuclear genomes, and remains very limited compared to sequence data.
Conclusions

- Gene-order data carries very good phylogenetic information—much better than sequence data!
Conclusions

- Gene-order data carries very good phylogenetic information—much better than sequence data!
- Current algorithmic approaches scale to significant sizes (1,000 for DCM-GRAPPA)—comparable to the best achievable with sequence data and with better results.
Conclusions

- Gene-order data carries very good phylogenetic information—much better than sequence data!
- Current algorithmic approaches scale to significant sizes (1,000 for DCM-GRAPPA)—comparable to the best achievable with sequence data and with better results.
- Current approaches remain unable to handle unequal gene content with duplications, but major progress has been made over the last 5 years.
Conclusions

- Gene-order data carries very good phylogenetic information—much better than sequence data!
- Current algorithmic approaches scale to significant sizes (1,000 for DCM-GRAPPA)—comparable to the best achievable with sequence data and with better results.
- Current approaches remain unable to handle unequal gene content with duplications, but major progress has been made over the last 5 years.
- Data availability is increasing rapidly for organellar genomes, slowly for nuclear genomes, and remains very limited compared to sequence data.
Laboratory for High-Performance Algorithm Engineering and Computational Molecular Biology

Includes all publications by our lab, GRAPPA source files, email addresses, and links to our main collaborators.