Computational Challenges from the Tree of Life

Bernard M.E. Moret

compbio.unm.edu

Department of Computer Science
University of New Mexico

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
- Postdocs
  - Tanya Berger-Wolf (now DIMACS & U. Illinois)
  - Tiffani Williams (now Harvard & Texas A&M)
  - Jijun Tang (now U. South Carolina)

Acknowledgments

- Frequent Collaborators:
  - *KTH Stockholm:* Jens Lagergren
  - *U Arizona:* Nancy Moran, Howard Ochman
  - *U Bologna:* Alberto Caprara
  - *UC Berkeley:* Brent Mishler, Gene Myers, Richard Karp, Christos Papadimitriou, Satish Rao, Russell Stewart
  - *UC San Diego:* Fran Berman, Pavel Pevzner
  - *U Montpellier:* Olivier Gascuel
  - *U Ottawa:* David Sankoff
  - *U Pennsylvania:* Junhyong Kim
  - *UT Austin:* Robert Jansen, Randy Linder
  - *Yale:* Michael Donoghue

Overview

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

Phylogenies: What?

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

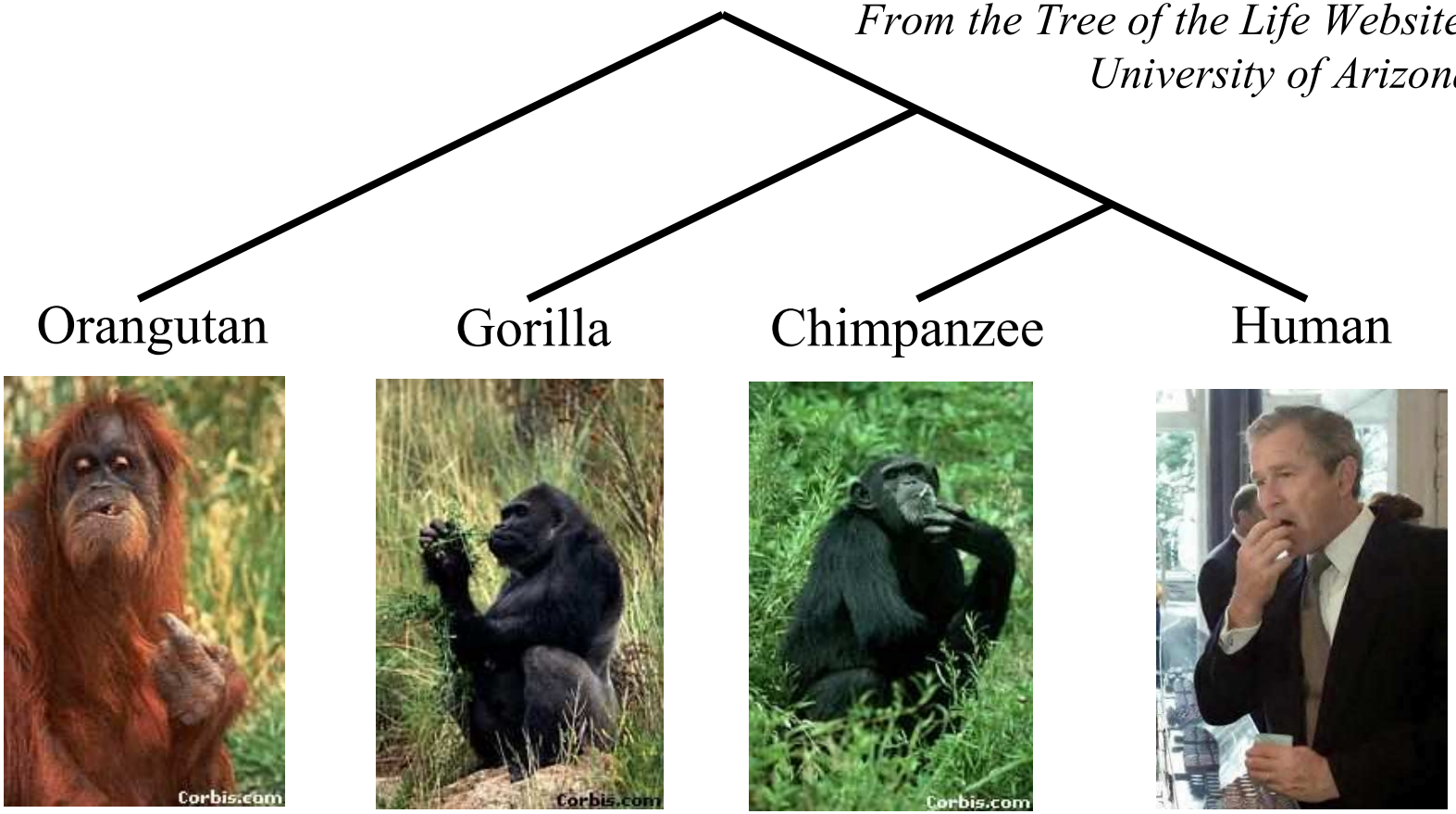
It usually takes the form of a tree.

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

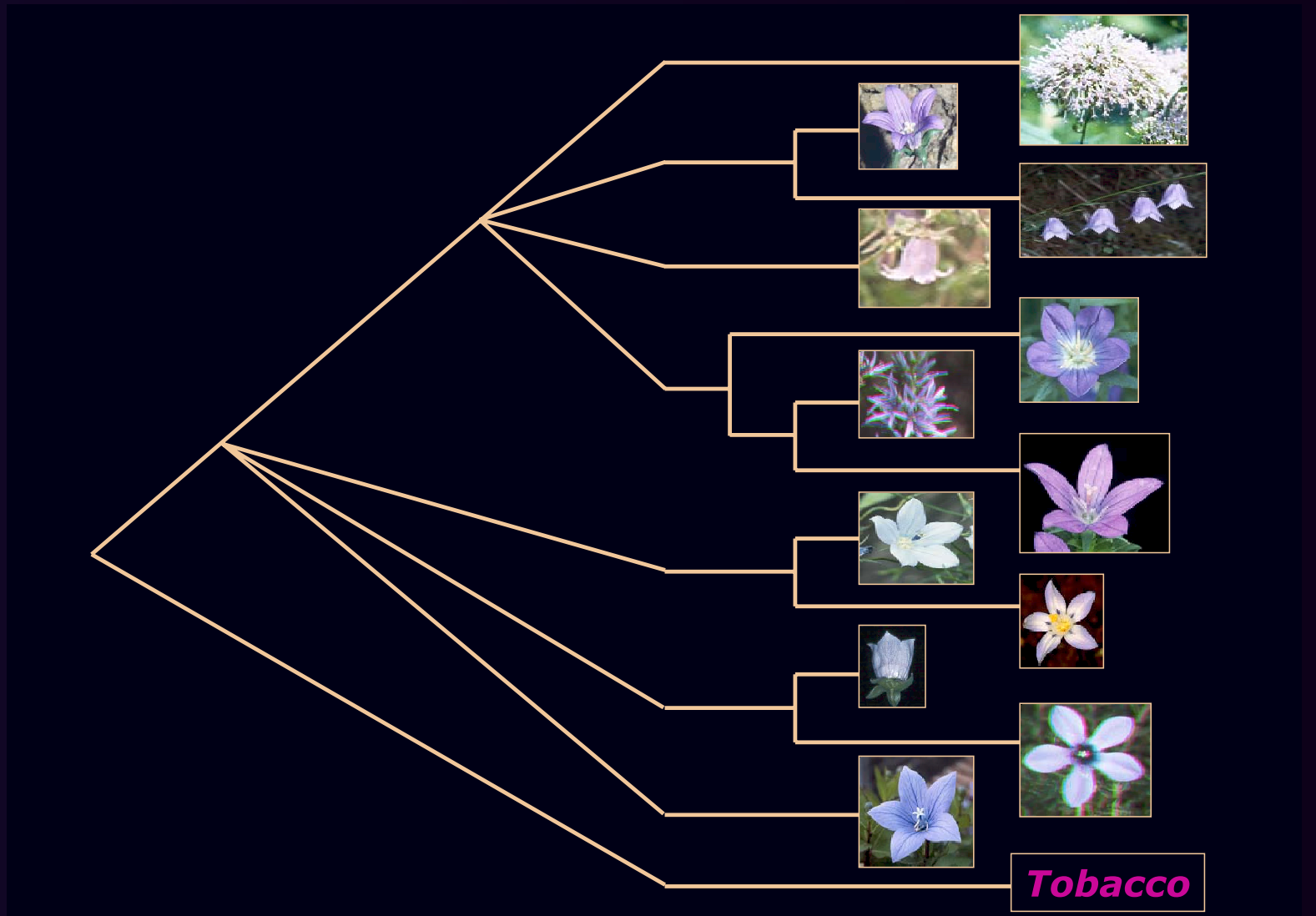
The Great Apes

Phylogeny

From the Tree of the Life Website,
University of Arizona



12 Species of Campanulaceae



By Robert Jansen & Linda Raubeson

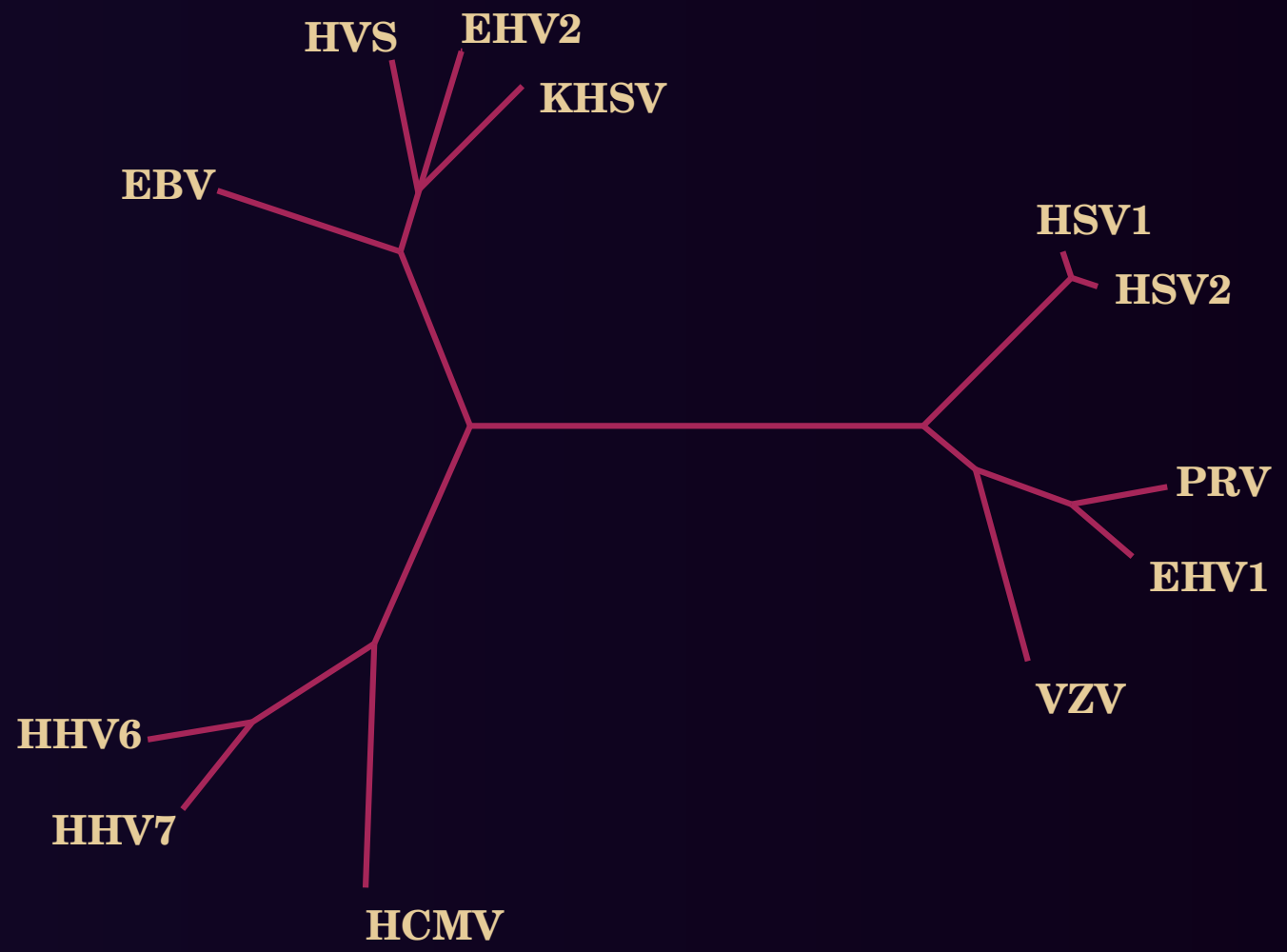
Phylogenies: Why?

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

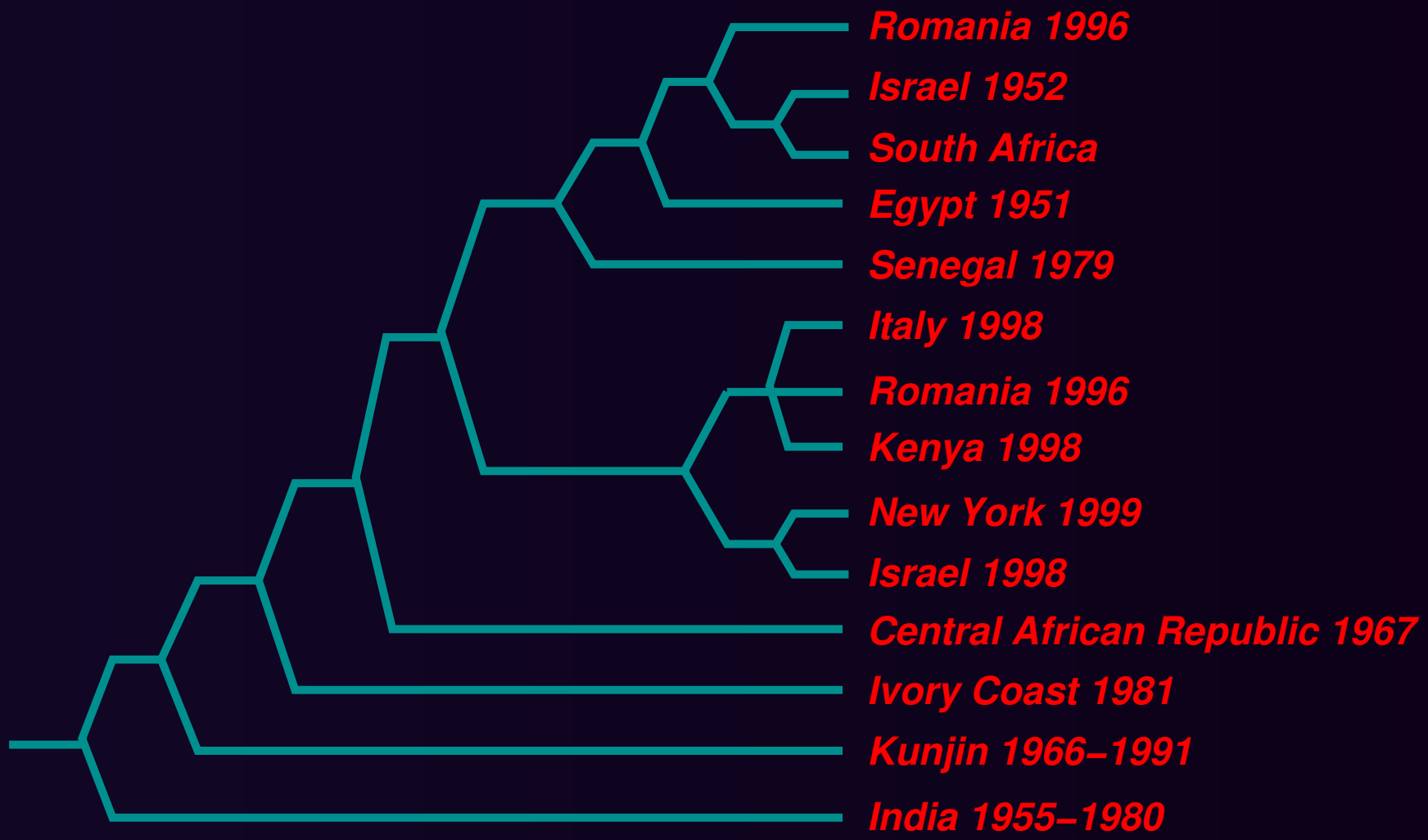
They help us understand and predict:

- 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

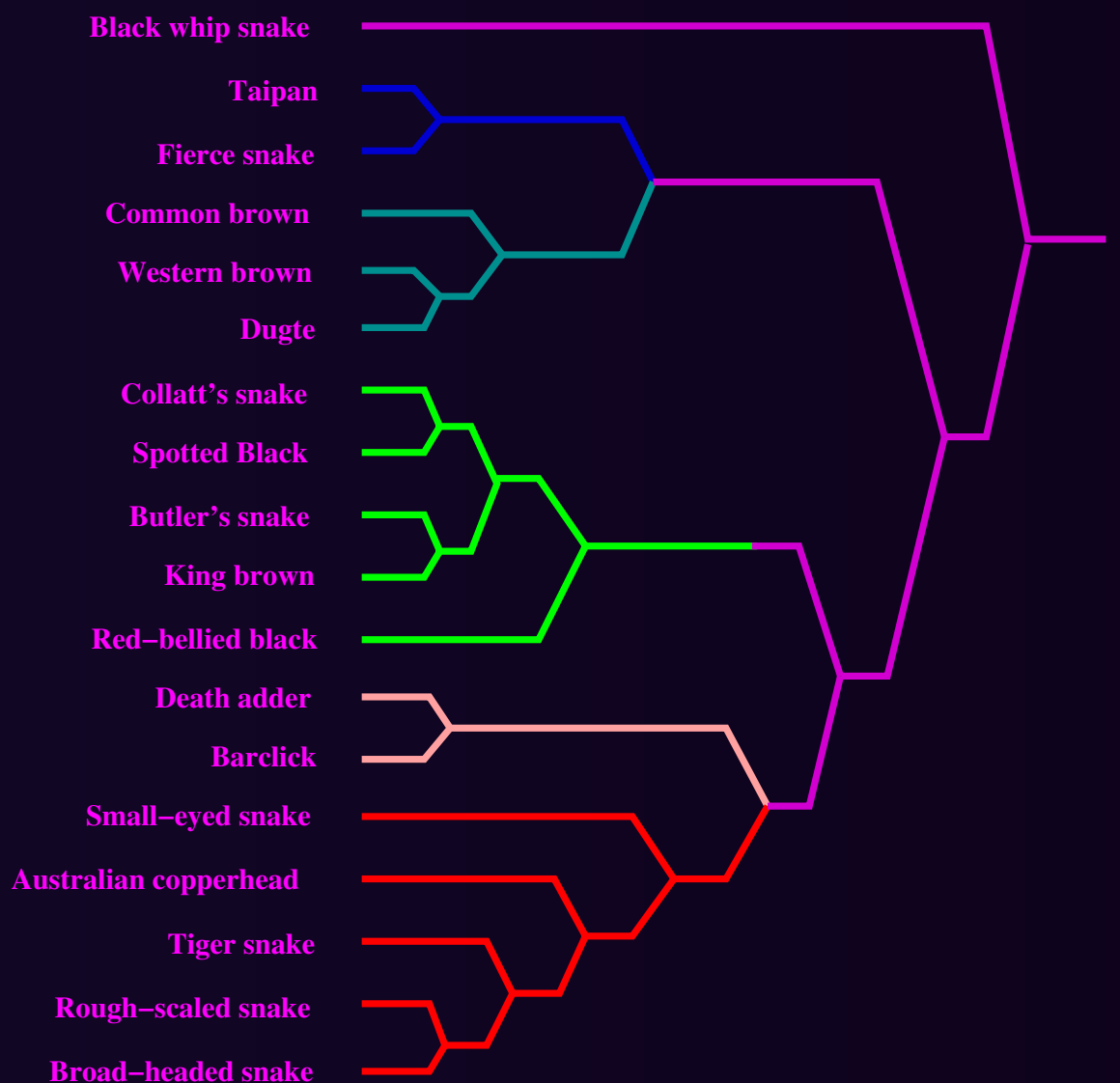
Herpes Viruses that Affect Humans



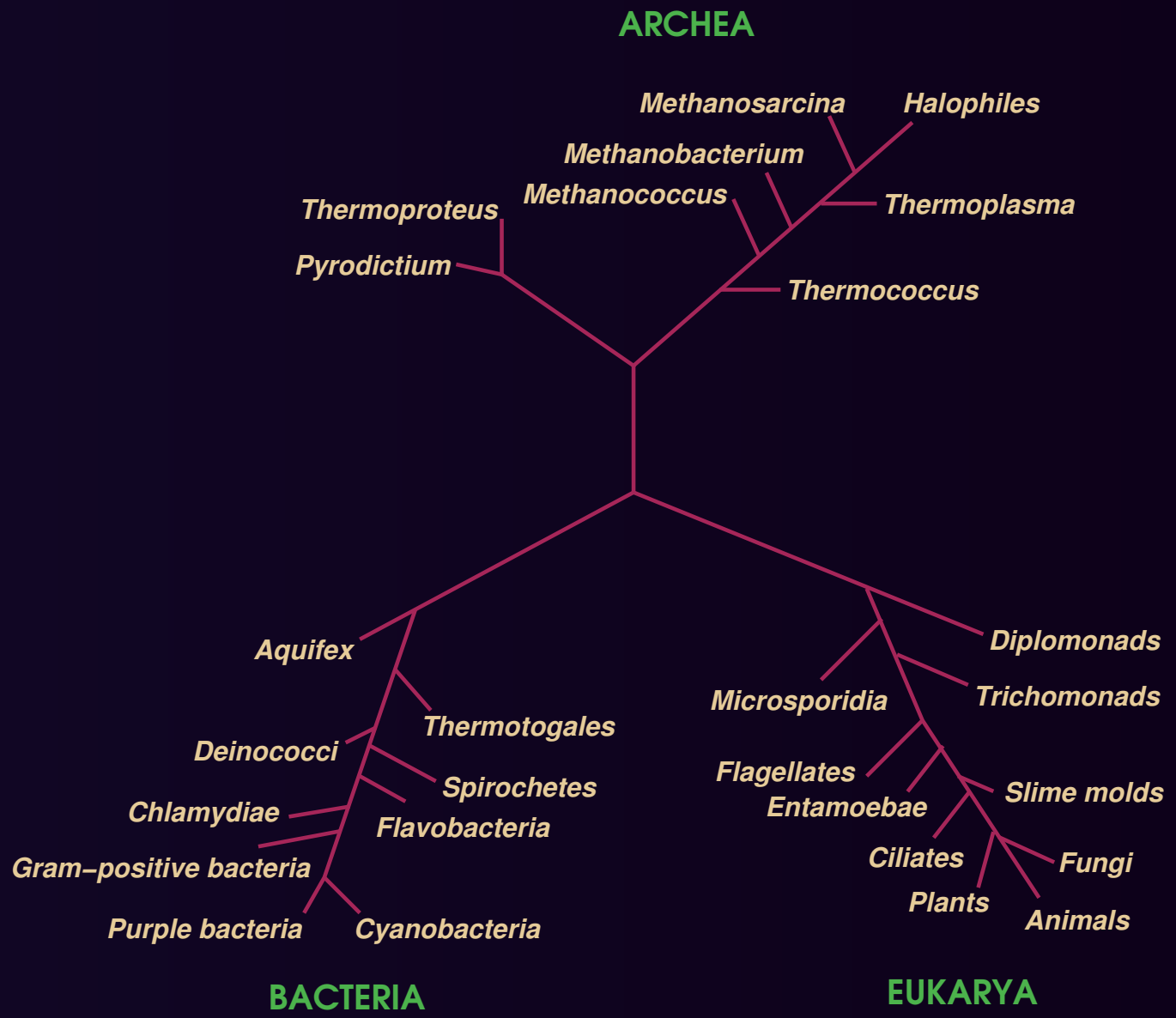
Epidemiology of West Nile Virus



Drug Design: Antivenins



The Tree of Life



The Tree of Life: Scale?

- 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 20–50organisms.
- Gene-based tools can handle 200–500 organisms.
- Both sets of tools scale exponentially with the amount of data.

Phylogenetic Reconstruction: How?

- **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.

Molecular Data

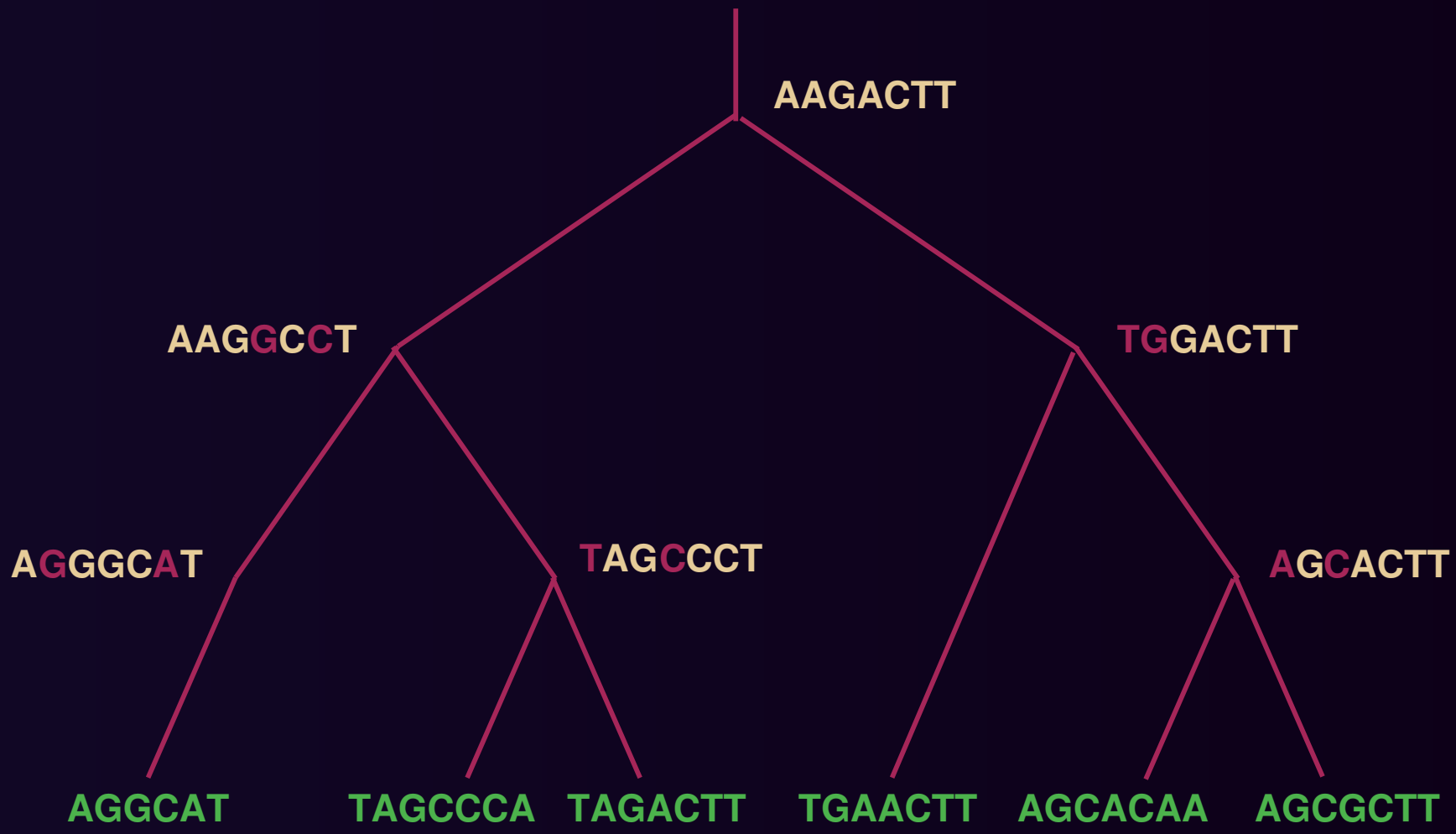
Typically the DNA sequence of a few genes.

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.

- 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



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

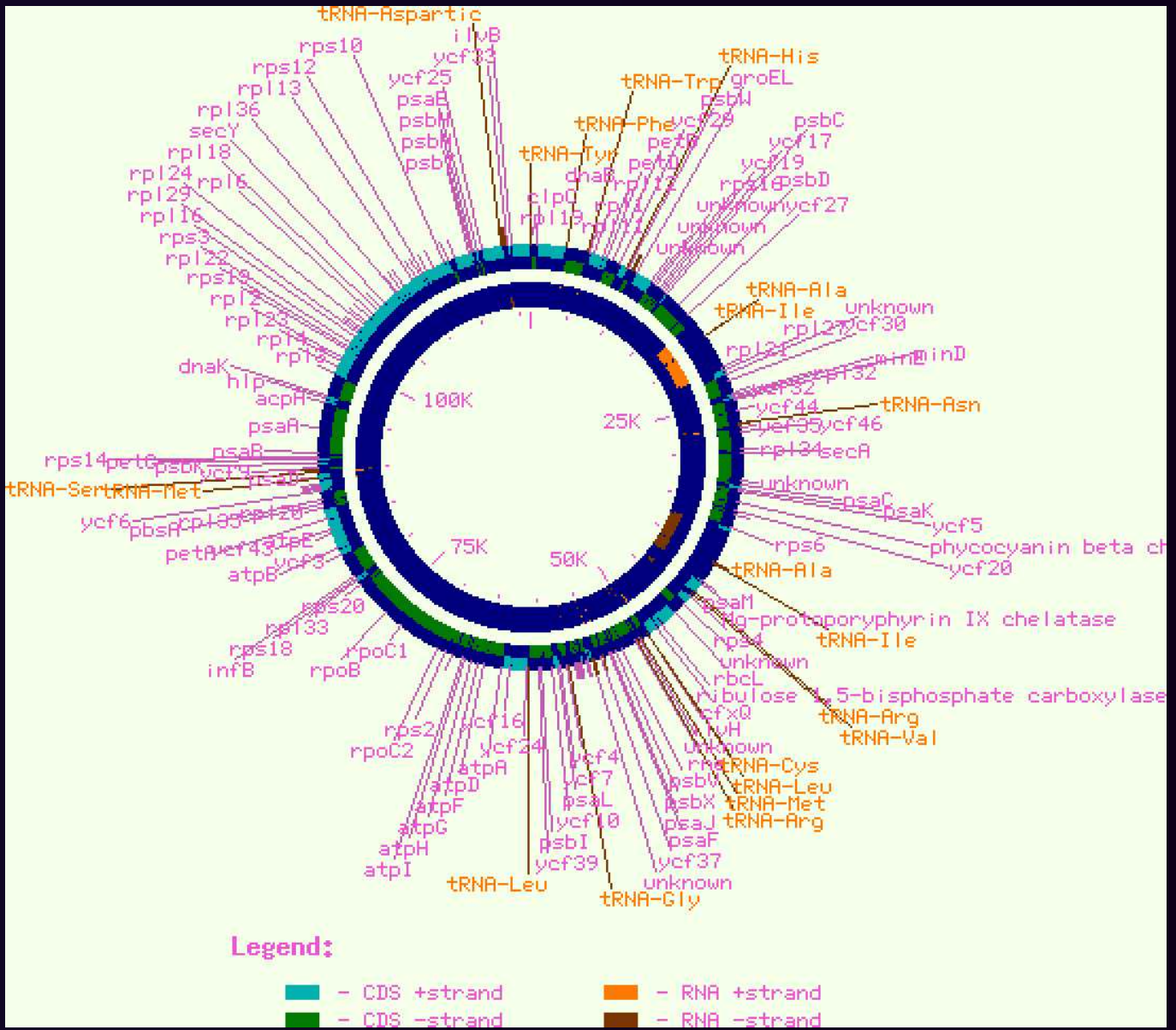
The ordered sequence of genes on one or more chromosomes.

Entire gene-order is a single character, which can assume a huge number of states.

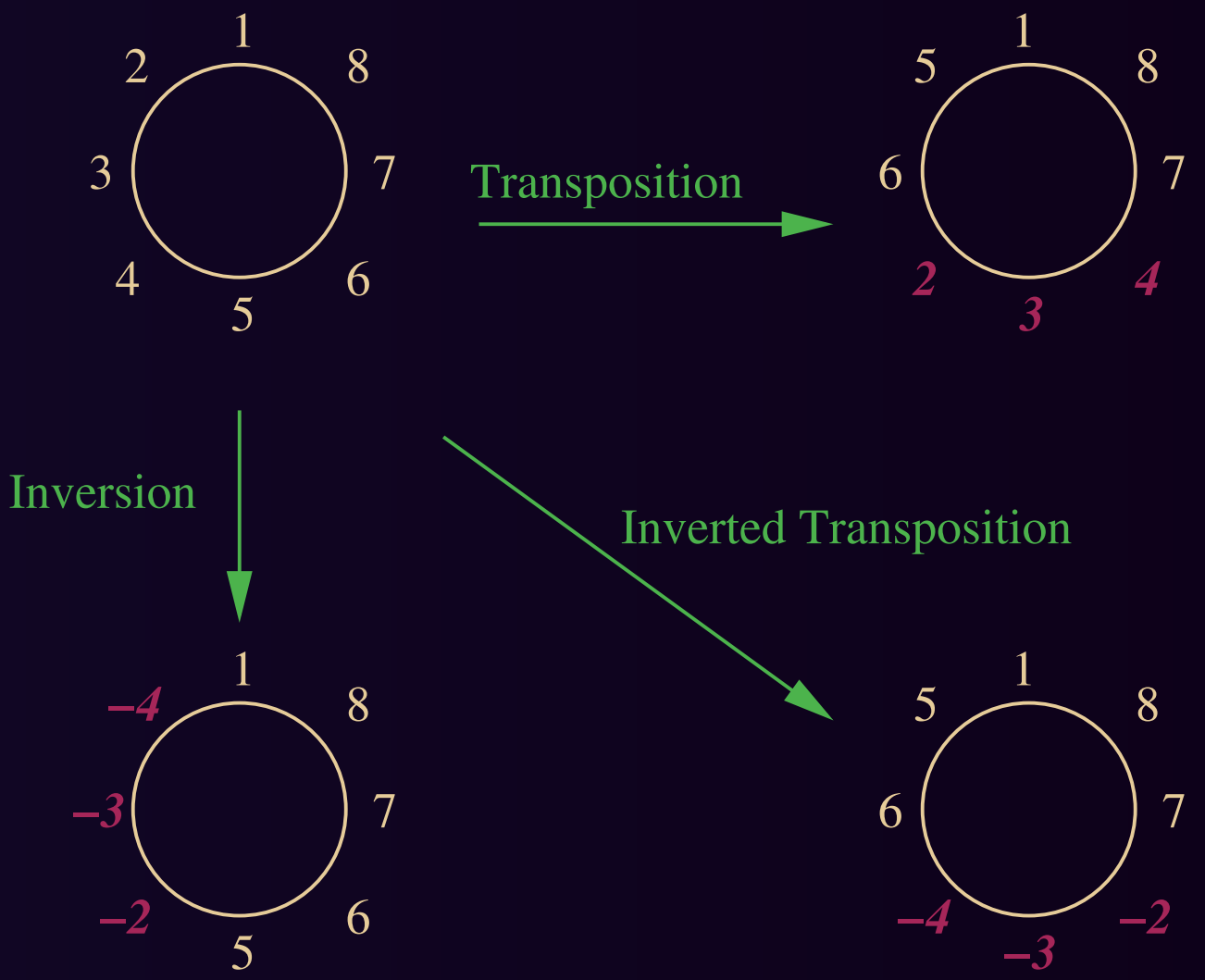
Evolves through inversions, insertions (incl. duplications), and deletions; also transpositions (in mitochondria) and translocations (between chromosomes).

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

Gene-Order: Guillardia Chloroplast



Gene-Order Data: Rearrangements



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.!**

Models

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.

Need more data and refinements:

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

Molecular Evolution Models

Based on large amounts of data, models build transition matrices (4×4 for nucleotides, 20×20 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.*

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

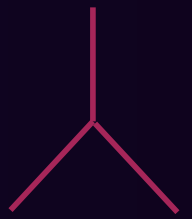
- **True evolutionary distance:**
  the actual number of permitted evolutionary events that took place to transform one datum into the other.
- **Edit distance:**
  the minimum number of permitted evolutionary events that can transform one datum into the other.
- **Expected true evolutionary distance:**
  obtained from the edit distance by correcting for the known (model or experiments) statistical relationship between true and edit distances.

Distance-Based Methods

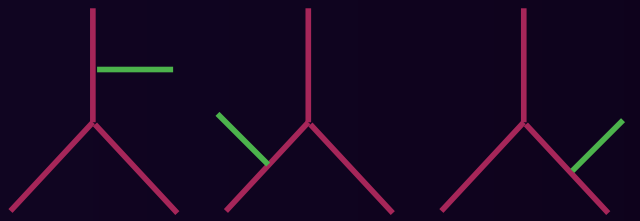
- Use edit or expected true evolutionary distances.
- Usually run in *low polynomial time*.
- Reconstruct *only topologies*: no ancestral data.
- Prototype is Neighbor-Joining.
- NJ is optimal on additive distances (where the distance along a path in the true tree equals the pairwise distance in the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

The Number of Trees for N Organisms

- **3 organisms: 1 tree**



- **4 organisms: 3 trees**



- **5 organisms: 15 trees**
- **13 organisms: 13.5 billion trees**
- ***n* organisms: (2*n*-5)!! trees**

  (2*n*-5)!! = (2*n*-5)*(2*n*-7)*...*5*3

Parsimony-Based Methods

- Aim to minimize total *number of character changes* (which can be weighted to reflect statistical evidence).
- Assume that characters are *independent*.
- Reconstruct *ancestral data*.
- Are known not to be statistically consistent with sequence data (but examples are fairly contrived).
- Finding most parsimonious tree is computational very expensive (NP-hard).
- Optimal solutions limited to sizes around 30; heuristic solutions appear fairly good to sizes of 500.

Likelihood-Based Methods

- Are based on a specific model of evolution and *must estimate all model parameters*.
- Produce *likelihood estimate* (prior or posterior conditional) for each tree.
- Are statistically consistent.
- Reconstruct *only topologies*.
- Are prone to numerical problems: likelihood of typical trees is infinitesimal.
- Are presumably NP-hard; even scoring one tree is very expensive.
- Optimal solutions limited to sizes below 10; heuristic solutions appear fairly good to sizes of 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.

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 from numerical problems)
- **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?

Database Challenges

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

Scaling Up

Distance-based methods scale poorly in accuracy,
so use criterion-based methods.
Criterion-based methods scale poorly for computation,
so use meta-methods.

Our latest findings:

- *To ensure 95% accuracy in reconstructing trees on n leaves, the criterion must be optimized with less than $\frac{1}{n}$ error!*
  (Unheard of in normal approximation problems!)
- *Using recursion and iteration, our latest Disk-Covering Method (Rec-I-DCM3) can handle datasets of 10,000–50,000 sequences as well as previous algorithms could handle 100–500.*
- *Another DCM approach can scale whole-genome analysis from 10 to over 1,000 genomes.*

Gene Rearrangement Phylogeny

Theory

1995
inversion distance
Hannenhalli & Pevzner

1997
breakpoint phylogeny
Blanchette, Bourque, & Sankoff

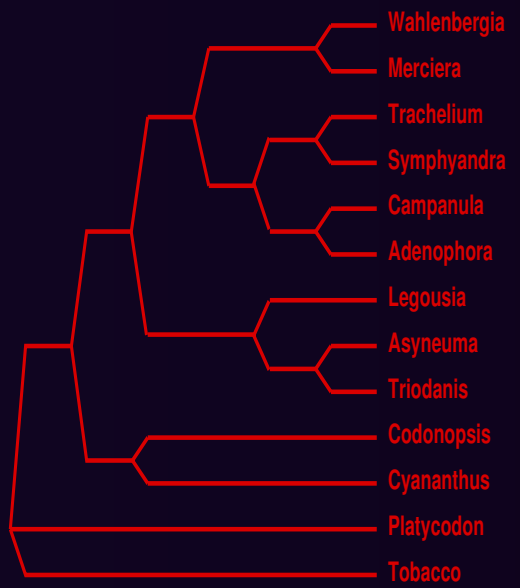
2000
inversion + deletion distance
El-Mabrouk

2001
distance correction
Wang, Warnow, & Moret

2003
inversion + deletion + insertion distance
Marron, Swenson, & Moret

Example

12 Campanulaceae + Tobacco
Jansen, Moret, & Warnow 2000



Reconstruction Software

1998 BPAnalysis
Sankoff
8 taxa ⇒ 1 day
13 taxa ⇒ 250 years

2000 GRAPPA
Moret, Bader, & Warnow
13 taxa ⇒ 1 day (512-proc.)
(200 serial, 100,000 parallel speedup)

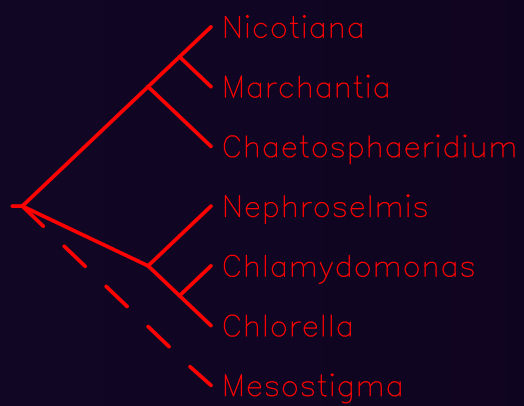
2001 GRAPPA
Moret, Tang, Wang, & Warnow
13 taxa ⇒ 1 hr (laptop)
(2,000,000 serial speedup)
20 taxa ⇒ 3 million years

2003 DCM-GRAPPA
Tang, Moret, & Warnow
1,000 taxa ⇒ 2 days
(effectively unbounded speedup)

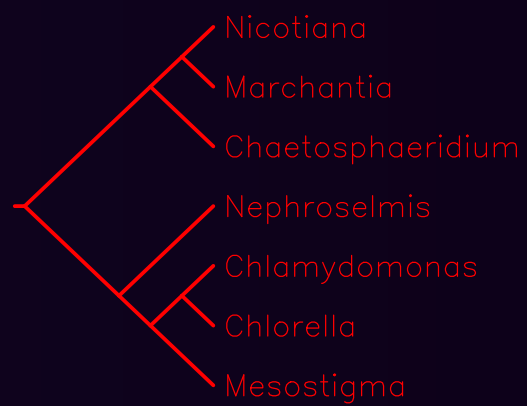
2004 DCM-GRAPPA
Tang & Moret
handles unequal gene content
(first method with that capability)

Unequal Gene Content

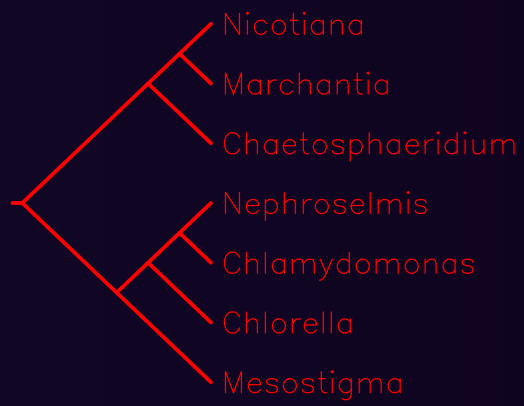
Tang/Moret/Cui/DePamphilis (2004): chloroplast data



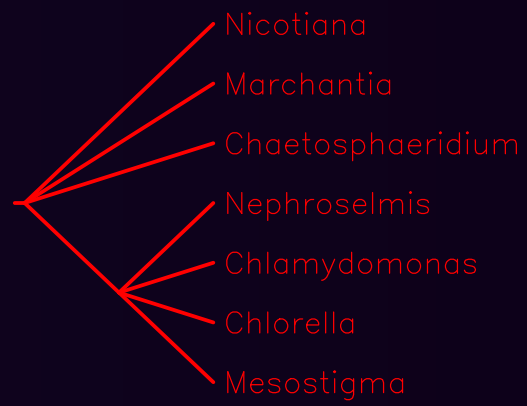
organismal



Tang/Moret GRAPPA



NJ (inv.)



breakpoint GRAPPA

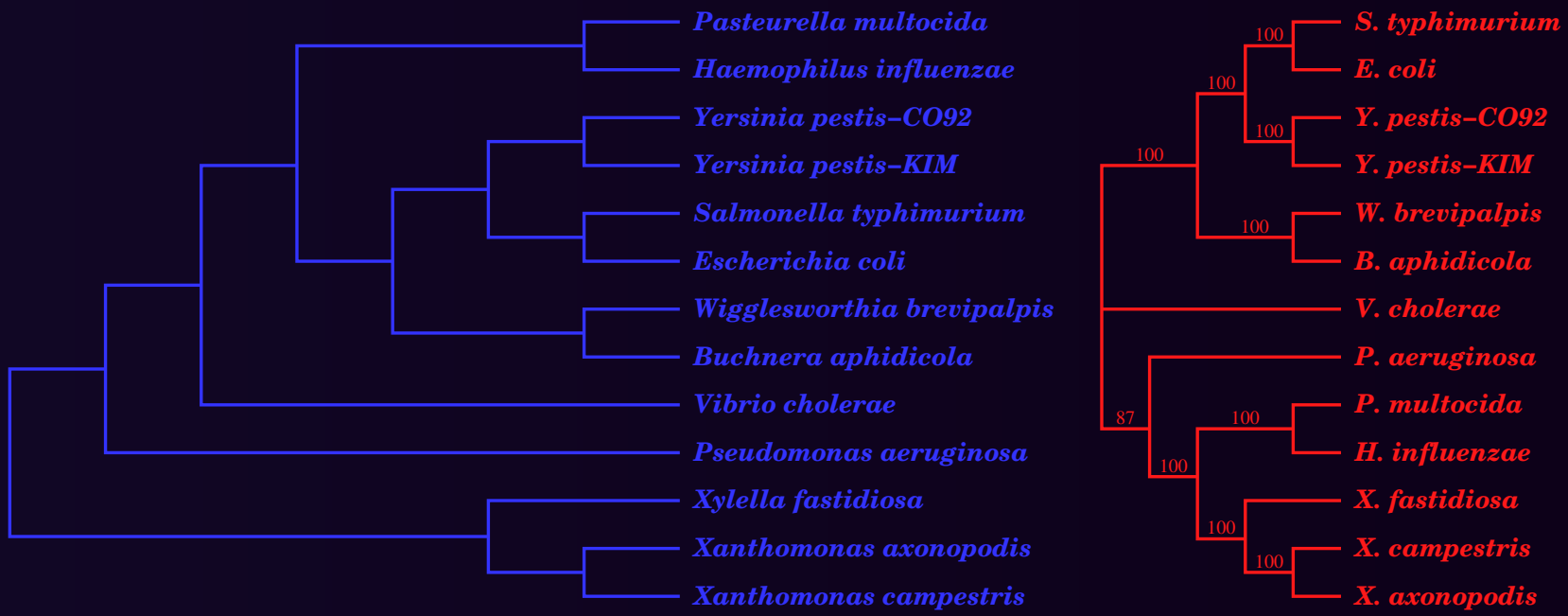
Large Genomes and Distances

(unpublished)

13 gamma proteobacteria (Lerat/Daubin/Moran 2003)

Only gene families occurring in at least 3 species.

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



Only one error in red tree: *{P. multocida/H. influenzae}* moved (long branch attraction in NJ).

The CIPRES Project

Cyber Infrastructure for Phylogenetic Research

www.phylo.org

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

- *over 15 institutions, including three museums*
- *over 40 researchers, evenly split between CS and Biology*
- *director: Bernard Moret*

CIPRES: Participants

CIPRES Members

University of New Mexico
Bernard Moret
David Bader
Tiffani Williams

UCSD/SDSC
Fran Berman
Alex Borchers
David Stockwell
Phil Bourne
John Huelsenbeck
Dana Jermanis
Mark Miller
Michael Alfaro
Tracy Zhao

University of Connecticut
Paul O Lewis

University of Pennsylvania
Junhyong Kim
Sampath Kannan

UT Austin
Tandy Warnow
David M. Hillis
Warren Hunt
Robert Jansen
Randy Linder
Lauren Meyers
Daniel Miranker
Usman Roshan
Luay Nakhleh

University of Arizona
David R. Maddison

University of British Columbia
Wayne Maddison

North Carolina State University
Spencer Muse

American Museum of Natural History
Ward C. Wheeler

UC Berkeley
Satish Rao
Joseph M. Hellerstein
Richard M Karp
Michael Levine
Brent Mishler
Elchanan Mossel
Eugene W. Myers
Christos M. Papadimitriou
Stuart J. Russell

SUNY Buffalo
William Piel

Florida State University
David L. Swofford
Mark Holder

Yale
Michael Donoghue
Paul Turner

Aventis Pharmaceuticals
Lisa Vawter

Conclusions

- Computational Molecular Biology is a marvelous playground for algorithm design, algorithm engineering, database design, etc.
- The computational challenges are truly awe-inspiring: scaling by at least four more orders of magnitude and ensuring 99.999999% accuracy!
- The Tree of Life project is active in Asia, New Zealand, Europe, and North and South America. Data are being collected at a rate that far exceeds Moore's law.
- Assembling the Tree of Life will be a major milestone in understanding life on Earth, and mankind in particular.

compbio.unm.edu

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.