Recent Advances in Phylogeny Reconstruction from Gene-Order Data

Bernard M.E. Moret

Department of Computer Science
University of New Mexico
Albuquerque, NM 87131

Collaborators and Support

- Collaborators:
  University of Texas, Austin:
  Tandy Warnow (Computer Science)
  David Hillis, *Robert Jansen*, *Randy Linder* (Biology)
  University of New Mexico:
  David Bader (Electrical & Comp. Eng.)
- Funding: *National Science Foundation*,
  at UNM: 6 grants for $2 million over 5 years
  with UT Austin: 10 grants for $8 million

Overview

- Phylogenies

Overview

- Phylogenies
- Gene-order data: mitochondrion and chloroplast genomes

Overview

- Phylogenies
- Gene-order data: mitochondrion and chloroplast genomes
- Inversion and other genomic distance measures

Overview

- Phylogenies
- Gene-order data: mitochondrion and chloroplast genomes
- Inversion and other genomic distance measures
- Estimating the true evolutionary distance

Overview

- Phylogenies
- Gene-order data: mitochondrion and chloroplast genomes
- Inversion and other genomic distance measures
- Estimating the true evolutionary distance
- Fast convergence for reconstruction methods

Overview

- Phylogenies
- Gene-order data: mitochondrion and chloroplast genomes
- Inversion and other genomic distance measures
- Estimating the true evolutionary distance
- Fast convergence for reconstruction methods
- GRAPPA news

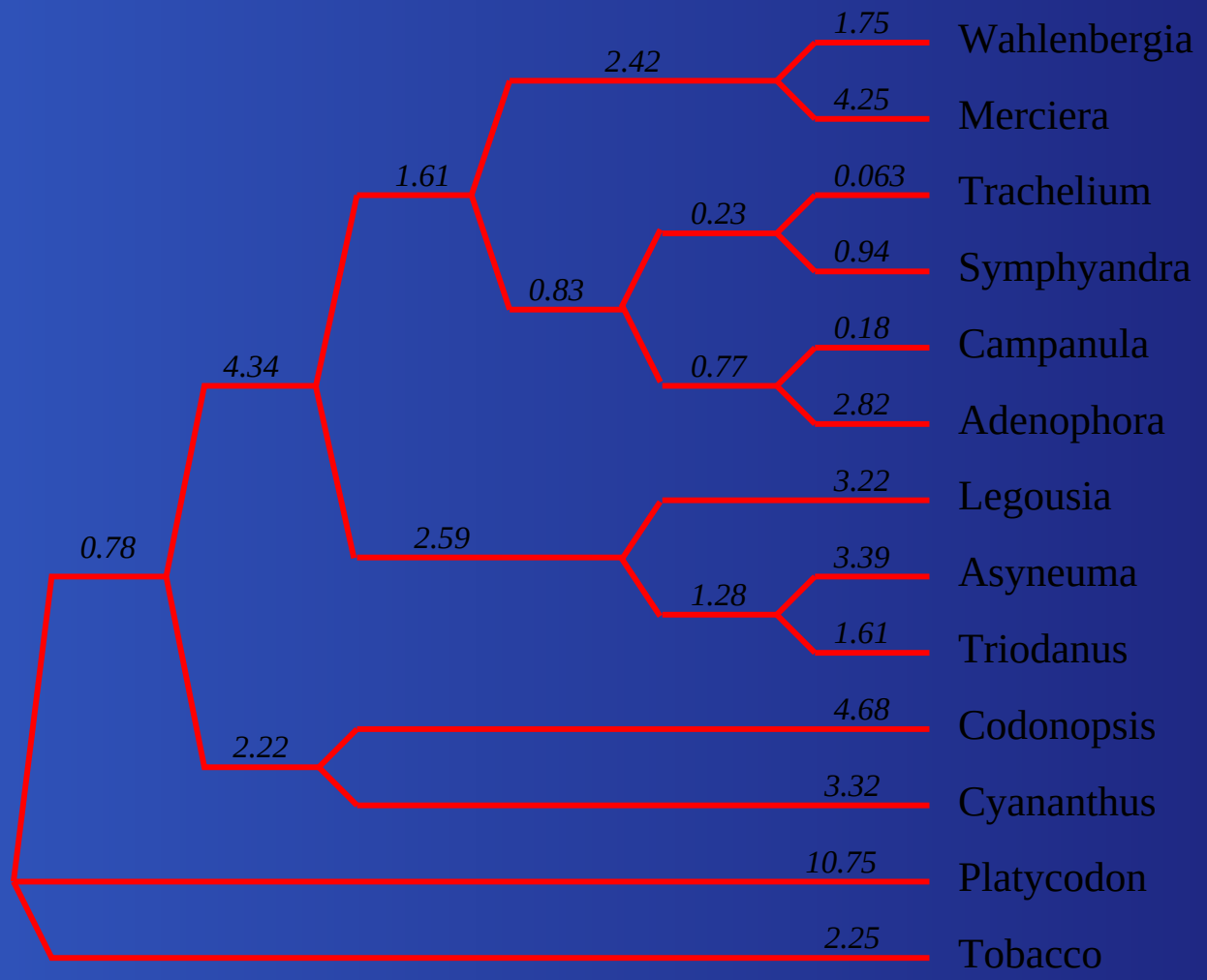
Phylogenies

A phylogeny is a reconstruction of the evolutionary history of a collection of organisms; it usually takes the form of a tree.

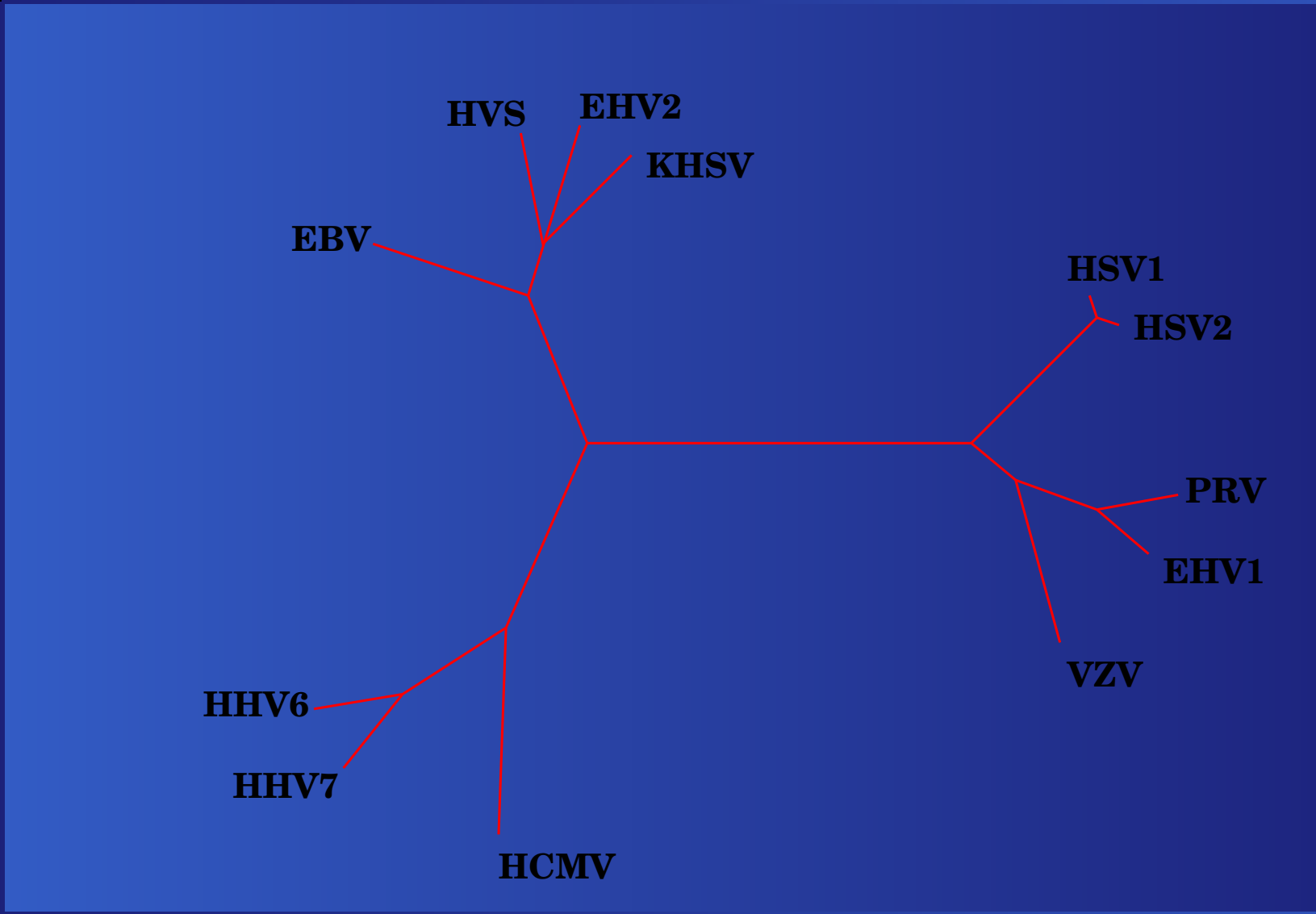
Modern organisms are placed at the leaves and ancestral organisms occupy internal nodes.

The edges of the tree denote evolutionary relationships.

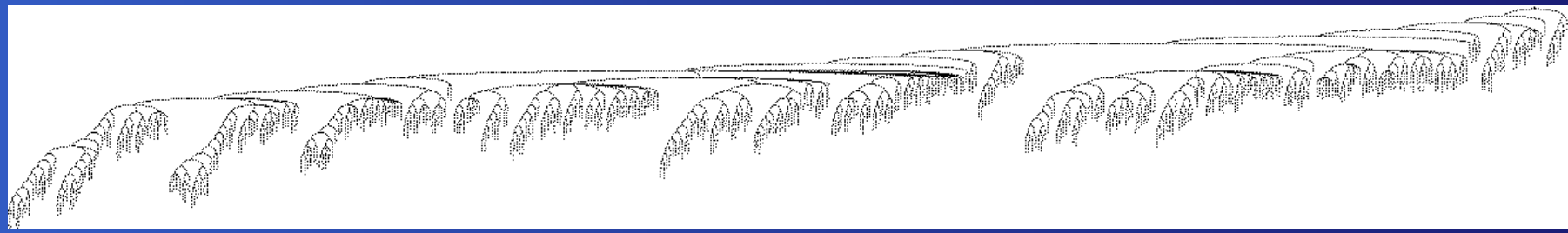
12 Species of *Campanulaceae*



Herpes Viruses that Affect Humans



A Large Phylogeny: 500 Green Plants



Reconstructing Phylogenies

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery (most famous is herbicide Roundup^{TM})*

Reconstructing Phylogenies

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery (most famous is herbicide Roundup™)*
- *understanding rapidly mutating viruses (HIV)*

Reconstructing Phylogenies

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery (most famous is herbicide Roundup™)*
- *understanding rapidly mutating viruses (HIV)*
- *designing enhanced organisms (rice, wheat)*

Reconstructing Phylogenies

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery (most famous is herbicide Roundup™)*
- *understanding rapidly mutating viruses (HIV)*
- *designing enhanced organisms (rice, wheat)*
- *explaining and predicting gene expression*

Reconstructing Phylogenies

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery (most famous is herbicide Roundup™)*
- *understanding rapidly mutating viruses (HIV)*
- *designing enhanced organisms (rice, wheat)*
- *explaining and predicting gene expression*
- *explaining and predicting ligands*

Reconstructing Phylogenies

Reconstructing phylogenies is a major component of modern research programs in many areas of biology and medicine:

- *pharmaceutical research for drug discovery (most famous is herbicide Roundup™)*
- *understanding rapidly mutating viruses (HIV)*
- *designing enhanced organisms (rice, wheat)*
- *explaining and predicting gene expression*
- *explaining and predicting ligands*
- *most centrally, understanding genomic evolution*

Reconstructing Phylogenies (cont'd)

- Requires a *model of tree evolution* (e.g., random or birth-death)
- Requires a *model of DNA/RNA/codon/gene order/etc. evolution* (e.g., Markov models with weights matrices such as Jukes-Cantor and Kimura)
- Requires an *optimization criterion* that relates to the previous two models (e.g., likelihood or parsimony)
- Requires data with sufficient *signal* (to recover defining information)

Computational Phylogenetics

- Is extremely computation-intensive.

  Is viewed very differently by biologists (one dataset only, accuracy first) and by computer scientists (efficiency first)

Computational Phylogenetics

- Is extremely computation-intensive.

  Is viewed very differently by biologists (one dataset only, accuracy first) and by computer scientists (efficiency first)
- Sequence data (RNA, DNA, and aminoacid) has been used for over 20 years and is fairly well understood, but methods do not scale up.

  Genomic data (gene order and content of whole genomes) provides new information, but is *much harder* to analyze than sequence data.

Gene-Order Data

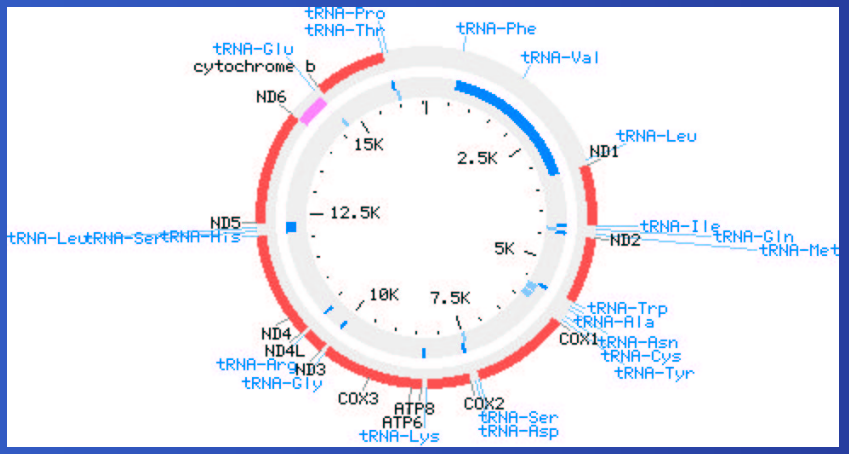
Certain genomes evolve mostly through rearrangement of the order of genes, with occasional gene duplication or gene loss.

A *chloroplast* is a semi-independent organism that lives within plant cells and allows them to photosynthesize.
Chloroplasts have one circular chromosome with $\sim$120 genes.

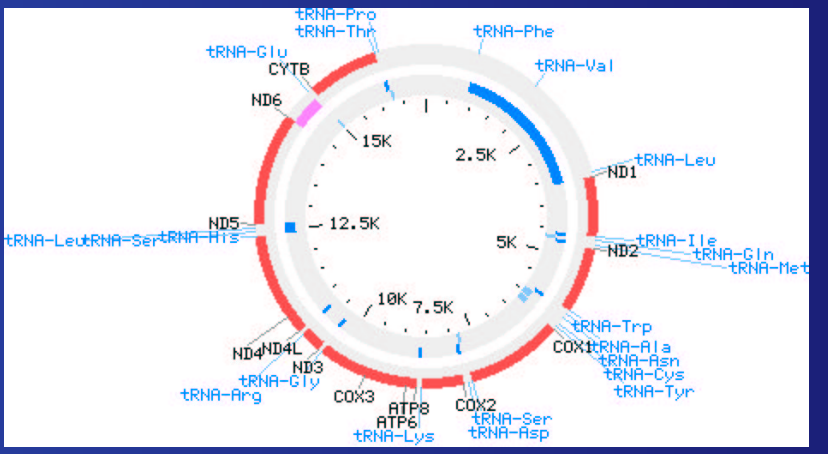
A *mitochondrion* is a semi-independent organism that lives within animal and some plant cells and supplies them with energy.

Mitochondria have one circular chromosome with $\sim$40 genes in animals, more in plants.

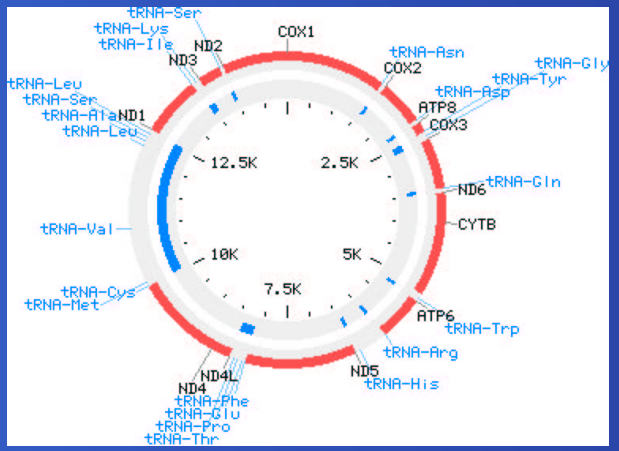
Mitochondria



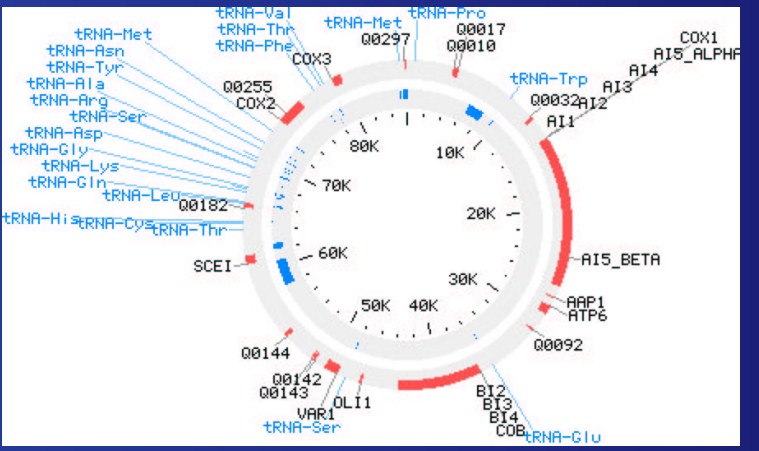
Homo sapiens



Felis catus

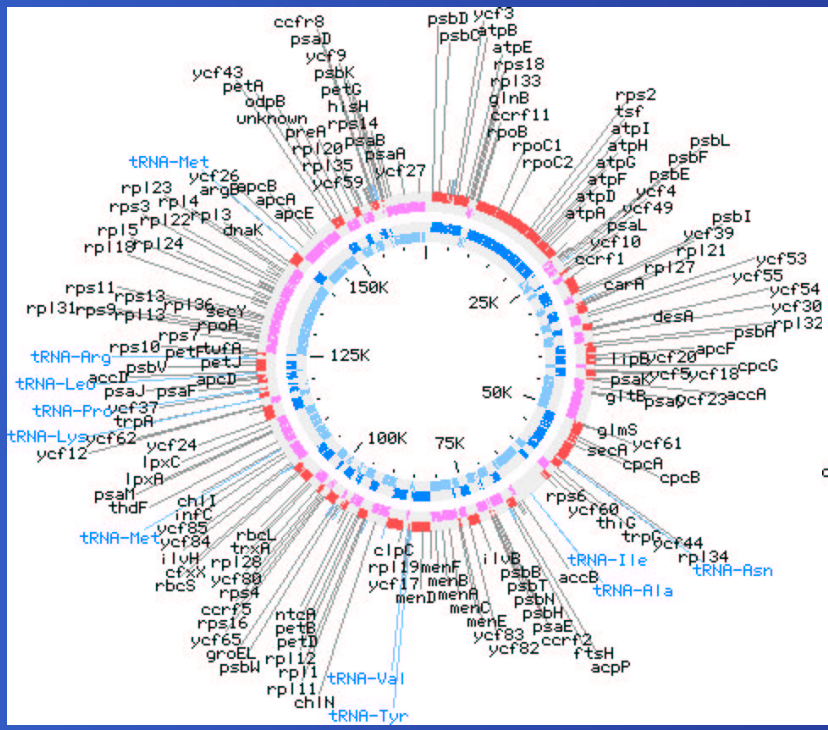


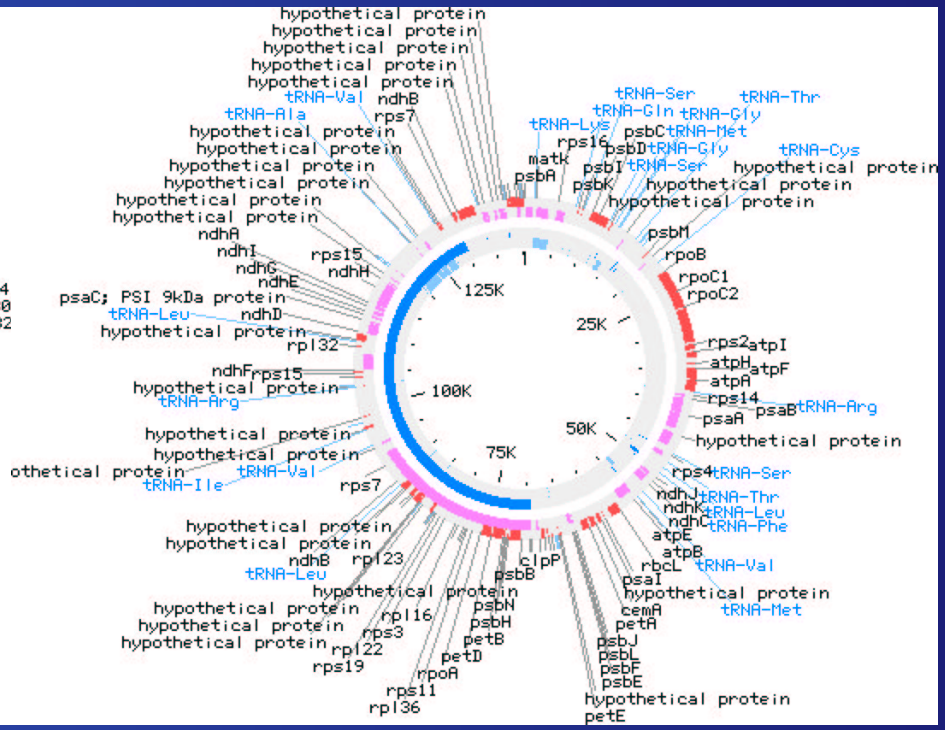
Lumbricus terrestris



Saccharomyces cerevisiae

Chloroplasts





Cyanidium caldarium *Zea mays*

Phylogenies from Gene-Order Data

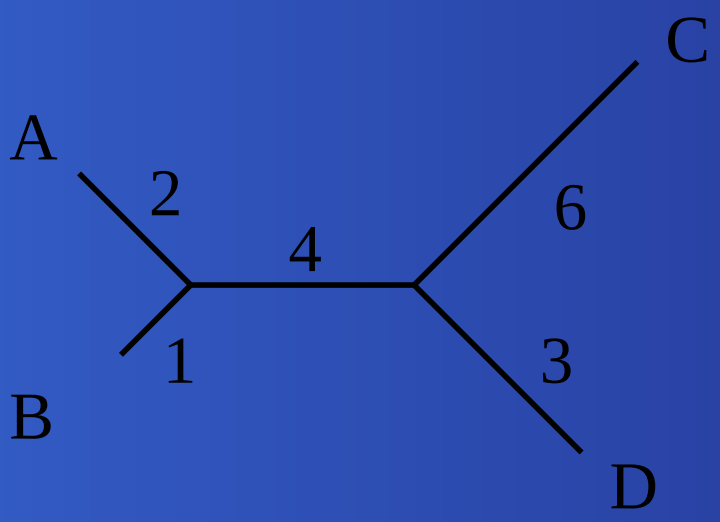
Optimization target: reconstruct the phylogeny with the least total number of genomic changes.

An application of Occam's razor; biologists call this the principle of parsimony.

True Evolutionary Distances

- *True Evolutionary Distance* (T.E.D.): actual number of events along an edge of the tree.
- *Edit Distance*: minimum number of events from one end of a tree edge to the other.
- We obtain better topological accuracy with T.E.D.s than with Edit Distances.
- T.E.D. can only be estimated.

True Evolutionary Distance

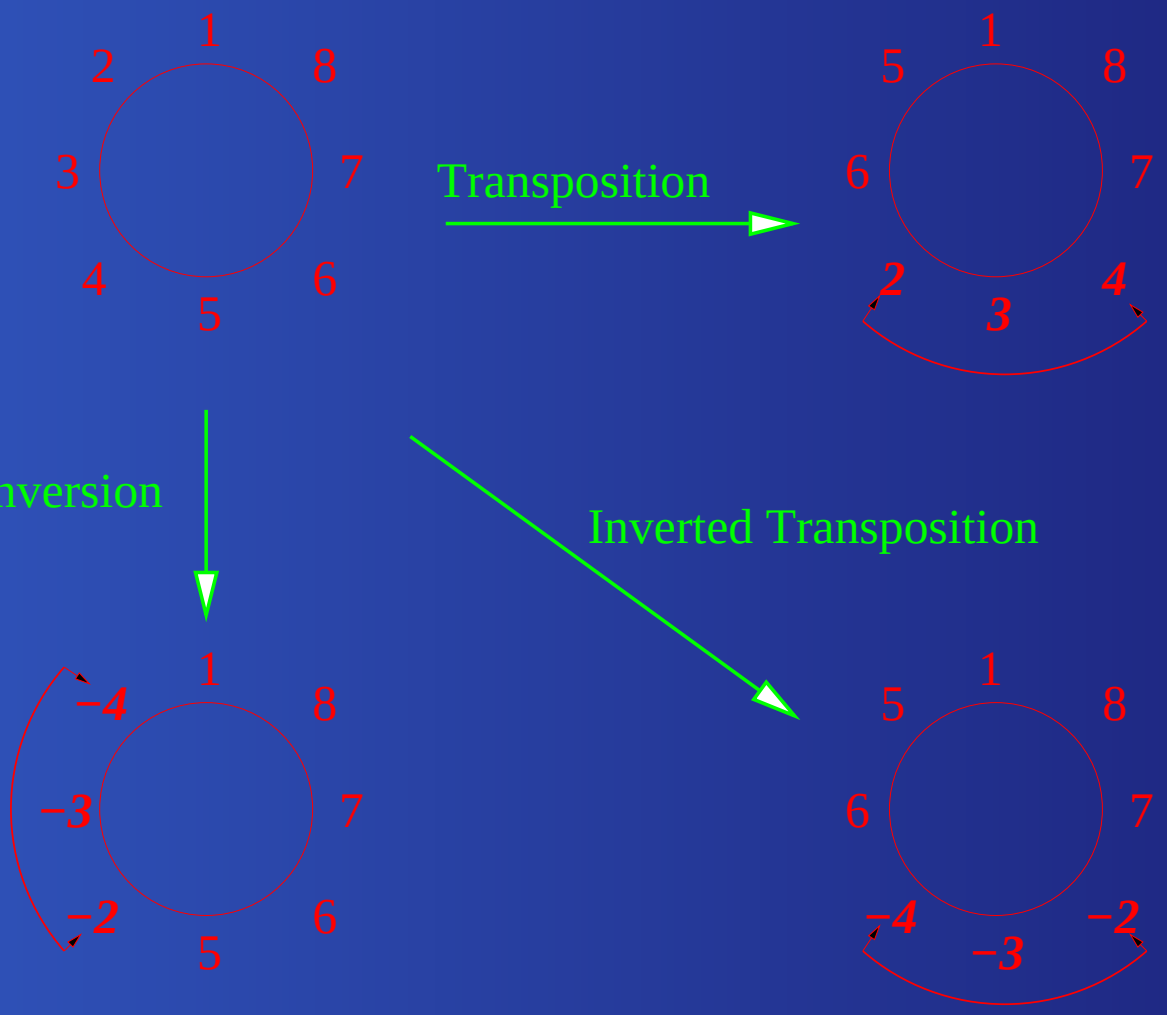


Polynomial Time

	A	B	C	D
A	0	3	12	9
B		0	11	8
C			0	9
D				0

The tree and, *a fortiori*, its edge lengths are not known.

Rearrangement Events

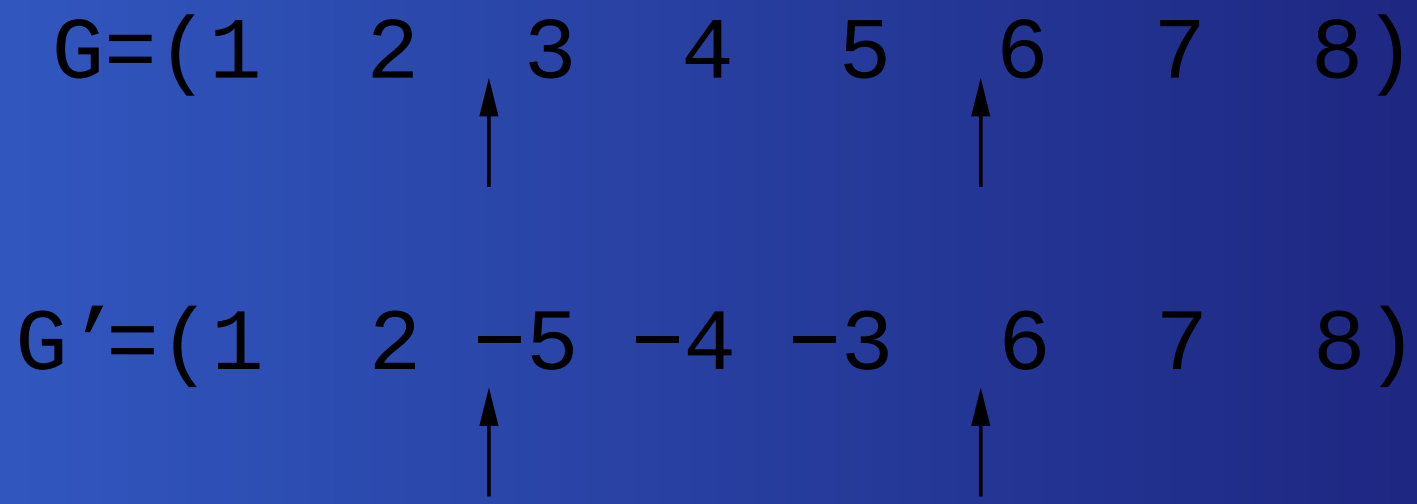


Generalized Nadeau-Taylor Model

- Inversions, Transpositions, and Inverted Transpositions
- All events of the same type are equiprobable
- Assign probabilities to different event types:
  - Transposition: α
  - Inverted Transposition: β
  - Inversion: $1 - \alpha - \beta$

Breakpoint Distance

$D_{BP}(G, G')$ = No. of breakpoints in G w.r.t G'



Genomic Distances

- BP: Breakpoint distance
- INV [Moret, Bader, Yan WADS 2001]:
  Minimum number of inversions required to transform one genome to another,
- IEBP [Wang, Warnow STOC'01]:
  Approximate the expected breakpoint distance with provable error.
- Exact IEBP [Wang WABI'01]:
  Invert the expected breakpoint distance
- EDE [Moret, Wang, Warnow, Wyman ISMB'01]:
  Estimate the expected inversion distance using simulation data.

Exact IEBP: Basic Idea

Let G_0 be the starting genome and G_k be the genome after k events.

- For every $k > 0$ compute $E[D_{BP}(G_k, G_0)]$, the expected number of breakpoints after k events.
- Return k that minimizes

$$|E[D_{BP}(G_k, G_0)] - D_{BP}(G, G')|.$$

The Counting Lemma

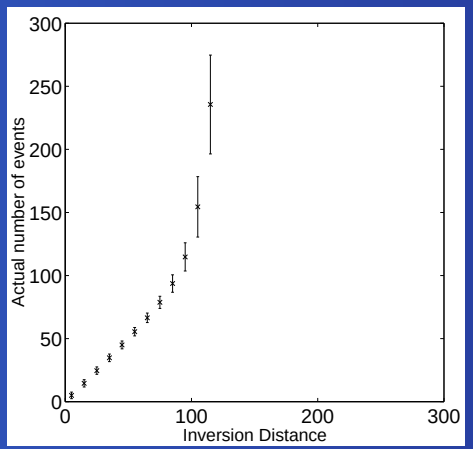
$$\iota_n(u,v) = \begin{cases} \min\{|u|-1, |v|-1, n+1-|u|, n+1-|v|\} & (\text{if } uv < 0) \\ 0 & (\text{if } u \neq v, uv > 0) \\ \binom{|u|-1}{2} + \binom{n+1-|u|}{2} & (\text{if } u = v) \end{cases}$$

$$\tau_n(u,v) = \begin{cases} 0 & (\text{if } uv < 0) \\ (\min\{|u|,|v|\}-1)(n+1-\max\{|u|,|v|\}) & (\text{if } u \neq v, uv > 0) \\ \binom{n+1-|u|}{3} + \binom{|u|-1}{3} & (\text{if } u = v) \end{cases}$$

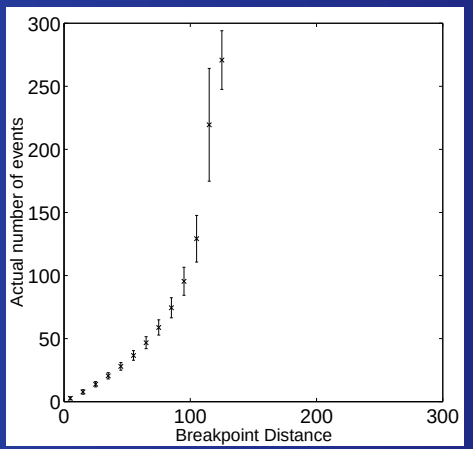
$$\nu_n(u,v) = \begin{cases} (n-2)\iota_n(u,v) & (\text{if } uv < 0) \\ \tau_n(u,v) & (\text{if } u \neq v, uv > 0) \\ 3\tau_n(u,v) & (\text{if } u = v) \end{cases}$$

Goodness of Fit of Distance Estimator

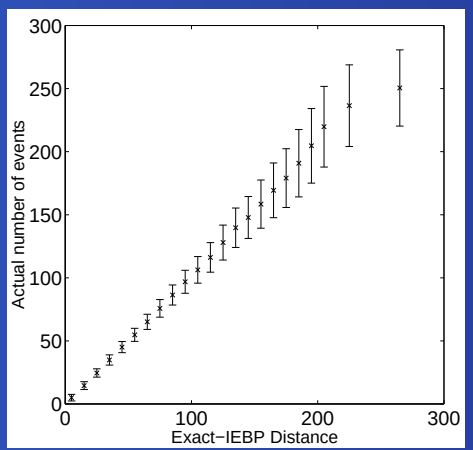
Inversion only on 120 genes



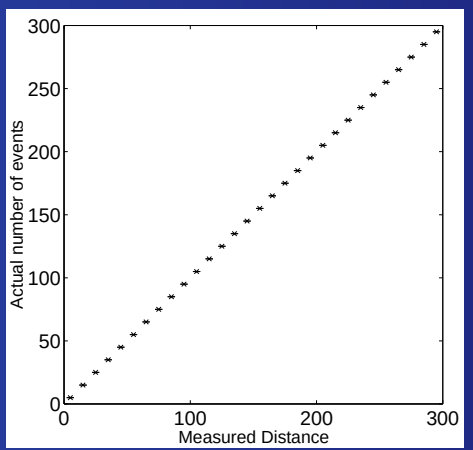
Inversion distance



Breakpoint distance



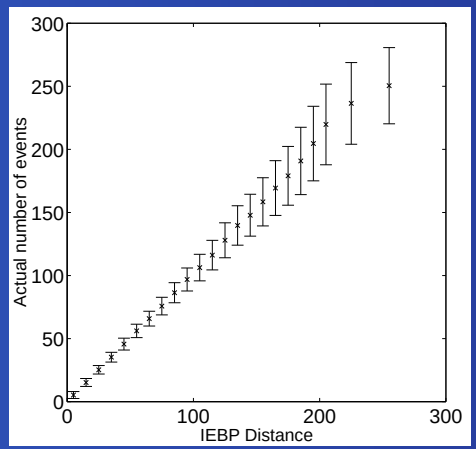
Exact-IEBP distance



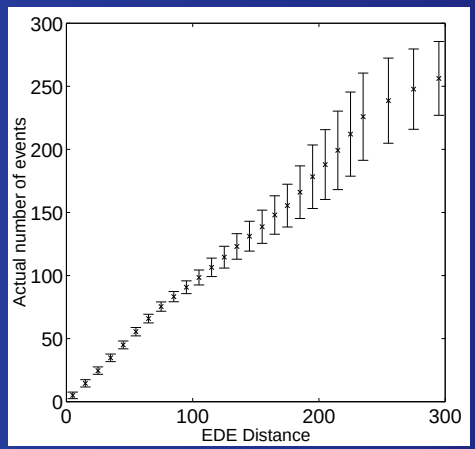
Ideal estimator

Goodness of Fit of Distance Estimator

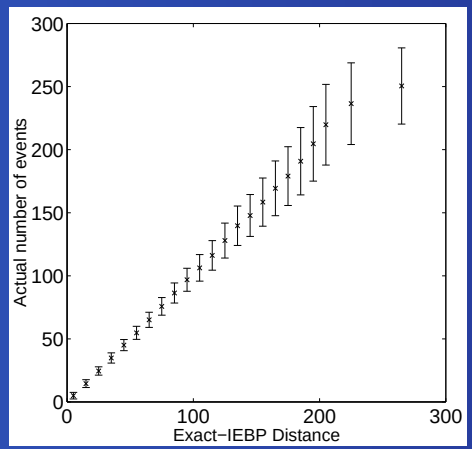
Inversion only on 120 genes



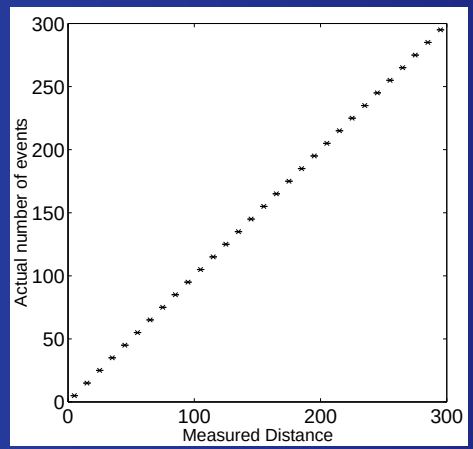
IEBP distance



EDE distance

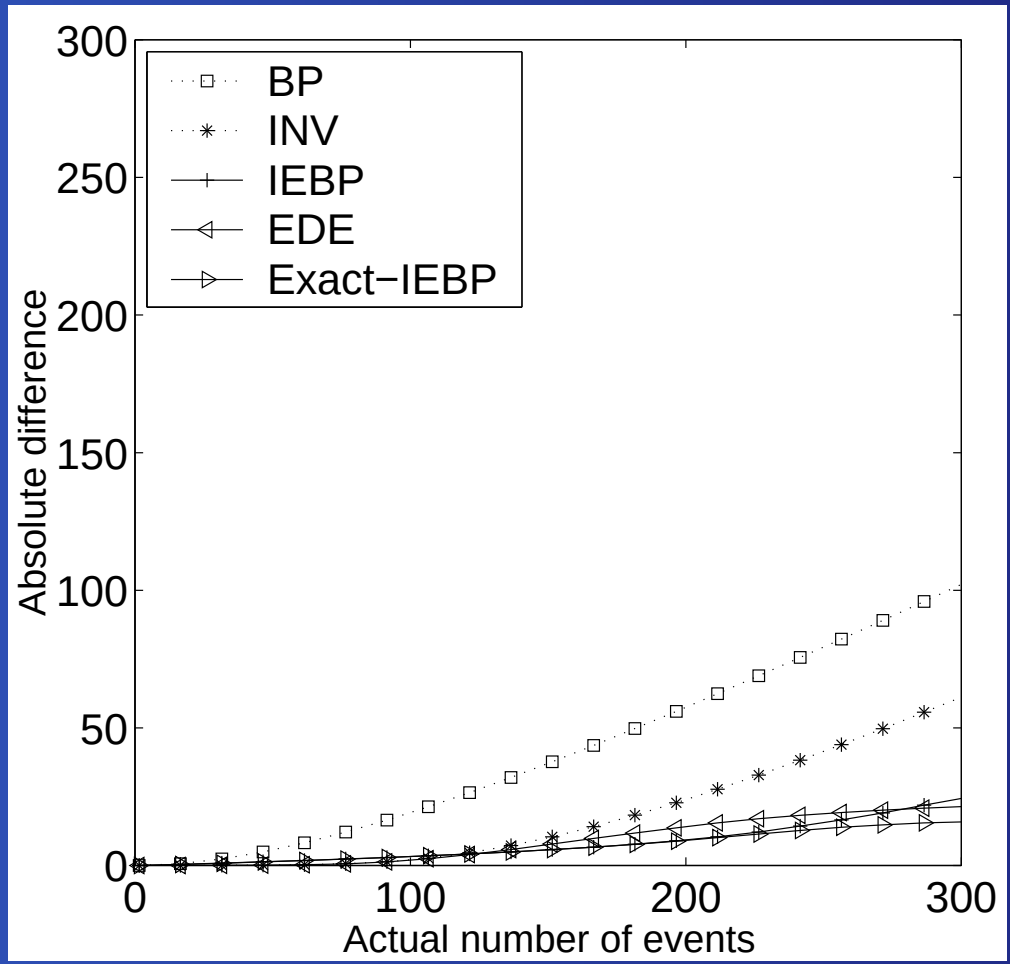


Exact-IEBP distance



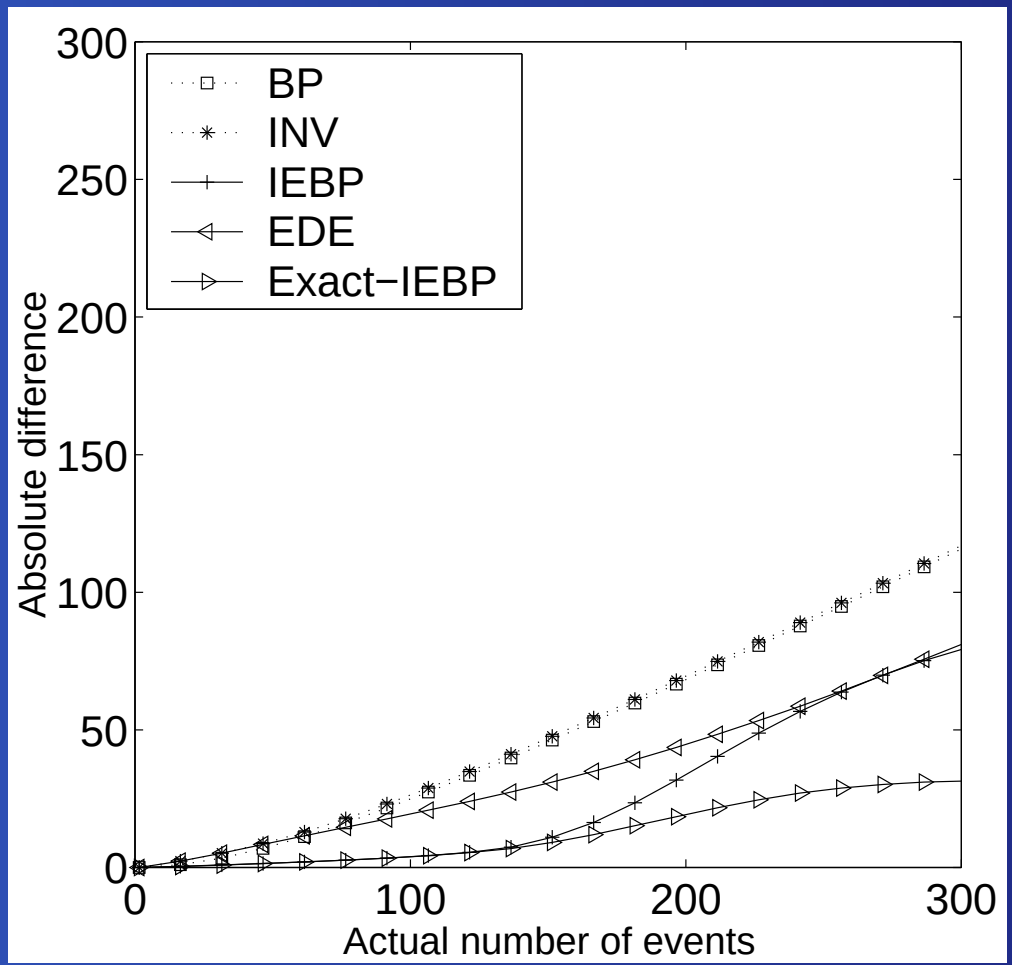
Ideal estimator

Absolute Error of Distance Estimators



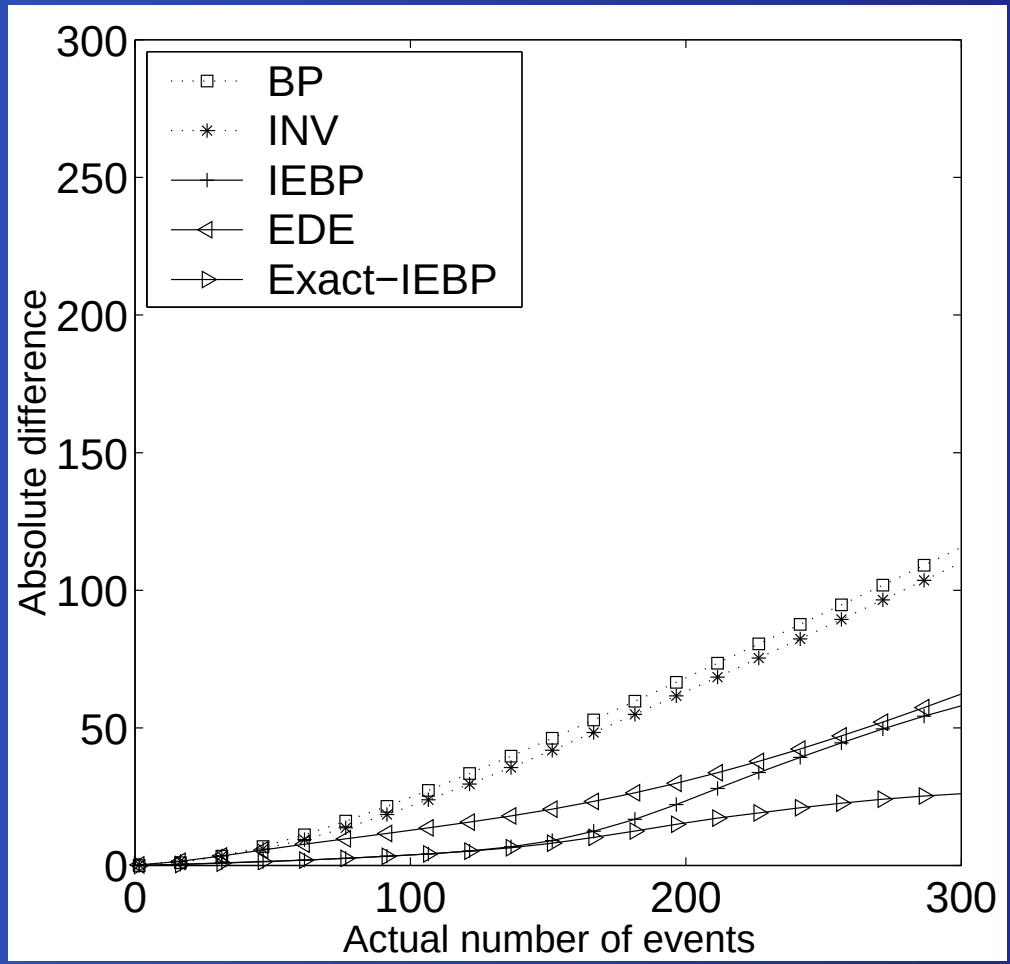
Inversion only

Absolute Error of Distance Estimators



Transpositions only

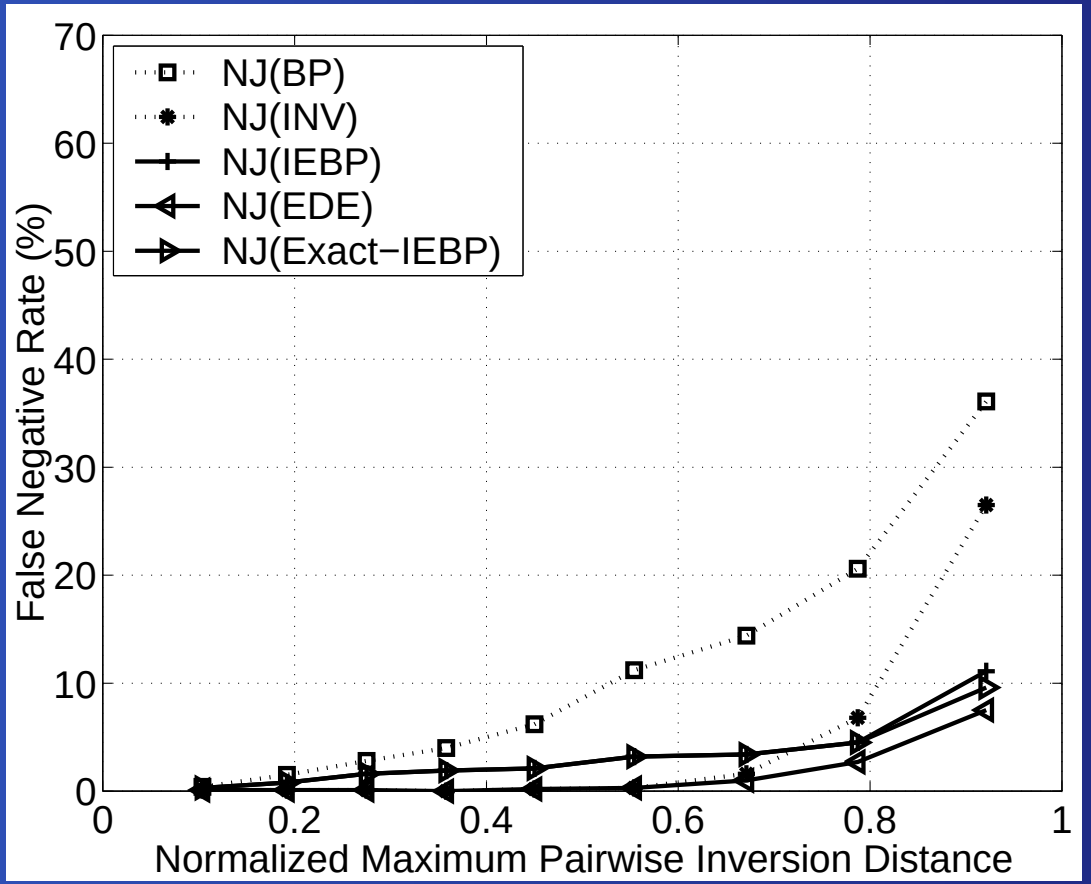
Absolute Error of Distance Estimators



All three classes equiprobable

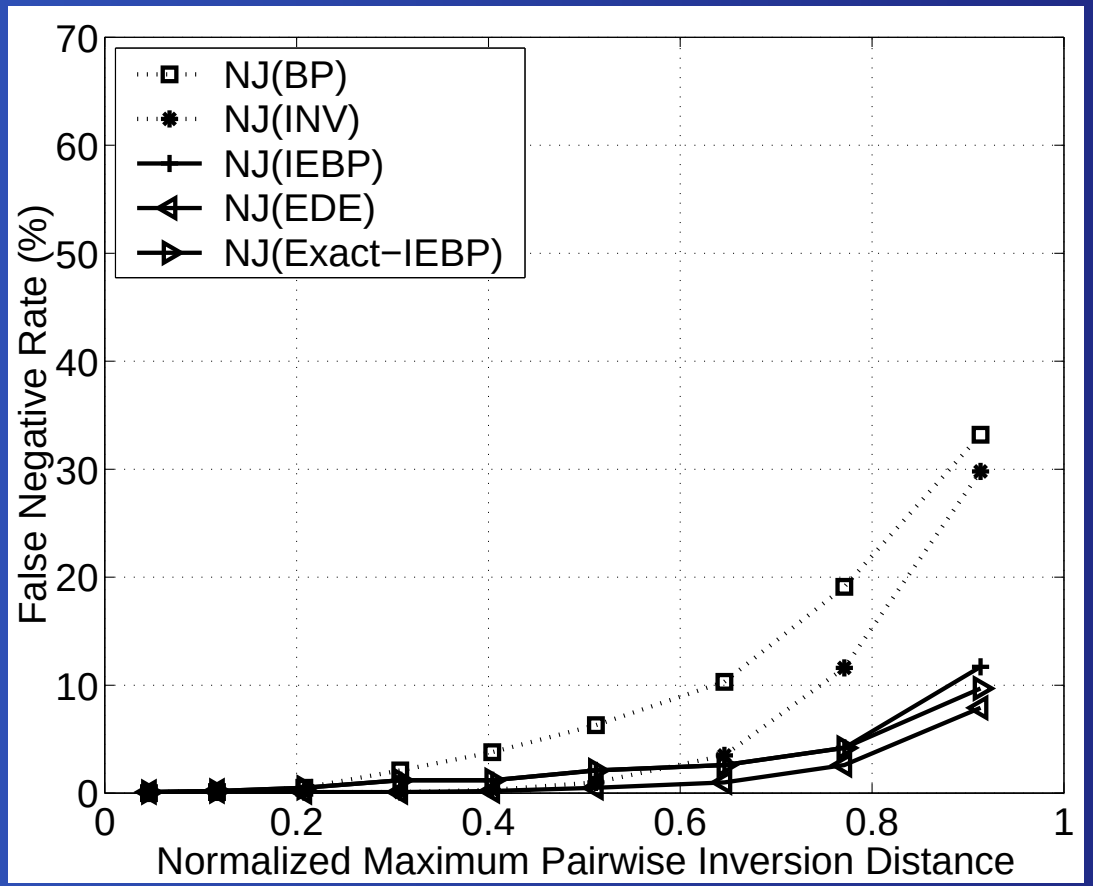
Accuracy of Neighbor Joining

120 genes, inversion only, 10/20/40/80/160 genomes



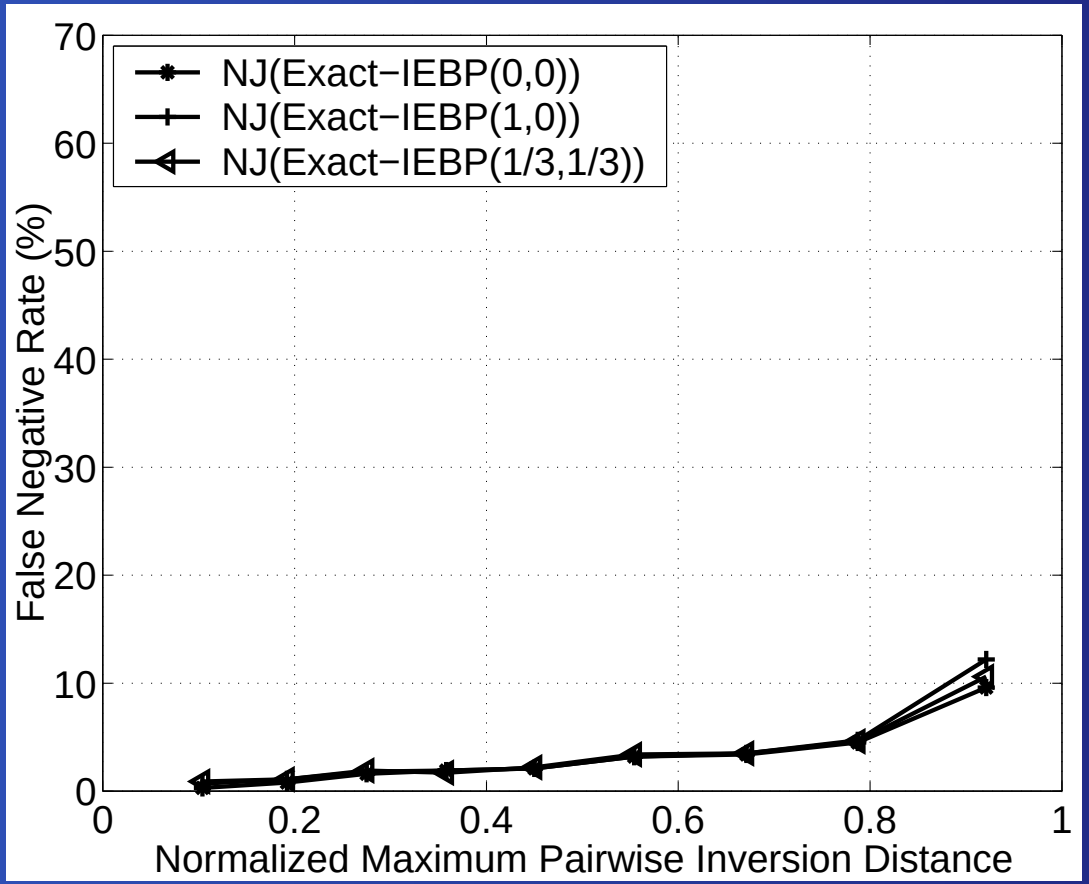
Accuracy of Neighbor Joining

120 genes, equiprobable events, 10/20/40/80/160 genomes



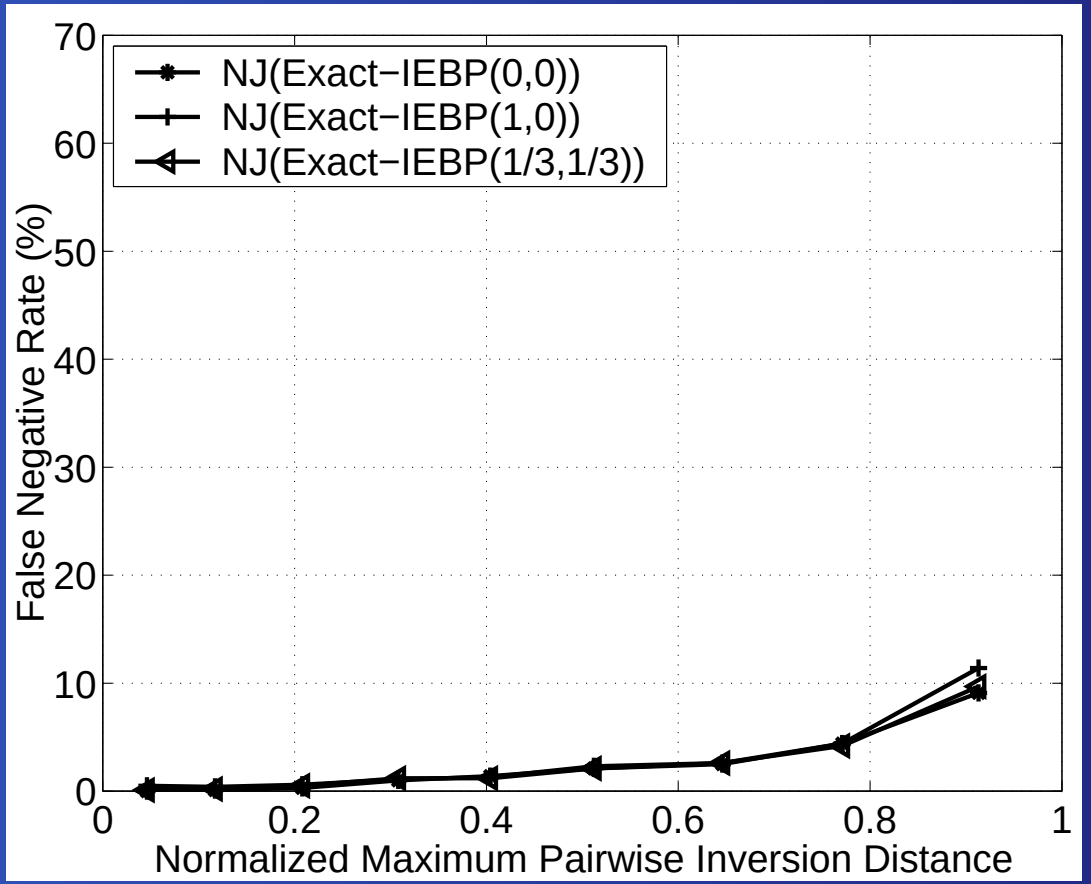
Robustness of Exact-IEBP

120 genes, inversion only, 10/20/40/80/160 genomes



Robustness of Exact-IEBP

120 genes, equiprobable events, 10/20/40/80/160 genomes



Convergence Rate

- A method is *statistically consistent* for a given model if, given long enough data sequences, it recovers the true tree with high probability.

Convergence Rate

- A method is *statistically consistent* for a given model if, given long enough data sequences, it recovers the true tree with high probability.
- Problem: “long enough" sequences may not exist in nature.

Convergence Rate

- A method is *statistically consistent* for a given model if, given long enough data sequences, it recovers the true tree with high probability.
- Problem: “long enough" sequences may not exist in nature.
- Solution: a method is *fast-converging* for a given model if, given sequences of polynomial length, it recovers the true tree with high probability.

Convergence Rate

- A method is *statistically consistent* for a given model if, given long enough data sequences, it recovers the true tree with high probability.
- Problem: “long enough" sequences may not exist in nature.
- Solution: a method is *fast-converging* for a given model if, given sequences of polynomial length, it recovers the true tree with high probability.
- Problem: the model conditions may not hold.

Convergence Rate

- A method is *statistically consistent* for a given model if, given long enough data sequences, it recovers the true tree with high probability.
- Problem: “long enough" sequences may not exist in nature.
- Solution: a method is *fast-converging* for a given model if, given sequences of polynomial length, it recovers the true tree with high probability.
- Problem: the model conditions may not hold.
- Solution: a method is *absolute fast-converging* if, given sequences of polynomial length, it recovers the true tree with high probability.

Known Fast-Converging Methods

- The short-quartet methods [Warnow *et al.*]: *absolute fast-converging*
- The disk-covering methods (DCM) [Warnow *et al.*]: *absolute fast-converging*
- The harmonic greedy triplet method [Kao *et al.*]
- The method of Cryan, Goldberg, and Golbderg
- DCM-boosted neighbor-joining [Warnow *et al.*]

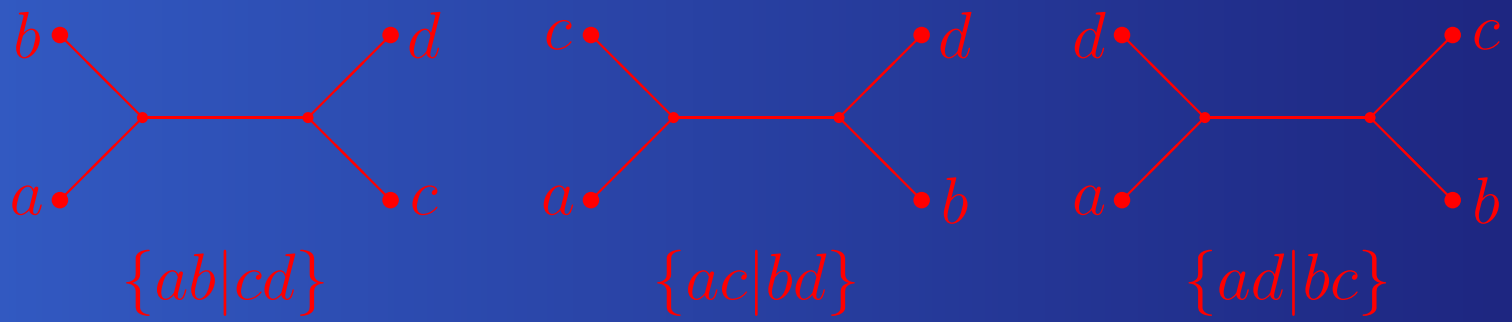
New Results

[Warnow, Moret, St. John SODA'01]

- New absolute fast-converging method: *weighted witness-antiwitness method (WIGWAM)*
- Decision procedure to turn fast-converging methods into absolute fast-converging methods: *short-quartet support (SQS)*
- Boosting method (DCM plus SQS) to turn many methods with exponential convergence (e.g., neighbor-joining) into absolute fast-converging ones
- Generalizations to families of boosting methods with same properties, but experimental behavior

What is a Quartet?

A *quartet* is an unrooted binary tree on four taxa—the smallest tree that induces a nontrivial bipartition.



A quartet $\{ab|cd\}$ *agrees* with a tree T if the subtree induced in T by the four taxa is the quartet itself.

Fast Convergence: Decision Problem

TRUE TREE SELECTION PROBLEM:

- **Input:**
  A set S of sequences over A, C, T, G generated on an unknown tree (T, M), and a collection $\mathcal{T} = \{T_1, T_2, \ldots, T_p\}$ of phylogenies on S.
- **Output:**
  The true tree T if T is in $\mathcal{T}$

Quartet Support

Let T be a fixed tree leaf-labelled by the set S
Let Q a fixed set of quartets on S
Let D be the distance matrix on S

The *support* of T with respect to Q is

$$\max\{l \mid (q \in Q \text{ and } diam_D(q) \leq l) \Longrightarrow q \in Q(T)\}$$

Short Quartet Support

PROCEDURE SQS($\mathcal{T}, S$)

- For each set of four taxa from S, compute the neighbor-joining quartet q; let $\mathcal{Q}$ be the set of all such quartets.
- Return T_i such that $s(T_i, \mathcal{Q})$ is maximum; if more than one such tree exists, return the one with the smallest index i.

SQS Theorem

For all $\varepsilon > 0$, there is a polynomial p such that, for all (T, M) in the model on set S of n sequences generated at random on T with length at least $p(n)$, we have

$$Pr[SQS(\mathcal{T}, S) = T] > 1 - \varepsilon$$

whenever T is in $\mathcal{T}$.

GRAPPA News: More Speed!

- Current release (1.03) runs from 2,000 to 10,000 times faster than the original tool, while also giving more capabilities.

GRAPPA News: More Speed!

- Current release (1.03) runs from 2,000 to 10,000 times faster than the original tool, while also giving more capabilities.
- Research version (1.1) runs from 10,000 to 500,000 times faster than the original tool, thanks to much better bounding.

GRAPPA News: More Speed!

- Current release (1.03) runs from 2,000 to 10,000 times faster than the original tool, while also giving more capabilities.
- Research version (1.1) runs from 10,000 to 500,000 times faster than the original tool, thanks to much better bounding.
- The 13-genome *Campanulaceae* now takes a few hours on a laptop instead of a few centuries on a large workstation.

GRAPPA News: More Speed!

- Current release (1.03) runs from 2,000 to 10,000 times faster than the original tool, while also giving more capabilities.
- Research version (1.1) runs from 10,000 to 500,000 times faster than the original tool, thanks to much better bounding.
- The 13-genome *Campanulaceae* now takes a few hours on a laptop instead of a few centuries on a large workstation.
- Speedup on Los Lobos is over 200,000,000!

Other Recent Results

- New sequence encodings for gene orders to enable classical parsimony searches.

Other Recent Results

- New sequence encodings for gene orders to enable classical parsimony searches.
- Combinations of fast-converging boosters with new encodings (i.e., use a new encoding and run a DCM+SQS booster on a classical parsimony optimizer): best accuracy to date.

Other Recent Results

- New sequence encodings for gene orders to enable classical parsimony searches.
- Combinations of fast-converging boosters with new encodings (i.e., use a new encoding and run a DCM+SQS booster on a classical parsimony optimizer): best accuracy to date.
- Combinations of fast-converging boosters with new encodings and fast heuristics (e.g., neighbor-joining): best speed/accuracy tradeoff to date.

Other Recent Results

- New sequence encodings for gene orders to enable classical parsimony searches.
- Combinations of fast-converging boosters with new encodings (i.e., use a new encoding and run a DCM+SQS booster on a classical parsimony optimizer): best accuracy to date.
- Combinations of fast-converging boosters with new encodings and fast heuristics (e.g., neighbor-joining): best speed/accuracy tradeoff to date.
- New results on computing inversion distances, inversion medians, etc.