Phylogenetic Reconstruction from Gene-Order Data: A Progress Report

Bernard M.E. Moret

compbio.unm.edu

Department of Computer Science
University of New Mexico

Acknowledgments

- *Main collaborators:*
  D. Bader (UNM), T. Warnow, R. Jansen, and C. Linder (UT Austin), N. Moran and H. Ochman (Arizona), C. DePamphilis (Penn. State), A. Caprara (Bologna)
- *Lab:*
  Work on large-scale phylogenetic reconstruction, reconstruction of reticulate evolution, and reconstruction from gene order and content.

  T. Berger-Wolf (now U. Illinois), T. Williams (now Texas A&M), J. Tang (now U. South Carolina), Z. Betz, J. Earnest-DeYoung, T. Liu, M. Marron, M. Morin, K. Swenson, A. Tholse, L. Zhao
- *Support:*

  NSF, NIH, Sloan Foundation, IBM Corporation

Overview

- Gene-Order Data vs. Sequence Data
- Phylogenetic Reconstruction:
  a fast review from a CS standpoint
- Computing with Gene-Order Data
- Reconstruction from Gene-Order Data
- Experimentation in Phylogeny
- Some Open Problems

Cultural Disclaimer...

- Oddities: CS folks do not show photos, draw their trees upside down, and misuse sensitive vocabulary.
- Literature: CS folks publish in conferences, rarely in journals.
  ISMB, RECOMB, WABI, CPM, CSB, PSB, and BIBE
- Data are: strings of symbols (sequences), ordered lists of symbols (gene orders for whole genomes), and operations on those (evolutionary models).
- Methodology: algorithms will operate on a large variety of instances, so must be tested through range of parameters: hence *tens of thousands* of test sets.
- Corollary (to methodology): CS folks show plots, rarely trees.

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 are poorly solved.
  - Gene evolution need not be identical to organism evolution. Recombination hampers lineage sorting.

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 (seen in mitochondria) and translocations (between chromosomes).

- Need to identify gene families.
- Need to refine rearrangement model for specific collections of organisms (e.g., to handle operons, exons, etc.).

Gene-Order Data: Attributes

- **Advantages:**
  - Rare genomic events (Rokas/Holland) and huge state space, so risk of homoplasy is very low.
  - No need for alignnments.
  - 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

Phylogenetic Reconstruction

Two categories of methods:

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

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)

Evolutionary 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 guess for the true evolutionary distance, often obtained by “correcting” the edit distance according to a model of evolution, but also derived heuristically.

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; BioNJ and Weighbor are two improvements for sequence data.
- 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 under the infinite sites assumption.

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 yield good results in most cases.
- Finding most parsimonious tree is computationally expensive (NP-hard).
- Optimal solutions limited to sizes around 30.
  Heuristic solutions fairly good to sizes of 500.

Likelihood-Based Methods

- 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.
- Are prone to numerical problems (likelihoods are extremely small numbers for almost all trees).
- Are presumably NP-hard; even scoring one tree is very expensive.
- Optimal solutions limited to specific sets of 4 taxa; heuristic search can be run to completion on at most 10 taxa, but appear fairly good to about 100 taxa (PhyML, TrExML).

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: 4 taxa); include Q*, Tree-Puzzle, Quartet-Cleaning.
  Slow and inherently *inaccurate* regardless of base method—consistently surpassed by NJ.
- Disk-Covering methods (DCMs): decompose the dataset into overlapping “disks" (tight subsets).
  High-powered machinery *succeeds*, especially when tree is imbalanced.

  Enables scaling up sequence-based MP analyses to tens of thousands of taxa.

Computing with Gene-Order Data

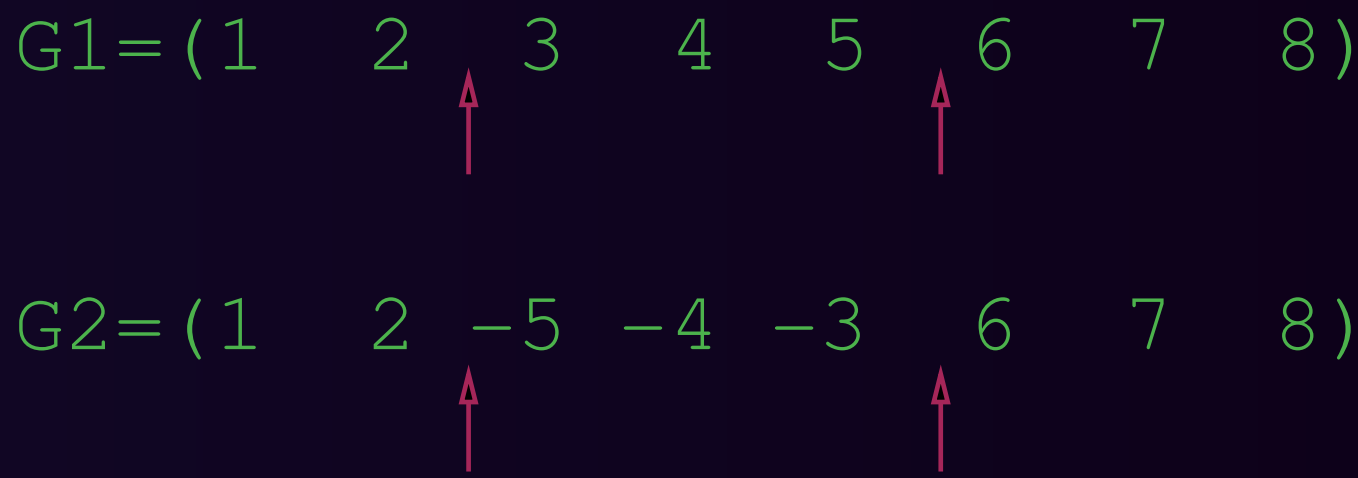
- Distances
- Evolutionary models and distance corrections
- Reconstructing ancestral genomes
- The median problem

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.



Inversion Distance

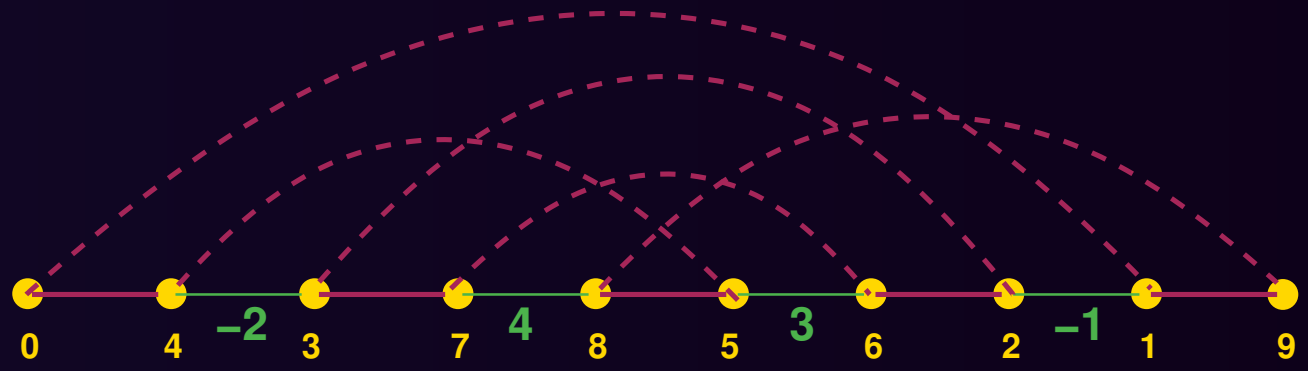
Given two signed gene orders of equal content, compute the inversion-only edit distance.

- The problem is NP-hard for *unsigned* permutations.
- Finding the shortest sequence of inversions can be done in polynomial time (Hannenhalli and Pevzner—very elaborate result).
  Current best algorithm (Tannier and Sagot 2004) runs in $O(n\sqrt{n \log n})$ time for n genes.
- Distance computation runs in $O(n)$ time (Bader/Moret/Yan 2001).

Inversion Distance

Algorithm is based on the breakpoint graph.
Assume one permutation is identity; each green edge is a single gene.

Solid red edges denote existing adjacencies and *dashed red* edges denote desired adjacencies.



Inversion distance is

n - #cycles + #hurdles + (fortress)

Gene-Order Distances in General

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

- Extension to translocations (transpositions between chromosomes) by Hannenhalli and Pevzner. Implemented by Tesler as GRIMM.
- Heuristic for duplications by Sankoff (exemplars, an NP-hard problem).
- Extension to inversions and deletions (no duplications allowed) by El-Mabrouk; linear-time distance computation by our group.
- Heuristic for unequal gene content by Bourque.
- Bounded approximation for unequal gene content and direct estimate of evolutionary distance by our group.

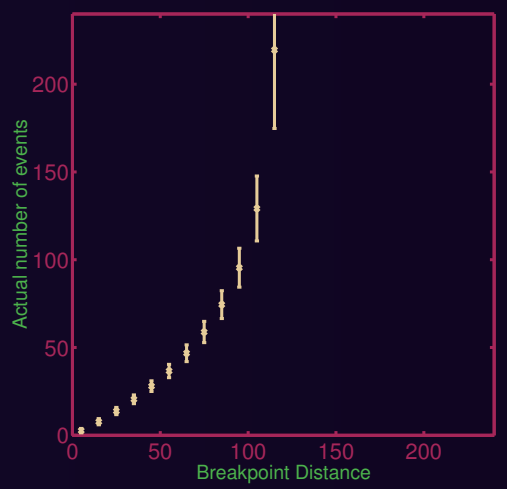
Evolutionary Models for Gene Order

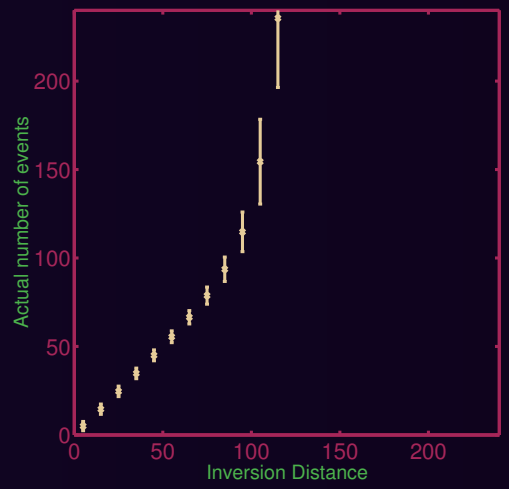
- Biological evidence for inversions in chloroplasts, transpositions in mitochondria, and translocations in eukaryotic genomes.
- Computational evidence (Sankoff 2001, Lefebvre et al. 2003) for short inversions in bacterial genomes.
- Hard radiation plus repair mechanisms could create more complex rearrangements.
- Respective probabilities unknown.

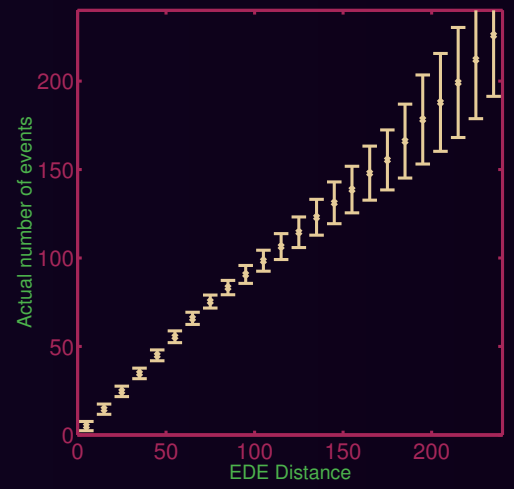
Distance Corrections for Gene Order

- Assume a distribution of events and compute relationship between number of events and edit distance.
- Wang/Warnow (2001) gave an exact derivation for correcting the breakpoint distance into an expected number of inversions (IEBP).
- Moret et al. (2002) gave an empirical derivation for correcting the inversion edit distance into an expected number of inversions (EDE).
- EDE correction substantially improves the performance of both distance-based and parsimony-based reconstruction methods.
- Direct estimate of Swenson et al. (2004) highly accurate even on large genomes with large distances.

EDE Distance Correction







breakpoint distance

inversion distance

EDE corrected distance

Vertical axis: actual number of inversions.
Horizontal axis: distance measure.

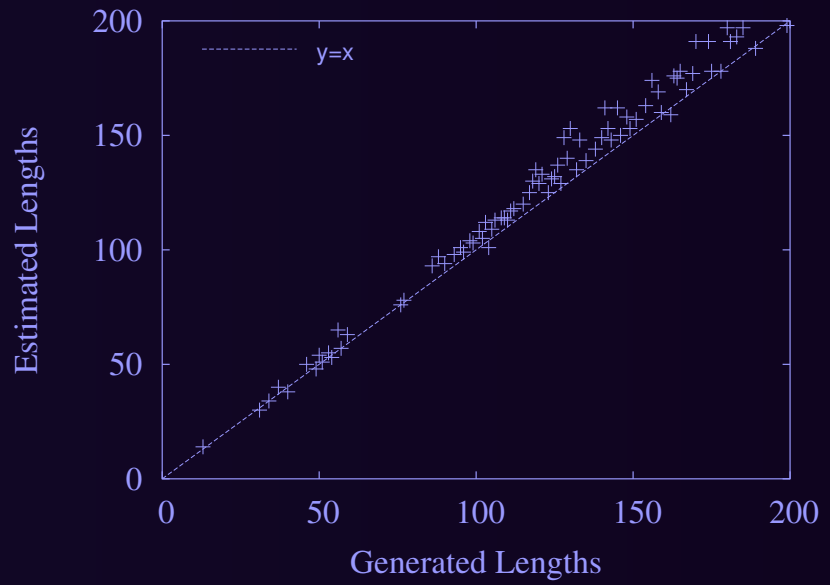
Direct Estimate of Distance

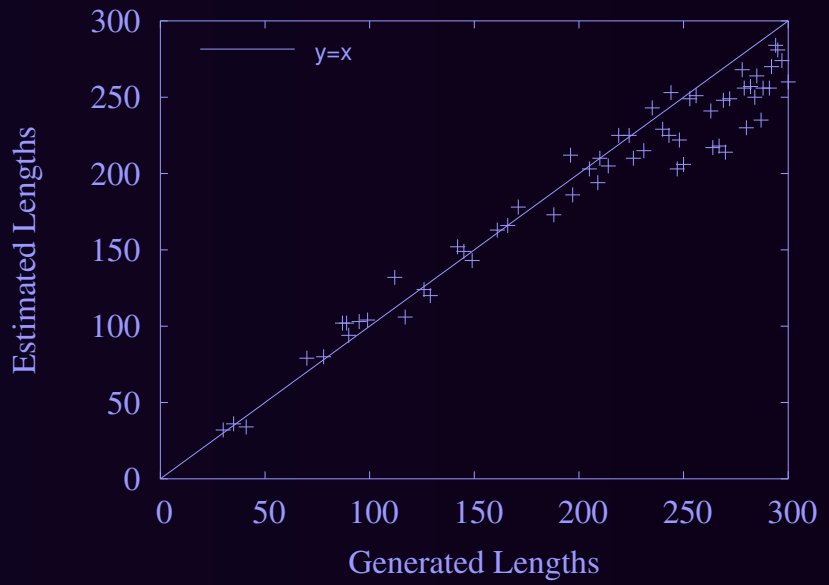
- Accounts for insertions, duplications, deletions, and inversions.
- Matches elements of gene families using a version of optimal covering; treats unmatched elements as insertions/deletions.
- Uses result of Marron et al.: *there always exists a shortest sequence that first does all insertions, then all inversions, and finally all deletions*.
- Uses result of El-Mabrouk: *exact algorithm for inversions plus 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





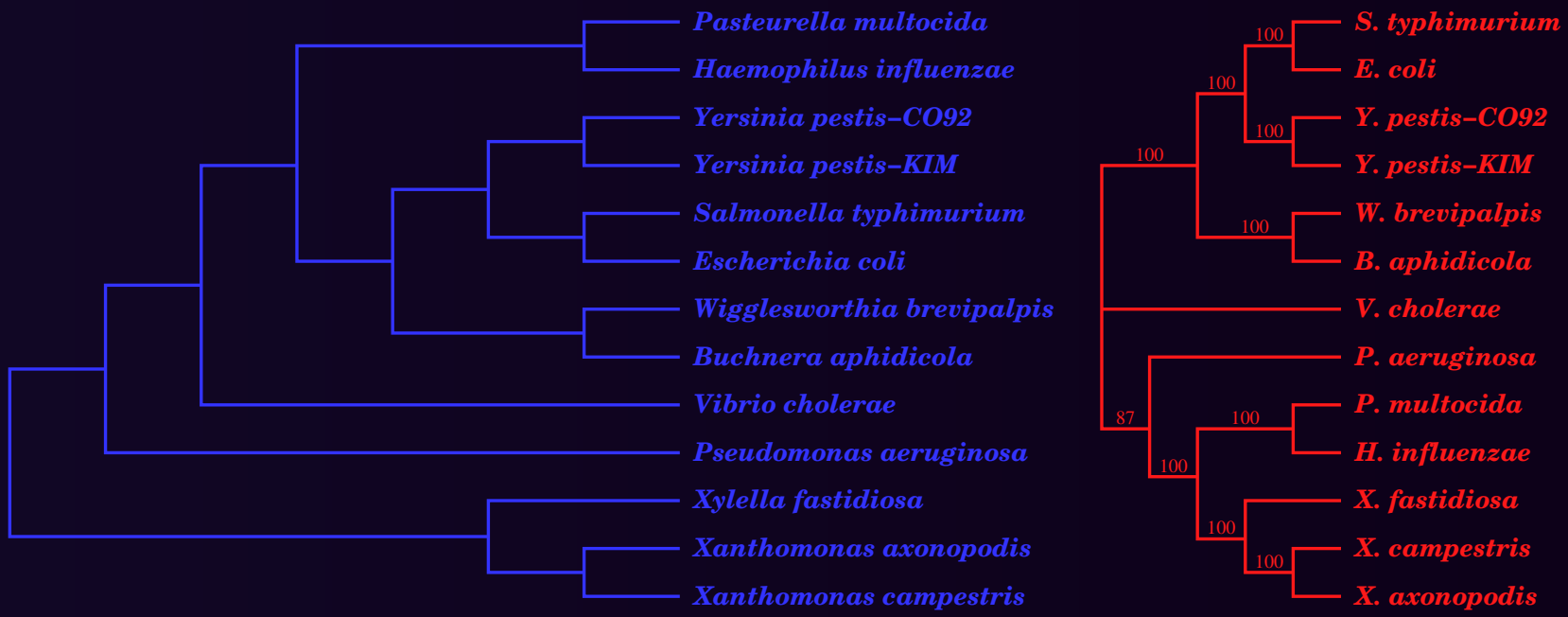
Using the Swenson et al. Estimate

(unpublished)

13 gamma proteobacteria (Lerat/Daubin/Moran 2003)

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).

Reconstructing Ancestral Genomes

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

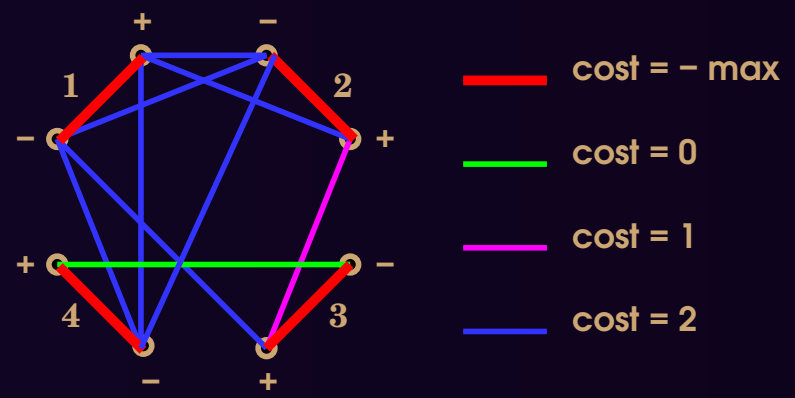
Problem is NP-hard even for just three leaves, no duplications, and simplest of distances (breakpoint, plain inversion)!

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.

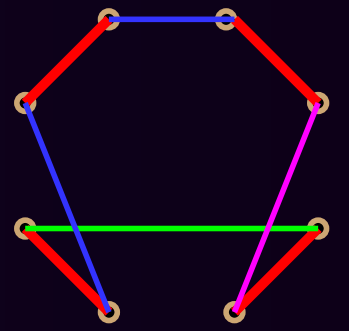
Median Problem for Breakpoints

Sankoff showed to to convert MPB to the Travelling Salesperson Problem when all three have identical gene content.

+1 −2 +4 +3
+1 +2 −3 −4
+2 −3 −4 −1



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 preferable to the breakpoint median (fewer ties, better trees).

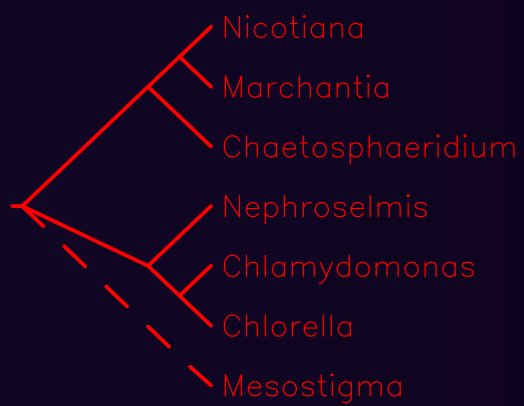
Median with Deletions/Duplications

Can be solved exactly for small numbers of deletions and duplications (Tang/Moret 2003).

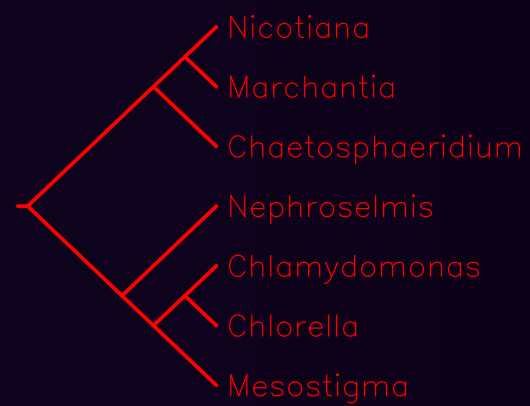
- Assume no change is reversed and that changes are independent and of low probability.
- Gene content of median can then be determined by preferring one event (e.g., one insertion) over two concurrent events (e.g., two deletions).
- Knowing gene content, all combinations of duplications or insertion locations can be examined—choice grows exponentially, but is manageable for organellar genomes.
- Accuracy is good (less than 5% error).

Small Example

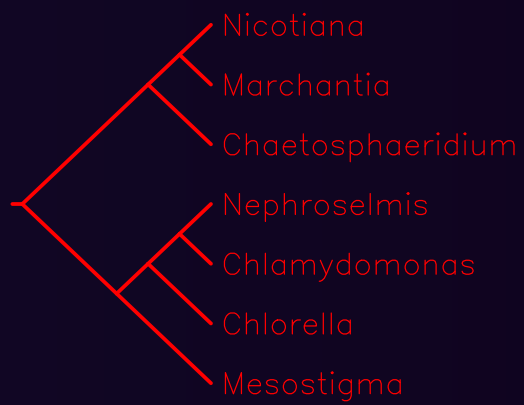
Tang/Moret/Cui/DePamphilis (2004): chloroplast data



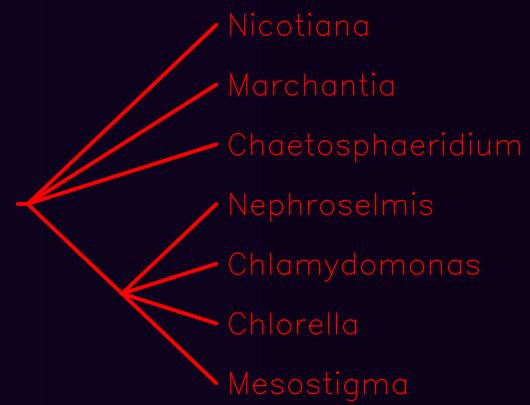
organismal



Tang/Moret GRAPPA



NJ (inv.)



breakpoint GRAPPA

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**

Distance Methods

Neighbor-joining and variants using BP, INV, IEBP, and EDE distances.

We showed that Weighbor-EDE is clearly better than other combinations:

- Low error rates up to a few hundred genomes.
- Robust against various models of genome rearrangements.
- Suffers for large tree diameters.

Improved by using DCM boosting, but surpassed by direct optimization methods (`GRAPPA`).

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
Parsimony & 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 . . .

MGR

Multichromosomal Genome Rearrangement (Bourque and Pevzner)

- Uses `GRAPPA` code for distance computations.
- Uses the median approach, but does not solve it exactly, so less accurate than `GRAPPA`.
- Can handle multiple chromosomes (translocations) using Hannenhalli-Pevzner's results.
- Scaling unknown—published results to date limited to very small datasets.

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%.

Tight Bounds

(unpublished)

We use selected triangle inequalities on a tree as linear programming constraints to compute a lower bound on the tree score.

- *With good selection, bound is extremely tight (99% or better).*
- *Avoids scoring tree with* `GRAPPA` *procedure—so no median computation and so very fast.*
- *Allows* `GRAPPA` *to handle much larger genomes.*
- *Does not yet help enough for reconstructing ancestral genomes.*

Likelihood Approaches

Only one so far, a Bayesian approach due to Larget et al.

- MCMC method with four moves (one redundant to help convergence).
- Untested (just two small datasets used in report).
- Promising results, unknown running time.
- Code not available.

Testing Algorithms

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.

Phylogenetic Considerations

- Tree shape plays a *very large role*.
  Challenge: the shape of the true trees is unknown; moreover, the shape depends on the selection of genomes—e.g., a single genus vs. a sampling of an entire kingdom.
- The evolutionary model is important.
  Challenge: devise an evolutionary model with few parameters that is easily manipulated analytically and computationally and that produces realistic data. (It would also be pleasant if it made biological sense. . .)
- Test a large range of parameters and use many runs for each setting to estimate variance.
  Challenge: even the simplest of models induces a huge parameter space—sampling it requires smoothness guarantees, but NP-hard search spaces lack smoothness.

Tree Topologies: Popular Models

Tree Topologies: Popular Models

- Birth-Death: start with root branch; in any Δt interval, there is probability p of branching along any of the current branches.
  Biologically motivated; but trees are well balanced—too well balanced compared to published phylogenies.

  All distance-based methods do very well on BD trees; DCM decompositions are mediocre, but not needed.

Tree Topologies: Popular Models

- Birth-Death: start with root branch; in any Δt interval, there is probability p of branching along any of the current branches.
  Biologically motivated; but trees are well balanced—too well balanced compared to published phylogenies.

  All distance-based methods do very well on BD trees; DCM decompositions are mediocre, but not needed.

- Uniform Random: all tree topologies equally likely.
  No biological process; trees are imbalanced— too imbalanced compared to published phylogenies.

  NJ does poorly on uniform trees; DCM helps a lot.

Tree Topologies: Newer Models

Tree Topologies: Newer Models

- Aldous' β-Splitting: parameter can be set to produce anything from a ladder (caterpillar), through uniform, birth-death, to perfect balance.
  No biological process; single parameter cannot localize structure.

  Recommended setting ($\beta = -1$) matches balance of published phylogenies, but algorithmic behavior does not match biological datasets.

Tree Topologies: Newer Models

- Aldous' β-Splitting: parameter can be set to produce anything from a ladder (caterpillar), through uniform, birth-death, to perfect balance.
  No biological process; single parameter cannot localize structure.

  Recommended setting ($\beta = -1$) matches balance of published phylogenies, but algorithmic behavior does not match biological datasets.

- S. Heard's: variable and inheritable evolutionary rates, inheritable branching traits, punctuated and gradual evolution.
  Strong biological motivation; too many parameters?

  Subtle differences cause large changes in performance of distance- and parsimony-based methods.

Some Open Problems

- Tree models (Heard's? better characterization?)
- 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

- Gene-order data carries a very strong phylogenetic signal.
- Current algorithmic approaches scale to significant sizes (1000s for DCM-GRAPPA)—comparable to the best achievable with sequence data and often with better results.
- Current approaches can just begin to handle significantly unequal gene content with duplications.
- Data availability is increasing rapidly for organellar genomes, slowly for nuclear genomes, and remains very limited compared to sequence data. (When will we get the $1,000.- genome?)

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.

With thanks to David Sankoff for creating this playground!