Introduction to Algorithmic Phylogenomics (Warnow Lab)

Tandy Warnow
University of Illinois at Urbana-Champaign
Phylogeny (evolutionary tree)

From the Tree of the Life Website, University of Arizona
Phylogeny + genomics = genome-scale phylogeney estimation.
Estimating the Tree of Life

Basic Biology:
How did life evolve?

Applications of phylogenies to:
protein structure and function
population genetics
human migrations
metagenomics

Figure from
https://en.wikipedia.org/wiki/Common_descent
Estimating the Tree of Life

Large datasets!
- Millions of species
- Thousands of genes

NP-hard optimization problems
- Exact solutions infeasible
- Approximation algorithms
- Heuristics
- Multiple optima

High Performance Computing:
- Necessary
- But not sufficient

Figure from https://en.wikipedia.org/wiki/Common_descent
Species sequenced by year

- Viruses
- Eukaryotes
- Prokaryotes

Muir, 2016
Computer Science Solving Problems in Biology and Linguistics

• Algorithm design using
  – Divide-and-conquer
  – Iteration
  – Heuristic search
  – Graph theory
• Algorithm analysis using
  – Probability Theory
  – Graph Theory
• Simulations and modelling
• Collaborations with biologists and linguists and data analysis
• Discoveries about how life evolved on earth (and how languages evolved, too)
Computational Phylogenetics (2005)

Current methods can use months to estimate trees on 1000 DNA sequences

Our objective:
More accurate trees and alignments on 500,000 sequences in under a week

Courtesy of the Tree of Life web project, tolweb.org
Computational Phylogenetics (2018)

1997-2001: Distance-based phylogenetic tree estimation from polynomial length sequences

2012: Computing accurate trees (almost) without multiple sequence alignments

2009-2015: Co-estimation of multiple sequence alignments and gene trees, now on 1,000,000 sequences in under two weeks

2014-2015: Species tree estimation from whole genomes in the presence of massive gene tree heterogeneity

2016-2017: Scaling methods to very large heterogeneous datasets using novel machine learning and supertree methods.
Divide-and-Conquer in Phylogenetics

All these algorithms are by CS PhD students working with me, and are the best-in-class:

• DCM1 (2001): Improving distance-based phylogeny estimation (and absolute fast convergence!)
• SATé (2009, 2012) and PASTA (2015): Co-estimation of multiple sequence alignments and gene trees (up to 1,000,000 sequences)
• DACTAL (2012): Almost alignment-free tree estimation
• NJMerge (2018) and INC-ML: divide-and-conquer without supertrees
Divide-and-Conquer in Phylogenetics

All these algorithms are by CS PhD students working with me, and are the *best-in-class*:

• **DCM1 (2001):** Improving distance-based phylogeny estimation (and absolute fast convergence!)

• **SATé (2009, 2012) and PASTA (2015):** Co-estimation of multiple sequence alignments and gene trees (up to 1,000,000 sequences)

• **DACTAL (2012):** Almost alignment-free tree estimation

• **NJMerge (2018) and INC-ML:** divide-and-conquer without supertrees
DNA Sequence Evolution

-3 mil yrs
-2 mil yrs
-1 mil yrs
today
Gene Tree Estimation

AGGGGCA
TAGCCCC
TAGACTT
TGCACAA
TGCGCTTT
Markov Model of Site Evolution

Simplest (Jukes-Cantor, 1969):

• The model tree T is binary and has substitution probabilities $p(e)$ on each edge $e$.
• The state at the root is randomly drawn from \{A,C,T,G\} (nucleotides)
• If a site (position) changes on an edge, *it changes with equal probability to each of the remaining states.*
• The evolutionary process is Markovian.

The different sites are assumed to evolve independently and identically down the tree (with rates that are drawn from a gamma distribution).

More complex models (such as the General Markov model) are also considered, often with little change to the theory.
Distance-based estimation

TRUE TREE

DNA SEQUENCES

S₁ ACAATTAGAAC
S₂ ACCCTTAGAAC
S₃ ACCATTCCAAC
S₄ ACCAGACCAAC

DISTANCE MATRIX

<table>
<thead>
<tr>
<th></th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>S₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>S₂</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>S₃</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>S₄</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

METHODS SUCH AS NEIGHBOR JOINING

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES
Quantifying Error

TRUE TREE

FN: false negative (missing edge)
FP: false positive (incorrect edge)

50% error rate

DNA SEQUENCES

INFERRED TREE

S1  ACAATTAGAAC
S2  ACCCTTAGAAC
S3  ACCATTCGAAC
S4  ACCAGACCAAC
S5  ACCAGACCGGA

S1  S4  S3
S2  S5

FP
Sequence length requirements

The sequence length (number of sites) that a phylogeny reconstruction method \( M \) needs to reconstruct the true tree with probability at least 1-\( \varepsilon \) depends on

- \( M \) (the method)
- \( \varepsilon \)
- \( f = \min p(e) \),
- \( g = \max p(e) \), and
- \( n \), the number of leaves

We fix everything but \( n \).
Statistical consistency, exponential convergence, and absolute fast convergence (afc)
Neighbor joining on large diameter trees

Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

DCMs: Divide-and-conquer for improving phylogeny reconstruction
DCM1 Decompositions

**Input:** Set $S$ of sequences, distance matrix $d$, threshold value $q \in \{d_{ij}\}$

1. Compute threshold graph
   \[ G_q = (V, E), V = S, E = \{(i, j) : d(i, j) \leq q\} \]

2. Perform minimum weight triangulation (note: if $d$ is an additive matrix, then the threshold graph is provably **triangulated**).

DCM1 decomposition:

- Compute maximal cliques
DCM1-boosting:
Warnow, St. John, and Moret,
SODA 2001

- The DCM1 phase produces a collection of trees (one for each threshold), and the SQS phase picks the “best” tree.
- For a given threshold, the base method is used to construct trees on small subsets (defined by the threshold) of the taxa. These small trees are then combined into a tree on the full set of taxa.
Neighbor joining on large diameter trees

Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

DCM1-boosting

Nakhleh et al., Bioinformatics 2001:
- DCM1-NJ has much better accuracy than NJ!

DCM1-boosting

Theorem (Warnow, St. John, and Moret, SODA 2001):
- DCM1-NJ converges to the true tree from polynomial length sequences

DCM1-boosting

DCM1-NJ:
- Graph theory
- Probability theory
- Implementation and evaluation on simulated datasets

“Boosters”, or “Meta-Methods”

- Meta-methods use divide-and-conquer and iteration (or other techniques) to “boost” the performance of base methods (phylogeny reconstruction, alignment estimation, etc)
Current projects in the lab

Research is a combination of
• theory (probability theory and graph theory)
• statistical inference and machine learning
• algorithm design and implementation
• evaluation on simulated and real datasets
• high performance computing

All good problems are NP-hard, so good heuristics are needed
Proofs of statistical consistency under stochastic models of evolution

Applications to real problems in biological sequence analysis, and collaboration with biologists around the world, for:
• Multiple Sequence Alignment
• Supertree estimation for divide-and-conquer
• Species tree and phylogenetic network estimation
• Microbiome analysis
Estimating the Tree of Life

Large datasets!
Millions of species
thousands of genes

NP-hard optimization problems
Exact solutions infeasible
Approximation algorithms
Heuristics
Multiple optima

High Performance Computing:
necessary
but not sufficient

Phylogenetic Tree of Life

Figure from https://en.wikipedia.org/wiki/Common_descent
Other Information

• Tandy’s office hours: Tuesdays 2-3 PM (and by appointment).
• Tandy’s email: warnow@illinois.edu
• Webpage: http://www.tandy.cs.illinois.edu
• Textbook: