CS 581
Algorithmic Computational Genomics

Tandy Warnow
University of Illinois at Urbana-Champaign
Course Staff

• Professor Tandy Warnow
  – Office hours Tuesdays after class (2-3 PM) in Siebel 3235
  – Email address: warnow@Illinois.edu

• T.A. Wei (Wesley) Qian
  – Email address: weiqian3
  – Office hours to be determined (stay tuned)
Today

• Explain the course
• Introduce some of the research in this area
• Describe some open problems
• Talk about course projects!
CS 581

• Algorithm design and analysis (largely in the context of statistical models) for
  – Multiple sequence alignment
  – Phylogeny Estimation
  – Metagenomics
  – Genome assembly

• I assume mathematical maturity and familiarity with graphs, discrete algorithms, and basic probability theory, equivalent to CS 374 and CS 361.

• No biology background is required!
Textbook

• **Computational Phylogenetics: An introduction to Designing Methods for Phylogeny Estimation**

• This may be available in the University Bookstore. If not, you can get it from Amazon.

• The textbook provides all the background you’ll need for the class, and homework assignments will be taken from the book.
Grading

• Homeworks: 25% (worst grade dropped)
• Take home midterm: 40%
• Final project: 25%
• Course participation and paper presentation: 10%

Note: no final exam
Course Project

• Course project is due the last day of class, and can be a research project or a survey paper.

• If you do a research project, you can do it with someone else (including someone in my research group); the goal will be to produce a publishable paper.

• If you do a survey paper, you will do it by yourself.
Some prior course projects that were published

- **ASTRAL**: genome-scale coalescent-based species tree estimation, Mirarab et al. 2014 (ECCB 2014 and Bioinformatics 2014)

- Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer, Davidson et al. 2014 (RECOMB-CG 2014 and BMC Genomics 2014)

- A comparative study of SVDquartets and other coalescent-based species tree estimation methods, Chou et al. (RECOMB-CG and BMC Genomics 2014)

- **ASTRID**: Accurate species trees from internode distances, Vachaspati and Warnow (RECOMB-CG 2015 and BMC Genomics 2015)

- Scaling statistical multiple sequence alignment to large datasets, Nute and Warnow (RECOMB-CG 2016 and BMC Genomics 2016)

- Are Profile Hidden Markov Models Identifiable?, Pattabiraman and Warnow (ACM-BCB 2018)

- **OCTAL**: Optimal Completion of Gene Trees in Polynomial Time, Christensen, Molloy, Vachaspati, and Warnow (WABI 2017 and Algorithms for Molecular Biology 2018)
Phylogeny (evolutionary tree)

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

Phylogenetic 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.

Courtesy of the Tree of Life web project, tolweb.org
The Tree of Life: *Multiple Challenges*

Scientific challenges:

- Ultra-large multiple-sequence alignment
- Gene tree estimation
- Metagenomic classification
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution

Techniques: applied probability theory, graph theory, supercomputing, and heuristics

Testing: simulations and real data
This talk

• Divide-and-conquer: Basic algorithm design technique

• Application of divide-and-conquer to:
  – Phylogeny estimation (and proof of “absolute fast convergence”)
  – Multiple sequence alignment
  – Almost alignment-free tree estimation

• Open problems
Divide-and-Conquer

• Divide-and-conquer is a basic algorithmic trick for solving problems!

• Three steps:
  – divide a dataset into two or more sets,
  – solve the problem on each set, and
  – combine solutions.
Objective: sort this list of integers from smallest to largest.

10, 3, 54, 23, 75, 5, 1, 25 should become 1, 3, 5, 10, 23, 25, 54, 75
MergeSort

Step 1: Divide into two sublists
Step 2: Recursively sort each sublist
Step 3: Merge the two sorted sublists
Key technique: Divide-and-conquer!

• Can be used to produce provably correct algorithms (as in MergeSort)
• But also – in practice, divide-and-conquer is useful for scaling methods to large datasets, because small datasets with not too much “heterogeneity” are easy to analyze with good accuracy.
Three divide-and-conquer algorithms

• **DCM1 (2001):** Improving distance-based phylogeny estimation

• **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
Three divide-and-conquer algorithms

- **DCM1 (2001)**: Improving distance-based phylogeny estimation
- **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
DNA Sequence Evolution

AGGGCCT

AGGGCAT

TAGCCCA

TAGACTT

AGCACAA

AGCGCCTT

AAGACTT

TGGACTT

AGCCT

today

-3 mil yrs

-2 mil yrs

-1 mil yrs
Gene Tree Estimation

AGGGCAT  TAGCCCA  TAGACTT  TGCACAA  TGCAGCCTT

Gene tree representation with nodes U, V, W, X, and Y connecting through branches.
Quantifying Error

FN: false negative
(missing edge)

FP: false positive
(incorrect edge)

50% error rate

TRUE TREE

DNA SEQUENCES

50% error rate

INFERRED TREE
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.
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)
Distance-based estimation

**TRUE TREE**

**DNA SEQUENCES**
- $S_1$: ACAATTAGAAC
- $S_2$: ACCCTTAGAAC
- $S_3$: ACCATTCCAAC
- $S_4$: ACCAGACCAAC

**METHODS SUCH AS NEIGHBOR JOINING**

**DISTANCE MATRIX**

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$S_2$</td>
<td></td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>$S_3$</td>
<td></td>
<td></td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$S_4$</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

**STATISTICAL ESTIMATION OF PAIRWISE DISTANCES**
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.
DCM1-boosting

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

Three divide-and-conquer algorithms

• DCM1 (2001): Improving distance-based phylogeny estimation
• 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
DNA Sequence Evolution

AAGACTT

-3 mil yrs

AAGGCCT

-2 mil yrs

AAGGGCAT

-1 mil yrs

TAGCCCT

today

TGGACTT

AGCACTT

AGGGCAT

TAGCCCA

TAGACTT

AGCACAA

AGCGCTT
Indels (insertions and deletions)

..ACGGTGCAGTTACCA...

..ACCCAGTCACCCA...
The true multiple alignment

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree
Gene Tree Estimation

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGGACCCGC
S3 = TAGCTGACCGGC
S4 = TCACGACCAGCA
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Phase 1: Alignment

\[ S1 = \text{AGGCTATCACCTGACCTCCA} \]
\[ S2 = \text{TAGCTATCACGACCGC} \]
\[ S3 = \text{TAGCTGACCGC} \]
\[ S4 = \text{TCACGACCGACA} \]
Phase 2: Construct tree

S1 = AGGCTATCAGCTGACCTCCA
S2 = TAGCTATCAGCGACCGC
S3 = TAGCTGACCGC
S4 = TCAGCGACGAGCA

S1

S4

S2

S3

S1 = -AGGCTATCAGCTGACCTCCA
S2 = TAG-CTATCAGCGACCGC--
S3 = TAG-CT---------GACCGC--
S4 = ---------TCACGACCGACGACA
Two-phase estimation

**Alignment methods**
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- Infernal (Bioinf. 2009)
- Etc.

**Phylogeny methods**
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.
# Two-phase estimation

<table>
<thead>
<tr>
<th>Alignment methods</th>
<th>Phylogeny methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Clustal</td>
<td>• Bayesian MCMC</td>
</tr>
<tr>
<td>• POY (and POY*)</td>
<td>• Maximum parsimony</td>
</tr>
<tr>
<td>• Probcons (and Probtree)</td>
<td>• Maximum likelihood</td>
</tr>
<tr>
<td>• Probalign</td>
<td>• Neighbor joining</td>
</tr>
<tr>
<td>• MAFFT</td>
<td>• FastME</td>
</tr>
<tr>
<td>• Muscle</td>
<td>• UPGMA</td>
</tr>
<tr>
<td>• Di-align</td>
<td>• Quartet puzzling</td>
</tr>
<tr>
<td>• T-Coffee</td>
<td>• Etc.</td>
</tr>
<tr>
<td>• Prank (PNAS 2005, Science 2008)</td>
<td></td>
</tr>
<tr>
<td>• Opal (ISMB and Bioinf. 2007)</td>
<td></td>
</tr>
<tr>
<td>• FSA (PLoS Comp. Bio. 2009)</td>
<td></td>
</tr>
<tr>
<td>• Infernal (Bioinf. 2009)</td>
<td></td>
</tr>
<tr>
<td>• Etc.</td>
<td></td>
</tr>
</tbody>
</table>
Two-phase estimation

Alignment methods
• Clustal
• POY (and POY*)
• Probcons (and Probtree)
• Probalign
• MAFFT
• Muscle
• Di-align
• T-Coffee
• Prank (PNAS 2005, Science 2008)
• Opal (ISMB and Bioinf. 2007)
• FSA (PLoS Comp. Bio. 2009)
• Infernal (Bioinf. 2009)
• Etc.

Phylogeny methods
• Bayesian MCMC
• Maximum parsimony
• Maximum likelihood
• Neighbor joining
• FastME
• UPGMA
• Quartet puzzling
• Etc.

RAxML: heuristic for large-scale ML optimization
1000-taxon models, ordered by difficulty (Liu et al., Science 19 June 2009)
Large-scale Co-estimation of alignments and trees

- SATé: Liu et al., Science 2009 (up to 10,000 sequences) and Systematic Biology 2012 (up to 50,000 sequences)
- PASTA: Mirarab et al., J. Computational Biology 2015 (up to 1,000,000 sequences)
Re-aligning on a tree

Decompose dataset

Align subproblems

Merge sub-alignments

Estimate ML tree on merged alignment
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

Tree

Use tree to compute new alignment

Alignment
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and return the alignment/tree pair with the best ML score
1000-taxon models, ordered by difficulty (Liu et al., Science 19 June 2009)
1000-taxon models, ordered by difficulty
(Liu et al., Science 19 June 2009)

24 hour SATé analysis, on desktop machines
(Similar improvements for biological datasets)
• Simulated RNAsim datasets from 10K to 200K taxa
• Limited to 24 hours using 12 CPUs
• Not all methods could run (missing bars could not finish)
• PASTA: J Comp Biol 22(5):377-386, Mirarab et al., 2015
Multiple Sequence Alignment (MSA):

* a scientific grand challenge*¹

\[
\begin{align*}
S_1 &= AGGCTATCACCTGACCTCCA & S_1 &= -AGGCTATCACCTGACCTCCA \\
S_2 &= TAGCTATCAGACCCG & S_2 &= TAG-CTATCAC--GACCGC-- \\
S_3 &= TAGCTGACCGC & S_3 &= TAG-CT---------GACCGC-- \\
\vdots & & \vdots \\
S_n &= TCACGACCGACA & S_n &= ---------TCAC--GACCGACA
\end{align*}
\]

**Novel techniques needed** for scalability and accuracy

NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

**Many important applications besides phylogenetic estimation**

¹ Frontiers in Massive Data Analysis, National Academies Press, 2013
Three divide-and-conquer algorithms

- **DCM1 (2001):** Improving distance-based phylogeny estimation
- **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
DACTAL

• Divide-And-Conquer Trees (Almost) without alignments
• Nelesen et al., ISMB 2012 and Bioinformatics 2012
• Input: unaligned sequences
• Output: Tree (but no alignment)
Unaligned Sequences

BLAST-based

Overlapping subsets

Decompose into small local subsets in tree

RAxML(MAFFT)

A tree for each subset

SuperFine

A tree for the entire dataset
Results on three biological datasets – 6000 to 28,000 sequences. We show results with 5 DACTAL iterations.
DACTAL

Start with any tree estimated on the data, decompose to any desired size.
“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)

Base method $M$ → Meta-method → $M^*$
Summary

- We showed three techniques to improve accuracy and scalability, that used divide-and-conquer (and sometimes iteration):
  - DCM1 (improves accuracy and scalability for distance-based methods)
  - SATe and PASTA (improves accuracy and scalability for multiple sequence alignment methods)
  - DACTAL (general technique for improving tree estimation)

- Innovative algorithm design can improve accuracy as well as reduce running time.
Open problems
(and possible course projects)

• Multiple Sequence Alignment:
  – Merging two alignments
  – Consensus alignments

• Supertree estimation
  – Need to scale to 100,000+ species with high accuracy
  – Approximation algorithms

• Species tree and phylogenetic network estimation
  – Address gene tree heterogeneity due to multiple causes
  – Theoretical guarantees under stochastic models
  – Scalability to large numbers of species with whole genomes

• Statistical Models of evolution
  – All models are seriously unrealistic
  – We need better statistical models
  – New theory

• Applications to historical linguistics, microbiome analysis, protein structure and function prediction, tumor evolution, etc.
The Tree of Life: **Multiple Challenges**

Scientific challenges:

- Ultra-large multiple-sequence alignment
- Gene tree estimation
- Metagenomic classification
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution

Techniques: applied probability theory, graph theory, supercomputing, and heuristics

Testing: simulations and real data
Other Information

• Tandy’s office hours: Tuesdays 2-3 PM (and by appointment).
• Tandy’s email: warnow@illinois.edu
• Homework will be submitted through Moodle.
• Course webpage:
  http://tandy.cs.illinois.edu/warnow-cs581-Fa2018.html