CS 581
Algorithmic Computational Genomics

Tandy Warnow
University of Illinois at Urbana-Champaign
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
  – 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%
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.
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

Phylogenetic Tree of Life

- Bacteria
  - Spirochetes
  - Proteobacteria
  - Cyanobacteria
  - Planctomycetes
  - Bacteroides
  - Cytophaga
  - Thermotoga
  - Aquifex

- Archaea
  - Green Filamentous bacteria
  - Methanobacterium
  - Methanococcus
  - T. celer
  - Thermopromus
  - Pyrodictium

- Eukaryota
  - Entamoeba
  - Slime molds
  - Animals
  - Fungi
  - Plants
  - Ciliates
  - Flagellates
  - Trichomonads
  - Microsporidia
  - Diplomonads

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.
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.
# Sorting

| 10 | 3 | 54 | 23 | 75 | 5 | 1 | 25 |

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

| 10 | 3  | 54 | 23 | 75 | 5  | 1  | 25 |

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

AAGACTT

AAGGCCT

AGGGCAT

AGGGCAT

TAGCCCA

TAGACTT

AGCACAA

AGCCGACTT

-3 mil yrs
-2 mil yrs
-1 mil yrs

today
Gene Tree Estimation
Quantifying Error

**FN**: false negative (missing edge)

**FP**: false positive (incorrect edge)

50% error rate
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 $n$. 

Sequence length requirements
Statistical consistency, exponential convergence, and absolute fast convergence (afc)
Distance-based estimation

TRUE TREE

S1 S2 S3 S4

INFERRED TREE

S1
S2 2
1
S3
2
S4

S1 ACAATTAGAAC
S2 ACCCTTAGAAC
S3 ACCATTCCAAC
S4 ACCAGACCAAC

DNA SEQUENCES

STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES

METHODS
SUCH AS
NEIGHBOR
JOINING

S1 S2 S3 S4
S1 0 3 6 5
S2 0 5 4
S3 0 5
S4 0

DISTANCE MATRIX
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

\textbf{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 \textit{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

TG

GACTT

AAG

G

C

C

T

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

A

G

GGC

A

T

T

AG

C

CCT

A

G

C

ACTT

AAGGCCT

TGGACTT

TAGCCC

TAGACTT

AGCACAA

AGCGCTT

AGGGCAT

TAGCCCA

AGGGCAT

AGGGCAT
Indels (insertions and deletions)

...ACGGTG CAGTTACCA...

Deletion

Mutation

...ACCAGT CACCA...
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 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACC CGC
S3 = TAGCTGACC CGC
S4 = TCACGACC CGACA
Phase 1: Alignment

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACC\_G\_GC
S3 = TAGCTGAC\_CCGC
S4 = TCACGACC\_CGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACC\_G\_GC
S3 = TAGCTGAC\_CCGC
S4 = TCACGACC\_CGACA
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACC CGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
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

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

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.

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

Estimate ML tree on merged alignment

Merge subalignments
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

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
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

S1 = AGGCTATCACCTGACCTCCA  S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGCCGCGC      S2 = TAG-CTATCAGCCGCGC--
S3 = TAGCTGACCGCGC          S3 = TAG-CT-------GACCGC--
...                          ...
Sn = TCACGACCGACA            Sn = ---------TCAC--GACCGACA

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)
DACTAL

Unaligned Sequences

BLAST-based

Decompose into small local subsets in tree

Overlapping subsets

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
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.
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)
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 3:15-4:15 (and by appointment).
• Tandy’s email: warnow@illinois.edu
• TA: Sarah Christensen (sac2@illinois.edu)
• Homework will be submitted through Moodle.
• Course webpage: http://tandy.cs.illinois.edu/581-2018.html