CS 394C
Algorithms for Computational Biology
Tandy Warnow
Spring 2012
Biology: 21st Century Science!

“When the human genome was sequenced seven years ago, scientists knew that most of the major scientific discoveries of the 21st century would be in biology.”

January 1, 2008, guardian.co.uk
Genome Sequencing Projects:

Started with the Human Genome Project
Whole Genome Shotgun Sequencing:

*Graph Algorithms and Combinatorial Optimization!*
Where did humans come from, and how did they move throughout the globe?

The 1000 Genome Project: using human genetic variation to better treat diseases
Other Genome Projects! (Neandertals, Wooly Mammoths, and more ordinary creatures…)

Neanderthals and humans

Anthropologists announced they have created a complete Neanderthal genome using ancient DNA samples. Neanderthals, the closest ancestor to modern humans, became extinct over 30,000 years ago.

How they compare to us

- Fossil evidence suggests that Neanderthals were muscular, with broad shoulders and strong limbs.
- Neanderthal (Homo neanderthalensis)
- Lower, larger skull
- Larger browridge
- Larger shoulder joint
- Larger, broader rib cage
- Larger elbow joint
- Shorter forearm
- Larger hip joint
- Larger, thicker knee
- Shorter, more flattened lower leg bone
- Larger ankle joint

© 2009 MCT
Source: Encyclopaedia Britannica, Incorporated
National History Channel & BBC Channel 4
For use, contact Lee Hulbert

The Mosquito Genome: Anopheles gambiae

Science

October 2002

Vol. 298

No. 5596

142

The Mosquito Genome: Anopheles gambiae
Metagenomics:

C. Ventner et al., Exploring the Sargasso Sea:

Scientists Discover One Million New Genes in Ocean Microbes
How did life evolve on earth?

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

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

We prove theorems using graph theory and probability theory, and our algorithms are studied on real and simulated data.
This course

• Fundamental mathematics of phylogeny and alignment estimation
• Applied research problems:
  – Metagenomics
  – Simultaneous estimation of alignments and trees
  – Ultra-large alignment and tree estimation
  – Phylogenomics
  – De novo genome assembly
  – Historical linguistics
Phylogenetic trees can be based upon morphology.
But some estimations need DNA!
DNA Sequence Evolution

-3 mil yrs
-2 mil yrs
-1 mil yrs
today
Phylogenetic reconstruction methods

1. Polynomial time distance-based methods (e.g., Neighbor-Joining)

2. Hill-climbing heuristics for NP-hard optimization criteria (Maximum Parsimony and Maximum Likelihood)

3. Bayesian methods
The neighbor joining method has high error rates on large 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.

[Nakhleh et al. ISMB 2001]
And solving NP-hard optimization problems in phylogenetics is … unlikely

<table>
<thead>
<tr>
<th># of Taxa</th>
<th># of Unrooted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10395</td>
</tr>
<tr>
<td>9</td>
<td>135135</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
</tr>
<tr>
<td>20</td>
<td>2.2 \times 10^{20}</td>
</tr>
<tr>
<td>100</td>
<td>4.5 \times 10^{190}</td>
</tr>
<tr>
<td>1000</td>
<td>2.7 \times 10^{2900}</td>
</tr>
</tbody>
</table>
Indels and substitutions at the DNA level

...ACGCTGCAGTTACCA...
Indels and substitutions at the DNA level

...ACGGTG\textcolor{red}{CAGTT}ACCA...
Indels and substitutions at the DNA level

Deletion  Mutation

...ACG\textcolor{teal}{\textbf{GTTG\textbf{CAGTT}}}ACCA...

...ACCA\textcolor{red}{\textbf{GTCACCA}}...
U  AGTGGAT
V  TATGCCCA
W  TATGACTT
X  AGCCCTA
Y  AGCCCGGCTT

U
V  W
X
Y
The true pairwise alignment is:

...ACGGTGCA GT TACCA...

...AC----CAGT CACCA...

The true multiple alignment on a set of homologous sequences is obtained by tracing their evolutionary history, and extending the pairwise alignments on the edges to a multiple alignment on the leaf sequences.
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGCGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCCGACA
Phase 1: Multiple Sequence Alignment

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = –AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCAGC
S4 = TCACGACCGACA

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

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

True tree and alignment

Unaligned Sequences

Estimated tree and alignment

Compare

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

Unaligned Sequences

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

True tree and alignment

Estimated tree and alignment

Compare
1000 taxon models, ordered by difficulty (Liu et al., 2009)
Problems

• Large datasets with high rates of evolution are hard to align accurately, and phylogeny estimation methods produce poor trees when alignments are poor.

• Many phylogeny estimation methods have poor accuracy on large datasets (even if given correct alignments)

• Potentially useful genes are often discarded if they are difficult to align.

These issues seriously impact large-scale phylogeny estimation (and Tree of Life projects)
Major Challenges

• Current phylogenetic datasets contain hundreds to thousands of taxa, with multiple genes.

• Future datasets will be substantially larger (e.g., iPlant plans to construct a tree on 500,000 plant species)

• Current methods have poor accuracy or cannot run on large datasets.
Theoretical Challenges:
- NP-hard problems
- Model violations

Empirical Challenges:
- Alignment estimation
- Data insufficient OR too much data
- Heuristics insufficient

The Tree of Life
Phylogenetic “boosters” (meta-methods)

Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

Examples:
• DCM-boosting for distance-based methods (1999)
• DCM-boosting for heuristics for NP-hard problems (1999)
• SATé-boosting for alignment methods (2009)
• SuperFine-boosting for supertree methods (2011)
• DACTAL-boosting for all phylogeny estimation methods (2011)
• SEPP-boosting for metagenomic analyses (2011)
Disk-Covering Methods (DCMs) (starting in 1998)
• DCMs “boost” the performance of phylogeny reconstruction methods.
The neighbor joining method has high error rates on large 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.

[Nakhleh et al. ISMB 2001]
DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]

**Theorem:**
DCM1-NJ converges to the true tree from polynomial length sequences.
Other “boosters”

• **SATé**: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009, and Liu et al. Systematic Biology, in press)

• **DACTAL**: Divide-and-Conquer Trees (Almost) without alignments (Nelesen et al., submitted)

• **SEPP**: SATé-enabled Phylogenetic Placement (Mirarab, Nguyen and Warnow, to appear, PSB 2012)
SATé Algorithm (Liu et al. Science 2009)

SATé = Simultaneous Alignment and Tree Estimation

Obtain initial alignment and estimated ML tree $T$

Estimate ML tree on new alignment

Use new tree ($T$) to compute new alignment ($A$)
One SATé iteration (really 32 subsets)

Decompose based on input tree

Estimate ML tree on merged alignment

Align subproblems

Merge subproblems
Results on 1000-taxon datasets

- 24 hour SATé analysis
- Other simultaneous estimation methods cannot run on large datasets
Limitations of SATé

- Decompose dataset
- Align subproblems
- Merge sub-alignments
- Estimate ML tree on merged alignment
Part II: DACTAL
(Divide–And–Conquer Trees (Almost) without alignments)

- Input: set S of unaligned sequences
- Output: tree on S (but no alignment)

(Nelesen, Liu, Wang, Linder, and Warnow, submitted)
New supertree method:

SuperFine

Existing Method:

RAxML(MAFFT)

Unaligned Sequences

BLAST-based

Overlapping subsets

pRecDCM3

A tree for each subset

A tree for the entire dataset
Average of 3 Largest CRW Datasets

CRW: Comparative RNA database,
Three 16S datasets with 6,323 to 27,643 sequences
Reference alignments based on secondary structure
Reference trees are 75% RAxML bootstrap trees

DACTAL (shown in red) run for 5 iterations starting from FT(Part)
FastTree (FT) and RAxML are ML methods
Observations

• DACTAL gives more accurate trees than all other methods on the largest datasets

• DACTAL is much faster than SATé (and can analyze datasets that SATé cannot)

• DACTAL is robust to starting trees and other algorithmic parameters
Taxon Identification in Metagenomics

- Input: set of shotgun sequences (very short)
- Output: a tree on the set of sequences, indicating the species identification of each sequence

- Issues: the sequences are not globally alignable, they are very short, and there are millions of them
Phylogenetic Placement

Input: **Backbone** alignment and tree on full-length sequences, and a set of **query** sequences (short fragments)

Output: Placement of query sequences on backbone tree

Applications:
- taxon identification of metagenomic data,
- phylogenetic analyses of NGS data.
Align Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-------GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = TAAAAC
Align Sequence

S1 = -AGGCTATCACCAGCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT--------GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = --------T-A--AAAC----------
Place Sequence

\[
\begin{align*}
S1 &= -AGGCTATCACCTGACCTCCA-AA \\
S2 &= TAG-CTATCAC--GACC\text{GC--GCA} \\
S3 &= TAG-CT------GACC\text{GC--GCT} \\
S4 &= TAC-----TCAC--G\text{ACC\text{GC}ACAGCT} \\
Q1 &= --------T-A--AAAC--------
\end{align*}
\]
HMMER vs. PaPaRa Alignments

Increasing rate of evolution
Divide-and-conquer with HMMER+pplacer
SEPP (10%-rule) on simulated data

Increasing rate of evolution
Historical linguistics

• Languages evolve, just like biological species.
• How can we determine how languages evolve?
• How can we use information on language evolution, to determine how human populations moved across the globe?
Questions about Indo-European (IE)

• How did the IE family of languages evolve?
• Where is the IE homeland?
• When did Proto-IE “end”?
• What was life like for the speakers of proto-Indo-European (PIE)?
Estimating the date and homeland of the proto-Indo-Europeans

- Step 1: Estimate the phylogeny
- Step 2: Reconstruct words for proto-Indo-European (and for intermediate proto-languages)
- Step 3: Use archaeological evidence to constrain dates and geographic locations of the proto-languages
“Perfect Phylogenetic Network” (Nakhleh et al., Language)
Reticulate evolution

• Not all evolution is tree-like:
  – Horizontal gene transfer
  – Hybrid speciation

• How can we detect reticulate evolution?
Course Details

• Phylogeny and multiple sequence alignment are the basis of almost everything in the course

• The first 1/3 of the class will provide the basics of the material

• The next 2/3 will go into depth into selected topics
Course details

• There is no textbook; I will provide notes.
• Homeworks: basic material and critical review of papers from the scientific literature
• Course project: either a research project (two students per project) or a literature survey (one student per project). The best projects should be submitted for publication in a journal or conference.
• Final exam: comprehensive, take home.
Grading

- Homework: 20%
- Class participation: 20%
- Final exam: 30%
- Class project: 30%
## Combined Analysis Methods

The table below provides a comprehensive overview of the combined analysis methods for three different genes:

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>TCTAATGGAA</td>
<td>TATTGATACACA</td>
</tr>
<tr>
<td>$S_2$</td>
<td>GCTAAGGGAA</td>
<td>TCTTGATACC</td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTAAGGGAA</td>
<td>TCTTGATACC</td>
</tr>
<tr>
<td>$S_4$</td>
<td>TCTAACGGAA</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_5$</td>
<td>TCTAACGGAA</td>
<td>CATTCATACC</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>TATGAGTGCA</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td>CATTGATACC</td>
</tr>
</tbody>
</table>
## Combined Analysis

<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>TCTAATGGAA</td>
<td>????????????</td>
<td>TATGATACA</td>
</tr>
<tr>
<td>$S_2$</td>
<td>GCTAAGGGAA</td>
<td>????????????</td>
<td>????????????</td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTAAGGGAA</td>
<td>????????????</td>
<td>TCTTGATACC</td>
</tr>
<tr>
<td>$S_4$</td>
<td>TCTAACGGAA</td>
<td>GGTAACCCTC</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_5$</td>
<td>???????????</td>
<td>GCTAAACCTC</td>
<td>????????????</td>
</tr>
<tr>
<td>$S_6$</td>
<td>???????????</td>
<td>GGTGACCATC</td>
<td>????????????</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>GCTAAACCTC</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td>????????????</td>
<td>CATTCATACC</td>
</tr>
</tbody>
</table>
Two competing approaches

Analyze separately

Combined Analysis

Supertree Method

Species

gene 1  gene 2  ...  gene k

Gene trees for each gene