Complexity and The Tree of Life

Tandy Warnow
The University of Texas at Austin
How did life evolve on earth?

An international effort to understand how life evolved on earth

Biomedical applications: drug design, protein structure and function prediction, biodiversity.

- Courtesy of the Tree of Life project
How did human languages evolve?
(Possible Indo-European tree, Ringe, Warnow and Taylor 2000)
DNA Sequence Evolution

AAGACTT

TG

GACTT

AAG

G

C

T

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

A

G

GGC

A

T

TAGCCCT

AGGCCT

TGGACTT

AAGACTT

TAGCCCT

AGGCCT

AGGGCAT

AGGGCAT

TAGCCCA

TAGACTT

AGCACAA

AGCGCTT
Standard Markov models

• Sequences evolve just with substitutions
• Sites (i.e., positions) evolve identically and independently, and have “rates of evolution” that are drawn from a common distribution (typically gamma)
• Numerical parameters describe the probability of substitutions of each type on each edge of the tree
Questions

• *Statistical consistency*: Is the given phylogeny reconstruction method guaranteed to reconstruct the model tree when infinitely long sequences are available?

• *Convergence rate* (sample size complexity): How long do the sequences need to be for the method to be accurate with high probability?
Quantifying Error

**TRUE TREE**

**DNA SEQUENCES**

$S_1$ ACAATTAGAAC
$S_2$ ACCCTTAGAAC
$S_3$ ACCATTCCAAC
$S_4$ ACCAGACCAAC
$S_5$ ACCAGACCGGA

**INFERRED TREE**

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

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

50% error rate
Statistical consistency, exponential convergence, and absolute fast convergence (afc)
Complexity \textit{viz.} The Tree of Life

- **Algorithmic complexity** (e.g., running time and NP-hardness)
- **Sample size complexity** (e.g. how long do the sequences need to be to obtain a highly accurate reconstruction with high probability?)
- **Stochastic model complexity** (i.e., how realistic are the models of evolution, and what are the consequences of making the models more realistic?)
Current state of knowledge

• We have established much of the statistical performance (consistency and convergence rates) of the major methods for phylogeny estimation.

• We have developed “fast converging” methods (guaranteed to reconstruct the true tree from polynomial length sequences) with excellent performance in practice.

• We have very fast methods for solving maximum likelihood and maximum parsimony, the major optimization problems, even for large datasets.
Distance-based Phylogenetic Methods
(polynomial time)

TRUE TREE

DNA SEQUENCES

S1 ACAAATAGAAC
S2 ACCCTTAGAAC
S3 ACCATTCCAAC
S4 ACCAGACCAAC

METHODS SUCH AS NEIGHBOR JOINING

INFERRED TREE

DISTANCE MATRIX

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

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES
Neighbor Joining’s sequence length requirement is exponential!

- Atteson: Let T be a General Markov model tree defining distance matrix D. Then Neighbor Joining will reconstruct the true tree with high probability from sequences that are of length at least $O(\lg n \, e^{\max \text{Dij}})$, where $n$ is the number of leaves in T.
Neighbor joining has poor performance on large diameter trees [Nakhleh et al. ISMB 2001]

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 distance-based methods

[Nakhleh et al. ISMB 2001]

Theorem: DCM1-NJ converges to the true tree from polynomial length sequences
Other “fast-converging” methods

• The “short quartet” methods (Erdös, Steel, Székéley and Warnow 1997) were the first fast-converging methods, published in RSA 1999 and TCS 1999.
• Csüros and Kao (SODA 1999)
• Cryan, Goldberg, and Goldberg (SICOMP 2001)
• Csüros (J Comp Bio 2002)
• Daskalakis et al. (RECOMB 2006)
• Daskalakis, Mossel and Roch (STOC 2006)
• Gronau, Moran and Snir (SODA 2008)
Maximum Likelihood (ML)

- Given: Set $S$ of aligned DNA sequences, and a parametric model of sequence evolution
- Objective: Find tree $T$ and numerical parameter values (e.g., substitution probabilities) so as to maximize the probability of the data.

NP-hard
Statistically consistent for standard models if solved exactly
Maximum Parsimony
(Hamming distance Steiner Tree problem)

Input: set of aligned sequences
Output: tree with minimum total length ("MP score")

MP score = 4

Not statistically consistent (even under simple models)
Finding the optimal MP tree is **NP-hard.**
Solving NP-hard problems exactly is … unlikely

- Number of (unrooted) binary trees on $n$ leaves is $(2n-5)!!$
- If each tree on 1000 taxa could be analyzed in 0.001 seconds, we would find the best tree in 2890 millennia

<table>
<thead>
<tr>
<th>#leaves</th>
<th>#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>
Problems with techniques for Maximum Parsimony

Shown here is the performance of a very good heuristic (TNT) for maximum parsimony analysis on a real dataset of almost 14,000 sequences. (“Optimal” here means *best score to date*, using any method for any amount of time.) Acceptable error is below 0.01%.

![Graph showing performance of TNT with time]
Rec-I-DCM3 significantly improves performance (Roshan et al. CSB 2004)

Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset. 
*Similar improvements obtained for RAxML (maximum likelihood).*
Current state of knowledge

• We have established much of the statistical performance (consistency and convergence rates) of the major methods for phylogeny estimation.

• We have developed “fast converging” methods (guaranteed to reconstruct the true tree from polynomial length sequences) with excellent performance in practice.

• We have very fast methods for solving maximum likelihood and maximum parsimony, the major optimization problems, even for large datasets.
But the **Standard Markov models are too simple!**

- Sequences evolve just with substitutions
- Sites (i.e., positions) evolve identically and independently, and have “rates of evolution” that are drawn from a common distribution (typically gamma)
- Numerical parameters describe the probability of substitutions of each type on each edge of the tree

*And all the positive results we’ve shown disappear under more realistic models*
The “tree of life” is not a tree

Reticulate evolution (horizontal gene transfer and hybridization) is also a problem
Languages also evolve with reticulation (Nakhleh et al., 2005)
Genome-scale evolution

(REARRANGEMENTS)

Inversion
Translocation
Duplication
indels (insertions and deletions) also occur!
The true pairwise alignment is:

...ACGGTGCA\text{GT}T\text{TACCA}...

...AC----CAGT\text{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 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Phase 1: Multiple Sequence Alignment

\[ S_1 = \text{AGGCTATCACCCTGACCTCCA} \]
\[ S_2 = \text{TAGCTATCAGACCCG} \]
\[ S_3 = \text{TAGCTGACCGC} \]
\[ S_4 = \text{TCACGACCACCA} \]

\[ S_1 = \text{-AGGCTATCACCCTGACCTCCA} \]
\[ S_2 = \text{TAG-CTATCAC--GACCCGC--} \]
\[ S_3 = \text{TAG-CT-------GACCGC--} \]
\[ S_4 = \text{-------TCAC--GACCGACA} \]
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCAGC
S3 = TAGCTGACCAGC
S4 = TCACGACCAGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAGACCAGC--
S3 = TAG-CT--------GACCAGC--
S4 = --------TCAC--GACCAGACA
DNA sequence evolution

Simulation using ROSE: 100 taxon model trees, models 1-4 have “long gaps”, and 5-8 have “short gaps”, site substitution is HKY+Gamma
SATé Algorithm (unpublished)

SATé keeps track of the maximum likelihood scores of the tree/alignment pairs it generates, and returns the best pair it finds.

Obtain initial alignment and estimated ML tree $T$

Estimate ML tree on new alignment

Use new tree ($T$) to compute new alignment ($A$)
Models 1-3 have 1000 taxa, Models 4-6 have 500 taxa
(gap length distributions: long, medium, short)
Complexity *viz.* The Tree of Life

- **Algorithmic complexity** (e.g., running time and NP-hardness)
- **Sample size complexity** (e.g. how long do the sequences need to be to obtain a highly accurate reconstruction with high probability?)
- **Stochastic model complexity** (i.e., how realistic are the models of evolution, and what are the consequences of making the models more realistic?)
Thoughts

- Current models of sequence evolution are clearly too simple, and more realistic ones are not identifiable.
- The relative performance between methods can change as the models become more complex or as the number of taxa increases.
- We do not know how methods perform under realistic conditions (nor how long we need to let computationally intensive methods run).
- Therefore, *simulations should be done under very realistic (sufficiently complex) models, even if estimations are done under simpler models (and it is likely that estimations are best done under more realistic models, too).*
Acknowledgements

• Funding: NSF, The David and Lucile Packard Foundation, The Program in Evolutionary Dynamics at Harvard, and The Institute for Cellular and Molecular Biology at UT-Austin.

• Collaborators:
  – Fast-converging methods: Peter Erdös, Daniel Huson, Bernard Moret, Luay Nakhleh, Usman Roshan, Katherine St. John, Michael Steel, and Laszlo Székely
  – Rec-I-DCM3: Usman Roshan, Bernard Moret, and Tiffani Williams
  – SATé: Randy Linder, Kevin Liu, Serita Nelesen, and Sindhu Raghavan
Simulated Model Conditions

<table>
<thead>
<tr>
<th>Model Condition</th>
<th>Taxa</th>
<th>Average gap length</th>
<th>ANHD</th>
<th>MNHD</th>
<th>Percent gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
<td>9.2 (7.2)</td>
<td>69.2 (.01)</td>
<td>76.7 (.01)</td>
<td>72.1 (.19)</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>5 (4.4)</td>
<td>68.0 (.02)</td>
<td>75.7 (.02)</td>
<td>70.4 (.10)</td>
</tr>
<tr>
<td>3</td>
<td>1000</td>
<td>2 (1.2)</td>
<td>69.1 (.01)</td>
<td>76.6 (.01)</td>
<td>41.7 (.14)</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>9.2 (7.2)</td>
<td>66.1 (.02)</td>
<td>74.3 (.01)</td>
<td>76.7 (.11)</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>5 (4.4)</td>
<td>66.3 (.02)</td>
<td>74.2 (.01)</td>
<td>64.7 (.14)</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>2 (1.2)</td>
<td>66.1 (.02)</td>
<td>74.2 (.02)</td>
<td>42.8 (.14)</td>
</tr>
</tbody>
</table>

- ANHD is the average normalized Hamming distance. MNHD is the maximum normalized Hamming distance. (Normalized Hamming distances are also known as \textit{p-distances}.)
- Standard deviations are given parenthetically for average gap length, and standard errors are given parenthetically for all other statistics.
Biological datasets

• We used 8 different biological datasets with curated alignments (produced by Robin Gutell (UT-Austin)) based upon secondary structures.
• We computed various alignments, and maximum likelihood trees on each alignment.
• We ran SATé for 24 hours, producing an alignment/tree pair.
• We evaluated alignments and trees in comparison to the curated alignment and to the reference tree (the 75% bootstrap maximum likelihood tree on the curated alignment), respectively.
Results for 23S rRNA dataset