Introduction to Phylogenetic Estimation Algorithms

Tandy Warnow
Questions

• What is a phylogeny?
• What data are used?
• What is involved in a phylogenetic analysis?
• What are the most popular methods?
• What is meant by “accuracy”, and how is it measured?
Phylogeny

From the Tree of the Life Website, University of Arizona
Data

- Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
- Molecular markers (e.g., SNPs, RFLPs, etc.)
- Morphology
- Gene order and content

These are “character data”: each character is a function mapping the set of taxa to distinct states (equivalence classes), with evolution modelled as a process that changes the state of a character.
Data

- Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
- Molecular markers (e.g., SNPs, RFLPs, etc.)
- Morphology
- Gene order and content

These are “character data”: each character is a function mapping the set of taxa to distinct states (equivalence classes), with evolution modelled as a process that changes the state of a character.
DNA Sequence Evolution

AAGACTT

TG

GACTT

AAG

G

C

C

T

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

A

GGC

A

T

T

AG

C

CCT

A

G

C

ACTT

AAGGCCT

TGGACTT

AAGGCCT

AGGGCAT

TAGCCCCA

TAGACTT

AGCACA

AGCGCTT

AGGGCAT

AGGGCAT
Phylogeny Problem

AGGGGCAT  TAGCCCA  TAGACTT  TGCACAA  TGCCTTT

U  V  W  X  Y

V  W

X  Y
Indels and substitutions at the DNA level

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

Deletion  Mutation

...ACG GTG CAG TTACCA...
Indels and substitutions at the DNA level

Deletion  Mutation

...ACG\text{\textcolor{blue}{G}}\text{\textcolor{green}{T}}\text{\textcolor{red}{G}}\text{\textcolor{blue}{C}}\text{\textcolor{red}{A}}\text{\textcolor{blue}{G}}\text{\textcolor{green}{T}}\text{\textcolor{red}{T}}\text{\textcolor{blue}{AC}}\text{\textcolor{green}{C}}...
The true pairwise alignment is:

```
...ACGGTGCA\textcolor{cyan}{G}T\textcolor{red}{T}ACCA...

...AC-----\textcolor{red}{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.
# Easy Sequence Alignment

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_WEAU160</td>
<td>ATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAAGTAGACAGG</td>
<td>45</td>
</tr>
<tr>
<td>A_U455</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>A_IFA86</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>A_92UG037</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>A_Q23</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_SF2</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_LAI</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_F12</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_HXB2R</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_LW123</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_NL43</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_NY5</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_MN</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_JRCSF</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_JRFL</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_NH52</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_OYI</td>
<td>......................................</td>
<td>45</td>
</tr>
<tr>
<td>B_CAM1</td>
<td>......................................</td>
<td>45</td>
</tr>
</tbody>
</table>
Harder Sequence Alignment

<table>
<thead>
<tr>
<th></th>
<th>ATGAGAGTGAAGGGGATCAGGAAGAATTATCAGCACTTTG</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_WEAU160</td>
<td>ATGAGAGTGAAGGGGATCAGGAAGAATTATCAGCACTTTG</td>
</tr>
<tr>
<td>A_U455</td>
<td>T.       ACA . G.         .   CTTG.</td>
</tr>
<tr>
<td>A_SF1703</td>
<td>T.       ACA . T.   C. G.   AA. A</td>
</tr>
<tr>
<td>A_92RW020.5</td>
<td>G.       ACA . C.   G.   GG. AA.</td>
</tr>
<tr>
<td>A_92UG031.7</td>
<td>G.A.    ACA . G.   GG.          A</td>
</tr>
<tr>
<td>A_92UG037.8</td>
<td>T.       AGA . G.         .   CTTG . G.</td>
</tr>
<tr>
<td>A_TZ017</td>
<td>G . A.  G.A. G.         .   A. A</td>
</tr>
<tr>
<td>A_UG275A</td>
<td>...........A. C. T.  CACA . T.   G. AA. G.</td>
</tr>
<tr>
<td>A_UG273A</td>
<td>........................................ACA . G.   GG........</td>
</tr>
<tr>
<td>A_DJ258A</td>
<td>........................................T.       ACA .........CA. T.  A</td>
</tr>
<tr>
<td>A_KENYA</td>
<td>........................................T.       CACA . G.   G.          A</td>
</tr>
<tr>
<td>A_CARGAN</td>
<td>........................................T.       ACA .........A.</td>
</tr>
<tr>
<td>A_CARSAS</td>
<td>........................................CACA .........CTCT . C........</td>
</tr>
<tr>
<td>A_CAR4054</td>
<td>........................................A.    CACA . G.   GG. CA.</td>
</tr>
<tr>
<td>A_CAR286A</td>
<td>........................................CACA . G.   GG. AA.</td>
</tr>
<tr>
<td>A_CAR4023</td>
<td>........................................A.       ---------------. A.</td>
</tr>
<tr>
<td>A_CAR423A</td>
<td>........................................A.       ---------------. A.</td>
</tr>
<tr>
<td>A_VI191A</td>
<td>........................................ACA . T.   GG. A.</td>
</tr>
</tbody>
</table>
Objective:
Estimate the “true alignment” (defined by the sequence of evolutionary events)

Typical approach:
1. Estimate an initial tree
2. Estimate a multiple alignment by performing a “progressive alignment” up the tree, using Needleman-Wunsch (or a variant) to align alignments
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCAGC
S3 = TAGCTGACCACG
S4 = TCACGACCAGACA
Phase 1: Multiple Sequence Alignment

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

S1 = AGGCTATCACCTGACCTCCA  S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCACGC  S2 = TAG-CTATCAC--GACCAGC--
S3 = TAGCTGACCCGC  S3 = TAG-CT-------GACCAGC--
S4 = TCACGACCCGACA  S4 = -------TCAC--GACCAGAC

\[\text{S1} \rightarrow \text{S2} \rightarrow \text{S3} \rightarrow \text{S4}\]
So many methods!!!

Alignment method
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Phylogeny method
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- UPGMA
- Quartet puzzling
- Etc.
So many methods!!!

Alignment method
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Phylogeny method
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- UPGMA
- Quartet puzzling
- Etc.
So many methods!!!

Alignment method
• Clustal
• POY (and POY*)
• Probcons (and Probtree)
• MAFFT
• Prank
• Muscle
• Di-align
• T-Coffee
• Satchmo
• Etc.

Blue = used by systematists

Purple = recommended by Edgar and Batzoglou for protein alignments

Phylogeny method
• Bayesian MCMC
• Maximum parsimony
• Maximum likelihood
• Neighbor joining
• UPGMA
• Quartet puzzling
• Etc.
Phylogenetic reconstruction methods

1. Polynomial time distance-based methods: UPGMA, Neighbor Joining, FastME, Weighbor, etc.

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

3. Bayesian methods
UPGMA

While $|S|>2$: 

- find pair $x,y$ of closest taxa;
- delete $x$;
- Recurse on $S\{-x\}$;
- Insert $y$ as sibling to $x$;
- Return tree.

```
  a
 / \  /  /
/   /   /   /
\  /   /   /   /
  b c d e
```
UPGMA

Works when evolution is "clocklike"
UPGMA

Fails to produce true tree if evolution deviates too much from a clock!
Performance criteria

• Running time.
• Space.
• Statistical performance issues (e.g., statistical consistency and sequence length requirements)
• “Topological accuracy” with respect to the underlying true tree. Typically studied in simulation.
• Accuracy with respect to a mathematical score (e.g. tree length or likelihood score) on real data.
Distance-based Methods

TRUE TREE

DNA SEQUENCES

S_1 ACAATTAGAAC
S_2 ACCCTTAGAAC
S_3 ACCATCCAAC
S_4 ACCAGACCAAC

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>0</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>S_3</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_4</td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

METHODS SUCH AS NEIGHBOR JOINING

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES
Additive Distance Matrices

<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>0</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

POLYTIME INVERTIBLE
Four-point condition

• A matrix D is additive if and only if for every four indices i,j,k,l, the maximum and median of the three pairwise sums are identical

\[ D_{ij} + D_{kl} < D_{ik} + D_{jl} = D_{il} + D_{jk} \]

The Four-Point Method computes trees on quartets using the Four-point condition.
Naïve Quartet Method

• Compute the tree on each quartet using the four-point condition

• Merge them into a tree on the entire set if they are compatible:
  – Find a sibling pair A,B
  – Recurse on S-{A}
  – If S-{A} has a tree T, insert A into T by making A a sibling to B, and return the tree
Better distance-based methods

• Neighbor Joining
• Minimum Evolution
• Weighted Neighbor Joining
• Bio-NJ
• DCM-NJ
• And others
Quantifying Error

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

50% error rate

DNA SEQUENCES

INFERRED TREE
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.
“Character-based” methods

- Maximum parsimony
- Maximum Likelihood
- Bayesian MCMC (also likelihood-based)

These are more popular than distance-based methods, and tend to give more accurate trees. However, these are computationally intensive!
Standard problem: Maximum Parsimony
(Hamming distance Steiner Tree)

- **Input:** Set $S$ of $n$ aligned sequences of length $k$
- **Output:** A phylogenetic tree $T$
  - leaf-labeled by sequences in $S$
  - additional sequences of length $k$ labeling the internal nodes of $T$

such that $\sum_{(i, j) \in E(T)} H(i, j)$ is minimized.
Maximum parsimony (example)

• **Input**: Four sequences
  – ACT
  – ACA
  – GTT
  – GTA

• **Question**: which of the three trees has the best MP scores?
Maximum Parsimony
Maximum Parsimony

MP score = 5

MP score = 7

MP score = 4

Optimal MP tree
Maximum Parsimony: computational complexity

Optimal labeling can be computed in linear time $O(nk)$

Finding the optimal MP tree is $NP$-hard
But solving this problem exactly 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>
Local search strategies

Cost

Phylogenetic trees

Local optimum

Global optimum
Local search strategies

• Hill-climbing based upon topological changes to the tree

• Incorporating randomness to exit from local optima
Evaluating heuristics with respect to MP or ML scores

Fake study

Score of best trees

Performance of Heuristic 1

Performance of Heuristic 2

Time
“Boosting” MP heuristics

- We use “Disk-covering methods” (DCMs) to improve heuristic searches for MP and ML
Rec-I-DCM3 significantly improves performance (Roshan et al.)

Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset
Current methods

- Maximum Parsimony (MP):
  - TNT
  - PAUP* (with Rec-I-DCM3)
- Maximum Likelihood (ML)
  - RAxML (with Rec-I-DCM3)
  - GARLI
  - PAUP*
- Datasets with up to a few thousand sequences can be analyzed in a few days
- Portal at www.phylo.org
But…

U  AGTGGAT
V  TATGCCCA
W  TATGACTT
X  AGCCCTA
Y  AGCCCGCTT
• Phylogenetic reconstruction methods assume the sequences all have the same length.

• Standard models of sequence evolution used in maximum likelihood and Bayesian analyses assume sequences evolve only via substitutions, producing sequences of equal length.

• And yet, almost all nucleotide datasets evolve with insertions and deletions ("indels"), producing datasets that violate these models and methods.

How can we reconstruct phylogenies from sequences of unequal length?
Basic Questions

• Does improving the alignment lead to an improved phylogeny?
• Are we getting good enough alignments from MSA methods? (In particular, is ClustalW - the usual method used by systematists - good enough?)
• Are we getting good enough trees from the phylogeny reconstruction methods?
• Can we improve these estimations, perhaps through simultaneous estimation of trees and alignments?
Simulation using ROSE: 100 taxon model trees, models 1-4 have “long gaps”, and 5-8 have “short gaps”, site substitution is HKY+Gamma
Results

A variety of models with 1000 sequences

Model difficulty