Phylogeny

Analysis of biological sequences 140.638
Phylogenetics

Phylon = tribe/race, genetikos = relative to birth

Phylogenetics: study of evolutionary relationships among organisms, sequences, or anything that can be assigned distances

Related to multiple sequence alignment: MSAs can be used to construct trees and phylogenetic trees are used in MSA construction
Phylogenetic tree anatomy

• Separate sequences are taxa [singular: taxon] — phylogenetically distinct units on the tree

• Branches connect nodes to nodes or nodes to leaves and represent evolutionary distance

• leaves == OTUs (Operational Taxonomic Unit)
Phylogenetic tree anatomy

- Trees are generally bifurcating (multifurcation possible for viruses)
- Trees may be rooted or unrooted
- Vertical layout arbitrary
Phylogenetic tree anatomy

• Trees are probably not strictly bifurcating
  • Some evidence for horizontal transfer, between-"species" transfer etc
  • other reasons?
  • “forest of life” or “reticulated tree of life”
homoplasy

• the same characteristic evolves independently in two branches of a tree

• tree may be longer than necessary, and possibly the wrong topology
Molecular Clock hypothesis

MC hypothesis: the rate of evolution is the same in all tree branches.

This is suitable for closely related species but isn’t always useful or appropriate.
Ultrametric distance

- Assumes that the rates of evolution are the same in all branches (molecular clock)

- If so: \( d_{AC} \leq \max(d_{AB}, d_{BC}) \) for all A, B, C

- This means that the maximum distance separating the three leaves is not unique
Ultrametric distance

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Ultrametric

yes

dog
  bear
  raccoon
  seal
  sea lion
  weasel
  cat
  monkey

no

dog
  raccoon
  bear
  sea lion
  seal
  monkey
  cat
  weasel
Phylogenetic analysis: methods

- Strong sequence similarity -> maximum parsimony

- Recognizable sequence similarity -> distance methods (UPGMA, WPGMA, Neighbor-joining)

- No? -> maximum likelihood methods
Distance methods

• Employ the number of changes between each pair in a group of sequences to create a phylogenetic tree

• Neighbors have the smallest number of sequence changes, so presumably they share their nearest common ancestor (minimize distance, minimize homoplasy)

• Pioneered by Feng and Doolittle
How to collect distance data

• Lab methods:
  • Mix single strands of DNA/cDNA from different species and measure association parameters (like a CoT curve)
  • Agglutination times
  • Many more!
• Sequence analysis methods:
  • Alignments
  • k-mer counting
  • Composition analysis
  • Shared domains
Tree-making

Goal: create a tree whose branch lengths reflect the distance metric

\[ d_{AB} = 2 \]
\[ d_{BC} = 1 \]
\[ d_{AC} = 2 \]
UPGMA / WPGMA

Unweighted/Weighted pair group method with arithmetic means
Sneath, PHA and Sokal, RR, in *Numerical Taxonomy* (1973)

Progressively cluster sequences by distance until a tree is formed
UPGMA

Algorithm:
Make sure distances are ultrametric
Choose i and j such that $d_{ij}$ is minimal and join them to form cluster k
Recompute distances for each of the other items (l) in the set:

$$d_{kl} = \frac{d_{ii} + d_{jj}}{2}$$

$$d_{kl} = \frac{N_i d_{ii} + N_j d_{jj}}{N_i + N_j}$$
UPGMA: an example

We have 5 sequences and all pairwise distances:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>
UPGMA: an example

\[
\begin{array}{cccc}
  & A & B & C & D \\
 B & 8 &   &   &   \\
 C & 8 & 2 &   &   \\
 D & 6 & 8 & 8 &   \\
 E & 2 & 8 & 8 & 6 \\
\end{array}
\]

\[
\begin{align*}
  d_{(AE)} &= 2 \\
  d_{(AE)B} &= \frac{d_{AB} + d_{EB}}{2} = 8 \\
  d_{(AE)C} &= \frac{d_{AC} + d_{EC}}{2} = 8 \\
  d_{(AE)D} &= \frac{d_{AD} + d_{ED}}{2} = 6
\end{align*}
\]
UPGMA: an example

\[
d_{(BC)} = 2 \\
d_{(BC)(AE)} = \frac{d_{B(AE)} + d_{C(AE)}}{2} = 8 \\
d_{(BC)D} = \frac{d_{BD} + d_{CD}}{2} = 8
\]
UPGMA: an example

\[ d_{(AE)D} = 6 \]

\[ d_{(AED)(BC)} = \frac{d_{(AED)B} + d_{(AED)C}}{2} = 8 \]
Additive distances satisfy the four-point metric condition, e.g.

\[ d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC}) \]
Not ultrametric?

Four point metric condition

\[ d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC}) \]

\[ d_{AB} = a+b; \ d_{CD} = c+d; \]
\[ d_{AC} = a+e+c; \ d_{BD} = b+e+d; \]
\[ d_{AD} = a+e+d; \ d_{BC}=b+e+c \]

this means that we can define and assign distances to inner points.
Additive trees don’t strictly require a molecular clock.

If the evolutionary distance is additive, we can use simple arithmetic to infer inner nodes.

\[ d(m, k) = \frac{(d(j, k) + d(i, k) - d(i, j))}{2} \]

Known:
- \( d(i, j) \)
- \( d(i, k) \)
- \( d(j, k) \)
Neighbor joining

- Good example of a minimum evolution method

- Goal is to minimize the sum of branch lengths (assumes this is the best estimate of phylogeny)

- In fact, rarely gives the shortest tree

- Requires additive distances (only additive distances can fit into an unrooted tree)

- Very fast

- No molecular clock required, so it’s good for real-life data
Neighbor Joining: example

1. Compute the net divergence for every node

\[ r_A = 5+4+7+6+8 = 30 \quad r_D = 38 \]
\[ r_B = 5+7+10+9+11 = 42 \quad r_E = 34 \]
\[ r_C = 32 \quad r_F = 44 \]

From The Phylogenetic Handbook, Salemi and Vandamme 2004
Neighbor Joining: example

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

2. Create a rate-corrected distance matrix where \( M_{ij} = d_{ij} - (r_i + r_j)/(N-2) \)

\[
M_{AB} = d_{AB} - (r_A + r_B)/(N-2) = 5 - (30 + 42)/4 = -13
\]

From The Phylogenetic Handbook, Salemi and Vandamme 2004
Neighbor Joining: example

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>-13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-11.5</td>
<td>-11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-10</td>
<td>-10</td>
<td>-10.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-10</td>
<td>-10</td>
<td>-10.5</td>
<td>-13</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-10.5</td>
<td>-10.5</td>
<td>-11</td>
<td>-11.5</td>
<td>-11.5</td>
</tr>
</tbody>
</table>

3. Define a new node for which $M_{ij}$ is minimal (either AB or DE here)

4. Compute the branch lengths from a new node, U, to A and B

\[
S_{AU} = \frac{d_{AB}}{2} + \frac{(r_A - r_B)}{2(N-2)} = 2.5 - \frac{12}{2*4} = 1
\]

\[
S_{BU} = d_{AB} - S_{AU} = 5 - 1 = 4 \text{ (alternatively, } S_{AU} = 4, S_{BU} = 1)\]

From The Phylogenetic Handbook, Salemi and Vandamme 2004
Neighbor Joining: example

5. Compute new distances from the new node U to all other nodes

\[ d_{CU} = \frac{d_{AC} + d_{BC} - d_{AB}}{2} = 3 \]
\[ d_{DU} = 6 \]
\[ d_{EU} = 5 \]
\[ d_{FU} = 7 \]

6. \( N = N - 1 \), repeat steps 1 through 5
Neighbor Joining: example
Evaluating the trees

Bootstrap analysis (Felsenstein 1985)

Obtain a new alignment from the original by randomly choosing columns; each column can be selected more than once or not at all.

For each new dataset, a tree is produced — look at the proportion of the trees that each branch is represented in; if < 70% for 200-2000 trees -> low confidence.

Long branch attraction will be supported at a high bootstrap level.

```
ACGCTGGT   A   GG   GTT
AGGCTGCT   A   GC   CTT
CCGGTCTG    C   GG   GTT
ACGGTCTG   A   CG   GTT
```

```
new tree
```

Diagram:

```
/  \
/   \n/    D
/      \n/        A
/          \nA         /  \n\    \     \  \\
B    /       \  \\
/     \     /  \\
C
```

```
Evaluating the trees

Jackknifing (another sampling approach)
  Repeatedly leave out sequences or positions
  Recompute tree
  Subtrees appearing at low frequency are suspect
Rooting trees

• NJ and UPGMA produce unrooted trees

• Finding the root means conferring evolutionary directionality and more meaning to the tree

• One method: add an outgroup (species that is more distantly related to the species already in the tree than they are to each other) — the point in the tree where the outgroup joins is a good candidate for the root

• Can also pick the midpoint of the longest chain of consecutive edges (will be OK if not too far from a molecular clock)
Parsimony

• Probably most widely used of tree-building algorithms
  • Finds tree that can explain the alignment/data with fewest substitutions
  • Assigns a cost to each possible tree
  • No explicit measure of distance
  • Character-based, not distance-based
  • easy to implement
• Best for sequences with very strong similarity
Parsimony

Suppose we have five species, identical except at one nucleotide, 3 Cs and 2 Ts.

Minimal tree has one evolutionary change:

```
ACGGATCAGCTTTAGCC
ACGGATCAGCTTTTAGCC
ACGGATCAGCTTTTAGCC
ACGGATCAGTTTTAGCC
ACGGATCAGTTTTAGCC
ACGGATCAGTTTTAGCC
```
Traditional Parsimony (Fitch 1971)

Example—four species, small sequences:

<table>
<thead>
<tr>
<th></th>
<th>(w)</th>
<th>AAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td>AAA</td>
<td></td>
</tr>
<tr>
<td>(y)</td>
<td>GGA</td>
<td></td>
</tr>
<tr>
<td>(z)</td>
<td>AGA</td>
<td></td>
</tr>
</tbody>
</table>

3 trees are possible if you have four species:

A

B

C

y

z

w

x

w

y

x

z
Traditional Parsimony (Fitch 1971)

First, examine tree layout “A”

If this tree is correct, how many changes happened in each position of the ancestral sequence?

- Position 1: 1
- Position 2: 2
- Position 3: 1
Traditional Parsimony (Fitch 1971)

First, examine tree layout “A”
If this tree is correct, how many changes happened in each position of the ancestral sequence?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

w AAG
x AAA
y GGA
z AGA

AGA or AAA
Traditional Parsimony (Fitch 1971)

Example:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>AAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>AAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>GGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>AGA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

most parsimonious
Parsimony

• Can be misleading when rates of sequence change vary in different branches of the tree

• Different sections of genes and different regions of a genome will evolve at different rates (why?)

• has particular problems with homoplasy (same sequence change in more than one branch of the tree)
Speeding up parsimony

• Branch-and-bound

• Greedy algorithms with branch swapping
  
  • Nearest-neighbor interchange
  
  • Subtree pruning and regrafting
  
  • Tree bisection and reconnection
Maximum likelihood methods

Look for the tree that maximizes the likelihood of the data that have been observed

- With a few sequences, can just go through all the trees and calculate likelihoods

- Use Felsenstein’s algorithm for larger # of sequences

- Large datasets, especially proteins, require new strategies (for example, sampling)
Sampling & Bayesian phylogeny

\[
P(T|D) = \frac{P(D|T)P(T)}{P(D)}
\]

\[
\frac{P(T_1|D)}{P(T_2|D)} = \frac{P(D|T_1)P(T_1)}{P(D|T_2)P(T_2)}
\]
Metropolis algorithm (MCMC)

You have one tree. Propose a new tree, with features taken from some distribution (or by perturbing current tree).

\[ P_1 = P(T_1|D) = \text{posterior probability of current tree} \]
\[ P_2 = P(T_2|D) = \text{posterior probability of new tree} \]

if \((P_2 \geq P_1)\) take new tree

if \((P_2 < P_1)\) take new tree with probability \(P_2/P_1\), else keep old tree
Metropolis algorithm (MCMC)

After some amount of time and computation you have a set of trees and probabilities.

Main idea: the frequency of property $f$ seen in a chosen tree will converge to the posterior probability of that property, given the model.
More realistic models

• We have used some drastic simplifications: ungapped alignments, each site independent and using the same substitution matrix, etc.

• Models exist for

  • Allowing different rates at different sites (Yang, Felsenstein and Churchill)

  • Allowing gaps (Mitchison and Durbin)

  • Allowing different probabilistic models
Comparison of methods

“diagnostic” tree:

right:

wrong:

MLE and neighbor-joining will get the correct tree

Parsimony picks the wrong tree
Comparison of methods

Maximum likelihood phylogeny

Advantages:
  - most flexible
  - good results using good evolutionary models
  - end up with likelihood/uncertainty for all suboptimal trees

Disadvantages
  - computationally intensive
  - bad results using bad evolutionary models
Comparison of methods

UPGMA assumes a molecular clock, so provides a rooted tree (may be too strong an assumption)

Neighbor-joining is good when evolutionary rates vary. Flexible. Proven to construct the correct tree under the proper circumstances.

Parsimony is good for closely related sequences and is generally fast

Likelihood method is the most general of all
The computational challenge

>1.7 million known species; # trees increases exponentially with each new species added

# unrooted trees for n species: \((2n-5)!! = 3 \times 5 \times 7 \times \ldots \times (2n-5)\)
Newer approaches

- creative ways to determine distance
- fast clustering methods
- no good way to draw really big trees though!
Complete composition vector

idea: we can count all dimers, trimers, tetramers etc in the sequences that we have
from there we can estimate the expected frequency of each 5-mers
if the prediction varies significantly from the final counts, there has been selection
For every trimer and tetramer count the occurrences of that sequence. Then the predicted number of counts of each 5-mer are defined by the observed frequencies of its trimer & tetramers

$$p^e(\alpha_1\alpha_2, \ldots, \alpha_k) = \frac{(p(\alpha_1\alpha_2, \ldots, \alpha_{k-1}) \cdot p(\alpha_2\alpha_2, \ldots, \alpha_k))}{p(\alpha_2\alpha_3, \ldots, \alpha_{k-1})}$$

For example, the expected frequency of AGCTT

is $(p(AGCT) \cdot p(GCTT))/p(GCT)$
Calculate the expected incidence of every kmer based on the observed frequencies of all of the (k-1)mers.

The observed frequency of a kmer may vary significantly from what’s expected, due to evolutionary pressures.

This O/E value can be a type of signature for an organism.
The differences in O/E for each kmer can be used as a distance metric.

Successful studies so far:

HIV

Influenza

Fungi
CCV for 86 fungi

A fungal phylogeny based on 82 complete genomes using the composition vector method

Hao Wang¹✉, Zhao Xu¹✉, Lei Gao²✉ and Bailin Hao¹,³,⁴✉
1 T-life Research Center, Department of Physics, Fudan University, Shanghai 200433, PR China
2 Department of Botany & Plant Sciences, University of California, Riverside, CA(92521), USA
3 Institute of Theoretical Physics, Academia Sinica, Beijing 100190, PR China
4 Santa Fe Institute, Santa Fe, NM(87501), USA
✉ author email  ✉ corresponding author email
