Evolutionary trees

Describe the relationship between objects, e.g. species or genes
Early evolutionary studies

- Anatomical features were the dominant criteria used to derive evolutionary relationships between species since Darwin till early 1960s

- The evolutionary relationships derived from these relatively subjective observations were often inconclusive, and some of them were later proved incorrect
Early evolutionary studies

- Anatomical features were the dominant criteria used to derive evolutionary relationships between species since Darwin till early 1960s.

- The evolutionary relationships derived from these relatively subjective observations were often inconclusive, and some of them were later proved incorrect.

The Giant Panda riddle

- For roughly 100 years scientists were unable to figure out which family the giant panda belongs to.

Giant pandas look like bears but have features that are unusual for bears and typical for raccoons, e.g., they do not hibernate.

In 1985, Steven O’Brien and colleagues solved the giant panda classification problem using DNA sequences and algorithms.
Evolutionary trees: DNA-based approach

- 50 years ago: Emile Zuckerkandl and Linus Pauling brought reconstructing evolutionary relationships with DNA into the spotlight

- In the first few years after Zuckerkandl and Pauling proposed using DNA for evolutionary studies, the possibility of reconstructing evolutionary trees by DNA analysis was hotly debated

- Now it is a dominant approach to study evolution …
Evolutionary trees: DNA-based approach

- 40 years ago: Emile Zuckerkandl and Linus Pauling brought reconstructing evolutionary relationships to the spotlight.
- In the first few years following their proposal, the possibility of reconstructing evolutionary trees by DNA analysis was hotly debated.
- Now it is a dominant approach to study evolution...

From the point of hemoglobin structure, it appears that gorilla is just an abnormal human, or man an abnormal gorilla, and the two species form actually one continuous population.

Emile Zuckerkandl,
Classification and Human Evolution, 1963

From any point of view other than that properly specified, that is of course nonsense. What the comparison really indicate is that hemoglobin is a bad choice and has nothing to tell us about attributes, or indeed tells us a lie.

Gaylord Simpson,
Science, 1964
Molecular evolution

100 million years ago

AGTAGCAGTACGATACG

Today

AGTTGCGAGTAGGATATG
The molecular clock
Genetic distances

- Human to Neanderthal
- Human to Chimpanzee
- Human to Bonobo
- Neanderthal to Chimpanzee
- Neanderthal to Bonobo
- Chimpanzee to Bonobo
A phylogenetic tree

Bonobo    Chimpanzee    Human    Neanderthal
Species trees and Gene trees

The species tree can differ from a tree built from homologous genes.
Species trees and Gene trees

The species tree can differ from a tree built from homologous genes.

Diagram showing the relationships between species and genes, highlighting homologs, orthologs, paralogs, and gene duplication events.
Species trees and Gene trees

Human speciation

[Diagram showing the evolutionary relationships between human species and their ancestors, including chimpanzees and orangutans.]
Species trees and Gene trees

Human speciation
Species trees and Gene trees
Species trees and Gene trees

Human speciation
Phylogenetic Tree of Life

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

Archaea
- Green Filamentous bacteria
- Gram positives
- Methanosarcina
- Methanobacterium
- Methanococcus
- T. celer
- Thermoproteus
- Pyrobaculum

Eucaryota
- Entamoebae
- Slime molds
- Animals
- Fungi
- Plants
- Ciliates
- Flagellates
- Trichomonads
- Microsporidia
- Diplomonads

A speculatively rooted tree for rRNA genes
**Terminology**

**Labeling / annotation**
leaves (and nodes) are labelled with species/objects (1, 2, ..., \(n\))
edge lengths reflect amount of evolution/time along the edge

**Tree structure / topology**
rooted vs. unrooted
binary (bifurcating) vs. non-binary (multifurcating)
Distances in trees

\[ d_{1,4} = 12 + 13 + 14 + 17 + 13 = 69 \]
... these two rooted binary trees are considered equal ...

Layout don’t t matter
Newick format

The following tree:

```
0.1
F
  0.2
  B
  0.5
  E
    0.3
    C
    0.4
    D
```

could be represented in several ways

1. `(),()`, no nodes are named
2. `([A, B, (C, D)])`, leaf nodes are named
3. `([A, B, (C, D)]E)F;`, all nodes are named
4. `((:0.1,:0.2,:0.3,:0.4):0.5);`, all but root node have a distance to parent
5. `([A:0.1, B:0.2, (C:0.3, D:0.4):0.5]);`, distances and leaf names (popular)
6. `([A:0.1, B:0.2, (C:0.3, D:0.4)E:0.5]F;`, distances and all names
7. `((E:0.2, (C:0.3, D:0.4)E:0.5)F:0.1)A;`, a tree rooted on a leaf node (rare)

http://en.wikipedia.org/wiki/Newick_format
http://biopython.org/wiki/Phylo
http://www.birc.au.dk/~mailund/newick.html
Reconstructing trees

Reconstruct the unknown true tree from info about the species

Character-based information

Each species is characterized by a number of characters which can be in different states, e.g. a multiple alignment:

- species 1: acgcta--cacacagt
- species 2: acgccca---acg-agt
- species 3: acg--attt-agc--t

Distance-based information

The distance $D(i,j)$ is given for each pair of species
Distance based reconstruction

The “ideal” case

The tree reflects the distance matrix, i.e. the distances induced by the tree are equal to the distances in the matrix.

If this is possible, then the distance matrix is called additive.

Reconstruction problem

Find both the tree topology and the edge lengths ...
Distance based reconstruction

**Goal:** Reconstruct an evolutionary tree from a distance matrix

**Input:** $n \times n$ distance matrix $D_{ij}$

**Output:** weighted tree $T$ with $n$ leaves fitting $D$

If $D$ is additive, this problem has an efficient solution (which we will come back to), otherwise we should aim at finding a tree which “a good as possible” …
A reconstruction method

**Least square methods**

Given a tree topology $T$, select edge lengths such that

$$Q(T) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (D_{ij} - d_{ij})^2$$

is minimized, where $D_{ij}$ is given by the matrix, $d_{ij}$ is the distance induced by the tree and the selected edge lengths, and $w_{ij}$ is the “confidence” we have in $D_{ij}$, e.g. $w_{ij}=1$.

Optimal edge lengths can be found by standard least square methods in time $O(n^3)$. The “large” version of the problem when the tree is unknown, e.g. we must also minimize over $T$, is NP-complete ...
Two practical algorithms

**UPGMA** constructs good trees (rooted binary trees) if a “molecular clock” is thought to be a reasonable assumption, if the distance matrix comes from a clocklike tree, then UPGMA finds the tree that is the exact representation of the matrix.

**Neighbor-Joining** constructs trees (unrooted binary trees) that can be seen as a rough approximations to least-square methods, if the distance matrix is additive, then Neighbor-Joining finds the tree that is the exact representation of the matrix.

Check Dendroscope for implementations of various tree reconstruction methods and visualizations:

http://ab.inf.uni-tuebingen.de/software/dendroscope/
## Counting rooted binary trees

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
</tr>
<tr>
<td>11</td>
<td>654,729,075</td>
</tr>
<tr>
<td>12</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>13</td>
<td>316,234,143,225</td>
</tr>
<tr>
<td>14</td>
<td>7,905,853,580,625</td>
</tr>
<tr>
<td>15</td>
<td>213,458,046,676,875</td>
</tr>
<tr>
<td>16</td>
<td>6,190,283,353,629,375</td>
</tr>
<tr>
<td>17</td>
<td>191,898,783,962,510,625</td>
</tr>
<tr>
<td>18</td>
<td>6,332,659,870,762,850,625</td>
</tr>
<tr>
<td>19</td>
<td>221,643,095,476,699,771,875</td>
</tr>
<tr>
<td>20</td>
<td>8,200,794,532,637,891,559,375</td>
</tr>
<tr>
<td>30</td>
<td>$4.9518 \times 10^{38}$</td>
</tr>
<tr>
<td>40</td>
<td>$1.00985 \times 10^{87}$</td>
</tr>
<tr>
<td>50</td>
<td>$2.75292 \times 10^{78}$</td>
</tr>
</tbody>
</table>
Counting rooted binary trees

A rooted binary tree with $n$ leaves has $2n-2$ edges. A new species can be added to any edge or "above" the root, i.e:

$$R(n) = R(n-1) (2(n-1)-2 + 1) = R(n-1) (2n - 3)$$

What about unrooted trees?

What about non-binary trees?