Introduction to Phylogenetics

Workshop on Molecular Evolution 2019

Marine Biological Lab, Woods Hole, MA. USA

Mark T. Holder
University of Kansas
1. phylogenetics is crucial for comparative biology

2. tree terminology

3. why phylogenetics is difficult

4. parsimony

5. distance-based methods

6. theoretical basis of multiple sequence alignment
Part #1: phylogenetics is crucial for biology

<table>
<thead>
<tr>
<th>Species</th>
<th>Habitat</th>
<th>Photoprotection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>terrestrial</td>
<td>xanthophyll</td>
</tr>
<tr>
<td>2</td>
<td>terrestrial</td>
<td>xanthophyll</td>
</tr>
<tr>
<td>3</td>
<td>terrestrial</td>
<td>xanthophyll</td>
</tr>
<tr>
<td>4</td>
<td>terrestrial</td>
<td>xanthophyll</td>
</tr>
<tr>
<td>5</td>
<td>terrestrial</td>
<td>xanthophyll</td>
</tr>
<tr>
<td>6</td>
<td>aquatic</td>
<td>none</td>
</tr>
<tr>
<td>7</td>
<td>aquatic</td>
<td>none</td>
</tr>
<tr>
<td>8</td>
<td>aquatic</td>
<td>none</td>
</tr>
<tr>
<td>9</td>
<td>aquatic</td>
<td>none</td>
</tr>
<tr>
<td>10</td>
<td>aquatic</td>
<td>none</td>
</tr>
</tbody>
</table>

slides by Paul Lewis
Phylogeny reveals the events that generate the pattern.

1 pair of changes. Coincidence?

5 pairs of changes. Much more convincing.
Many evolutionary questions require a phylogeny

- Determining whether a trait tends to be lost more often than gained, or vice versa
- Estimating divergence times (Tracy Heath)
- Distinguishing homology from analogy
- Inferring parts of a gene under strong positive selection (Joe Bielawski and Belinda Chang)
Part 2: Tree terminology

- **interior node** (or vertex, degree 3+)
- **terminal node** (or leaf, degree 1)
- **branch** (edge)
- **root node of tree** (degree 2)
- **split** (bipartition)
  - also written $AB|CDE$
  - or portrayed **---**
Monophyletic groups ("clades"): the basis of phylogenetic classification

black state = a synapomorphy
white state = a plesiomorphy
Paraphyletic

grey state is an autapomorphy

Polyphyletic

(images from Wikipedia)
Branch rotation does not matter
\textit{ingroup}: the focal taxa

\textit{outgroup}: the taxa that are more distantly related. Assuming that the ingroup is monophyletic with respect to the outgroup can root a tree.
Warning: software often displays unrooted trees like this:

```
/--------------------------------- Chara
  |
  |
  /----------------------------- Chlorella
  |
  /------16
  |
  "--------------------- Volvox
+---------------------17
  |
  "------------------ Anabaena

/------------------------ Conocephalum
  |
  |
  /--------------------- Bazzania
  |
  "------------------ Anthoceros

  |
  |
  /----------------- Osmunda
  |
  /------18
  |
  "-------------------- Asplenium

  |
  |
  /------- Ginkgo
  |
  /------19
  |
  "------------------ Picea

  |
  |
  /------- Iris
  |
  /-----20
  |
  "------------------ Zea

  |
  |
  /------- Nicotiana
  |
  /------21
  |
  "------------------ Lycopodium
```
Part 3: Phylogenetics is difficult

a. Many types of trees - species trees vs “gene trees” – coalescents or “gene family trees”

b. Many sources of error

c. No clean sampling theory that gives us clean hypothesis tests

d. Computational + statistical difficulties
(3a) Many types of trees: cellular genealogies

Figure 1 from DeWett et al. 2018
(3a) Many types of trees: genealogies in a population

Present

Past
(3a) Many types of trees: genealogies in a population

Present

Past
(3a) Many types of trees: genealogies in a population

Present

Past
(3a) Many types of trees: genealogies in a population
Biparental inheritance would make the picture messier, but the genealogy of the gene copies would still form a tree (if there is no recombination).
It is tempting to refer to the tips of these gene trees as alleles or haplotypes.

- **allele** – an alternative form a gene.
- **haplotype** – a linked set of alleles

But both of these terms require a differences in sequence.

The gene trees that we draw depict genealogical relationships – regardless of whether or not nucleotide differences distinguish the “gene copies” at the tips of the tree.
(thanks to Peter Beerli for the images - next 3 slides)
(3a) A “gene tree” within a species tree
• coalescence – merging of the genealogy of multiple gene copies into their common ancestor. “Merging” only makes sense when viewed backwards in time.

• “deep coalescence” or “incomplete lineage sorting” refer to the failure of gene copies to coalesce within the duration of the species – the lineages coalesce in an ancestral species
(3a) Inferring a species tree while accounting for the coalescent

Figure 2 from Heled and Drummond (2010) *BEAST
See also the recent work by Huw Ogilvie and colleagues on StarBEAST2.
(3a) Considering coalescent effects without modeling gene trees

PoMo model

SVDQuartets
(Kubatko + Swofford)

Figure 1 from De Maio et al. (2015)
Many types of tree: A “gene family tree”

Opazo, Hoffmann and Storz
“Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals”
PNAS 105(5) 2008
Materials and Methods

specific transcriptional regulation. parallel or convergent evolution of protein function and stage-
mentally regulated system of hemoglobin synthesis entailed be important to assess whether the reinvention of a develop-
of therian mammals. If this proves to be the case, then it will also

sequences from fish (\textit{Danio rerio}) and placental mammals, a pair of

and 1 amphibian species. We also included sequences from shorter records

of therian mammals. The

DNA Sequence Data.

that these genes are not 1:1 orthologs of the

subclass Prototheria. We use the ''P'' superscript to acknowledge

-globin gene family from the High Throughput Genomic Se-

-globin gene that occurred before the divergence of birds and mammals but after the amniote/amphibian split. The

-globin gene cluster in 36 mammalian species, 1 bird species,

We obtained genomic DNA sequences for structural

and 1 amphibian species. We also included sequences from shorter records

of therian mammals.

We estimated phylogenetic relationships among the

server (http://phylogenomics.berkeley.edu), with manual adjustments per-

gram MUSCLE (39) as implemented in the Berkeley Phylogenomics group web

nih.gov/blast/bl2seq). Sequence alignment was carried out by using the pro-

the program Genscan (37) and by comparing known exon sequences to

the program BLAST 2, version 2.2 (38), available from

the National Center for Biotechnology Information web site (www.ncbi.nlm.

frequencies remained

trees sampled after convergence. The average standard deviation of split

iterations of a Markov Chain Monte Carlo algorithm, with eight simultaneous

position. Two simultaneous independent runs were performed for 30

independent GTR

approach as implemented in Mr.Bayes v3.1.2 (40). The program was used to

We identified globin genes in unannotated genomic sequences by using

- and

- globin) originated via duplication of a proto

- and

P

- globin) originated via duplication of a proto

- globin and the adult-expressed

- globin gene is expressed during adulthood, and, on the basis of positional

We estimated phylogenetic relationships among the

and

opposite strand. Two simultaneous independent runs were performed for 30

independent GTR

approach as implemented in Mr.Bayes v3.1.2 (40). The program was used to

Phylogenetic Analyses.

The\(\varepsilon\)-globin and\(\beta\)-globin genes arose via duplication of a proto\(\beta\)-globin gene in the ancestor of therian mammals.

The\(\varepsilon^P\)- and\(\beta^P\)-globin genes arose via duplication of a proto\(\beta\)-globin gene in the monotreme lineage.

Opazo, Hoffmann and Storz “Genomic evidence for independent origins of\(\beta\)-like globin genes in monotremes and therian mammals” PNAS \textbf{105}(5) 2008
• duplication – the creation of a new copy of a gene within the same genome.

• homologous – descended from a common ancestor.

• paralogous – homologous, but resulting from a gene duplication in the common ancestor.

• orthologous – homologous, and resulting from a speciation event at the common ancestor.

(Emily Jane McTavish’s and Laura Eme’s lectures)
Joint estimation of gene duplication, loss, and species trees using PHYLDOG

Figure 2A from Boussau et al. (2013)
(3a) Many types of trees:

<table>
<thead>
<tr>
<th></th>
<th>The cause of splitting</th>
<th>Important caveats</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Gene tree” or “a coalescent”</td>
<td>DNA replication</td>
<td>recombination is usually ignored</td>
</tr>
<tr>
<td>Species tree Phylogeny</td>
<td>speciation</td>
<td>recombination, hybridization, lateral gene transfer, and deep coalescence cause conflict in the data we use to estimate phylogenies</td>
</tr>
<tr>
<td>Gene family tree</td>
<td>speciation or duplication</td>
<td>recombination (e.g. domain swapping) is not tree-like</td>
</tr>
</tbody>
</table>
(3a) Joint estimation of gene duplication, loss, and coalescence with DLCoalRecon

Figure 2A from Rasmussen and Kellis (2012)
(3a) DL models and coalescence

Figure 2B from Rasmussen and Kellis (2012)
Many types of trees: Lateral Gene Transfer

*tree* - a graph without cycles (loops)

*network* - general graph; cycles allowed

Cycles can represent

- lateral ("horizontal") gene transfer,
- hybridization between species,
- introgression between populations.
(3a) Many types of trees: Lateral Gene Transfer

Figure 2c from Szöllősi et al. (2013)
They used 423 single-copy genes in $\geq 34$ of 36 cyanobacteria

They estimate:

2.56 losses/family

2.15 transfers/family

$\approx 28\%$ of transfers between non-overlapping branches
Figure 4 from Noutahi et al. (2016)
Part 3: Phylogenetics is difficult

a. Many types of trees - species trees vs "gene trees" – coalescents or "gene family trees"

b. Many sources of error

c. No clean sampling theory that gives us clean hypothesis tests

d. Computational + statistical difficulties
(3b) sources of error cartoon

- Errors modeling multiple hits
- Assembly + asc. bias errors
- Alignment errors
- Paralogy errors
- Errors from deep coalescence
- No signal
- Deep time
- Now
- Amount of anxiety

Diagram showing the relationship between different sources of error and their effects over time.
Figure 1 from Liu et al. (2010)

Figure 2 from Hahn and Nakhleh (2016)
Hypothesis testing in phylogenetics is tricky

- complex literature on frequentists tests of topology (Holder last day)
- bootstrapping - examining effects of sampling error using resampling via computer
- Bayesian methods (Paul Lewis, John Huelsenbeck, Tracy Heath, and Michael Landis)
The bootstrap

(unknown) true value of $\theta$

(unknown) true distribution

empirical distribution of sample

estimate of $\theta$

Distribution of estimates of parameters

Bootstrap replicates

Slide from Joe Felsenstein
The bootstrap for phylogenies

Original Data

Bootstrap sample #1

Bootstrap sample #2

sample same number of sites, with replacement

T

T(1)

T(2)

(and so on)

Slide from Joe Felsenstein
The majority-rule consensus tree

Trees:

How many times each partition of species is found:

<table>
<thead>
<tr>
<th>Partition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE I BCDF</td>
<td>4</td>
</tr>
<tr>
<td>ACE I BDF</td>
<td>3</td>
</tr>
<tr>
<td>ACEF I BD</td>
<td>1</td>
</tr>
<tr>
<td>AC I BDEF</td>
<td>1</td>
</tr>
<tr>
<td>AEF I BCD</td>
<td>1</td>
</tr>
<tr>
<td>ADEF I BC</td>
<td>2</td>
</tr>
<tr>
<td>ABCE I DF</td>
<td>3</td>
</tr>
</tbody>
</table>

Slide from Joe Felsenstein
(3c) bootstrapping

- http://phylo.bio.ku.edu/mephytis/boot-sample.html
- http://phylo.bio.ku.edu/mephytis/bootstrap.html
Phylogenetics is computationally difficult

Problems:

- Huge number of trees
- Strange geometry of tree space
- Large number of numerical parameters that need to be considered.

Some strategies:

- Pragmatic computational heuristics for tree searching – Bui Quang Minh (Tuesday)
- Markov chain Monte Carlo (Paul Lewis, John Huelsenbeck, Tracy Heath, and Michael Landis - this Sunday and the last Saturday)
Optimality criteria

A rule for ranking trees (according to the data). Each criterion produces a score.

Examples:

• Parsimony (Maximum Parsimony, MP)
• Maximum Likelihood (ML)
• Minimum Evolution (ME)
• Least Squares (LS)
Species 1  C  G  A  C  C  A  G  G  T  .  .  .
Species 2  C  G  A  C  C  A  G  G  T  .  .  .
Species 3  C  G  G  T  C  C  C  G  G  T  .  .  .
Species 4  C  G  G  C  C  T  G  G  T  .  .  .
One of the 3 possible trees:

Species 1  Species 3
Species 2  Species 4
One of the 3 possible trees:

Species 1  Species 3
Species 2  Species 4

Same tree with states at character 6 instead of species names

A  C
A  T
Unordered Parsimony
Things to note about the last slide

• 2 steps was the minimum score attainable.

• Multiple ancestral character state reconstructions gave a score of 2.

• Enumeration of all possible ancestral character states is **not** the most efficient algorithm.
Each character (site) is assumed to be independent

To calculate the parsimony score for a tree we simply sum the scores for every site.

| Species 1 | C | G | A | C | C | A | G | G | T |
| Species 2 | C | G | A | C | C | A | G | G | T |
| Species 3 | C | G | G | T | C | C | G | G | T |
| Species 4 | C | G | G | C | C | T | G | G | T |
| Score     | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 |

Tree 1 has a score of 4
**Considering a different tree**

We can repeat the scoring for each tree.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Species 2</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Species 3</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Species 4</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
</tbody>
</table>

| Score | 0 | 0 | 2 | 1 | 0 | 2 | 0 | 0 | 0 |

Tree 2 has a score of **5**
One more tree

Tree 3 has the same score as tree 2

<table>
<thead>
<tr>
<th>Score</th>
<th>0</th>
<th>0</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>2</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Species 2</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Species 3</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Species 4</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Tree 3 has a score of 5
Tree 1 required the fewest number of state changes (DNA substitutions) to explain the data.

Some parsimony advocates equate the preference for the fewest number of changes to the general scientific principle of preferring the simplest explanation (Ockham’s Razor), but this connection has not been made in a rigorous manner.
Parsimony terms

- *homoplasy* multiple acquisitions of the same character state
  - parallelism, reversal, convergence
  - recognized by a tree requiring more than the minimum number of steps
  - minimum number of steps is the number of observed states minus 1

The parsimony criterion is equivalent to minimizing homoplasy.

Homoplasy is one form of the multiple hits problem. In pop-gen terms, it is a violation of the infinite-alleles model.
In the example matrix at the beginning of these slides, only character 3 is parsimony informative.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Species 2</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Species 3</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Species 4</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Max score</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Min score</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Qualitative description of parsimony

• Enables estimation of ancestral sequences.
• Even though parsimony always seeks to minimize the number of changes, it can perform well even when changes are not rare.
• Does not “prefer” to put changes on one branch over another
• Hard to characterize statistically
  – the set of conditions in which parsimony is guaranteed to work well is very restrictive (low probability of change and not too much branch length heterogeneity);
  – Parsimony often performs well in simulation studies (even when outside the zones in which it is guaranteed to work);
  – Estimates of the tree can be extremely biased.
Long branch attraction


The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).
Long branch attraction

Parsimony is almost guaranteed to get this tree wrong.

```
true
1 3
2 4
inferred
```
Inconsistency

- Statistical Consistency (roughly speaking) is converging to the true answer as the amount of data goes to $\infty$.

- Parsimony based tree inference is *not* consistent for some tree shapes. In fact it can be "positively misleading":
  - "Felsenstein zone" tree

- Methods for assessing confidence (e.g. bootstrapping) will indicate that you should be very confident in the wrong answer.
Long branch attraction tree again

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).
If the data is generated such that:

\[
\Pr \begin{pmatrix} A \\ A \\ G \\ G \end{pmatrix} \approx 0.0003 \quad \text{and} \quad \Pr \begin{pmatrix} A \\ G \\ G \\ A \end{pmatrix} \approx 0.008
\]

then how can we hope to infer the tree \(((1,2),3,4)\)?
Note: \(((1,2),3,4)\) is referred to as Newick or New Hampshire notation for the tree.

You can read it by following the rules:

• start at a node,
• if the next symbol is ‘(’ then add a child to the current node and move to this child,
• if the next symbol is a label, then label the node that you are at,
• if the next symbol is a comma, then move back to the current node’s parent and add another child,
• if the next symbol is a ‘)’, then move back to the current node’s parent.
If the data is generated such that:

\[
\begin{pmatrix}
A \\
A \\
G \\
G
\end{pmatrix}
\approx 0.0003
\quad \text{and} \quad
\begin{pmatrix}
A \\
G \\
G \\
A
\end{pmatrix}
\approx 0.008
\]

then how can we hope to infer the tree \(((1,2),3,4)\)?
Looking at the data in “bird’s eye” view (using Mesquite):
Looking at the data in “bird’s eye” view (using Mesquite):

We see that sequences 1 and 4 are clearly very different.
Perhaps we can estimate the tree if we use the branch length information from the sequences...
Why doesn’t simple clustering work?

Step 1: use sequences to estimate pairwise distances between taxa.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>-</td>
<td>0.46</td>
<td>0.4</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>-</td>
<td>0.7</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
Why doesn’t simple clustering work?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>0.46</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td></td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
Why doesn’t simple clustering work?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.2</td>
<td>0.5</td>
<td><strong>0.4</strong></td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td><strong>0.4</strong></td>
<td><strong>0.4</strong></td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

```
A  B  D
|   |   |   |
A B
```

```
A  B  D
|   |   |   |
A   B
```
Why doesn’t simple clustering work?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>0.46</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>D</td>
<td>-</td>
<td>-</td>
<td></td>
<td>0.7</td>
</tr>
</tbody>
</table>

Tree from clustering
Why doesn’t simple clustering work?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.2</td>
<td>0.46</td>
<td>0.4</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>0.46</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>D</td>
<td>0.4</td>
<td>0.4</td>
<td>0.7</td>
<td>0</td>
</tr>
</tbody>
</table>

Tree from clustering
### Why doesn’t simple clustering work?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
<td>0.</td>
<td>0.46</td>
<td>0.4</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>0.46</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>D</td>
<td>0.4</td>
<td>0.4</td>
<td>0.7</td>
<td>0</td>
</tr>
</tbody>
</table>

Tree from clustering:

![Tree diagram](image)

Tree with perfect fit:

![Tree diagram](image)
Simple clustering methods are sensitive to...

1. differences in the rate of sequence evolution.

2. The “multiple hits” problem. – some sites are affected by more than 1 mutation
Distance-based approaches to inferring trees

• Convert the raw data (sequences) to a pairwise distances

• Try to find a tree that explains these distances.

• *Not* simply clustering the most similar sequences.
<table>
<thead>
<tr>
<th></th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
<th>Species 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Species 2</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Species 3</td>
<td>0.3</td>
<td>0.3</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>Species 4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0</td>
</tr>
</tbody>
</table>
Note that the distance matrix is symmetric.

<table>
<thead>
<tr>
<th></th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
<th>Species 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Species 2</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Species 3</td>
<td>0.3</td>
<td>0.3</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>Species 4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0</td>
</tr>
</tbody>
</table>
so we can just use the lower triangle.

<table>
<thead>
<tr>
<th></th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species 3</td>
<td>0.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Species 4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Can we find a tree that would predict these observed character divergences?
<table>
<thead>
<tr>
<th></th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species 3</td>
<td>0.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Species 4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Can we find a tree that would predict these observed character divergences?
\[
\begin{align*}
p_{12} &= a + b \\
p_{13} &= a + i + c \\
p_{14} &= a + i + d \\
p_{23} &= b + i + c \\
p_{23} &= b + i + d \\
p_{34} &= c + d
\end{align*}
\]

<table>
<thead>
<tr>
<th>data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
If our pairwise distance measurements were error-free estimates of the *evolutionary distance* between the sequences, then we could always infer the tree from the distances.

The evolutionary distance is the number of mutations that have occurred along the path that connects two tips.

We hope the distances that we measure can produce good estimates of the evolutionary distance, but we know that they cannot be perfect.
Intuition of sequence divergence vs evolutionary distance

This can’t be right!

$\text{Evolutionary distance}$

$p$-dist

slide from Paul Lewis
Sequence divergence vs evolutionary distance

\[ p\text{-dist} \]

Evolutionary distance

the \( p\)-dist “levels off”
“Multiple hits” problem (also known as saturation)

- Levelling off of sequence divergence vs time plot is caused by multiple substitutions affecting the same site in the DNA.
- At large distances the “raw” sequence divergence (also known as the \( p \)-distance or Hamming distance) is a poor estimate of the true evolutionary distance.
- Statistical models must be used to correct for unobservable substitutions Paul Lewis (tomorrow)
- Large \( p \)-distances respond more to model-based correction – and there is a larger error associated with the correction.
Number of substitutions simulated onto a twenty-base sequence.
Distance corrections

• applied to distances before tree estimation,
• converts raw distances to an estimate of the evolutionary distance

\[ d = -\frac{3}{4} \ln \left( 1 - \frac{4c}{3} \right) \]

<table>
<thead>
<tr>
<th>“raw” p-distances</th>
<th>corrected distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>( c_{12} )</td>
<td>( d_{12} )</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>( c_{13} )</td>
<td>( d_{13} )</td>
</tr>
<tr>
<td>( c_{23} )</td>
<td>( d_{23} )</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>( c_{14} )</td>
<td>( d_{14} )</td>
</tr>
<tr>
<td>( c_{24} )</td>
<td>( d_{24} )</td>
</tr>
<tr>
<td>( c_{34} )</td>
<td>( d_{34} )</td>
</tr>
</tbody>
</table>
\[ d = -\frac{3}{4} \ln \left( 1 - \frac{4c}{3} \right) \]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

“raw” \( p \)-distances

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.383</td>
<td>0.383</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.233</td>
<td>0.233</td>
<td>0.383</td>
</tr>
</tbody>
</table>

corrected distances
Least Squares Branch Lengths

Sum of Squares = \( \sum_i \sum_j \frac{(p_{ij} - d_{ij})^2}{\sigma_{ij}^k} \)

- minimize discrepancy between path lengths and observed distances

- \( \sigma_{ij}^k \) is used to "downweight" distance estimates with high variance
Least Squares Branch Lengths

Sum of Squares = \[\sum_i \sum_j \frac{(p_{ij} - d_{ij})^2}{\sigma_{ij}^k}\]

- in unweighted least-squares (Cavalli-Sforza & Edwards, 1967): \(k = 0\)

- in the method Fitch-Margoliash (1967): \(k = 2\) and \(\sigma_{ij} = d_{ij}\)
## Poor fit using arbitrary branch lengths

<table>
<thead>
<tr>
<th>Species</th>
<th>$d_{ij}$</th>
<th>$p_{ij}$</th>
<th>$(p - d)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu-Ch</td>
<td>0.09267</td>
<td>0.2</td>
<td>0.01152</td>
</tr>
<tr>
<td>Hu-Go</td>
<td>0.10928</td>
<td>0.3</td>
<td>0.03637</td>
</tr>
<tr>
<td>Hu-Or</td>
<td>0.17848</td>
<td>0.4</td>
<td>0.04907</td>
</tr>
<tr>
<td>Hu-Gi</td>
<td>0.20420</td>
<td>0.4</td>
<td>0.03834</td>
</tr>
<tr>
<td>Ch-Go</td>
<td>0.11440</td>
<td>0.3</td>
<td>0.03445</td>
</tr>
<tr>
<td>Ch-Or</td>
<td>0.19413</td>
<td>0.4</td>
<td>0.04238</td>
</tr>
<tr>
<td>Ch-Gi</td>
<td>0.21591</td>
<td>0.4</td>
<td>0.03389</td>
</tr>
<tr>
<td>Go-Or</td>
<td>0.18836</td>
<td>0.3</td>
<td>0.01246</td>
</tr>
<tr>
<td>Go-Gi</td>
<td>0.21592</td>
<td>0.3</td>
<td>0.00707</td>
</tr>
<tr>
<td>Or-Gi</td>
<td>0.21466</td>
<td>0.2</td>
<td>0.00021</td>
</tr>
<tr>
<td><strong>S.S.</strong></td>
<td><strong>0.26577</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram:

```
       Go
      /  \
    0.1 0.1
   Ch  Gi
  /  \
Hu  Or
```

Values represent branch lengths and fit indices.
Optimizing branch lengths yields the least-squares score

<table>
<thead>
<tr>
<th>Species</th>
<th>$d_{ij}$</th>
<th>$p_{ij}$</th>
<th>$(p - d)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu-Ch</td>
<td>0.09267</td>
<td>0.09267</td>
<td>0.0000000000</td>
</tr>
<tr>
<td>Hu-Go</td>
<td>0.10928</td>
<td>0.10643</td>
<td>0.000008123</td>
</tr>
<tr>
<td>Hu-Or</td>
<td>0.17848</td>
<td>0.18026</td>
<td>0.000003168</td>
</tr>
<tr>
<td>Hu-Gi</td>
<td>0.20420</td>
<td>0.20528</td>
<td>0.000001166</td>
</tr>
<tr>
<td>Ch-Go</td>
<td>0.11440</td>
<td>0.11726</td>
<td>0.000008180</td>
</tr>
<tr>
<td>Ch-Or</td>
<td>0.19413</td>
<td>0.19109</td>
<td>0.000009242</td>
</tr>
<tr>
<td>Ch-Gi</td>
<td>0.21591</td>
<td>0.21611</td>
<td>0.000000040</td>
</tr>
<tr>
<td>Go-Or</td>
<td>0.18836</td>
<td>0.18963</td>
<td>0.000001613</td>
</tr>
<tr>
<td>Go-Gi</td>
<td>0.21592</td>
<td>0.21465</td>
<td>0.000001613</td>
</tr>
<tr>
<td>Or-Gi</td>
<td>0.21466</td>
<td>0.21466</td>
<td>0.000000000</td>
</tr>
</tbody>
</table>

S.S. 0.000033144
Least squares as an optimality criterion

SS = 0.00034

SS = 0.0003314 (best tree)
Failure to correct distance sufficiently leads to poor performance

“Under-correcting” will underestimate long evolutionary distances more than short distances
Failure to correct distance sufficiently leads to poor performance

The result is the classic “long-branch attraction” phenomenon.
Distance methods: pros

- Fast – the FastTree method Price et al. (2009) can calculate a tree in less time than it takes to calculate a full distance matrix!

- Can use models to correct for unobserved differences

- Works well for closely related sequences

- Works well for clock-like sequences
Distance methods: cons

- Do not use all of the information in sequences
- Do not reconstruct character histories, so they not enforce all logical constraints
Outline

1. phylogenetics is crucial for comparative biology

2. tree terminology

3. why phylogenetics is difficult

4. parsimony

5. distance-based methods

6. theoretical basis of multiple sequence alignment
• The goal of MSA is to introduce gaps such that residues in the same column are homologous (all residues in the column descended from a residue in their common ancestor).
Expressing homology detection as a bioinformatics challenge

• The problem is recast as:
  – reward matches (+ scores)
  – penalize rare substitutions (− scores),
  – penalize gaps (− scores),
  – try to find an alignment that maximizes the total score
• Pairwise alignment is tractable

• Most MSA programs use progressive alignment:
  – this reduces MSA to a series of pairwise operations.
  – these algorithms are heuristic. They are not guaranteed to return the optimal solution.
  – the criteria used are not ideal from an evolutionary standpoint (and this has implications for tree inference).
Simultaneous inference of MSA and tree is the most appropriate choice (see Hossain et al., 2015), but is computationally demanding. See: Poisson Indel Process (Bouchard-Côté and Jordan, 2013), Bali-Phy, Handel, AliFritz, and POY software.

Many people filter the automatically generated alignments: GUIDANCE2 (and similar tools) cull ambiguously aligned regions to lower the chance that misalignment leads to errors in downstream analyses.
### BLOSUM 62 Substitution matrix

|     | A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A   | 4  |    |    |    | 5  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| R   | -1 | 5  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| N   | -2 | 0  | 6  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| D   | -2 | -2 | 1  | 6  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| C   | 0  | -3 | -3 | -3 | 9  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Q   | -1 | 1  | 0  | 0  | -3 | 5  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| E   | -1 | 0  | 2  | -4 | 2  | 5  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| G   | 0  | -2 | 0  | -1 | -3 | -2 | -2 | 6  |    |    |    |    |    |    |    |    |    |    |    |    |    |
| H   | -2 | 0  | 1  | -1 | -3 | 0  | 0  | -2 | 8  |    |    |    |    |    |    |    |    |    |    |    |    |
| I   | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4  |    |    |    |    |    |    |    |    |    |    |    |
| L   | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2  | 4  |    |    |    |    |    |    |    |    |    |    |
| K   | -1 | 2  | 0  | -1 | -3 | 1  | -2 | -1 | -3 | -2 | 5  |    |    |    |    |    |    |    |    |    |    |
| M   | -1 | -1 | -2 | -3 | -1 | 0  | -2 | -3 | -2 | 1  | 2  | -1 | 5  |    |    |    |    |    |    |    |    |
| F   | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0  | 0  | -3 | 0  | 6  |    |    |    |    |    |    |    |
| P   | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7  |    |    |    |    |    |    |    |
| S   | 1  | -1 | 1  | 0  | -1 | 0  | 0  | 0  | -1 | -2 | -2 | 0  | -1 | -2 | -1 | 4  |    |    |    |    |    |
| T   | 0  | -1 | 0  | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1  | 1  | 5  |    |    |    |
| W   | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -3 | -1 | 1  | 1  | 4  | -3 | -2 | 11 |    |    |    |
| Y   | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2  | -1 | -1 | -2 | 1  | 3  | -3 | -2 | -2 | -2 | 7  |    |    |
| V   | 0  | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3  | 1  | -2 | 1  | -1 | -2 | -2 | 0  | -3 | -1 | 4  |
|     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
### Scoring an alignment with the BLOSUM 62 matrix

<table>
<thead>
<tr>
<th>Pongo</th>
<th>V</th>
<th>D</th>
<th>E</th>
<th>V</th>
<th>G</th>
<th>G</th>
<th>G</th>
<th>E</th>
<th>L</th>
<th>G</th>
<th>R</th>
<th>L</th>
<th>F</th>
<th>V</th>
<th>V</th>
<th>P</th>
<th>T</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla</td>
<td>V</td>
<td>E</td>
<td>V</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>L</td>
<td>G</td>
<td>R</td>
<td>L</td>
<td>L</td>
<td>I</td>
<td>V</td>
<td>Y</td>
<td>P</td>
<td>S</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>4</td>
<td>2</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td>-6</td>
<td>-3</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>-1</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

The score for the alignment is

\[ D_{ij} = \sum_{k} d_{ij}^{(k)} \]

If \( i \) indicates *Pongo* and \( j \) indicates *Gorilla*. \((k)\) is just an index for the column.

\[ D_{ij} = 12 \]
Scoring an alignment with gaps

If we were to use a gap penalty of -8:

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>D</th>
<th>E</th>
<th>V</th>
<th>G</th>
<th>G</th>
<th>E</th>
<th>L</th>
<th>G</th>
<th>R</th>
<th>L</th>
<th>F</th>
<th>V</th>
<th>V</th>
<th>P</th>
<th>T</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pongo</td>
<td></td>
<td>V</td>
<td>D</td>
<td>E</td>
<td>V</td>
<td>G</td>
<td>G</td>
<td>E</td>
<td>L</td>
<td>G</td>
<td>R</td>
<td>L</td>
<td>F</td>
<td>V</td>
<td>V</td>
<td>P</td>
<td>T</td>
</tr>
<tr>
<td>Gorilla</td>
<td></td>
<td>V</td>
<td>-</td>
<td>E</td>
<td>V</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>L</td>
<td>G</td>
<td>R</td>
<td>L</td>
<td>L</td>
<td>I</td>
<td>V</td>
<td>Y</td>
<td>P</td>
</tr>
<tr>
<td>Score</td>
<td>4</td>
<td>-8</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>-8</td>
<td>0</td>
<td>4</td>
<td>-1</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

By introducing gaps we have improved the score:

\[ D_{ij} = 40 \]
Gap Penalties

Penalizing gaps more heavily than substitutions avoids alignments like this:

Pongo VDEVGGE–LGRLFVVPTQ
Gorilla VDEVGG–DLGRLFVVPTQ
Affine gap penalties are often used to accommodate multi-site indels:

\[ GP = GO + (l)GE \]

where:

- \( GP \) is the gap penalty.
- \( GO \) is the “gap-opening penalty”
- \( GE \) is the “gap-extension penalty”
- \( l \) is the length of the gap
Pairwise alignment costs

- Paul Lewis will explain likelihood tomorrow,
- Additive costs can be justified as approximations to the log of likelihoods if:
  - we can identify the events that must have occurred in generate the data, and
  - we can assign (relative) probabilities based on whether these events are rare or common.
<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>D</th>
<th>E</th>
<th>V</th>
<th>G</th>
<th>G</th>
<th>E</th>
<th>L</th>
<th>G</th>
<th>R</th>
<th>L</th>
<th>-</th>
<th>F</th>
<th>V</th>
<th>V</th>
<th>P</th>
<th>T</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla</td>
<td>V</td>
<td>-</td>
<td>E</td>
<td>V</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>L</td>
<td>G</td>
<td>R</td>
<td>L</td>
<td>L</td>
<td>I</td>
<td>V</td>
<td>Y</td>
<td>P</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td>Score</td>
<td>4</td>
<td>-8</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>-8</td>
<td>0</td>
<td>4</td>
<td>-1</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

**Pongo**

**Gorilla**
\begin{align*}
\text{Pongo} & \quad V \quad D \quad E \quad V \quad G \quad G \quad E \quad L \quad G \quad R \quad L \quad - \quad F \quad V \quad V \quad P \quad T \quad Q \\
\text{Gorilla} & \quad V \quad - \quad E \quad V \quad A \quad G \quad D \quad L \quad G \quad R \quad L \quad L \quad I \quad V \quad Y \quad P \quad S \quad R \\
\text{Score} & \quad 4 \quad -8 \quad 5 \quad 5 \quad 0 \quad 6 \quad 2 \quad 4 \quad 6 \quad 5 \quad 4 \quad -8 \quad 0 \quad 4 \quad -1 \quad 7 \quad 4 \quad 1 
\end{align*}

\text{Pongo} \quad \downarrow \quad \text{Gorilla}

V \leftrightarrow V \quad \mathbb{P}(\text{pos. 1}) = \mathbb{P}(V \leftrightarrow V)
<table>
<thead>
<tr>
<th>Pongo</th>
<th>V D E V G G E L G R L - F V V P T Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla</td>
<td>V - E V A G D L G R L L I V Y P S R</td>
</tr>
<tr>
<td>Score</td>
<td>4 -8 5 5 0 6 2 4 6 5 4 -8 0 4 -1 7 4 1</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{Pongo} & \quad \text{Gorilla} \\
V \leftrightarrow V & \quad \text{P}(\text{pos. } 1 - 2) = \text{P}(V \leftrightarrow V) \\
D \leftrightarrow - & \quad \times \text{P}(D \leftrightarrow -)
\end{align*}
\]
<table>
<thead>
<tr>
<th>Pongo</th>
<th>V</th>
<th>D</th>
<th>E</th>
<th>V</th>
<th>G</th>
<th>G</th>
<th>E</th>
<th>L</th>
<th>G</th>
<th>R</th>
<th>L</th>
<th>-</th>
<th>F</th>
<th>V</th>
<th>V</th>
<th>P</th>
<th>T</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla</td>
<td>V</td>
<td>-</td>
<td>E</td>
<td>V</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>L</td>
<td>G</td>
<td>R</td>
<td>L</td>
<td>L</td>
<td>I</td>
<td>V</td>
<td>Y</td>
<td>P</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td>Score</td>
<td>4</td>
<td>-8</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>-8</td>
<td>0</td>
<td>4</td>
<td>-1</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
P(\text{pos. } 1 - 3) = P(V \leftrightarrow V) \\
\times P(D \leftrightarrow -) \\
\times P(E \leftrightarrow E)
\]
<table>
<thead>
<tr>
<th>Pongo</th>
<th>V</th>
<th>D</th>
<th>E</th>
<th>V</th>
<th>G</th>
<th>G</th>
<th>E</th>
<th>L</th>
<th>G</th>
<th>R</th>
<th>L</th>
<th>-</th>
<th>F</th>
<th>V</th>
<th>V</th>
<th>P</th>
<th>T</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla</td>
<td>V</td>
<td>-</td>
<td>E</td>
<td>V</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>L</td>
<td>G</td>
<td>R</td>
<td>L</td>
<td>L</td>
<td>I</td>
<td>V</td>
<td>Y</td>
<td>P</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td>Score</td>
<td>4</td>
<td>-8</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>-8</td>
<td>0</td>
<td>4</td>
<td>-1</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\ln \mathbb{P}(\text{pos. } 1 - 3) = \ln \mathbb{P}(V \leftrightarrow V) + \ln \mathbb{P}(D \leftrightarrow -) + \ln \mathbb{P}(E \leftrightarrow E)
\]

Pongo \hspace{1cm} Gorilla

\[
\begin{align*}
V & \leftrightarrow V \\
D & \leftrightarrow - \\
E & \leftrightarrow E
\end{align*}
\]
Multiple sequence alignment is an ugly topic in bioinformatics

- Clever programming tricks help, but we still have to rely on *heuristics* – approaches that provide good solutions, but are not guaranteed to find the best solution.

- The additive scoring system suffers from the fact that we do not observe ancestral sequences.
A  PEEKSAVTALWGGKVN--VDEVGG
B  GEEKAAVLALWDKVNEEEEVGG
C  PADKTNVKAAWGKVGAHAGEYGA
D  AADKTNVKAAWSKVGGHAGEYGA
E  EHEWQLVHLHVWAKVEADVAGHGQ

pairwise alignment

A  -
B  .17  -
C  .59  .60  -
D  .59  .59  .13  -
E  .77  .77  .75  .75  -

tree inference

A  PEEKSAVTALWGGKVN--VDEVGG
B  GEEKAAVLALWDKVNEEEEVGG
C  PADKTNVKAAWGKVGAHAGEYGA
D  AADKTNVKAAWSKVGGHAGEYGA
E  EHEWQLVHLHVWAKVEADVAGHGQ

alignment stage
Aligning multiple sequences

B

D

A

C

E

0.1

0.2

0.27

0.15

0.1

0.12

0.09

0.1

Seq-Seq

Seq-Seq

Seq-Group

Group-Group
Imperfect scoring system. Consider one position in a group-to-group alignment:

\[(A, A, G) \leftrightarrow (A, A, L)\]

The sum-of-pairs score for aligning would be:

\[
\frac{4}{9}(A \leftrightarrow A) + \frac{2}{9}(A \leftrightarrow L) + \frac{2}{9}(G \leftrightarrow A) + \frac{1}{9}(G \leftrightarrow L)
\]
But in the context of the tree we might be pretty certain of an \( A \leftrightarrow A \) event

\[
\begin{array}{c}
G \\
\downarrow \\
A \\
\downarrow \\
A \\
\downarrow \\
A \leftrightarrow A \\
\downarrow \\
A \\
\downarrow \\
L
\end{array}
\]

Note: weighted sum-of-pairs would help reflect the effect of ancestry better (but still not perfectly; sum-of-pairs techniques are simply not very sophisticated forms of ancestral sequence reconstruction).
Löytynoja and Goldman (2005) showed most progressive alignment techniques were particularly prone to compression because of poor ancestral reconstruction:
Flagging inserted residues allows PRANK to effectively skip over these positions in the ancestor, producing more phylogenetically-sensible alignments:
Greedy choices leading to failure to find the best alignment

Consider the scoring scheme:
match = 0  mismatch = -3  gap = -7

Guide Tree:  

Sequences:

Sp1  Sp2  Sp3  
GACCGTG  GCCGTAG  GACCGTAG

Sp1  GACCGTG  
Sp2  GCCGTAG  
Sp3  GACCGTAG
Greedy choices leading to failure to find the best alignment

match = 0  mismatch = -3  gap = -7

ungapped1 vs 2

\[
\begin{array}{cccccccc}
\text{Sp1} & G & A & C & C & G & T & G \\
\text{Sp2} & G & C & C & G & T & A & G \\
\end{array}
\]

Score 0 -3 0 -3 -3 -3 0 Total= -12

would be preferred over gapped1 vs 2:

\[
\begin{array}{cccccccc}
\text{Sp1} & G & A & C & C & G & T & - & G \\
\text{Sp2} & G & - & C & C & G & T & A & G \\
\end{array}
\]

Score 0 -7 0 0 0 0 -7 0 Total= -14
Adding a *Sp3* to ungapped1 vs 2:

<table>
<thead>
<tr>
<th>Sp1</th>
<th>G - A</th>
<th>C</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sp2</td>
<td>G - C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>Sp3</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
</tbody>
</table>

This implies 1 indel, and 4 substitutions. Score = -19 *

**If** we had been able to use gapped1 vs 2 then we could have:

<table>
<thead>
<tr>
<th>Sp1</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>-</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sp2</td>
<td>G</td>
<td>-</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>Sp3</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
</tbody>
</table>

score = -14 *

* = “sort of...”
Score = -19 if we count events, but sum of pairs score would differ
Score = -14 if we count events, but sum of pairs score would differ.
Polishing (aka “iterative alignment” can correct some errors caused by greedy heuristics)

1. break the alignment into 2 groups of sequences (often by breaking an edge in the merge tree).

2. realign those 2 groups to each other

3. keep the realignment if it improves the score

Opal also uses random 3-group polishing.
References


