

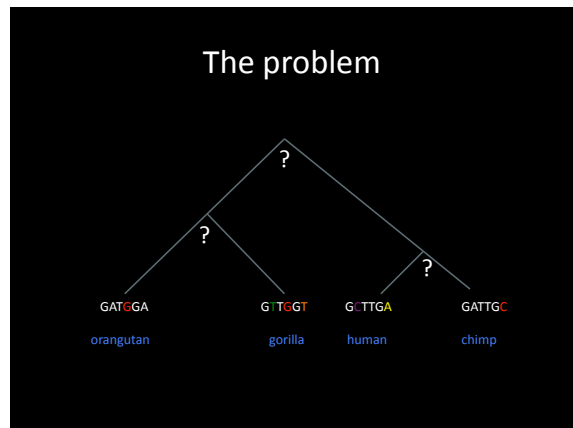
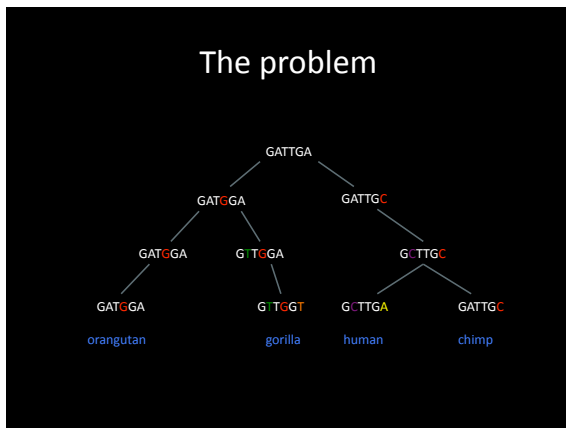
BIOL3004 PSE Elective

Phylogenetic Inference

Steve Chenoweth
116 Goddard building
s.chenoweth@uq.edu.au

TRY TO EXPERIMENT!

- Phylogenetics is a huge field rich in both methods and, for better or for worse, strong opinions.
- Historical arguments were particularly vicious where it perhaps matters the least: classification and taxonomy.
- Felsenstein's fourth great school of classification the: *"It-Doesn't-Matter-Very-Much School"*
- In reality phylogenetic reconstruction applied in a wide range of fields




Methods

1. Distance Methods
 - UPGMA (assumes molecular clock)
 - Neighbour-joining (fast, relatively accurate)
2. Parsimony Methods
 - Slower, topology only
3. Likelihood Methods
 - Powerful BUT computationally demanding
4. Bayesian Methods
 - Powerful BUT computationally demanding

Differences between methods

<p>Multiple Sequence Alignment</p> <p>GAACTGCAA GAAACGCAA GACAGCGA GACAGCAA</p>	<p>Model of Sequence Evolution (pairwise distance matrix)</p> <table border="1" style="margin: auto;"> <tr><td>0</td><td>1</td><td>3</td><td>2</td></tr> <tr><td>1</td><td>0</td><td>2</td><td>1</td></tr> <tr><td>3</td><td>2</td><td>0</td><td>1</td></tr> <tr><td>2</td><td>1</td><td>1</td><td>0</td></tr> </table>	0	1	3	2	1	0	2	1	3	2	0	1	2	1	1	0	<p>Tree-clustering algorithm</p> <p>UPGMA Neighbour-Joining</p>
0	1	3	2															
1	0	2	1															
3	2	0	1															
2	1	1	0															
<p>↓</p>																		
<p>Informative vs. noninformative sites search space of possible trees</p> <table border="0" style="margin: auto;"> <tr> <td style="padding-right: 20px;">Parsimony</td> <td>(no assumed model of sequence evolution)</td> </tr> <tr> <td>Likelihood/Bayesian</td> <td>(assume model of sequence evolution)</td> </tr> </table>			Parsimony	(no assumed model of sequence evolution)	Likelihood/Bayesian	(assume model of sequence evolution)												
Parsimony	(no assumed model of sequence evolution)																	
Likelihood/Bayesian	(assume model of sequence evolution)																	

Using PHYLIP to build a NJ tree

- Multiple Sequence Alignment (MAS)
 - GAAGTCAA
 - GAACAACAA From CLUSTALX
 - GACACGCA
 - GACAGCAA
- Model of Protein Evolution
 - 0 1 3 2
 - 1 0 2 1
 - 3 2 0 1
 - 2 1 1 0
 Protodist.exe
- Tree-clustering algorithm
 - 
 Neighbor.exe
- Statistical confidence bootstrapping
 - 
 Seqboot.exe

PHYLIP Issues

- Phylip always outputs a file called **outfile** and sometimes an **outtree** from each program. This will overwrite and replace any previous files with the same name.
- After running any phylip program, immediately rename the outfile (and outtree) files to something unique, and useful; eg mydata.dst, mydata.nei, mydatatree.prs, mydatatree.nei
- These files are all text files, and you can read them using Wordpad or any other text editor.
- The **outtree** files are Newick format you can open in Treeview.

PHYLIP Protodist.exe

```

C:\test\phylip\3.63\exc\protodist.exe
Protein distance algorithm, version 3.63
Settings for this run:
P Use JTT, PAM, PDM, Kimura, categories model? Dayhoff PAM matrix
C Gamma distribution of rates among positions? No
G One category of substitution rates? Yes
W Use weights for positions? No
M Analyze multiple data sets? No
I Input sequence in interleaved? Yes
J Terminal type (IBM PC, RMC, none)? IBM PC
1 Print out the data at start of run No
2 Print indications of progress of run Yes
3
Are those settings correct? (Type Y or the letter for one to change)
    
```

```

4
Sakai_EC51 0.000000 0.162909 0.128193 1.850381
Crod_NleH 0.162909 0.000000 0.173424 1.85800
Sakai_EC50 0.128193 0.173424 0.000000 1.842367
Sflx_OspG 1.850381 1.858000 1.842367 0.000000
    
```

Protein evolution models (PAM)

- Probability of Point Mutation
- Margret Dayhoff (1979)
- Use empirical frequencies of observed amino acid substitutions in real data.
- Expresses distances in units of expected fraction of amino acids changed. 1.0 is 100 PAMS.
- More models and reading Chapter 14 of Felsenstein 2004 (in library)
- Understanding Bioinformatics

PHYLIP – neighbor.exe

```

C:\test\phylip\3.63\exc\neighbor.exe
Neighbor-Joining/UPGMA method version 3.63
Settings for this run:
N Neighbor-joining or UPGMA tree? Neighbor-joining
O Outgroup root? No, use an outgroup species 1
L Lower-triangular data matrix? No
E Upper-triangular data matrix? No
J Subreplicates? No
M Randomize input order of species? Yes (random number seed = 545)
I Analyze multiple data sets? No
P Terminal type (IBM PC, RMC, none)? IBM PC
1 Print out the data at start of run No
2 Print indications of progress of run Yes
3 Print out tree Yes
4 Write out trees onto tree files? Yes
Y to accept these or type the letter for one to change
    
```


PHYLIP NJ tree output

```

myouttree - Notepad
File Edit Format View Help
{Sakai_EC51:1.0,0.06472,(CF)X_OspG:1.76810,Crod_NleH:0.08990):0.01437,Sakai_EC51:1.0,0.06347);
    
```

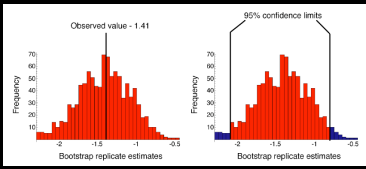

A simple example

Critical Assumption: sample is unbiased



Sage cricket males sometimes offer their hind-wings to females to eat during mating.
Do females who eat hind-wings wait longer to re-mate?

Real data: $T_1 = T_2 = 1.41$		Bootstrap data: $T_1 = T_2 = 1.78$	
Min-estimates	Max-estimates	Min-estimates	Max-estimates
0	1.4	0.7	1.4
0.7	1.6	1.4	2.8
1.4	2.3	1.4	2.8
1.6	2.6	1.8	2.8
1.8	2.9	1.8	3.1
1.9	2.9	1.8	3.1
1.9	3.1	1.9	4.5
2.2	3.9	2.1	4.7
2.1	3.9	2.1	4.7
2.1	4.5	2.1	4.7
	6.7		6.7

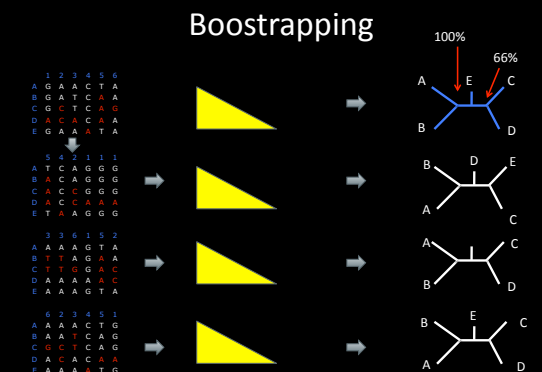


Source Mike Whitlock: <http://www.zoology.ubc.ca/~whitlock/bio300/overheads.html>

Bootstrapping Phylogenies

- From the original multiple alignment, "pseudoreplicate" multiple sequence alignments are created by randomly selecting columns.
- Statistically, these pseudoreplicates are similar, but not identical, to the original.
- For each pseudoreplicate, the tree is calculated.
- For each branch in the original tree, we count how many times the pseudoreplicate trees have the same branch.
- Note that we are repeating the complete analysis multiple times - this can be slow!

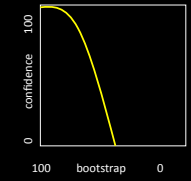
Bootstrapping



100%
66%

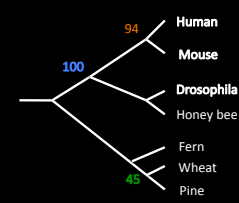
Bootstrapping

- Bootstrap values for phylogenetic trees do not follow typical statistical behaviour
- Bootstrap value 95% : actually close to 100% confidence in that branch
- Bootstrap value 75% : often close to 95% confidence
- Bootstrap value 60% : much lower confidence
- Less than 50% bootstrap: no confidence in that branch over an alternative



Newick format: topology, branch lengths and bootstrap values

```
((Human:0.05, Mouse:0.06)*94*:0.14, (Drosophila:0.11, Honey bee:0.11)*0.09*100*:0.10, (Fern:0.13, (Wheat:0.06, Pine:0.07)*45*:0.07):0.17);
```



NOTE: If using PHYLIP – you may have to edit these manually before displaying

Bootstrapping in PHYLIP

- http://compbio.chemistry.uq.edu.au/mediawiki/index.php/Phylogenetic_tree
- I can help in the workshops
- Easier in MEGA

Examining the alignment for unusual regions

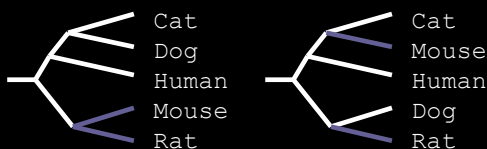
- Expect “families” or clusters of sequences with similar patterns from closely related species. These are often visible in the CLUSTAL alignment.
- However, some sequences may appear part of one family in one region of the alignment, and part of another in another part.
- To test the difference between regions, prepare smaller alignments containing the separate regions, and construct the distance tree with bootstrap values for each.

Some Nasty Stuff

- Homoplasy
 - Recurrent mutation back to ancestral state
- Long Branch Attraction – “The Felsenstein Zone”
 - Check levels of divergence, sanity checks
- Among-site evolutionary rate variation
 - Correct distances using gamma distribution

Parsimony analysis

- The principle of parsimony is to find the tree which has the smallest number of changes required to get the multiple sequence alignment.
- Eg a position where cat, dog and human have Glycine and mouse and rat have Threonine



Parsimony in Phylip

- You need an output file from clustal for phylip input (.phy)
- The phylip program is called **protpars** (**dnapars**)
- The program calculates the parsimony score for a given tree, then tries other trees and sees if the score is improved.
- Parsimony (particularly for many sequences) is typically slower than distance methods.

```
Protein parsimony algorithm, version 3.6
Setting for this run:
U          Search for best tree? Yes
J          Randomize input order of sequences? No, use input order
O          Outgroup root? No, use as outgroup species 1
T          Use Threshold parsimony? No, use ordinary parsimony
```

Other Tools

- Mega4 - Nei and Kumar menu driven, all in one installed on ILC computers <http://www.megasoftware.net>
- TREE-PUZZLE – free “quartet puzzling” by maximum likelihood
- RAXXL <http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm>
- Comprehensive list at Joe Felsenstein’s page: <http://evolution.genetics.washington.edu/phylip/software.html>
- Inferring Phylogenies 2004. Joe Felsenstein – an excellent reference