**Phylogenetics**

BIOL3004 electives

**Evolution**

The aim is to reconstruct the evolutionary history and relationships among the sequences you have collected using blast and clustal

- Human
- Mouse
- Drosophila
- Honey bee
- Fern
- Wheat
- Pine

**Sequences evolve over time**

- Changes include point mutations and insertions and deletions.
- Changes are subject to natural selection - the differences you observe are the combination several processes
  - Speciation creates branches - the sequences are now in two different species, and can evolve separately
  - Genes can be duplicated within the genome, and the two copies become distinct - eg specialised functions.
  - Genes can be lost from a lineage
  - Genes are sometimes transferred between lineages

**Sequence evolution terminology**

- Sequence similarity is a measure of identical or conserved amino acids
- Sequence homology indicates descent from a shared common ancestor
- Sequences are orthologous if they not only descend from a common ancestor, but have the same function, unaltered during that time
- Sequences are paralogous if they descend from a common ancestor, but have altered function or are the result of gene duplication.
- (rare) A xenologue is transferred from another species.

**Constructing a phylogenetic tree from a MSA**

- There are three main classes of methods:
  - distance matrix methods (ClustalX, Phylip)
  - parsimony (Phylop)
  - maximum likelihood - too slow and complex for an introductory project
- Additionally:
  - testing reliability of tree (bootstrapping)
  - drawing and presenting trees (TreeView)

**Rooted and unrooted (radial) trees**

The rooted tree includes information about time and origin
**Distance matrix methods**

- The sequences in the MSA are compared to each other pairwise and it is determined how different they are to each other.
- The pairwise distance matrix contains the results of all the comparisons.

<table>
<thead>
<tr>
<th>Species</th>
<th>Distance Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>GTTATCTACATGTATA</td>
</tr>
<tr>
<td>Mouse</td>
<td>ACTGTCACGCGCA</td>
</tr>
<tr>
<td>Rat</td>
<td>ACTGTCACGCGCA</td>
</tr>
<tr>
<td>Dog</td>
<td>AGCTGCACTGACATA</td>
</tr>
<tr>
<td>Cat</td>
<td>AGCTGCACTGACATA</td>
</tr>
</tbody>
</table>

**Constructing a distance tree in ClustalX**

- Make sure the sequences in the window are aligned (either just recently aligned, or by importing a saved alignment).
- Go to the “Tree” menu.
- Turn on the “Correct for multiple substitutions” option (your sequences are unlikely to be similar enough not to need this).
- Run “Draw Tree”
- Output is filename.ph, a tree in New Hampshire (Newick) Format.
- You can also try the other options and compare the trees you get.

**New Hampshire (Newick) notation**

- Evolutionary trees are easiest to work with and understand as diagrams.
- However, not all evolutionary programs can draw trees and diagrams can’t be transferred between programs.
- New Hampshire format is the format used for text description of trees and to allow easy movement between programs.
- It is useful to be familiar with this format and be able to translate between it and tree diagrams.

**New Hampshire (Newick) format for trees**

```
(((Human, Mouse), (Drosophila, Honey bee)),
(Fern, (Wheat, Pine)));
```

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

[Tree Diagram]
**TreeView**

- Use TreeView to draw a tree in the New Hampshire format.
- TreeView draws four types of trees: radial (unrooted), cladogram, square cladogram and phylogram. Use radial and phylogram, as only these two show branch lengths correctly.

**Parsimony in Phylip**

- You need an output file from clustal for phylip input (*.phy).
- The phylip program is called protpars (dnaps).
- 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.

Proto 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 an outgroup species 1
- T: Use Threshold parsimony? No, use ordinary parsimony
- C: Use which genetic code? Universal
- W: Sites weighted? No
- A: Analyze multiple data set? No
- I: Input sequences interleaved? Yes
- E: Terminal type (IBM PC, ANSI, none)? ANSI
- P: Print indications of progress of run? Yes
- 5: Print output tree? Yes
- 5: Print out steps in each site? No
- P: Print sequences at all nodes of tree? No
- 5: Write out trees onto tree file? Yes

Are these settings correct? (type Y or the letter for one to change)

**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.
- E.g., a position where cat, dog, and human have Glycine and mouse and rat have Throneine

**Phylip in general**

- Phylip doesn’t have a fancy user interface. In windows, you will get a plain blue window with some text options. Firstly, to enter the name of the input data file (e.g., mydata.phy)
- When you run a phylip program, you will see a menu list of options, and you type in the letter on the left to change that option.
- When there are multiple options for a letter, keep entering that letter to cycle through the options.
- Type “y” once you have changed the required options to run the phylip program.

**Phylip output**

- 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; e.g., mydata.dat, mydata.dat, mydata.dat, mydata.dat.
- 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.
**Distance methods in Phylip**

- You can also construct a neighbor-joining tree using phylip programs.
- The same *.phy* input file is used, but the program to run is protdist (dndist). The outfile contains the distance matrix (intermediate step).
- This outfile becomes the infile to neighbor, which calculates the neighbour-joining tree.
- Phylip allows a number of options, such as the distance method used.

**Testing the quality of a tree**

- You can now compare the distance matrix tree to the parsimony tree. They may not be the same, but are the differences important?
- Is there enough evidence in the MSA to support one branching pattern over another?
- You can “bootstrap” your distance tree to see how reliable different parts of the tree are

---

**Bootstrapping**

- From the original multiple sequence 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!

---

**Bootstrapped tree**

---

**What do the bootstrap values mean?**

- 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

---

**Bootstrapping in ClustalX**

- First, calculate your tree as before (“Draw tree”) producing filename.ph
- Then, select the “Bootstrap tree” option in the Tree menu. Try 100 bootstraps first. If that goes quickly, rerun with 1000 bootstraps.
- The output is called filename.phb, and is also a Newick tree, but it includes the bootstrap values.
- Be careful when bootstrapping. Make sure you are bootstrapping with the same options you drew the tree with, and that the tree is in filename.ph.
- Bootstrapping in Phylip: full details at http://foo.maths.uq.edu.au/twiki/bin/view/Know/PhylipBootstrapping
**Treeview with bootstrap values**
- In ClustalW, Tree menu, Output tree format options, change “Phylip bootstrap position” from BRANCH to NODE.
- Bootstrap tree (filename.phb)
- Import filename.phb into Treeview, displaying as phylogram
- Select display option “Show internal edge labels”
- Ignore the “TRICHOTOMY” at the base of the tree, or edit filename.phb first, deleting it.

**Examining the alignment for unusual regions**
- Expect there to be “families” or clusters of sequences with similar patterns from close 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 in ClustalX.

**References**
- Phylogenetics is a huge field, with a large number of techniques and software
- Methods are covered in more depth in MATH210 and BIOL3014
- Commonly-used software: Phylip  PAUP*  MEGA
- Index to phylogenetic software at http://evolution.genetics.washington.edu/phylip/software.html
- Books: Felsenstein; Li & Graur; Nei & Kumar