The aim of this handout is to present you with some theoretical background that might be helpful for building phylogenetic trees. This handout also provides an step-by-step example of how to build a phylogenetic tree\(^1\).

## Contents

1. What is a phylogenetic Tree

2. What kind of data can be used to build a phylogenetic tree?

3. What methods for building phylogenetic trees exists?

4. Advantaged and disadvantages of specific methods
   - 4.1 Distance methods
     - 4.1.1 UPGMA (Unweighted pair-group method using arithmetic averages)
     - 4.1.2 NJ (Neighbour Joining)
   - 4.2 Character-based methods

5. Output

6. Comparing the methods

7. Testing the Topology

8. An Example
   - 8.1 Finding homologous sequences
   - 8.2 Producing a multiple sequence alignment
   - 8.3 Building the phylogenetic tree

9. Further Reading

10. Available tools for building phylogenetic trees

\(^1\)This text has been adopted from

http://www.bioinfo.no/cookbook/sequences/basic-phylogenetic-sequence-analysis
1 What is a phylogenetic Tree

A phylogenetic tree is a statement about the evolutionary relationship between a set of homologous characters of one or several organisms. Homology according to Fitch is the relationship of two characters that have descended, usually with divergence, from a common ancestral character. The characters can be any genic (gene sequence, protein sequence), structural (i.e. morphological) or behavioural feature of an organism.

2 What kind of data can be used to build a phylogenetic tree?

Before the advent of genomic data, it was common to use morphological data. Here, however, we will only consider the building of phylogenetic data based on genic characters i.e. nucleotide or protein sequence data. The user has to first make sure that the sequences that he/she is intending to build a phylogenetic tree for are homologous. In the case of genic data, this means using BLAST or FASTA. The BLAST and FASTA tools, which can be used for searching for similar sequences, are available online via the EBI website and can also be downloaded. It should be emphasised that similarity does not imply homology because of the possibility of homoplasy. However, when similarity is high the similarity implies homology paradigm can be considered to apply. The recipe for BLAST searches contains some rules of thumb for assessing whether two sequences might be homologous given a similarity measure.

Once it has been established that the sequences are homologous, the next step is to compute a multiple sequence alignment.

3 What methods for building phylogenetic trees exists?

There are two main criteria that can be used for classifying phylogenetic tree building methods: the type of data that they use as input and the algorithm that they use for computing the tree.

The data can either be molecular data or it can be distance data:

- Molecular data is essentially the aligned sequences (nucleotide or amino acids)

![Image of aligned sequences]

Figure 1: Example for aligned sequences.

- Distance data is a matrix in which a measure of the evolutionary distance between each pair of sequences in the multiple alignment has been calculated. The main advantage of distance data is that it enables rapid calculation of phylogenetic trees, the main disadvantages are that there is a loss of information when going from molecular data to distance data and that it is not trivial to choose a good distance measure among the many distance measures that exist. The distance measures must among other things compensate for the possibility that there may be multiple substitutions at a particular site of the alignment which, if not corrected for, would result in an underestimation of the evolutionary distance between two sequences.

The algorithms used for building a tree fall into one of three main categories:

\[\text{EBI: http://www.ebi.ac.uk/Tools/homology.html}\]
Figure 2: Example for the computation of a distance matrix from a multiple sequence alignment.

- Distance methods, which can only be used on distance data, are based on the idea of grouping (clustering) together the two sequences that are closest first, recalculating the distances, and the grouping again in an iterative process which stops when all the sequences have been clustered. The order in which the sequences are clustered determines the tree topology.

- Character-based methods, which can be used on both sequence and distance data, are based on the idea of generating all the possible tree topologies with the input sequences and then searching among these trees for the tree that best matches the data given some criteria. The disadvantage of these algorithms are that they are very time consuming since the number of tree topologies grows very rapidly with the number of sequences (see box). There are fortunately heuristic search methods that avoid having to evaluate all the trees but, nevertheless, the need to evaluate many different trees makes the optimality methods considerably more time consuming than the clustering methods. Considerably more time consuming means that for clustering methods the result is ready within seconds whereas for optimality methods the result may easily take several minutes and maybe hours (depending on the number of sequences and their length).

4 Advantaged and disadvantages of specific methods

This section briefly describes the basic concept of each method and any particular limitations or advantages.

4.1 Distance methods

4.1.1 UPGMA (Unweighted pair-group method using arithmetic averages)

- This is the simplest tree building method based on the most simple clustering algorithm.

- Main limitation: the method assumes a constant rate of evolution of the sequences in all branches of the tree (molecular clock assumption). This assumption is very unlikely to hold in particular if the sequences are separated by large evolutionary distances.

- Main advantage: outputs a rooted tree and is extremely fast.

4.1.2 NJ (Neighbour Joining)

- The Neighbour Joining method essentially applies the Minimum Evolution principle (ME) at each step in the clustering process. The name Neighbour Joining refers to the way in which the clustering is performed.

- Main advantage: does not assume that the rate of evolution is the same in all branches of the tree.
4.2 Character-based methods

The main generic advantages of character-based methods is that: they enable the computation of ancestral sequences (i.e. the sequence at interior nodes of the tree) and there is no loss of information (they operate directly on the multiple sequence alignment data).

- **MP (Maximum Parsimony)** The theoretical basis of this method is William of Ockham’s philosophical idea that the best hypothesis to explain a process is the one that requires the smallest set of assumptions, i.e. the most parsimonious is best. Maximum parsimony applied to phylogenetic tree building involves computing the minimum number of substitutions over all sites for each topology. This is referred to as the tree length. The MP tree is the tree with the minimum tree length.
  
  Main advantage: MP is good with very distantly related sequences (information in the few conserved sites tends to disappear if one uses distance methods with distantly related sequences).

- **ML (Maximum Likelihood)** The likelihood of observing a given set of sequence data for a specific substitution model is maximised for each topology and the topology that gives the highest maximum likelihood is chosen as the ML tree.
  
  Main advantage: accurately reconstructing the relationships between sequences that have been separated for a long time, or are evolving rapidly, requires a method that corrects for multiple mutational events at the same site. ML carries out such a correction whereas MP does not. To make this correction ML uses a model of sequence evolution that describes the relative probability of various evolutionary events.
  
  Main limitation: very time consuming, considerably more than MP which is itself more time consuming than clustering methods.

5 Output

The basic output of all methods will be a text file containing a description of the tree. The most widely used format for describing phylogenetic trees is phylip (also called newick or new hampshire - nh). This format is described in detail on the phylip webpages\(^3\). See the boxes for examples.

All the methods with the exception of the UPGMA method will output an unrooted tree. The reason that the tree is unrooted is that the methods are not able to determine where in the tree the evolution that the tree represents started. However, the tree will appear rooted in the output. This rooting is random i.e. the program will just have randomly rooted the tree along one of the branches so as to be able to represent it in newick format (the "Unrooted tree" box illustrates the five different ways of representing in newick format an unrooted tree with 4 sequences).

There are several methods for determining the root of a tree, the most widely accepted of which is to use an outgroup i.e. a sequence that does not belong to the group of interest. For example, a chimpanzee sequence could be included in a set of homologous human DNA sequences, this data can be inputted to a tree building method and then the unrooted output can be correctly rooted with the chimpanzee sequence.

All methods produce some measure of the branch lengths i.e. a measure of the amount of evolution that is estimated to have taken place along each branch of the tree. A simple example of branch lengths in newick format is:((A:1.5,B:1.0):0.5,(C:2.3,D:1.4):2.4);.

For slightly more complex examples of newick format and to get an idea of how these translate into a tree graphically refer to [http://iubio.bio.indiana.edu/treeapp/treeprint-form.html](http://iubio.bio.indiana.edu/treeapp/treeprint-form.html).

6 Comparing the methods

There are five desirable properties a tree-building method should have:

- **Efficiency**: how fast is the method?

• power: how much data does the method need to produce a reasonable result?
• consistency: will it converge on the right answer given enough data?
• robustness: to what extent will violations of the methods assumptions result in poor phylogenies?
• falsifiability: will the method tell us when its assumptions are violated?

There are two of these desirable properties that one can say something immediate about. In terms of efficiency the distance methods (and in particular those that use clustering) tend to be superior to the character-based methods. This is due to the fact that character-based methods have to search among all possible trees for the tree that best fits the data. In terms of consistency, maximum likelihood has been shown in most simulations to be the best method but maximum parsimony and neighbour joining also perform well.

7 Testing the Topology

Once you have built a tree you will want to get an answer to the question: "how strongly does the data support each of the relationships in the tree?". The way to get an answer to this question is to apply a technique called bootstrapping. The columns from the original multiple sequence alignment are randomly re-sampled with replacement to produce pseudo-replicate data sets. For each of these pseudo-replicate data sets a tree is calculated and the topology of this tree is then compared with the original tree. Any interior branch of the original tree which gives the same partition of sequences as that of the bootstrap tree is given value 1 whereas other branches are given 0. The process is repeated for all bootstrapped trees and a percentage is calculated for each branch.

The exact interpretation of the bootstrap is unclear. A higher value is clearly better, but what is a reasonable cut-off? Some workers have concluded that bootstrap proportions are conservative measures of support, so a value of $\geq 70\%$ might indicate strong support for the group.
8 An Example

This section guides you through, step by step, the search for homologous sequences and the building of a phylogenetic tree. If you already have a set of homologous sequences for which you want to build a tree, simply skip the initial steps.

8.1 Finding homologous sequences

We begin by building a set of homologous proteins sequences for which we intend to build a tree. I pick the cellular tumor antigen p53 sequence from rat and feed it into the blastp web interface at the NCBI by pasting it into the main search box and clicking on the "BLAST" button.

>gi|129372|sp|P10361|P53_RAT Cellular tumor antigen p53 (Tumor suppressor p53) MEDOSDMSIELPQSEITCLWKLPPDILPTATGSPVSNMDELFLPOQVAIELEPEGEAQVSAPAA QEPGTEAPVAYAPASATPWP55VSVPQYQNGYFHGFLQSGTAKSVMTYSLINKLFCQLAKTCP VQLVWSTPPPPGTRVAMAYKKQQMTEVVRRCPHERCSDGDDGALPPQHLVREGNYAEYLDROTF RHSVVVEYPEPVEVGSDITTTHYKYMNCN55CMGMNRRPILFITLDEDSSGNLLGDSFENVACPGRDR RTNEENFRKKEHECPELPPSSAKRALPTSTSSSPQKKP2L2EDGEYFTLKRGERFEMFRELNEALELKD ARAAESGD5RAH5SYPTKGGQTSRHKPMMKQVGPSO

Figure 3: A single sequence in fasta format.

We obtain a list of hits (the full list).

<table>
<thead>
<tr>
<th>Score</th>
<th>E</th>
<th>Sequences producing significant alignments:</th>
<th>(bits)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>129372</td>
<td>sp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>13591878</td>
<td>ref</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>19338639</td>
<td>gi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>1963247</td>
<td>gi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>129371</td>
<td>sp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>6755881</td>
<td>ref</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>200201</td>
<td>gi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>5081783</td>
<td>gi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>1537502</td>
<td>gb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>35571</td>
<td>emb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>223382</td>
<td>prf</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>1813451</td>
<td>gi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gi</td>
<td>2897527</td>
<td>gb</td>
</tr>
</tbody>
</table>

Figure 4: Results from a BLAST search.

We scroll down the page and select the sequences that we want to build a phylogenetic tree for (mammalian p53 proteins: P53_RAT, P53_MOUSE, P53_RABIT, P53_MESAU, P53_CRIGR, P53_CERAE, P53_MACMU, P53_HUMAN, P53_MARMO, P53_MACFA, P53_PIG, P53_TUPGB, P53_CAVPO, P53_FELCA, P53_SHEEP, P53_CANFA, P53_BARBU) and click on the "Get selected sequences" button. You don’t need all of these sequences but make sure you include P53_BARBU since this is a teleost fish sequence and we will use it to root the tree.

We obtain a list of the selected sequences. Change the "Display" pop-down from "Summary" to "FASTA" and click on "Display". We obtain the sequences we are interested in FASTA format. Change the "Send to" pop-down from "File" to "Text" and click on "Send to". We obtain a list of the sequences in FASTA format (the full list).

5http://www.bioinfo.no/cookbook/sequences/resolveuid/d6706d1b0af4112fa70fe5c0f31da705
6http://www.bioinfo.no/cookbook/sequences/resolveuid/bc225a2364066d90e7a335f2b2a76b32
8.2 Producing a multiple sequence alignment

Copy the sequences in FASTA format, go to the URL http://www.ebi.ac.uk/clustalw and paste the text into the relevant box. We produce an alignment in Clustal by modifying the "Output format" setting to "phylip", and clicking on the "Run" button. We obtain an alignment (the full file).

Figure 6: Aligned sequences (Only partly shown).

It is a good idea to inspect the alignment before building the tree. We see that p 53 is extremely well conserved. Notice also that the last sequence aligns well with the other sequences but has suffered from some inserts/deletions. This is not surprising since this is a p53 sequence from a fish whereas all the other sequences are from mammals.

You can reuse this alignment with other tree building programs.

8.3 Building the phylogenetic tree

Return to the ClustalW input page and change "Tree type" setting from "none" to "nj". Also modify the "Correct dist." setting from "off" to "on". And then click on the "Run" button.

You get an output divided in several sections. The "Neighbor-Joining Tree" section gives a summary of the distances between the sequences and the order in which the NJ algorithm clustered the sequences to produce the tree. The "Phylip Tree" section gives the tree in newick format.

Copy this text, go to http://iubio.bio.indiana.edu/treeapp/treeprint-form.html, paste the text into the main box and click "Submit". This should produce an unrooted tree.

First note the very long branch leading the P53_BARB node with is the p53 sequence from a teleost fish. This branch is so long due to the high amount of sequence divergence. Remember that it is because of this that we included the sequence in the analysis and it is along this branch that we should root the tree. Second, note the
overlapping names, in particular those in the bottom left hand corner of the picture. The reason for this is that these represent p53 sequences from primates which are so closely related that the branch lengths leading to these nodes are very close to zero. Finally, note how mouse and rat sequence do not appear to be each other's closest relative. This is an indication that p53 is not a very suitable sequence for retrieving the phylogenetic relationship of the species, there has probably been some convergent evolution. It could also be that we would have obtained a tree that better reflected the species relationships had we used another method.

Rooting the tree along the branch leading to the P53_BARB node.
9 Further Reading

- Molecular Evolution, Page and Holmes, Blackwell Science 2002
- Inferring phylogenies, Felsenstein, Sinauer Associates 2003

10 Available tools for building phylogenetic trees

Online phylogenetic tree tools:

- The institut Pasteur has a large amount of phylogenetic tree building software available via the web\(^8\)
- Standard tools for building alignments, can also be used for building phylogenetic trees\(^9\)

Downloadable software for phylogenetic tree building:

- There are very many phylogenetic tree building tools that can be installed on your computer. The two most popular ones are probably: PAUP (http://paup.csit.fsu.edu/index.html) and Phylip (http://evolution.genetics.washington.edu/phylip.html)
- A comprehensive listing of phylogenetic tools (both tree building tools and other phylogenetic analysis tools) is maintained by Joseph Felsenstein\(^{10}\)

---

\(^8\)Phylogeny Software: http://bioweb.pasteur.fr/seqanal/phylogeny/intro-uk.html
\(^9\)Alignments: http://www.ebi.ac.uk/Tools/sequence.html
\(^{10}\)Felsenstein: http://evolution.genetics.washington.edu/phylip/software.html