Phylogenetics

Phylogenetics describes the taxonomical classification of organisms based on their evolutionary history i.e. their *phylogeny*. Phylogenetics is therefore an integral part of the science of *systematics* that aims to establish the phylogeny of organisms based on their characteristics. Furthermore, phylogenetics is central to evolutionary biology as a whole as it is the condensation of the overall paradigm of how life arose and developed on earth.

The phylogenetic tree

A phylogenetic tree is a statement about the evolutionary relationship between a set of homologous characters of one or several organisms. Homology 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. The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 1 shows a proposed phylogeny for the great apes, *Hominidae*, taken in part from Purvis [Purvis, 1995]. The tree consists of a number of nodes (also termed vertices) and branches (also termed edges). These nodes can represent an individual, a species, or a higher grouping and are thus broadly termed taxonomical units. In this case, the terminal nodes (also called leaves or tips of the tree) represent extant species of *Hominidae* and are the *operational taxonomical units* (OTUs). The internal nodes, which here represent extinct common ancestors of the great apes, are termed *hypothetical taxonomical units* since they are not directly observable.

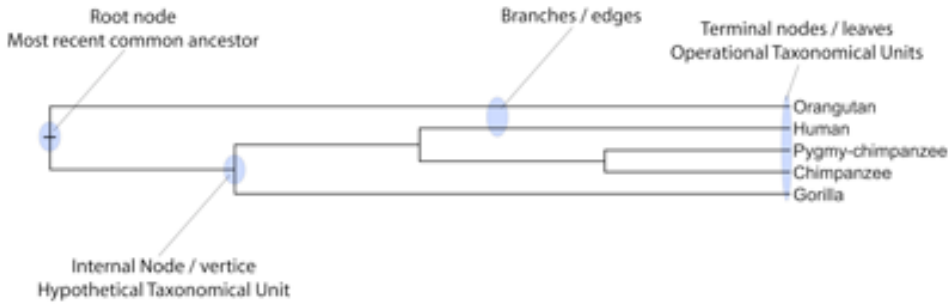


Figure 1: *A proposed phylogeny of the great apes (Hominidae). Different components of the tree are marked, see text for description.*

The ordering of the nodes determine the tree *topology* and describes how lineages have diverged over the course of evolution. The branches of the tree represent the amount of evolutionary divergence between two nodes in the tree and can be based on different measurements. A tree is completely specified by its topology and the set of all edge lengths.

The phylogenetic tree in figure 2.1 is rooted at the most recent common ancestor of all *Hominidae* species, and therefore represents a hypothesis of the direction of evolution e.g. that the common ancestor of gorilla, chimpanzees and human existed before the common ancestor of chimpanzees and human. If this information is absent trees can be drawn as unrooted.

Modern usage of phylogenies

Besides evolutionary biology, and systematics the inference of phylogenies is central to other areas of research.

As more and more genetic diversity is being revealed through the completion of multiple genomes, an active area of research within bioinformatics is the development of comparative machine learning algorithms that can simultaneously process data from multiple species [Siepel and Haussler, 2004]. Through the comparative approach, valuable evolutionary information can

be obtained about which amino acid substitutions are functionally tolerant to the organism and which are not. This information can be used to identify substitutions that affect protein function and stability, and is of major importance to the study of proteins [Knudsen and Miyamoto, 2001]. Knowledge of the underlying phylogeny is, however, paramount to comparative methods of inference as the phylogeny describes the underlying correlation from shared history that exists between data from different species.

In molecular epidemiology of infectious diseases, phylogenetic inference is also an important tool. The very fast substitution rate of microorganisms, especially the RNA viruses, means that these show substantial genetic divergence over the time-scale of months and years. Therefore, the phylogenetic relationship between the pathogens from individuals in an epidemic can be resolved and contribute valuable epidemiological information about transmission chains and epidemiologically significant events [Leitner and Albert, 1999], [Forsberg et al., 2001].

What kind of data do I need 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 (http://www.ebi.ac.uk/Tools/homology.html) 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.

Definition: homoplasy

Similarity caused by convergent evolution from different ancestors, rather than divergent evolution from a common ancestor (homology).

Once it has been established that the sequences are homologous, the next step is to compute a multiple sequence alignment. A standard tool for computing multiple sequence alignments is ClustalW (http://www.ebi.ac.uk/clustalw/) available from EBI both online and for download.

Reconstructing phylogenies from molecular data

Traditionally, phylogenies have been constructed from morphological data but following the growth of genetic information it has become common practice to construct phylogenies based on molecular data, known as *molecular phylogeny*. The data is most commonly in the form of DNA or protein sequences but can also be in the form of e.g. restriction fragment length polymorphism (RFLP).

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)

- 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.

Methods for constructing molecular phylogenies can be distance based or character based.

Distance based methods
Two common algorithms, both based on pairwise distances, are the UPGMA and the Neighbor Joining algorithms. Thus, the first step in these analyses is to compute a matrix of pairwise distances between OTUs from their sequence differences. To correct for multiple substitutions it is common to use distances corrected by a model of molecular evolution such as the Jukes-Cantor model [Jukes and Cantor, 1969].

UPGMA: A simple but popular clustering algorithm for distance data is Unweighted Pair Group Method using Arithmetic averages (UPGMA).[Michener and Sokal, 1957], [Sneath and Sokal, 1973]. This method works by initially having all sequences in separate clusters and continuously joining these. The tree is constructed by considering all initial clusters as leaf nodes in the tree, and each time two clusters are joined, a node is added to the tree as the parent of the two chosen nodes. The clusters to be joined are chosen as those with minimal pairwise distance. The branch lengths are set corresponding to the distance between clusters, which is calculated as the average distance between pairs of sequences in each cluster.

The algorithm assumes that the distance data has the so-called *molecular clock* property i.e. the divergence of sequences occur at the same constant rate at all parts of the tree. This means that the leaves of UPGMA trees all line up at the extant sequences and that a root is estimated as part of the procedure.

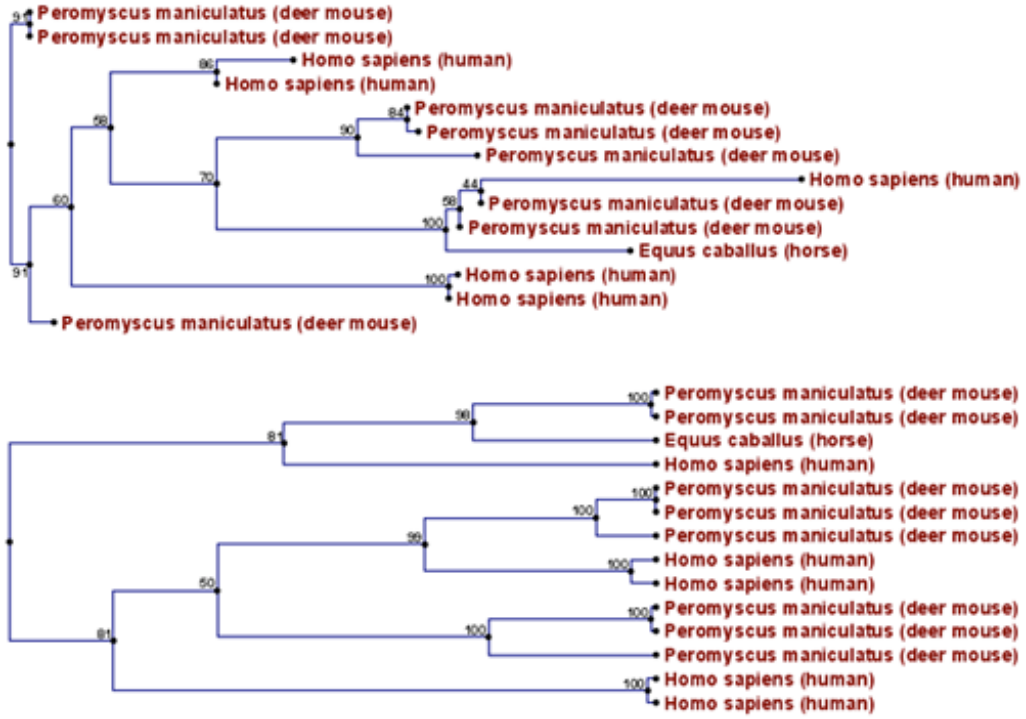


Figure 2.2: *Algorithm choices for phylogenetic inference. The top shows a tree found by the neighbor joining algorithm, while the bottom shows a tree found by the UPGMA algorithm. The latter algorithm assumes that the evolution occurs at a constant rate in different lineages.*

Neighbor Joining: The neighbor joining algorithm,[Saitou and Nei, 1987], on the other hand, builds a tree where the evolutionary rates are free to differ in different lineages. I.e., the tree does not have a particular root. Some programs always draw trees with roots for practical reasons, but

for neighbor joining trees, no particular biological hypotheses is postulated by the placement of the root. The method works very much like UPGMA. The main difference is that instead of using pairwise distance, this method subtracts the distance to all other nodes from the pairwise distance. This is done to take care of situations where the two closest nodes are not neighbors in the ''real'' tree. The neighbor join algorithm is generally considered to be the fairly good and is widely used. Algorithms that improves its cubic time performance exist. The improvement is only significant for quite large datasets.

Character based methods
Whereas the distance based methods compress all sequence information into a single number, the character based methods attempt to infer the phylogeny based on all the individual characters (nucleotides or amino acids).

Parsimony: In parsimony based methods a number of sites are defined which are informative about the topology of the tree. Based on these, the best topology is found by minimizing the number of substitutions needed to explain the informative sites. Parsimony methods are not based on explicit evolutionary models.

Maximum Likelihood: Maximum likelihood and Bayesian methods (see below) are probabilistic methods of inference. Both have the pleasing properties of using explicit models of molecular evolution and allowing for rigorous statistical inference. However, both approaches are very computer intensive.

A stochastic model of molecular evolution is used to assign a probability (likelihood) to each phylogeny, given the sequence data of the OTUs. Maximum likelihood inference [Felsenstein, 1981] then consists of finding the tree which assign the highest probability (likelihood) to the data.

Bayesian inference: The objective of Bayesian phylogenetic inference is not to infer a single "correct" phylogeny, but rather to obtain the full posterior probability distribution of all possible phylogenies. This is obtained by combining the likelihood and the prior probability distribution of evolutionary parameters. The vast number of possible trees means that bayesian phylogenetics must be performed by approximative Monte Carlo based methods.[Larget and Simon, 1999], [Yang and Rannala, 1997].

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 (http://evolution.genetics.washington.edu/phylip/newicktree.html). 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 inputed 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.

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 optimality criterion methods. This is due to the fact that optimality 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.

Interpreting phylogenies

Bootstrap values
A popular way of evaluating the reliability of an inferred phylogenetic tree is bootstrap analysis. The first step in a bootstrap analysis is to re-sample the alignment columns with replacement. I.e., in the re-sampled alignment, a given column in the original alignment may occur two or more times, while some columns may not be represented in the new alignment at all. The re-sampled alignment represents an estimate of how a different set of sequences from the same genes and the same species may have evolved on the same tree.

If a new tree reconstruction on the re-sampled alignment results in a tree similar to the original one, this increases the confidence in the original tree. If, on the other hand, the new tree looks very different, it means that the inferred tree is unreliable. By re-sampling a number of times it is possibly to put reliability weights on each internal branch of the inferred tree. If the data was bootstrapped a 100 times, a bootstrap score of 100 means that the correspond branch occurs in all 100 trees made from re-sampled alignments. Thus, a high bootstrap score is a sign of greater reliability.

What are the tools that are available for building a phylogenetic tree?

Online phylogenetic tree tools:

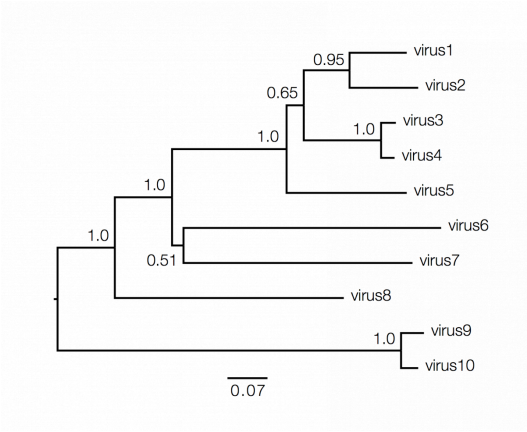
- The institut Pasteur has a large amount of phylogenetic tree building software available via the web (http://bioweb.pasteur.fr/seqanal/phylogeny/intro-uk.html)
- ClustalO, a standard tool for building alignments, can also be used for building phylogenetic trees (http://www.ebi.ac.uk/clustalw/)

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 (http://evolution.genetics.washington.edu/phylip/software.html)

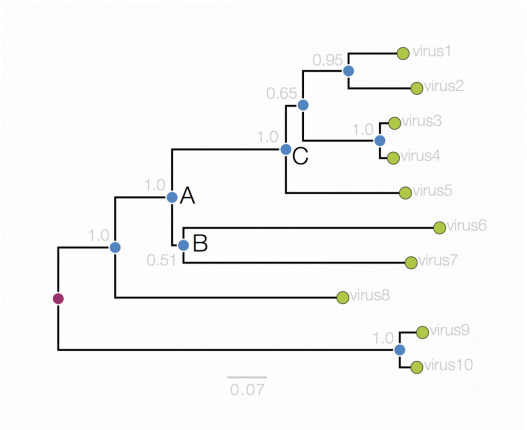
How to read a phylogenetic tree

Phylogenetics trees contain a lot of information about the inferred evolutionary relationships between entities such as a set of viruses. Decoding that information is not always straightforward and requires some understanding of the elements of a phylogeny and what they represent. Here is an example (fictional) phylogeny as it may be presented in a journal article:



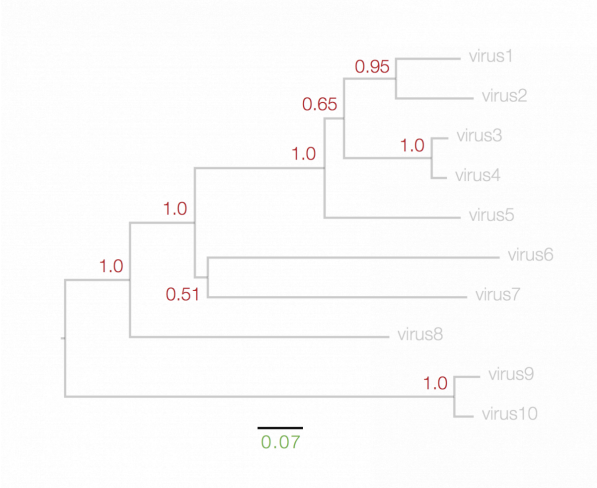
What information does the tree contain?

We can start with the dimensions of the figure. In this figure the horizonal dimension gives the amount of genetic change. The horizonal lines are branches and represent evolutionary lineages changing over time. The longer the branch in the horizonal dimension, the larger the amount of change. The bar at the bottom of the figure provides a scale for this. In this case the line segment with the number '0.07' shows the length of branch that represents an amount genetic change of 0.07. The units of branch length are usually nucleotide substitutions per site – that is the number of changes or 'substitutions' divided by the length of the sequence (although they may be given as % change, i.e., the number of changes per 100 nucleotide sites). The vertical dimension in this figure has no meaning and is used simply to lay out the tree visually with the labels evenly spaced vertically. The vertical lines therefore simply tell you which horizontal line connects to which and how long they are is irrelevent.



Next, we will consider tree structure itself. This can be broken down into nodes (represented in the tree, above, as circles) and branches (the lines connecting them). There are two types of nodes; external nodes, also called 'tips' or 'leaves' (you can only take the tree metaphor so far and I prefer the term 'tip'), and internal nodes. The tips are shown here with green circles and these represent the actual viruses sampled and sequenced. These are our data and we usually know information about these, beyond the actual sequence, such as when they were collected, what host they were in, where that host was found, clinical features of the disease. The internal nodes are represented by blue circles and these represent putative ancestors for the sampled viruses. Ancestors in this context is an infected host at

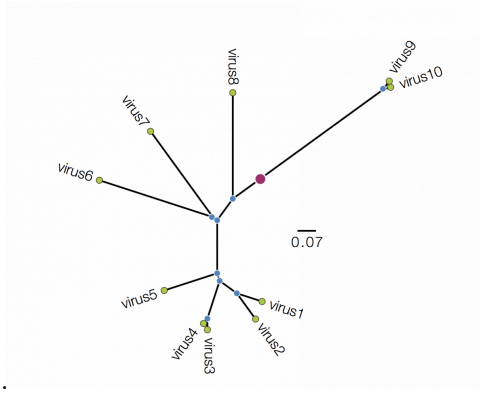
sometime in the past that in turn infected 2 or more new hosts producing chains of infections that lead to the sampled viruses. The branches then represent this chain of infections. This tree is rooted which suggests we know where the ultimate common ancestor of all the sampled viruses was (the red circle). Knowing this gives the tree an order of branching events in the horizonal dimension: Ancestor 'A' exists prior to ancestors 'B' and 'C' and time is approximately flowing from left to right. I say 'approximately' because in this tree the horizonal axis is measured as genetic change and to convert this into actual time we need to make some assumptions about the relationship between genetic change and time. These assumptions are referred to as the 'molecular clock' and I will discuss this below.



The numbers next to each node, in red, above, represent a measure of support for the node. These are generally numbers between 0 and 1 (but may be given as percentages) where 1 represents maximal support. These can be computed by a range of statistical approaches including 'bootstrapping' and 'Bayesian posterior probabilities'. The details of what technique was used will be in the figure legend. A high value means that there is strong evidence that the sequences to the right of the node cluster together to the exclusion of any other.

Trees are sometimes drawn in other ways. Both these figures are representations of the same underlying tree as above:

A.

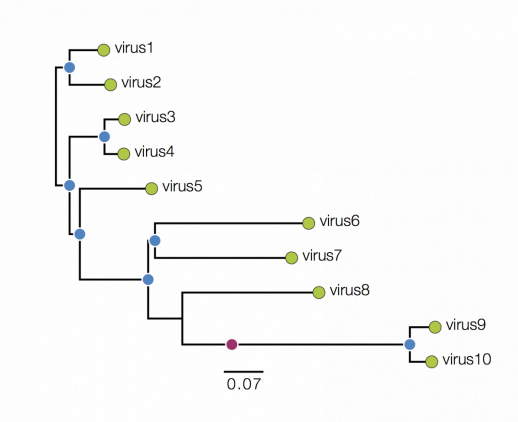


B.

Tree A is in polar format (often called a circle tree). This is basically the same as the trees above but in polar coordinates. The vertical dimension is now the angle of the circle and the horizonal dimension is the distance from the centre point. These trees are generally used to make a big visual impact in papers but generally have reduced readability - it is difficult to compare how far nodes are from the centre. Generally best avoided. Tree B is a radial format tree. This is often used when the rooting of the tree is not known (although I have marked with a red circle the equivalent position of the root in trees above). This format tends to clump closely related sequences together making their precise relationships difficult to see. Generally best avoided too. I will not mention these formats again.

The root of the tree

I mentioned above that if we know the root of the tree then that provides information about the order of nodes in the tree. What do we do if we don't know? How can we work out where the root is? Many methods of reconstructing phylogenies from gene sequences do not explicitly estimate the root of the tree. When the tree is generated it will often have an arbitrary root. For example, here is the tree, above, rooted in an arbitrary place:



This is exactly the same underlying tree as those above. I have marked the previous rooting position with a red circle. What is important to note is that it no longer holds that the left to right order of the internal nodes (the blue circles) can be interpreted as the order of common ancestors. Figures that are arbitrarily rooted should mention this in the legend but they often don't.

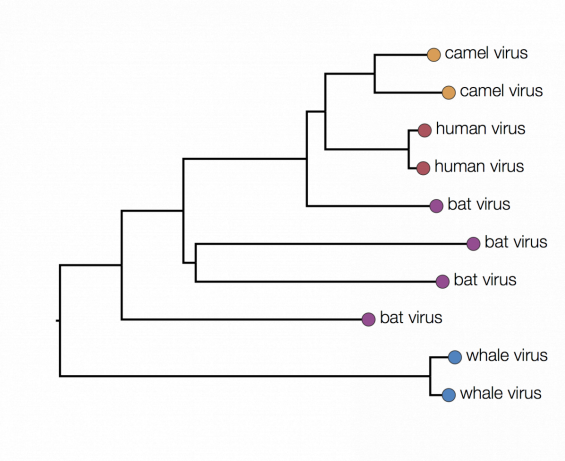
How do we work out where the root is?

There are two ways of finding the root of a phylogenetic tree. The first is to include one or more sequences in the data set that are *known* to lie outside the diversity of the sequences of interest. These sequences are usually referred to as the 'outgroup'. For example, in the trees above, the pair of tips labelled 'virus9' and 'virus10' could be the outgroup allowing us to root the tree at the red circle. How do we know the the outgroup is an outgroup? It is possible the outgroup has signficant genomic differences suggesting they are a different group of viruses. However, this might also mean the outgroup viruses are extremely divergent from the ones we are interested in. If the outgroups are too divergent from the sequences of interest then the root position will be unreliable. Alternatively it might be possible to assume that one or more sequences are the outgroup simply because they are the most divergent (virus9 and virus10, above, might be an example of this).

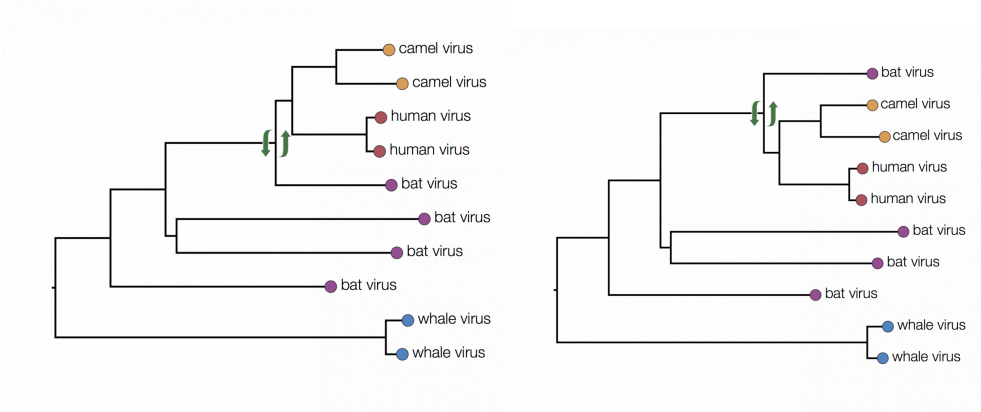
The second approach to rooting the tree is to use a method that implicitly assumes a time scale – a molecular clock model – as described below.

Reconstructing epidemiology

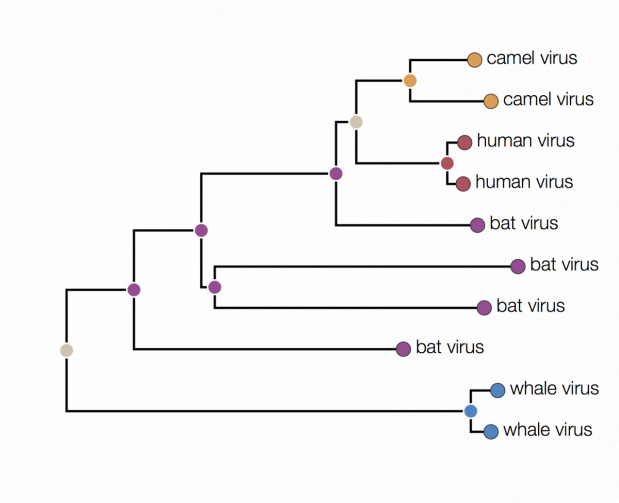
Here is the same tree as above but with the tips labeled by the type of host they were isolated from:



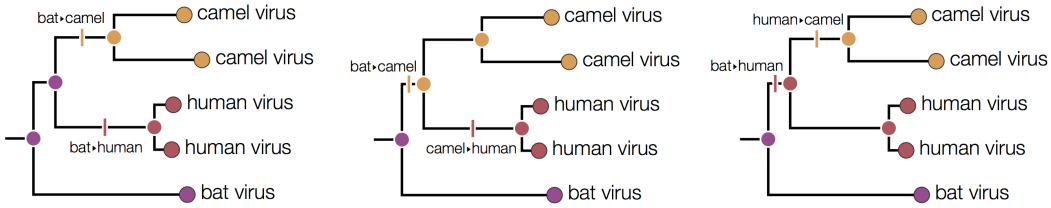
You can immediately see that there is some structure there with viruses grouping by host. For example the two viruses from humans have a closer common ancestor with each other than they do with any other virus. At first glance it may seem that human viruses are more closely related to bat viruses than camel viruses because they sit next to each other but remember that the vertical dimension is meaningless. In fact the viruses can be swapped round at any internal node and the tree is the same:



In fact the human and camel viruses are more closely related to each other and equally related to the bat viruses. This means we can't say from this tree if camels are the source of the human viruses or vice-versa, or just as likely, bats are independently the source of both human and camel outbreaks. We can however suggest that bats were the ultimate source of both camel and human viruses because of the much greater diversity of bat viruses. Another way to look at this is that the common ancestors of the human and camel viruses lie within the diversity of all the bat viruses.



In this tree the internal nodes are labelled with the reconstructed host species based on the principle of parsimony. This is the reconstruction that requires the fewest jumps between host species. The grey nodes are those that cannot be unambiguously reconstructed. For example, the common ancestor of the human and camel viruses could equally well be in humans, bats or camels with all three possibilities only requiring 2 host jumps:



Distinguishing these three possibilities generally requires additional data perhaps with a denser sampling of viruses.

A concrete 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.

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 (http://www.ncbi.nlm.nih.gov/BLAST/) 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)
MEDSQSDMSIELPLSQETFSCLWKLLPPDDILPTTATGSPNSMEDLFLPQDVAELLEGPEEALQVSAPAA
QEPGTEAPAPVAPASATPWPLSSSVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSISLNKLFCQLAKTCP
VQLWVTSTPPPGTRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNPYAEYLDDRQTF
RHSVVVPYEPPEVGSDYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDR
RTEEENFRKKEEHCPELPPGSAKRALPTSTSSSPQQKKKPLDGEYFTLKIRGRERFEMFRELNEALELKD
ARAAEESGDSRAHSSYPKTKKGQSTSRHKKPMIKKVGPDSD
```

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

```
Score    E
Sequences producing significant alignments:                      (bits) Value

gi|129372|sp|P10361|P53_RAT  Cellular tumor antigen p53 (Tum...   697  0.0
gi|13591878|ref|NP_112251.1|  tumor protein p53; tumor prote...   694  0.0
gi|1938365|gb|AAB80959.1|  mutant p53 [Rattus norvegicus]         692  0.0
gi|2961247|gb|AAC05704.1|  tumor suppressor p53 [Mus musculus]    611  e-174
gi|129371|sp|P02340|P53_MOUSE  Cellular tumor antigen p53 (T...   607  e-172
gi|6755881|ref|NP_035770.1|  transformation related protein ...   605  e-172
gi|200201|gb|AAA39882.1|  p53                                     605  e-172
gi|5081783|gb|AAD39535.1|  tumor suppressor p53 [Mus musculu...   603  e-171
gi|15375072|gb|AAK94783.1|  transformation related protein 5...   603  e-171
gi|53571|emb|CAA25323.1|  unnamed protein product [Mus muscu...   602  e-171
gi|223827|prf||1001197A  antigen p53,tumor                        601  e-171
gi|1813451|gb|AAB41831.1|  p53                                    598  e-170
gi|28975327|gb|AAO60156.1|  tumor suppressor p53; p53as [Mus...   563  e-159
...
```

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).

```
>gi|129372|sp|P10361|P53_RAT Cellular tumor antigen p53 (Tumor suppressor p53)
MEDSQSDMSIELPLSQETFSCLWKLLPPDDILPTTATGSPNSMEDLFLPQDVAELLEGPEEALQVSAPAA
QEPGTEAPAPVAPASATPWPLSSSVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSISLNKLFCQLAKTCP
VQLWVTSTPPPGTRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNPYAEYLDDRQTF
RHSVVVPYEPPEVGSDYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDR
RTEEENFRKKEEHCPELPPGSAKRALPTSTSSSPQQKKKPLDGEYFTLKIRGRERFEMFRELNEALELKD
ARAAEESGDSRAHSSYPKTKKGQSTSRHKKPMIKKVGPDSD

>gi|129371|sp|P02340|P53_MOUSE Cellular tumor antigen p53 (Tumor suppressor p53)
MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEALRVSGAPAAQ
DPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFCQLAKTCPV
QLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNLYPEYLEDRQTFR
HSVVVPYEPPEAGSEYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRR
TEEENFRKKEVLCPELPPGSAKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDA
```

```
HATEESGDSRAHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

>gi|2842741|sp|Q95330|P53_RABIT Cellular tumor antigen p53 (Tumor suppressor p53)
MEESQSDLSLEPPLSQETFSDLWKLLPENNLLTTSLNPPVDDLLSAEDVANWLNEDPEEGLRVPAAPAPE
APAPAAPALAAPAPATSWPLSSSVPSQKTYHGNYGFRLGFLHSGTAKSVTCTYSPCLNKLFCQLAKTCPV
QLWVDSTPPPGTRVRAMAIYKKSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRAEYLDDRNTFR
HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRR
TEEENFRKKGEPCPELPPGSSKRALPTTTTDSSPQTKKKPLDGEYFILKIRGRERFEMFRELNEALELKD
AQAEKEPGGSRAHSSYLKAKKGQSTSRHKKPMFKREGPDSD
......
```

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).

```
17    421
gi|129372| ---MEDSQ-- SDMSIELPLS QETFSCLWKL LPPDDILPTT ATGSP-NSME
gi|129371| MTAMEESQ-- SDISLELPLS QETFSGLWKL LPPEDILP-- ---SP-HCMD
gi|129370| ---MEEPQ-- SDLSIELPLS QETFSDLWKL LPPNNVLSTL P--SS-DSIE
gi|2499428 ---MEEPQ-- SDLSIELPLS QETFSDLWKL LPPNNVLSTL P--SS-DSIE
gi|2842741 ---MEESQ-- SDLSLEPPLS QETFSDLWKL LPENNLLTTS LN----PPVD
gi|1072019 ---MEEAQ-- SDLSIEPPLS QETFSDLWNL LPENNVLSPV LS----PPMD
gi|3024332 ---MEEPQ-- SDPSIEPPLS QETFSDLWKL LPENNVLSPL PS----QAVD
gi|3024331 ---MEEPQ-- SDPSIEPPLS QETFSDLWKL LPENHVLSPL PS----QAVD
gi|129367| ---MEEPQ-- SDPSIEPPLS QETFSDLWKL LPENNVLSPL PS----QAVD
gi|129369| ---MEEPQ-- SDPSVEPPLS QETFSDLWKL LPENNVLSPL PS----QAMD
gi|1072019 ---MEEPQ-- SDPSVEPPLS QETFSDLWKL LPENNVLSPL PS----QAMD
gi|1072019 ---MEEPH-- SDLSIEPPLS QETFSDLWKL LPENNVLSDS LS----PPMD
gi|1072018 ---MEESQ-- SELGVEPPLS QETFSDLWKL LPENNLLSSE LS---LAAVN
gi|1709531 ---MEESQ-- AELGVEPPLS QETFSDLWNL LPENNLLSSE LS----APVD
gi|1171969 ---MQEPP-- LELTIEPPLS QETFSELWNL LPENNVLSSE LS----SAMN
gi|6093639 ---MEESQ-- SELNIDPPLS QETFSELWNL LPENNVLSSE LC----PAVD
gi|1072019 ---MAESQEF AELWERNLIS TQEAGTCWEL INDEYLPSSF DP----NIFD

....
```

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 sufferred 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 builing programs.

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.

```
(((((((((gi|129372|sp|P10361|P53_RAT:0.65449,gi|10720197|sp|Q9WUR6|P53_CAVP:-
0.31392):0.27918,
gi|10720190|sp|O36006|P53_MARM:-
0.18911):0.63596,(gi|129371|sp|P02340|P53_MOUSE:1.77497,
gi|2842741|sp|Q95330|P53_RABIT:-
1.18496):1.78731):0.84230,(gi|129370|sp|Q00366|P53_MESAU:1.31394,
((gi|10720186|sp|Q9TUB2|P53_PIG:0.08695,gi|1171969|sp|P41685|P53_FELCA:0.10481):1.4660
9,
(gi|1709531|sp|P51664|P53_SHEEP:1.17669,(gi|6093639|sp|Q29537|P53_CANFA:1.06148,
gi|10720195|sp|Q9W678|P53_BARB:6.68852):0.22831):0.55803):0.50856):0.35590):1.14706,
gi|2499428|sp|O09185|P53_CRIGR:-0.02921):0.23514,gi|10720194|sp|Q9TTA1|P53_TUPG:-
0.02106):0.07596,
gi|129369|sp|P04637|P53_HUMAN:0.00463):0.03787,gi|3024332|sp|P56424|P53_MACMU:-
0.00206):0.00205,
gi|129367|sp|P13481|P53_CERAE:0.00393,gi|3024331|sp|P56423|P53_MACFA:0.00379);
```

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 others 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.

```
(((((gi|1171969|sp|P41685|P53_FELCA:0.10481,gi|10720186|sp|Q9TUB2|P53_PIG:0.08695):1.4
6609,
(((((gi|10720197|sp|Q9WUR6|P53_CAVP:0.0,gi|129372|sp|P10361|P53_RAT:0.65449):0.27918,
gi|10720190|sp|O36006|P53_MARM:0.0):0.63596,(gi|2842741|sp|Q95330|P53_RABIT:0.0,
gi|129371|sp|P02340|P53_MOUSE:1.77497):1.78731):0.8423,(((((gi|3024331|sp|P56423|P53_M
ACFA:0.00379,
gi|129367|sp|P13481|P53_CERAE:0.00393):0.00205,gi|3024332|sp|P56424|P53_MACMU:0.0):0.0
3787,
gi|129369|sp|P04637|P53_HUMAN:0.00463):0.07596,gi|10720194|sp|Q9TTA1|P53_TUPG:0.0):0.2
3514,
gi|2499428|sp|O09185|P53_CRIGR:0.0):1.14706):0.3559,gi|129370|sp|Q00366|P53_MESAU:1.31
):0.50):0.55,
gi|1709531|sp|P51664|P53_SHEEP:1.17669):0.22831,gi|6093639|sp|Q29537|P53_CANFA:1.06148
):3.34426,
gi|10720195|sp|Q9W678|P53_BARB:3.34426);
```

Bayesian methods - The new kid on the block

Bayesian methods have only recently been applied to phylogenetic tree building. Bayesian methods are related to Maximum Likelihood methods but have two main advantages:

- Among the traditional methods, maximum likelihood is considered the conceptual most appealing way to compute phylogenitic trees. However, bayesian methods are considered to be even more conceptually appealing: they calculate the probability of the tree given the data, rather than the probability of the data given the tree (as in ML)
- Maximum likelihood calculations are very time consuming and prohibitevely so if the user wishes to carry out bootstrapping. Bayesian methods can usually compute the equivalent of the bootstrapping in a shorter time.

For more details on bayesian methods refer to the further reading.

Further reading

- Phylogeny estimation: traditional and bayesian approaches, Holder and Lewis, Nature, Vol 4, April 2003.
- Molecular Evolution, Page and Holmes, Blackwell Science 2002
- Molecular Evolution and Phylogenetics, Nei and Kumar, Oxford University Press 2000.
- Inferring phylogenies, Felsenstein, Sinauer Associates 2003