Part V - Phylogenetics

**Slide 1:** Phylogenetics

For this slide deck, we are giving only a brief introduction to this complex topic, defining a few terms and talking in general about how one does this. For more information, we highly recommend “Phylogenetic Trees Made Easy: A How To Manual” by Barry Hall (ISBN-13: 978-0878936069).

**Slide 2:** For this slide deck, participants will get to know the basics of phylogenetics and recognize the different kinds of phylogenetic trees. Finally, participants will do an exercise where they learn how to construct phylogenetics trees using MEGA.

**Slide 3:** We first need to define the term “Phylogenetics.” Phylogenetics entails the determination of evolutionary relations among a group of biological entities. We have used a definition where “entities” means that classification can be done at any taxonomic level. For example, here we can see the evolutionary relationship between the three domains of life- bacteria, archaea and eucarya. Evolutionary relationships are illustrated by the lines on the tree.

**Slide 4:** This slide shows an advanced tree of life generated by Hug and colleagues in 2016 in Nature Microbiology. Here we see that the largest diversity of life is within the domain of bacteria, where the red dots represent lineages which have not been cultured in lab. And if you notice where humans lie on this tree, we’re just a small little branch in the domain Eukaryotes.

**Slide 5:** When talking about phylogenetic trees, we need to understand some of the nomenclature terms. The vertices of the tree are referred to as nodes. The groups of organisms placed at the external nodes, in this case A, B, C and D, are referred to as taxa. A clade is a group of taxa that share a common ancestor and the root here is the common ancestor of all taxa under consideration. So, phylogenetic trees are a visual representation of the evolutionary relationship between biological entities.

**Slide 6:** There are two basic types of phylogenetic trees. One way to represent the data is either as a phylogram or a cladogram. In a phylogram, the branch lengths are proportional to the distance of divergence between organisms, so the branch length is a measure of divergence times. Cladograms are simple representation of the evolutionary relationships and the branch lengths are not proportional to the distance.

**Slide 7:** Another important term to define is the outgroup. An outgroup is the lineage that falls outside the clades of interest that are being studied. This outgroup has information common to all the taxa that you are considering but every member of the clades is more closely related to each other than the outgroup. The outgroup in the slide are represented by the ovals.

**Slide 8:** Another distinction among trees is whether they are rooted or unrooted. The root here represents the common ancestor of all the taxa under consideration and when that common ancestor is known we can generate a rooted tree. An unrooted tree on the other hand can tell us about evolutionary relationships among taxa but not the evolutionary path. The tree of life that we saw in slide 4 is an example of an unrooted tree.

**Slide 9:** Phylogenetic trees are formed on the basis of homology. Groups that share a common ancestor are considered to be homologous. The homology can be inferred using sequence identity or sequence similarity and thus sequences that are significantly similar will be homologous.

**Slide 10:** How do we begin to construct a phylogenetic tree?

The first step is database search and sequence retrieval based on BLAST to identify potential homologs. These data can be accessed from NCBI as we discussed in Part II - Databases.

After we have our sequences we need to create a multiple sequence alignment based on the comparative genomics techniques we learned in Part IV – Comparative Genomics.

Then the next step is to identify a method of analysis. Here we will discuss three methods that can be used.

The last step is construction and evaluation of the tree. Is it useful for our hypothesis? Can it be refined?

**Slide 11**: Some additional details about collecting sequences. Many of the details for obtaining sequences were described earlier in this series, but it is important to consider the reason for tree construction at the outset. What hypothesis are you testing? That will drive the kinds of sequences you plan to obtain – for example: should they be closely or distantly related?

**Slide 12:** Once you have obtained your sequences and aligned them, then what? You need to think about what method you would like to use to construct your tree.

Distance based methods measure the dissimilarity between two sequences and constructs a tree based on a distance matrix. Two examples are the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method and the neighbor joining method.

The maximum parsimony method chooses the best phylogeny based on the smallest number of changes between each character state.

The maximum likelihood method uses a substitution model that assesses the probability of certain mutations and uses a substitution model to generate a tree.

Over the next few slides we’ll discuss each of these in a little more detail using nucleotide sequences as the examples. The exercise that follows this slide deck will allow you to explore proteins.

**Slide 13:** UPGMA refers to Unweighted Pair Group Method with Arithmetic Mean. Here we have three organisms, A, B and C which are placed in a distance matrix. This model calculates the dissimilarity between AB, AC, and BC. Here we see that A and B are most closely related to each other, so those two taxa are placed on the same clade. In the next step C is compared to the mean of A and B where they are collapsed into a single composite taxon. This step can be repeated multiple times for more taxa.

**Slide 14:** For the Neighbor-joining method the taxa under consideration are arranged into a star tree shape. Now the model separates every pair of taxa and calculates the total branch length based on some dissimilarity parameter. Then the pair of taxa that minimizes the total branch length is picked and this branch is collapsed into a single taxon. Then the process starts again with this collapsed taxon part of the star tree formation, until all the taxa are joined together.

**Slide 15:** The maximum parsimony method puts an emphasis on the least number of changes from the original character state as being the best phylogeny. Here in this example we have 4 sequences, and they are placed in three different arrangements. And then the model looks for the one case in which the least number of point mutations can explain their placement on the phylogenetic tree. Then this tree is picked as the output. In this example, the top tree has the fewest number of changes.

**Slide 16:** For maximum likelihood method, the HKY85 method which was discussed previously in Part IV – Comparative Genomics is used as the substitution model and then can be used to create the phylogenetic tree.

**Slide 17:** This slide shows a comparison of all three methods and lists their strengths and weaknesses. The Distance Method has the fastest computational speed but the sequences need to be evolutionarily close for this method to work. Also by compressing sequence information into a single number we lose out on information.

The Maximum Parsimony method is faster than Maximum Likelihood and gives information on ancestral relationships between the taxa. One flaw of this method is that the minimum number of evolutionary changes used to calculate the tree may give biased results.

Maximum Likelihood is thought to be the most accurate method and generates multiple trees before picking the best one. It takes more computational time than the other methods and is impractical for large datasets.

Based on the type of data that you want to compare. you can make a decision about which method will be suitable to generating your tree. An important feature of MEGA, which we use as the basis for the exercise that follows this slide deck, is that it tells you the best method to use based on your data.

**Slide 18:** Finally, to assess the reproducibility of your phylogenetic tree you can use the process of bootstrapping. The bootstrap test resamples the alignments and constructs trees from the re-sampled data. You can then generate a number which tells you how many times a particular arrangement of taxa was generated out of multiple simulations. In the example on the slide 89 out of 100 times taxa 1 and 2 and taxa 3 and 4 were on the same clade, respectively. The bootstrapping number corresponds to accuracy of the arrangement of the clades in the phylogenetic tree, numbers approaching 100 are more accurate than lower numbers.

**Slide 19**: A few notes about MEGA – MEGA is a free software package that is useful for creating phylogenetic trees. It can be downloaded from here: <https://www.megasoftware.net/>. The current version is MEGA7. Be sure to download the version that matches your operating system. You also have the option of downloading a GUI version or a command-line version. The version used in the exercise following this slide deck assumes the GUI version. One caution – MEGA generally works better on Windows machines, although the Mac iOS version of MEGA7 isn’t nearly as problematic as older versions of MEGA.

**Slide 20**: In the exercise about phylogenetics that follows this slide deck, we tell you to choose the model that leads to the lowest BIC value. This slide explains why we suggest this.

**Slide 21:** In summary, we learned about phylogenetic trees and the concept of homology. Now we will use the data that we have accessed from NCBI in the Comparative Genomics exercise to construct a phylogenetic tree using MEGA software.