Phylogenetics: An introduction

Laura Emery [1]

- DNA & RNA
- Beginner
- 1 hour

This course provides a basic introduction to the field of phylogenetics, with an emphasis on how to read and interpret phylogenetic trees.

An undergraduate-level knowledge of biology would be an advantage.

Learning objectives:

- Describe several applications of phylogenetics
- Explain how to read simple trees
- Identify major stages in phylogenetic analyses
- Access bioinformatics tools for phylogenetics

What is phylogenetics?

Phylogenetics is the study of evolutionary relationships among biological entities - often species, individuals or genes (which may be referred to as taxa [2]). The major elements of phylogenetics are summarised in Figure 1 below.
Figure 1 Elements of phylogenetics.

Typically phylogeneticists study one of the following types of question:
• What are the evolutionary relationships or histories among my species/individuals/genes of interest?
• How do sequences evolve?
• Can I better describe processes of sequence evolution with a mathematical model?

We can reconstruct a phylogenetic tree by looking at the nucleotide or protein sequences and combining this with our understanding of sequence evolution, which is described using an evolutionary model. This enables us to infer evolutionary events that happened in the past, and also provides more information about the evolutionary processes operating on sequences. Thus we may refine our understanding of how evolution works and develop better mathematical models of evolution.

Why use molecular data?

Today almost all evolutionary relationships are inferred from molecular sequence data. This is because:
Phylogenetics: An introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

DNA is the inherited material;
We can now easily, quickly, inexpensively and reliably sequence genetic material;
Sequences are highly specific and are often information rich.

In rare cases, where it is not possible to obtain genetic material (e.g. in the case of certain ancient fossil samples), morphological measurements can be used to infer evolutionary relationships. However, this approach is less reliable than using molecular data because we know that sometimes the same morphological trait can arise from multiple independent evolutionary lineages (i.e. occurrences of the trait are analogous [3]).

This course will focus on molecular phylogenetics.

Why is phylogenetics important?

Phylogenetics is important because it enriches our understanding of how genes, genomes, species (and molecular sequences more generally) evolve. Through phylogenetics, we learn not only how the sequences came to be the way they are today, but also general principles that enable us to predict how they will change in the future. This is not only of fundamental importance but also extremely useful for numerous applications (Figure 2).

Applications of phylogenetics

Figure 2 A summary of some applications of phylogenetics.
Classification: Phylogenetics based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before the advent of molecular sequencing. Phylogenetics now informs the Linnaean [4] classification of new species.

Forensics: Phylogenetics is used to assess DNA evidence presented in court cases to inform situations, e.g. where someone has committed a crime, when food is contaminated, or where the father of a child is unknown.

Identifying the origin of pathogens: Molecular sequencing technologies and phylogenetic approaches can be used to learn more about a new pathogen outbreak. This includes finding out about which species the pathogen is related to and subsequently the likely source of transmission. This can lead to new recommendations for public health policy.

Conservation: Phylogenetics can help to inform conservation policy when conservation biologists have to make tough decisions about which species they try to prevent from becoming extinct.

Bioinformatics and computing: Many of the algorithms developed for phylogenetics have been used to develop software in other fields.

Coming soon...? With the advent of newer, faster sequencing technologies, it is now possible to take a sequencing machine out to the field and sequence species of interest in situ. Phylogenetics is needed to add biological meaning to the data.

What is a phylogeny?

A phylogeny, also known as a tree, is an explanation of how sequences evolved, their genealogical [5] relationships, and therefore how they came to be the way they are today.

One the first sketches of a phylogenetic tree was made by Charles Darwin (Figure 3). This famous diagram is one of his earlier sketches from a series of notes that he used to develop the idea.
Figure 3 Sketches by Charles Darwin who is considered to be the first person to use the metaphor of a tree to represent evolutionary relationships. *Image source: Wikimedia Commons.*

The example of a family tree

Here is a familiar example that everyone will know about - a family tree (Figure 4). You will probably already know how to interpret patterns of relatedness on a family tree, and it turns out that the same principles apply to phylogenetic trees more generally.

![Family Tree Diagram]

Figure 4 A simple family tree to illustrate a phylogenetic tree.

In Figure 4 we can see that you are more closely related to your brother than you are to your first cousin. Trace the tree back in time to your ancestors and you will see why. The most recent common ancestor (MRCA) that you share with your brother is your mother. In contrast, the MRCA you share with your first cousin is your maternal grandmother.

These principles also apply when you are reading more complex phylogenies – remember to trace back to MRCAs to examine patterns of relatedness.

A true tree?

The case of a family tree is a rare example where we typically know part of the true biological tree; we have a complete explanation of the genealogical relationships among family members. This is because we know:

(i) all of the ancestors and descendants who are or were recently living;

(ii) the identities of the biological parents of each child.

Unfortunately, when we estimate phylogenies from present day sequences that are not so closely related, we seldom know the true tree. This is because we do not know what sequences occurred in the ancestors, and therefore what genetic changes occurred to make the sequences the way they are today. This is why we need methods of phylogenetic reconstruction to infer the true tree.

Aspects of phylogenies
Before we discuss how to construct phylogenies, it is important to learn how to read and interpret them. This can be particularly challenging if you are familiar with the field of network biology, as elements of trees resemble networks, and yet have a very different meaning.

In the next section we will look at some major aspects of phylogenies and how they can be interpreted.

Specifically we will focus on:

1. **Topology** [7]
2. **Branches** [8]
3. **Nodes** [9]
   - (i) Tips
   - (ii) Internal nodes
   - (iii) **Root** [10]
4. **Confidence** [11]

You might also want to look at [Andrew Rambaut's tutorial](https://www.ebi.ac.uk/training/online) on how to read trees, and then test your skills with the [tree thinking test](https://www.ebi.ac.uk/training/online).

## Topology

The topology is the branching structure of the tree. It is of particular biological significance because it indicates patterns of relatedness among **taxa** [2], meaning that trees with the same topology and **root** [10] have the same biological interpretation.
Figure 5 Examples of trees with the same (top box) and different (lower box) topologies.

The three trees in the top box of Figure 5 have topologies in common. This is evident from the statement "A and B are more closely related to one another than they are to C", which is true for all three trees. In the lower box, all three trees have different topologies from one another. For example, in the left most tree, A and B are more closely related to one another than they are to C. However, in the middle tree, it is C and A who are each other’s closest relatives, and in the right hand tree it is B and C which are most closely related.

Making a model of a tree

If you find it difficult to imagine which trees have the same topology, it can be useful to make a model of a tree using either string or pipe cleaners, and a pin. An example of one we made earlier is shown below (Figure 6).

![Figure 6](image)

Figure 6 A model of a tree made with pipe cleaners and a pin to illustrate examples of the same topology.

The pipe cleaners represent branches and the pin indicates the root position. Rotating any combination of branches around the root always results in the same topology, and thus patterns of relatedness.

Why not try making your own?

Branches

Branches show the path of transmission of genetic information from one generation to the next. Branch lengths indicate genetic change i.e. the longer the branch, the more genetic change (or divergence) has occurred. Typically we measure the extent of genetic change by estimating the average number of nucleotide or protein substitutions per site.
Figure 7 Branch length representations.

Informative branch lengths are typically drawn to scale and indicate the number of substitutions per site (Figure 7). Branch lengths are occasionally shown on the phylogeny (left), but it is far more common to see branch lengths represented by a scale bar (right). It can therefore be useful to keep a ruler close to hand for interpreting phylogenies that you see in the literature!

How do we estimate genetic change?

Estimating the extent of genetic change is not a trivial task. A naïve method is to align pairs of sequences, count up the number of differences and divide by the sequence length.

Figure 8 Simple sequence alignment.

In the simple alignment above (Figure 8), we can see that there is one site that is different between the two sequences, and we could say that based upon this tiny sample there are $1/10 = 0.1$ substitutions per site. However this assumes that we have observed every substitution that has happened, and therefore does not account any multiple substitutions [15] that have occurred at any of the sites. We have also assumed that every substitution (e.g. from T>C, or A>G) is equally likely to have occurred, and we now know that this is unrealistic. To overcome these issues, it is now commonplace to use an evolutionary model to infer the genetic change that has occurred.

Beware of very long branches!

To get a value of one substitution per site using the simple method above would require the pair of sequences to be completely different to each other at all 10/10 sites. It is unlikely you would align such sequences since two random nucleotide sequences are likely to be 25% identical. So if you see figures in the literature with branches longer than ~3 substitutions per site then you might want to worry about the confidence we have in those estimates!
Nodes

Nodes are the points at the ends of branches which represent sequences or hypothetical sequences at various points in evolutionary history.

The three types of node and their positions in the example phylogeny are indicated in Figure 9, below.

![Diagram of a tree indicating types of nodes.](image)

**Figure 9** A cartoon diagram of a tree indicating types of nodes. *Image courtesy of Andrew Rambaut.*

The tips

The sequences that we sampled and used to construct our phylogeny occur on single terminal branches, known as the tips or external nodes.

Internal nodes

Internal nodes occur at the points where more than one branch meet and represent the (usually inferred) ancestral sequences. For example, in Figure 9, the internal node indicated by a blue arrow is the hypothetical common ancestor [6] of sequences A and B.

The root

The root is a very important internal node representing the most recent common ancestor of all sequences in the phylogeny. We will talk more about the root on the next page...

Root

The root is the most recent common ancestor [6] of all of the taxa [2] in the tree. It is therefore the oldest part of the tree and tells us the direction of evolution, with the flow of genetic information moving from the root, towards the
tips with each successive generation.

Most methods of phylogenetic reconstruction do not estimate the position of the root, in part because this increases the number of possible trees, and therefore time that it takes to calculate the tree. An example of an unrooted tree which has been subsequently rooted is shown below in Figure 10.

Figure 10 Rooting a tree: before (left) and after (right). NB: The branch lengths are not drawn to scale.

If you have difficulty visualising the rooting process (shown above in Figure 10) then imagine that the tree was made from string, and that you are pushing a pin into the string to rotate the remaining branches around the pin-point. The arrow indicates the direction of evolution as implied by the root position.

Rooting a tree affects its meaning

Deciding upon an appropriate root position is critical for phylogenetic interpretation because the root tells us the direction of evolution and so affects statements that we make about patterns of relatedness. For example, in the unrooted tree above (Figure 10, left) we cannot make statements such as "A is more closely related to B than it is to C" because this would not be true if the root occurred anywhere on the branches that connect A and B.

Where to root a tree?

There are two main approaches that we can use to root a tree:

**Outgroup rooting**: A preferred approach is generally to include one or more sequences in our analysis that we know are definitely more distantly related to our sequences of interest than they are to one another. These sequence are usually referred to as 'outgroups'. The root estimate is then simply the point at which our outgroup(s) join the rest [16] of our tree of interest. The best possible outgroups are those available which are most closely related to our sequences of interest. If outgroups are too distantly related then they can be unreliable as they may be difficult to align reliably or have become saturated with substitutions.

**Midpoint rooting**: This method requires you to make the assumption that all of your sequences are evolving at the same rate - you should do so cautiously because this assumption does not hold for many biological datasets. In this case, the root is positioned at the midpoint between the two longest branches. If you have taxa that were not sampled at the same time point then a slight modification of this method would be required to take into account the time elapsed between samples.
Uncertainty in phylogenetics

Inferring phylogenies is an inherently uncertain process because we usually have no more information to guide us than the sequences from our present day taxa [2]. Some sequences are more informative than others, and so provide us with better estimates of genealogical [5] relationships.

Estimating our confidence in trees is itself a difficult problem. In other areas of bioinformatics we can examine metrics such as sensitivity [17] and specificity [18], which assess the estimated or inferred results of a method against empirically known true positive and negative examples. However this is not possible in case of phylogenetics because we do not have examples of 'known' ancestral sequences and phylogenies.

Approaches for estimating confidence

There are several approaches that are commonly used to estimate our confidence in the inferred tree topology including bootstraps, likelihood [19] and Bayesian [20] approaches. Whilst it is beyond the scope of this course to explain how these are estimated, we will explore basic aspects of their interpretation using the most widely used example of the (non-parametric [21]) bootstrap [22].

Learn more about how bootstraps are computed in Sandra Baldauf's published tutorial [23] (4 [13]).

How to read confidence estimates on a tree

Confidence estimates on a tree refer to the internal branches that they are shown on. An example is provided in Figure 11.

![Figure 11](image)

**Figure 11** Two typical representations of confidence on part of a phylogeny.

Typically you will see bootstrap (or other confidence estimate) values shown on a phylogeny. For example in the diagram above (Figure 11; left), bootstrap values are shown from 100 replicates. Sometimes asterisks (*) are used instead to indicate branches where there is greater than 80% (or 90%) bootstrap support (e.g. Figure 11; right). In this example, the bootstrap value of 63 out of 100 (63%) is not represented by an asterisk because it is less than 80% support.
Interpreting confidence estimates on a tree

Interpreting the exact meaning of confidence values on a phylogeny is still an area of debate, but experts generally agree that we can accept branches with >80% bootstrap support (or 90% depending on who you ask!), provided that an appropriate evolutionary model was used to estimate the phylogeny.

Taking the above approach enables us to make the following statements about Figure 11:

(1) "there is consistent (100% bootstrap) support that taxa C and D are more closely related to each other than they are to B";

(2) "from these data it is unclear that B, C and D are each other's closest relatives".

Relating distance, rate and time

Evolutionary rate and time are confounded

Earlier you learned that branch lengths tell us the genetic change that has occurred. The genetic change is calculated by a combination of the rate of substitution and the time that has elapsed, as shown in the equation below:

\[
\text{Genetic change} = \text{Evolutionary rate} \times \text{Divergence time} \\
\text{(substitutions/site)} \times \text{(substitutions/site/year)} \times \text{(years)}
\]

This means that without any other prior information about the tree, we do not know whether branch D shown below (Figure 12) represents an extinct or frozen lineage that was sampled a long time ago; a lineage with a relatively slow evolutionary rate; or some combination of both.

Figure 12 Hypothetical rooted phylogeny to illustrate a relatively short branch (D).

This makes it very important for you to consider the biological information and context that you know about when interpreting trees. If you are lucky enough to know the dates when your samples were isolated, then there are
some methods available that can help you to disentangle these processes.

**Alternative representation of phylogenies**

The same topology can be drawn in lots of different ways. Some of the most common formats are shown in the diagram below (Figure 13).

![Alternative representations of phylogenies](image)

**Figure 13** Alternative representations of the same topology. Red lines indicate the same branch in each representation. Trees can be rotated on the page and still depict the same tree. NB: The trees are not drawn to the same scale.

Diagonal and rectangular formats are most commonly used for publication. Rectangular format has the advantage that it is easy for the human eye to quickly assess the relative branch lengths; however you should note that the length of the lines that are perpendicular to the branch lengths have no meaning.

Curved format is often used in review papers to represent summaries of phylogenies, where precise knowledge of branch lengths may be less important. Radial format is typically used to show unrooted trees, and circular format is often used to represent large phylogenies with many **taxa** [2] (for the simple reason that it is easy to fit onto a printed page).

**Interpreting patterns of relatedness**

The key to interpreting patterns of relatedness in an evolutionary context is to trace back to the point in the tree where **taxa** [2] share their **most recent common ancestor** (MRCA). You can practice this by finding two species of
interest in a tree and tracing back along their branches to identify the node where they meet - this is their MRCA.

Below we will explore examples from the vertebrate phylogeny to see how we can describe evolutionary relationships (Figure 14).

![Vertebrate phylogeny with ancestral nodes highlighted by coloured spots.](image)

**Figure 14** Vertebrate phylogeny with ancestral nodes highlighted by coloured spots.

Looking at Figure 14, we can make the following statements:

**Example 1:**

"Humans (red) are more closely related to mice (bright blue) than they are to lizards (green)." This is because humans share a common ancestor [6] more recently with mice (ancestor = purple spot) than they do with lizards (ancestor = dark blue spot).

**Example 2:**

"Frogs (purple) are more closely related to lizards (green) than they are to fish (pink)." This is because frogs share a common ancestor with lizards more recently (ancestor = orange spot) than they do with fish (ancestor = black spot)."

**Example 3: equal relatedness**

"Fish (pink) are equally related to mice (bright blue) as they are to frogs (purple)." This is less intuitive, but if you trace back to the MRCAs you will see why: mice and frogs share the same common ancestor (black spot) with fish, so neither species is more closely related to fish.

**Remember:**

- These interpretations rely on our tree being rooted because the root is needed to define the direction of evolution and therefore what is 'more recent' in evolutionary time.
If we rotate the branches to change the topology of the tree then the tree still has the same biological meaning and evolutionary relationships.

You can read more about common misconceptions when interpreting patterns of relatedness in this tutorial [24] (3 [13]).

Major stages in phylogenetic analyses

There is no recipe available to complete a phylogenetic analysis as the approaches that you take are highly context and situation specific, and ultimately depend upon your biological question of interest. There are however some major stages that you are very likely to encounter and so we'll look at an overview of these below (Figure 15).

Figure 15 Outline of major common stages in many phylogenetic analyses.

Figure 15

You can read a more detailed overview of phylogenetic analysis in this review [25] (4 [13]).

A comprehensive list of phylogenetic software is maintained here [26] (6 [13]).
Phylogenetics resources at EMBL-EBI

EMBL-EBI offers a range of tools and resources that are relevant to the field of phylogenetics:

1. **Ensembl** [27]: is a vast resource for vertebrate genome data, where (among other things) you can download nucleotide sequence data and run **BLAST** [28] searches to inform the identification of **homologous** [29] sequences. For many genes, you can also download the **orthologous** [30] sequences that we have identified, as well as interspecific alignments that we have pre-computed. To do so, simply search for your favourite gene (e.g. ADH1B), and explore the ‘genomic alignments’ options within the comparative genomics menu.

   **Ensembl genomes** [31]: extends Ensembl across the tree of life, making genome data publically available for Bacteria, Plants, Fungi, Protists and **Metazoa** [32]. This includes pre-computed alignments and orthologues.

   **Ensembl compara** [33]: offers pre-computed phylogenies for visualisation and download. For further information see the documentation page [33].

2. **ClustalW2 Phylogeny** [34]: is a basic tool for estimating evolutionary trees from multiple sequence alignments. It uses the **Neighbour Joining method** [35] with the option of a very simple model of sequence evolution (Jukes and Cantor, 1969).

3. **Clustal Omega** [36] and **Prank** [37]: are multiple sequence alignment programs which are useful as part of a phylogenetics workflow.

4. **EMBOSS Seqret** [38]: is a file format conversion tool that can be useful at multiple stages of a phylogenetics workflow.

---

**Summary**

- Phylogenetics is the study of evolutionary relationships among nucleotide or protein sequences.
- There are many applications of phylogenetics, including forensics, pathogen surveillance, conservation and bioinformatics.
- There are several aspects of phylogenies that you need to understand in order to interpret your trees: topology, branch lengths, nodes and confidence.
- Careful interpretation is critical to understanding the biological meaning of phylogenies.
- The same phylogenetic tree can be visualised in many different ways.
- Evolutionary relationships can be unraveled by identifying the most recent common **ancestor** [6] (MRCA) shared by species.
EMBL-EBI has several resources and tools that are relevant to the field of phylogenetics including: Ensembl [27], ClustalW2 Phylogeny [34], Clustal Omega [36] and Prank [37]. [39]

Quiz: Introduction to phylogenetics

Questions: 10
Attempts allowed: Unlimited
Available: Always
Pass rate: 75 %
Backwards navigation: Allowed

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Find out more

Review articles


Published teaching materials


Books


Recommended courses

Online
Phylogenetics: An introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

- Andrew Rambaut's How to read a phylogeny tutorial [12]
- Understanding Evolution tutorials and resources [47], University of California Museum of Paleontology
- Joe Felsenstein's comprehensive collection of phylogenetics software [26]

References

1. Andrew Rambaut's How to read a phylogeny tutorial [12].


5. Modeling nucleotide sequence evolution [48], College of Computer, Mathematical and Natural Sciences.


Contributors

Laura Emery [1]
EMBL-EBI
Scientific Training Officer - Training

Laura completed her undergraduate degree in Genetics (BSc first class hons) at the University of Nottingham. After dabbling in experimental work on a summer placement at the Wellcome Trust Sanger Institute, she found joy in using computer-based analyses to explore biological questions. Laura completed her PhD in Evolutionary Genetics at the University of Edinburgh’s Institute of Evolutionary Biology. There she used a bioinformatics approach to investigate patterns of codon usage in Archaea while under the supervision of Prof Paul Sharp FRS.

She also worked as a postdoc at the University of Manchester, where she explored molecular co-evolution among interacting proteins. Having obtained qualified teacher status (QTS), Laura joined EMBL-EBI as a Scientific Training Officer in 2012 and is responsible for the development and delivery of training courses.

Acknowledgements

Many thanks to Melissa Ward [49] for her helpful comments.