Phylogenetic analysis

Plant Genetic Resources (3502-470)

23 May 2025

Table of contents

1 Motivation

Understanding the evolutionary relationships helps to elucidate patterns of crop biodiversity, understand the process of domestication and to identify genetic resources that are suitable and available for the improvement of a particular crop species.

If you recall the diversity of maize landraces in Peru from the introduction (Figure 1) one may ask when and where this diversity originated.

Figure 1: A sample of diversity of maize diversity in Peru. Photo: Karl Schmid

Phenotypic differences between individual plants can be caused by the environment (this variation is not heritable) or by genetic differences, which are heritable. One may safely assume that a large proportion of phenotypic variation that differentiates not individual plants, but landraces or plant varieties is controlled by genetic variation. It is therefore of interest to investigate the extent to which genetic differences are responsible for phenotypic variation in plant genetic resources.

Genetic differences ultimately originate from mutation and recombination. To identify useful genetic variation in plant genetic resources, the following questions are frequent objectives of scientific investigations:

1. When and where did an interesting mutation, which causes a particular phenotype, originate?
2. Did it originate once or several times?
3. How did the phenotype that is caused by a mutation spread geographically?

One approach to address these questions is to trace back the evolutionary history of a mutation, a gene, a landrace, or a species using phylogenetics, which we define as the study of evolutionary relationships.

2 Learning goals

- Understand the importance of *phylogenetic tree thinking* in the investigation of plant genetic resources
- Know the key terms that describe phylogenies
- Be able to explain the main approaches to construct phylogenies
- Be able to interpret phylogenetic trees

3 Tree thinking in biology

One approach to address these questions is to trace back the evolutionary history of a mutation, a gene, a landrace, or a species.

Methods of phylogenetic analysis or **phylogenetic inference**, which is the determination of the evolutionary relationship of taxa allow to investigate these questions.

Phylogenetics is an area of evolutionary biology that investigates the evolutionary relationship of species.

Its key metaphor is the **evolutionary tree**, or **phylogeny**, that expresses different degrees of evolutionary relationships and states the principle of **common ancestry**, i.e., that all individuals of a species, or different species of a genus, derive from a common ancestor.

The concept of the evolutionary tree was invented by Charles Darwin who first drew a tree in his notebook (Figure 2) and then put a tree as the only figure in his 1859 book *On the Origin of Species* (Figure 3). Due to its importance in the analysis of biodiversity, phylogenetics has developed into a mature and highly sophisticated field of biology. Many tools and methods are available for analysing the evolutionary relationship of species but also of individuals within species.

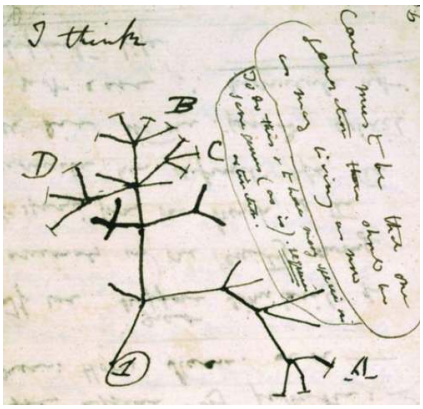
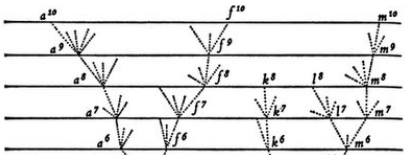


Figure 2: A depiction of a phylogenetic tree in Charles Darwin's notebook. Source: Wikipedia label:fig:darwinthinks



Darwin's tree already demonstrates three important principles of phylogenetics:

1. **Using present information to reconstruct the past:** Key concepts are **ancestral** and **derived** traits. When considering genetic data, a trait is an allele, i.e. a specific haplotype in a genetic region. Ancestral traits were shared by the ancestors and represent the old state, whereas derived traits are evolutionary changes that occurred more recently and represent new states.
2. **Origin of new lineages:** New lineages originate from ancestral ones. A **bifurcation** is the origin of two new lineages, and a **multifurcation** the origin of multiple lineages from a **common ancestor**.
3. **Extinction of evolutionary lineages:** Based on the analysis of the fossil record, Darwin concluded that for each lineage we observe in the present, many more lineages likely have gone extinct. This not only refers to the evolution of species, but also to the lineages within species, which will be relevant in the context of coalescence theory.

In summary, phylogenetic analysis is a method to group entities that are more similar to each others than to others by reconstructing the ancestral tree. There are numerous questions in the context of plant genetic resources that can be analyzed with phylogenetic methods. They include

- the reconstruction of domestication history
- the identification of extended gene pools of close and distant relatives to be used for introgression
- the identification of gene flow between wild and domesticated species, between modern varieties and land races, and between conventional cultivars and genetically modified organisms (GMOs)
- the analysis of the varieties released by commercial competitors

4 The anatomy of a phylogenetic tree

A phylogenetic tree consists of **nodes** and **branches** (Figure 4). A node connects two branches in strictly **bifurcating** tree.

In a **multifurcating** tree, three or more branches can be connected in a node.

The terminal branches are called leaves, or **operational taxonomic units (OTUs)**. They are the unit of analysis such as species, individuals or genes.

Trees can be **rooted** or **unrooted**. In rooted trees, the node with the **most recent common ancestor (MRCA)** can be identified (Figure 5)

To root a tree, an **outgroup** OTU is needed. An outgroup is an OTU known (or believed) to be ancestral to all other OTUs in the tree. In crop species, the wild ancestors of cultivated species are often suitable outgroups.

Branches in a tree can be presented **scaled** or **unscaled**. If they are scaled, the length of the branches is proportional to the distance of the OTUs that are connected by the branches. The evolutionary distance of OTUs is quantified by a measure. Such measures can be 'Million years', or a more abstract measure, such es 'Mutational distance' (i.e., the number or

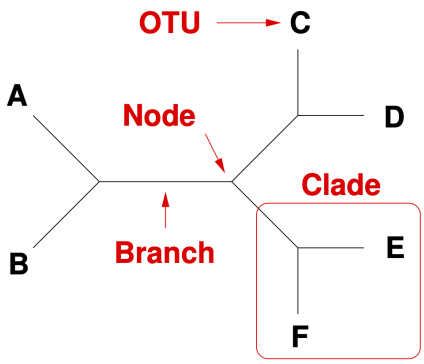
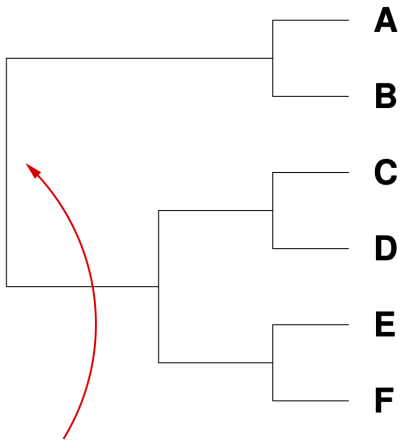


Figure 4: Example of an unrooted tree label:fig:unrooted_tree_annotate



Most recent common ancestor (MRCA)

Figure 5: Example of a rooted tree

proportion of mutations that occurred between two taxa). In coalescence theory, a frequent measure is 'Number of generations'.

There are two types of common ancestry in a tree. **Monophyletic** OTUs have a common ancestor in the tree whereas **polyphyletic** OTUs do not have one.

Phylogenetic trees can be represented as a tree, or in a format that can be read by computers. The most popular computer-readable format is called **Newick format**. The Newick notation for the tree in Figure 4 is

```
((A,B),(C,D),(E,F))
```

The Newick format is also able to express the branch lengths. The same tree as above but with different branch lengths can be expressed as:

```
((A:2.0,B:1.0):2.3,(C:0.5,D:0.5):4,(E:3.14,F:2.7):0.5)
```

Phylogenetic trees that were constructed from data such as DNA sequences are **inferred trees** because it is a tree that was *estimated* from a particular data set with a certain method. It may or not be the true tree.

Simple code to plot the above tree is shown in the following code[1]

[1] If you want to run this code, you need to install the package `phytools` first.

```
library(phytools)
tree <- "((A:2.0,B:1.0):2.3,(C:0.5,D:0.5):4,(E:3.14,F:2.7):0.5);"
phy <- read.newick(text=tree)
plot(phy)
```

As a first example of a phylogeny in the context of plant genetic resources, Figure 6 shows the phylogeny of barley landraces (Zeng et al., 2018). It was taken from a study that was aimed at elucidating the evolutionary history of Tibetan barleys, which are morphologically distinct from other barley landraces. The phylogeny shows that Tibetan barleys are also genetically distinct from barley varieties from the region of domestication.

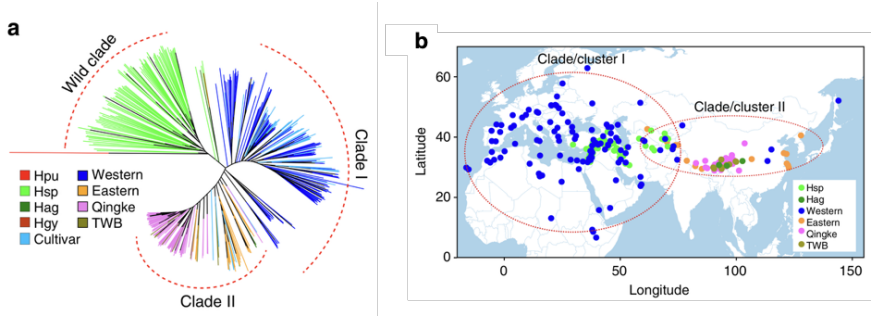


Figure 6 – *Phylogenetic relationship of wild and domesticated barley. Wild clade are accessions of wild barley* Hordeum spontaneum*, Clade I are barley landraces from the Near East (center of domestication, Europe and North Africa. Clade II are barley landraces and subspecies from Eastern regions. Abbreviations: Hpu* H. pubiflorum*, Hsp* H. spontaneum*, Hag* H. agriocrithon*, western: landraces in clade/cluster I, eastern landraces in clade/cluster II, TWB Tibetan weedy barley. (a) Neighbor-joining clustering based on genetic distance. (b) Global distribution of barley clade/clusters revealed by (a). Source: Zeng et al. (2018)*

5 Methods for calculating phylogenetic trees

In the following we only consider the construction of phylogenetic trees from SNP markers or DNA sequences. Phylogenies can also be created from other types of data such as phenotypic data.

There are four main methods for constructing phylogenetic trees:

- Distance-based
- Parsimony-based
- Maximum likelihood
- Bayesian method

Common to all methods is that they take some form of sequence alignment or data matrix as an input and then calculate a phylogenetic tree.

6 Distance based methods

Distance trees are calculated in two steps. First, a **distance matrix** is calculated from a multiple sequence alignment. All possible pairwise comparisons of sequences are made and then the pairwise distance is calculated. The result is a $n \times n$ matrix of pairwise distances. The **pairwise distance** can be calculated with several methods. In a second step, a tree is calculated from the distance matrix. Frequently used methods are **UPGMA (Unweighted Pair Group Method with Arithmetic Mean)** or **Neighbor Joining** (NJ).

6.1 Constructing the pairwise distance matrix

If one compares different pairs of sequences, it is obvious that two sequences that differ by 30% of their sequence are less closely related than two sequences that differ by only 10% of their sequences.

Usually, this means that the common ancestor of two divergent sequences is more distant in the past than of two highly similar sequences.

Several methods are available to calculate the distance between a pair of sequences.

They depend on whether the distance is calculated from a matrix of markers (e.g., a 0/1 matrix for a set of markers) or whether DNA or protein sequences are being compared.

```
SEQ1 GATCTATCTACTACT
SEQ2 ..G.....C....A.
```

The simplest distance measure is the **Hamming distance** (Wikipedia), which corresponds to the proportion of nucleotide differences between two sequences. In the above example 3 out of 15 nucleotides differ between the two sequences, and the Hamming distance is therefore 3/15 or 0.2.

Other methods such as **Kimura's distance** (Wikipedia) take into account rates of evolution between different nucleotides and are therefore

biologically more realistic. For example, it has been observed that usually there is a higher proportion of change from GC to AT nucleotides than vice versa, which can be taken into account into the models. For example the distance accoring two Kimura's two parameter model is calculated as

$$K = -\frac{1}{2}\ln((1-2p-q)\sqrt{1-2q} \quad (1)$$

with p as proportion of sites with transitional differences (purine to purine, $A \leftrightarrow G$, or pyrimidine to pyrimidine, $C \leftrightarrow T$ and q as proportion of sites with transversional differences, (pyrimidine to purine or purine to pyrimidine). In the above example, 1 of 15 sites show a transition and two of 15 sites a transversion, therefore $p = 1/15 = 0.0666$ and $q = 2/15 = 0.1333$. The Kimura distance can then be calculated as

```
K = -0.5*log((1-2*(1/15) - (2/15))*sqrt(1-2*(2/15)))
K
```

```
[1] 0.2326162
```

6.2 Clustering with the UPGMA method

UPGMA is the simplest method for clustering. The algorithm finds the pair of taxa with the smallest distance between them and defines the branching point between them als half of that distance. Effectively, a midpoint is placed on the branch. The two taxa are combined into a cluster, which is the substitute for the two taxa. The number of individuals is reduced by one and a new pairwise distance matrix is calculated with all the other sequences. The process is repeated until the matrix consists of a single entry. Then, a tree is reconstructed from the distances.

The following steps are used to construct a UPGMA tree:

1. Find the two OTUs i and j with the least distance K_{ij}.
2. Join i and j by creating a new node. Both branches that link i to (i, j) and j to (i, j) get the length $K_{ij}/2$.
3. Calculate the distances between the new group (i, j) and all other groups $k (k \neq i$ and $k \neq j)$ as:

The rows and columns of the distance matrix that correspond to j and i are deleted and replaced by a row and column of group i, j. If there is only a single cell in the matrix left, the algorithm is finished, otherwise start again. At later stages i and j are the newly defined groups and taxa.

6.3 Example of a UPGMA calculation

The following steps are necessary to calculate a UPGMA tree from the following sequence alignment.

First, get a sequence alignment:

OTU	1	2	3	4	5	6	7	8
A	G	C	A	G	T	A	C	T

OTU	1	2	3	4	5	6	7	8
B	G	T	A	G	T	A	C	T
C	A	C	A	A	T	A	C	C
D	A	C	A	A	C	A	C	T

Second, calculate a pairwise distance matrix:

	A	B	C	D
A	0			
B	1	0		
C	3	4	0	
D	3	4	2	0

Then select the pair with the lowest distance and calculate a new distance matrix as arithmetic mean with the following equations.

The pair with the lowest distance is A, B.

$$\begin{aligned} K_{C(A,B)} &= \frac{1}{2}K_{C,A} + \frac{1}{2}K_{C,B} \\ &= \frac{1}{2} \times 3 + \frac{1}{2} \times 4 \\ &= 3.5 \end{aligned}$$

Analogous,

$$\begin{aligned} K_{D(A,B)} &= \frac{1}{2}K_{D,A} + \frac{1}{2}K_{D,B} \\ &= \frac{1}{2} \times 3 + \frac{1}{2} \times 4 \\ &= 3.5 \end{aligned}$$

From these calculations, we construct a new table:

	(AB)	C	D
(AB)	0		
C	3.5	0	
D	3.5	2	0

We then select the next pair with the smallest distance, which is C and D, and calculate the new matrix as

$$\begin{aligned} K_{(A,B),(C,D)} &= \frac{1}{2}K_{(A,B),C} + \frac{1}{2}K_{(A,B),D} \\ &= \frac{1}{2} \times 3.5 + \frac{1}{2} \times 3.5 \\ &= 3.5 \end{aligned} \tag{2}$$

The new matrix is then

	(AB)	(CD)
(AB)	0	
(CD)	3.5	0

The resulting unrooted tree is shown in Figure 7.

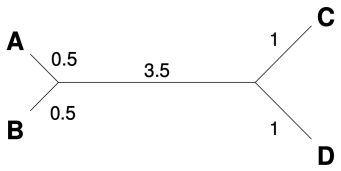


Figure 7: UPGMA tree calculated from the above data. Numbers indicate the distances branch lengths.

6.4 Neighbor-joining (NJ) clustering method

The **Neighbor-joining (NJ)** method (Wikipedia) is similar to UPGMA because it utilizes a pairwise matrix and reduces the size at each step. However, it calculates distances to internal (rather than external) nodes.

NJ is more frequently used than UPGMA because it does not assume that all taxa are equidistant from a root, which is a biologically very unrealistic assumption.

7 Maximum Parsimony

The **maximum parsimony** (MP) method attempts to find a tree that is constructed with the least number of substitutions to explain the observed differences in a tree. It is based on the philosophical principle of **Occam's razor** that states that simpler scientific explanations are to be preferred over complex ones.

The approach is to compare all possible trees with a given number of species and then to test which tree requires the smallest amount of changes (i.e., mutations) to make the data consistent with the tree. This is an application of the **parsimony criterion**.

Parsimony analysis has the disadvantage that branch lengths can not be calculated. The resulting tree provides only the **topology** of a phylogenetic tree.

8 Maximum Likelihood

This method is based on an explicit evolutionary model of sequence evolution and tries to find the tree that has the **highest likelihood** to be consistent with a given data set.[2] Maximum likelihood is a statistical approach that was developed by R. A. Fisher and attempts to maximise the probabilitiy of the data given a model:

The term $P(\text{Data}|\text{Model})$ is called **likelihood**. The likelihood L of a tree is the probability P that the observed data can be achieved with a given tree and a chosen matrix of transition probabilities between transitions and transversions of nucleotides.

The ML method tries to find the tree with the highest value of L. To calculate the total likelihood of a tree, individual likelihoods of each nucleotide position in the alignment are calculated. The total likelihood is then the sum of the individual likelihoods (Figure 8).

A key assumption of ML is that different alignment positions evolve independently. A second assumption is independent evolution of different lineages. Then, the total likelihood of the tree for the whole alignment is the product of the likelihoods of the individual nucleotides, L_i:

The important step is to calculate the likelihood for the individual nucleotides. To do this, all possible trees for a single alignment position are considered. In the case of a sequence alignment with four sequences, this is 16 different trees (Figure 8). The probabilities are calculated by using an evolutionary

[2] The term likelihood is closely related to probability and values also range from 0 to 1. It can be described as the probability of a phylogenetic tree model with a certain set of parameters (topology and branch length) given the observed data, i.e., a sequence alignment.

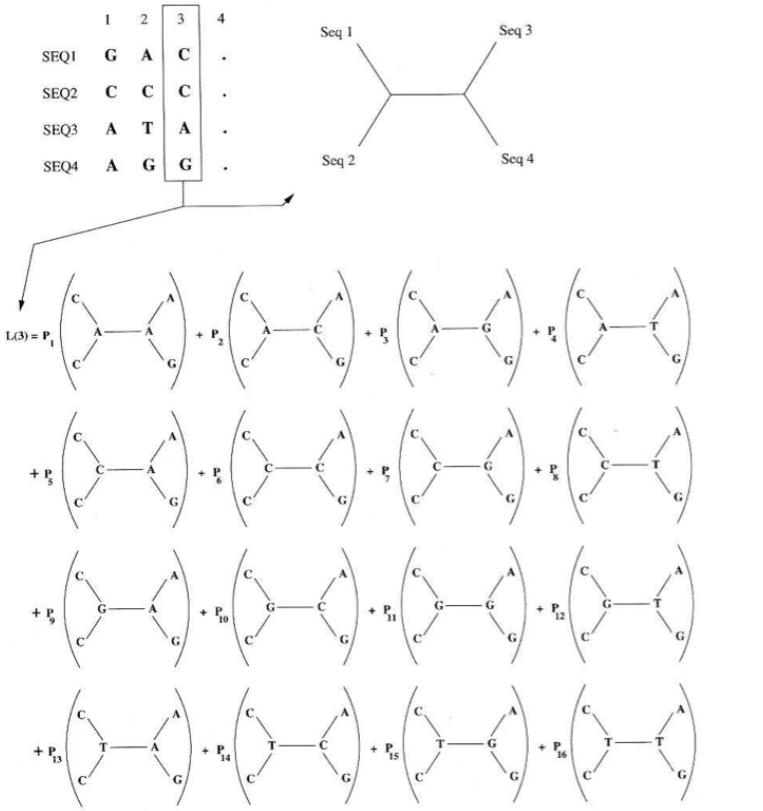


Figure 8 – *Calculation of the individual probabilities for a position in the alignment of four operational taxonomic units.. Source: Halliburton and Halliburton (2004)*

model of sequence evolution that gives probabilities of mutations between the different nucleotides of a tree. The evolutionary model results in a matrix that describes the mutation probabilities between the four nucleotides. One example of such a matrix is shown in the following table.

	A	C	G	T
A	0.7	0.1	0.1	0.1
C	0.1	0.7	0.1	0.1
G	0.1	0.1	0.7	0.1
T	0.1	0.1	0.1	0.7

This matrix is used to calculate the probability that the four observed nucleotides result by mutation from the 16 possible ancestral configurations. The following steps are taken for calculating the total probability of a tree:

1. Calculate the probability of an individual configuration of ancestral and current nucleotides by multiplying the mutation probabilities between nucleotides.
2. Repeat this calculation for each of the 16 possible configurations (in an alignment with 4 sequences)
3. Sum up all 16 probabilities. This is the likelihood $L_{(i)}$ for nucleotide position i.
4. Multiply the likelihoods for all individual nucleotide positions. The result is called composite likelihood.
5. Repeat this for a different possible tree. In an alignment of four sequences, three different trees are possible.

6. Choose the tree with the highest composite likelihood as the best tree.

Note that the maximum likelihood procedure for reconstructing phylogenetic trees explicitly assumes a model of evolution to calculate the matrix of mutation probabilities (e.g., the PAM-1 matrix for protein sequences or the Jukes and Cantor model for DNA sequences).

Therefore, the quality of the tree will depend on how *biologically realistic* the model of evolution is. ML methods are *computationally very expensive*, but many computer programs were developed that allow to calculate ML trees efficiently. The maximum likelihood method is able to estimate tree lengths. The best and most widely used methods for phylogenetic inference such as IQ-TREE[3] are based on Maximum Likelihood.

[3] http://www.iqtree.org

9 Bootstrapping: Confidence in Reconstructed Trees

Bootstrapping is a technique commonly used for estimating statistics or parameters when the distribution is difficult to derive analytically, i.e., with an explicit mathematical expression (Efron and Tibshirani, 1991). For example, phylogenetic inference from data in the following table results in a tree in which A and B are children of one ancestor and C and D are children of one another.

Taxon	1	2	3	4	5	6	7	8
A	G	C	A	G	T	A	C	T
B	G	T	A	G	T	A	C	T
C	A	C	A	A	T	A	C	C
D	A	C	A	A	C	A	C	T

It is interesting to obtain a measure of confidence that A and B belong together and that C and D belong together.

Bootstrapping creates pseudo-samples by choosing a site at random and placing it in column 1. This sampling with replacement process is repeated until the number of random columns generated equals the number of sites in the original sequences. Note that because of the sampling with replacement a given column can be sampled repeatedly and each pseudo sample likely omits different columns from the original alignment as shown in the following table, where columns 5 and 1 have been sampled twice each.

Taxon	1	2	*5*	4	5	6	*1*	8
A	G	C	T	G	T	A	G	T
B	G	T	T	G	T	A	G	T
C	A	C	T	A	T	A	A	C
D	A	C	C	A	C	A	A	T

After sampling, the pseudo-sample is used to construct a tree, and for each node in the original tree it is checked, whether it also is present in the bootstrapped sample. The whole process of bootstrapping and analysis of the bootstrapped samples is repeated many times (usually $> 1,000$ times). The confidence for each pair clustering together (i.e., A clustering with B) is

then the fraction of times they appear together in the trees generated from many pseudo-samples.

A bootstrap value of 95 then indicates that in 95% of the pseudo-samples OTUs A and B are clustered together, which is quite a good support. For robust inference, bootstrapping requires hundreds or thousands of samples.

10 Which method should one use?

The choice of the four very different approaches presented above[4] for solving the same problem makes it necessary to select between methods. There is no objective criterion which method to choose, and there have been almost religious wars between the adherents of the different approaches, in particular between ML and Bayesian phylogenicists. One can, however, follow a pragmatic approach:

- Among distance-based methods, UPGMA is not optimal because it assumes a stricht **molecular clock** (rates of evolution are the same in all branches of the phylogeny). In contrast, Neighbor-Joining corrects for different rates of evolution.
- Use at least two different methods (e.g., NJ and ML) and compare the resulting trees. If they are similar (especially at the important nodes), they can be considered to be robust.
- Popular programs for these analyses are RAxML (https://cme.h-its.org/exelixis/web/software/raxml/) or BEAST/SNAPP (http://www.beast2.org/). They can be difficult to use, but some of the simpler algorithms are also implemented in R packages.
- Use different models of sequence evolution to check whether the results are robust.
- There are methods available to formalize **model selection** of different models of sequence evolution. Computer programs are available for model selection, e.g., jMODELTEST at http://darwin.uvigo.es.
- Use bootstrapping to check whether nodes are supported by each method.
- With large data sizes, use faster methods. In particular, ML and Bayesian approaches can have long running times.

[4] We did not present the Bayesian approach to constructing phylogenies. See the review by Nascimento et al. (2017) or the Wikipedia article.

11 Applications to crop species

11.1 Phylogeny of the rye genus Secale

The phylogeny of the rye genus *Secale* was calculated using AFLP markers and the UPGMA method (Figure 9). Numbers on branches indicate percent bootstrap support. For interpretation, the taxon *Secale cereale* is domesticated rye, or other species are either the ancestors or crop wild relatives (CWR) of domesticated rye.

The tree provides several insights:

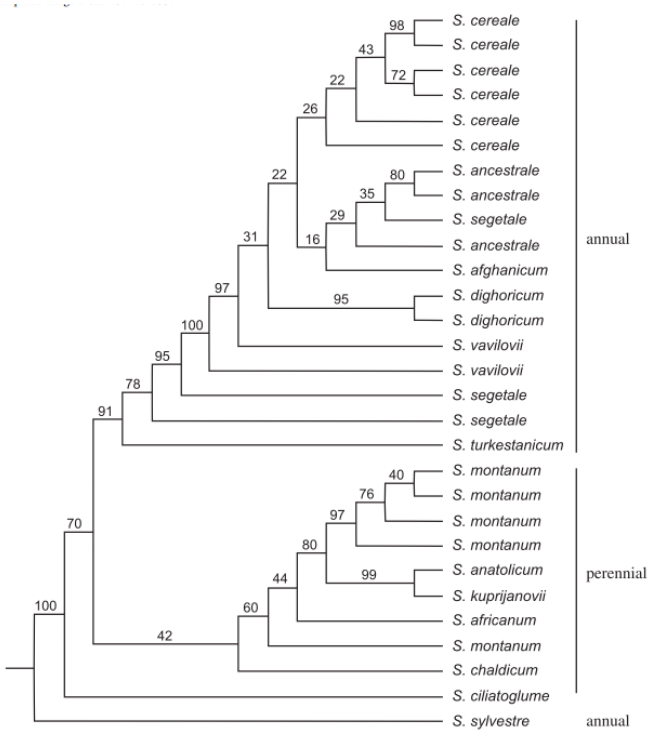


Figure 9 *– AFLP-based phylogeny of species from the genus Secale. Source: Chikmawati et al. (2005).*

Bootstrap values vary widely: Some branches are supported very well, whereas others are not. For example the clustering of the various lines of cultivated rye, *Secale cereale* is expected, but the data do not support a strong separation of closely related forms such as *Secale ancestrale*.

Polyphyletic origin of taxa: Some taxa can be found at different positions in the tree, such as *S. segetale* or *S. montanum*. This suggests that they do not have a common ancestor even though they have the same name, and they are therefore **polyphyletic**. However, an alternative explanation is that the markers used are not powerful enough to resolve the tree, or that individual plants have been misclassified and given incorrect names.

11.2 Phylogeny of rice landraces and varieties

The phylogeny is based on 160,000 SNP markers that were detected by microarray hybridizations (Figure 10). The three major groups of rice (*indica, japonica, austral*) are shown in different colors. The tree was reconstructed with an unweighted neighbor-joining method.

Note that bootstrap values (or another measure of statistical support) is not shown, but the branch lengths are indicated.

The tree shows that the three major groups of rice are well separated, i.e., have an independent phylogenetic history, and that the *aus* and *indica* groups are more closely related with each other.

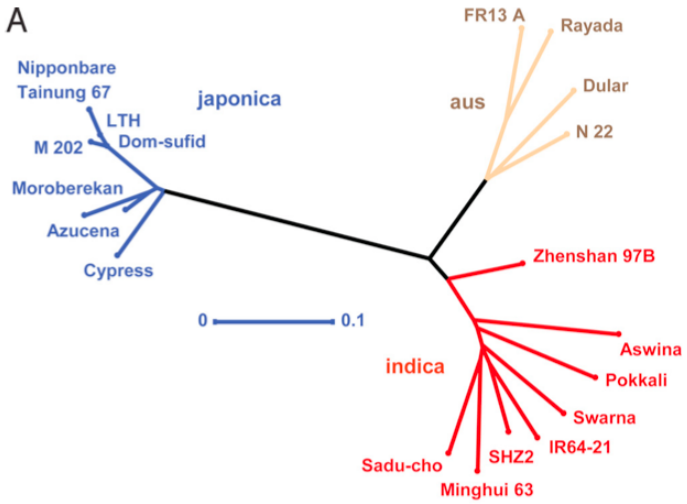


Figure 10 *– Genome-wide variation of rice landraces and varieties based on a SNP genotyping array. Source: McNally et al. (2009).*

12 Key concepts

- ☐ OTU
- ☐ Bifurcating vs. multifurcating tree
- ☐ MRCA
- ☐ Clade
- ☐ Newick format
- ☐ Bootstrapping

13 Summary

- Phylogenetic methods reconstruct the past from present data. However, with the advent of the sequencing of ancient DNA, phylogenetic analysis does not depend on present data anymore.
- There are four major methods for reconstructing phylogenies: Distance-based, maximum parsimony, maximum likelihood and Bayesian methods that differ in their assumptions, models and statistical approaches.
- There are no objective criteria for selecting a particular method, but by comparing methods and models, it is possible to evaluate the robustness of phylogenetic inference.
- Phylogenetic analysis is a very important method for characterising plant genetic resources.

14 Further reading

Books:

- Hartl and Clark, *Principles of Population Genetics*, 4th edition, Chapter 7.8 - An introduction into phylogenetic analyis
- Barry Hall: *Phylogenetics Trees made easy*, 5th edition. Sinauer Associates – A How-To manual of phylogenetics

- Joseph Felsenstein: *Inferring Phylogenies*. Sinauer Associates – The standard book on phylogenetic methods.

Articles:

- Gregory (2008) – A very accessible introduction to *tree thinking* and phylogenetics
- Kapli et al. (2020) – A state of the art review on phylogenetics for the genomics age.
- Smith et al. (2020) – A review on using phylogenetics to understand the evolution of phenotypic traits
- Zuntini et al. (2024) – Currently one of the largest efforts to understand the phylogeny of seed plants

15 Study questions

1. What is the difference between a bifurcating and a multifurcating tree? What are biological processes may contribute to bifurcating and multifurcating trees?
2. Which examples of ancestral and derived traits do you know in the context of crop plants and their wild ancestors, and in plants generally?
3. What are the different steps in the construction of a distance-based tree?
4. What are the differences between distance-based and maximum likelihood methods?
5. Why is it an important assumption that sites evolve independently in maximum likelihood trees (Hint: Think about the addition and multiplication rules of probability theory)
6. What are ancestral and derived traits?
7. Why is it difficult to select a correct method for tree reconstruction? What would be the best method for selecting the best tree?
8. Where are discrepancies between marker-based inference of phylogeny and the taxonomic classification (species names) in Figure 9? What are possible explanations for the differences?

16 Problems

16.1 Problem 1

Search for the figure Figure 10 of this handout in the paper of McNally et al. (2009) and try to find out what a distance of 0.1 means in this figure.

16.2 Problem 2

Consider the following distances, based on the allele sharing distance, between different accessions of maize landraces BKN* and wild teosinte TIL* from a paper on maize evolution.

Based on the first 10,000 SNPs in a 3 Mb region

	BKN009	BKN029	TIL01	TIL07
BKN009				
BKN029	0.08			
TIL01	0.33	0.34		
TIL07	0.21	0.20	0.37	

Based on the 30,001th - 40,000th SNPs

	BKN009	BKN029	TIL01	TIL07
BKN009				
BKN029	0.11			
TIL01	0.32	0.34		
TIL07	0.21	0.21	0.25	

Reconstruct the phylogeny of the OTUs using the UPGMA method.

16.3 Review: The anatomy of a phylogenetic tree

Review the following technical terms which are important when describing and discussing a phylogenetic tree: branch, node, leave, OTU, unrooted phylogeny, rooted phylogeny, clade (monophyletic, paraphyletic, polyphyletic).

16.4 The Newick format

Phylogenetic trees are usually expressed in the Newick format which is computer-readable. 1. Write the phylogenetic tree drawn by the instructor in Newick format. Is an alternative possible? 2. A phylogenetic tree is given in Newick format: ((((X,D),F),(S,Y)),W). Draw the tree.

16.5 Number of possible phylogenetic trees.

1. Assume you have three OTUs. How many different unrooted phylogenies can you get with three OTUs?
2. Assume you have four OTUs. How many different unrooted phylogenies can you get with four OTUs?
3. Assume you have five OTUs of which one is an outgroup. How many different rooted phylogenies can you get?

You may want to search the web for a formula that allows to model the number of possible trees given a number of taxa.

16.6 Phylogenies in scientific studies (In class exercise)

Phylogenies are often not that straightforward to read and interpret. In this exercise you will investigate some phylogenies from scientific publications in order to familiarise yourself with them.

Examples for studies that used phylogenetic methods in the context of plant genetic resources:

- Lychee genome
- Tomato relatives
- Cauliflower diversity
- Soybean ERF gene family

You will work in groups. Each group is given a phylogeny from a scientific publication. Afterwards you will present your findings to the other groups.

Note: There is no need to read the publication. It is enough to skim through it to collect the relevant information. Also, it is very possibly that not all information is given.

Answer the following questions: - What is shown in this phylogeny? What data was used? - Which method was used to make this phylogeny? Which software was used? - Is it a rooted or an unrooted phylogeny? - Did the authors use an outgroup sequence? - Are the genetic distances shown, i.e. the branch lengths? - Are there any bootstrap values and if yes, how well do they support the nodes? - Is there any clustering of samples? If yes, which are the clusters, how many are there and are all samples in clusters? - Try to interpret and discuss the phylogeny (e.g. which samples are ancestral, which ones are derived).

References

Chikmawati T, Skovmand B, Gustafson J. 2005. Phylogenetic relationships among secale species revealed by amplified fragment length polymorphisms. *GENOME* **48**:792–801.

Efron B, Tibshirani R. 1991. Statistical data analysis in the computer age. *Science* **253**:390–395.

Gregory TR. 2008. Understanding Evolutionary Trees. *Evolution: Education and Outreach* **1**:121–137. doi:10.1007/s12052-008-0035-x

Halliburton R, Halliburton R. 2004. Introduction to population genetics. Pearson/Prentice Hall Upper Saddle River.

Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* **21**:428–444. doi:10.1038/s41576-020-0233-0

McNally K, Childs K, Bohnert R, Davidson R, Zhao K, Ulat V, Zeller G, Clark R, Hoen D, Bureau T, Stokowski R, Ballinger D, Frazer K, Cox D, Padhukasahasram B, Bustamante C, Weigel D, Mackill D, Bruskiewich R, Rätsch G, Buell C, Leung H, Leach J. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci USA*. doi:10.1073/pnas.0900992106

Nascimento FF, Reis M dos, Yang Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution* **1**:1446–1454. doi:10.1038/s41559-017-0280-x

Smith SD, Pennell MW, Dunn CW, Edwards SV. 2020. Phylogenetics is the New Genetics (for Most of Biodiversity). *Trends in Ecology & Evolution* **35**:415–425. doi:10.1016/j.tree.2020.01.005

Zeng X, Guo Y, Xu Q, Mascher M, Guo G, Li S, Mao L, Liu Q, Xia Z, Zhou J, Yuan H, Tai S, Wang Y, Wei Z, Song L, Zha S, Li S, Tang Y, Bai L, Zhuang Z, He W, Zhao S, Fang X, Gao Q, Yin Y, Wang J, Yang H, Zhang J, Henry RJ, Stein N, Tashi N. 2018. Origin and evolution of qingke barley in Tibet. *Nature Communications* **9**:1–11. doi:10.1038/s41467-018-07920-5

Zuntini AR, Carruthers T, Maurin et al Olivier. 2024. Phylogenomics and the rise of the angiosperms. *Nature* 1–8. doi:10.1038/s41586-024-07324-0