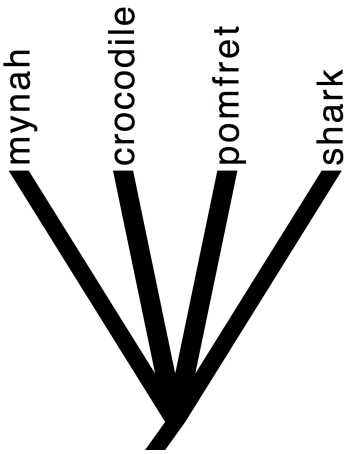
Estimating phylogenetic trees

Introduction

We want to know the relationships among four species – mynah, crocodile, shark and pomfret. Because evolution is a branching process (due to speciation) we will estimate relationships as a phylogenetic tree. The phylogenetic tree is a graph in which we show closely related species as being closely connected, while more distantly related species are more distantly connected on the graph.



In this tree, we do not know the relationships among the four species. Are mynahs more closely related to crocodiles than to pomfrets? Are sharks more closely related to pomfrets than to crocodiles? Can we answer these questions? How to get answers?

This is what we want to find out.

1 Character, character state and data matrix

We assess relationships based on similarities and differences in characters. *Characters* are features of organisms that are heritable and may differ between species. Different forms of a character are character states (compare to the concept of the gene and its different forms, or alleles).

Why do species look similar to each other?
Similarity may occur between two species because the same character state is inherited from the common ancestor of the two closely related species, and remains similar during evolution (*homology*). Similarity may also arise in two distantly related species because of selection (*homoplasy: parallelism, convergence*) and other factors. This similarity is due to these evolutionary processes, but not due to common (shared) evolutionary history. Not all similarities are the same.

"Good" taxonomic characters are homologous. We use homologous characters to compare character states in different species to make up a *data matrix* of characters and species, then use the data matrix to classify the species. [Units in a phylogenetic analysis may also be individuals, genes etc.]

Why do species differ from one another?
After they arise due to a split in an ancestral species, two new species may follow different evolutionary trajectories and diverge from each other in some characters, but not in others. This allows us to recognise different species and also to classify them based in these shared similarities and differences.

Data matrix

Character / Species	Vertebrae	Bony skeleton	Limbs
Shark	Present	Absent	Absent
Pomfret	Present	Present	Absent
Crocodile	Present	Present	Absent
Mynah	Present	Present	Present

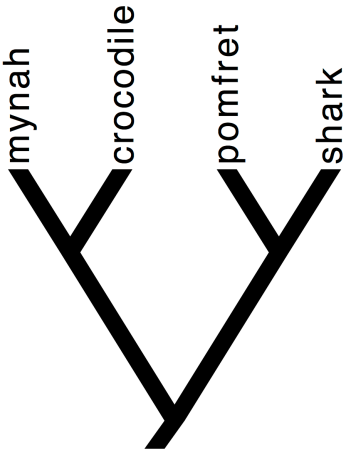
Convert the data matrix into a coded form:
The character vertebrae has 2 character states: absent (0) and present (1)
The character bony skeleton has 2 character states: absent (0) and present (1)
The character limbs has 2 character states: absent (0) and present (1)

Character / Species	Vertebrae	Bony skeleton	Limbs
Shark	1		0
Pomfret		1	
Crocodile	1		
Mynah			1

2 Methods to analyse evolutionary relationships

2.1 Grouping based on overall similarity and differences (e.g., UPGMA)

- Sharks and pomfrets share 2 character states; sharks and crocodiles share 1 character state; sharks and mynahs share 1 character state.
- Pomfrets and sharks share 2 character states; pomfrets and crocodiles share 2 character states; pomfrets and mynahs share 2 character states.
- Crocodiles and mynahs share 3 character states; crocodiles and pomfrets share 2 character states; crocodiles and sharks share 1 character state. Mynahs and sharks share 1 character state; mynahs and pomfrets share 2 character states; crocodiles and mynahs share 3 character states.



Sharks and pomfrets share more character states with each other than with either crocodiles or mynahs.

So sharks and pomfrets are more closely related to each other than to crocodiles or mynahs.

Crocodiles and mynahs share more character states with each other than to either sharks or pomfrets.

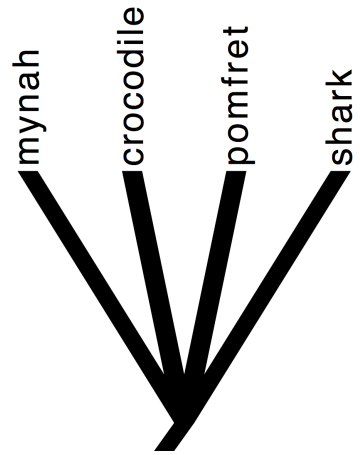
So crocodiles and mynahs are more closely related to each other to either shark or pomfret.

2.2 Classification based on evolutionary similarities and differences (e.g., maximum parsimony, MP)

A character in two species may be homologous because it was present in the common ancestor of four species; some species keep the character state (shared ancestral character), e.g., absence of limbs in fish. This does not indicate a closer relationship between the fish compared to the other species, because their older ancestors (not shown) too did not have limbs.

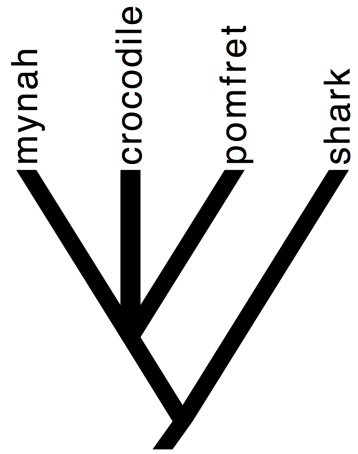
Another character state in two species may be homologous because it arises within the group and was not present in the common ancestor of all the species (shared derived character), e.g., bony skeleton in pomfret, crocodile and mynah. This trait does point to the close relationships of these species, because neither sharks nor the older ancestors had a bony skeleton.

Not all homologies are the same or of equal importance as evidence of shared evolutionary history. The second type of homology is thought to provide better evidence that two species share a recent common ancestor. Such traits are given more importance in maximum parsimony (MP) analysis.



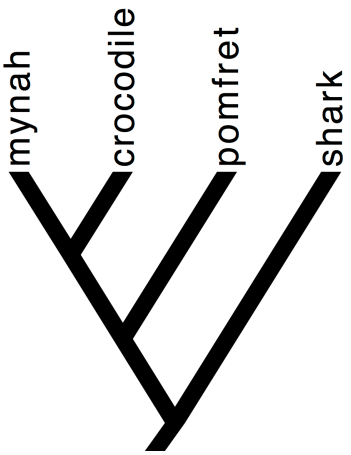
Vertebrae are present in all the species in this set of species. This indicates that all these species have a common ancestor who had vertebrae. However, this trait carries no information about relationships of the species with each other.

So the tree cannot be resolved any further using the trait, vertebrae.



The bony skeleton (a derived trait) is shared by pomfrets, mynahs and crocodiles.

This trait indicates that the three species share a common ancestor with each other but not with sharks. So we can put pomfrets, mynahs and crocodiles in one group separate from sharks.



Limbs are present in both mynahs and crocodiles. Presence of limbs is a shared derived character that was not present in pomfrets or the common ancestor they share with pomfrets.

This trait indicates that mynahs and crocodiles share a more recent common ancestor with each other than with pomfrets.

So we put mynahs and crocodiles in one group.

The tree is fully resolved–the position and relationships of every species is known.

Describe the relationships of these species.

Do you see how the overall similarity criterion can sometimes lead us astray? Because sharks and pomfrets are so similar to each other and look like what we know as 'fish' there seemed to be no question that they should be most closely related to each other. In fact, pomfrets are more closely related to birds and mammals!

Because of the interesting and unpredictable ways in which evolution proceeds, we often may be misled by this criterion of overall similarity in trying to establish evolutionary relationships.

Nowadays we try to take evolutionary processes into account when assessing evolutionary relationships. Maximum parsimony (MP) is an intuitive and easily understood criterion that is a good introduction to phylogenetic analysis. Other, more frequently used methods include maximum likelihood (ML), maximum posterior probability (MPP, Bayesian) and neighbour joining (NJ) methods.

3 A (hypothetical) data-matrix of sequences

Molecular data can be analyzed exactly like morphological characters. First, the DNA (or RNA or protein) sequences of different species are aligned to match up nucleotides to maximize the overall similarity. Each site is a character (homologous across the species) and the four nucleotides are different character states. This is equivalent to the morphological data matrix we used above.

Below is a very small molecular data matrix of short (10 bases) hypothetical sequences.

Character / Taxa	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Site 10
Human	A	C	C	T	T	T	G	T	A	G
Mynah	A	C	C	T	T	A	G	T	A	G
Mushroom	A	C	C	T	T	C	G	T	T	G
Mango	A	C	C	T	A	G	G	T	C	C

You can analyse this data matrix just as you did the morphological one of animals above.

To make things easier, first identify the characters that do not vary.

How many characters do not vary? Which ones? Can you use them to group any of the species together?

Character 6 is different in each taxon – can it be used to group the taxa?

In the remaining characters, Follow the same method as you used above for the vertebrate tree. Look for a character that separates the largest group possible (3) from smallest number possible (1). Then look for character/s that can further separate the species in the group of three. Do you get a fully resolved tree?

Draw this fully resolved phylogenetic tree and describe it.

4 Maximum Parsimony – an example

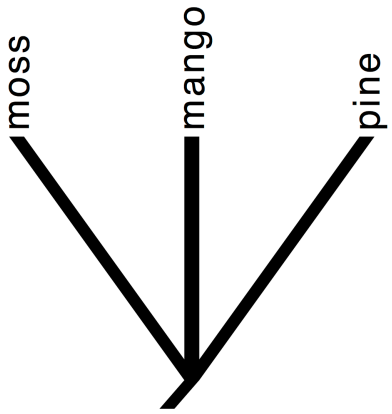
We can do the kind of argumentation described above for a few taxa and characters, but it can quickly become tedious with more characters and taxa. Usually, computers are used to analyse large data sets.

This introductory exercise will use an approach used by computers, but using a small data matrix.

Fill in the data matrix using these data:
Cell walls are present in mosses, pines and mangoes. [code absent = 0; present =1]
Vascular tissue is present in pines and mangos but not mosses. [code absent = 0; present =1]
Seeds are present in pines and mangoes but not in mosses. [code absent = 0; present =1]

Character / Species	Cell Walls	Vascular tissue	Seed
Moss			
Pine			
Mango			

Choosing among different possible trees using maximum parsimony:



This is an unresolved tree of three taxa moss, mango and pine.

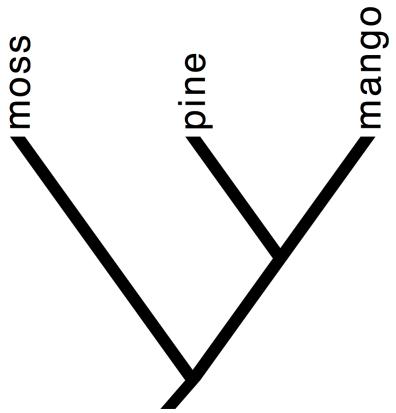
How are these three taxa related to each other?

There are three possible hypotheses. Draw them all.

There are two basic principles of maximum parsimony:

- If two taxa share a derived character state the change must have occurred in the common ancestor (instead of independently in the two descendents, by parsimony)
- Choose the tree that requires the fewest number of changes (steps) in characters. (Because different characters may group taxa differently, we try to minimize such conflict by parsimony.)

Now determine where the character states change on the tree. Mark the branch where the change must have occurred. In all cases, assume that presence is a derived state and absence is an ancestral state.



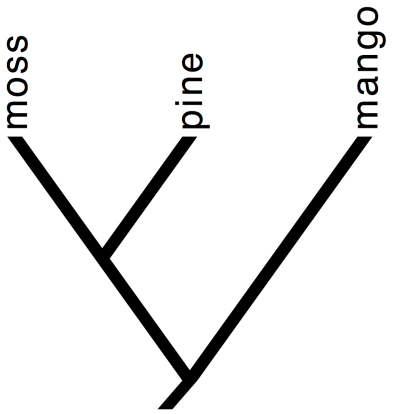
Cell walls are present in all the taxa, so there was no change in this set of taxa.

Vascular tissue is present in pines and mangoes but not mosses. It must originate in the common ancestor of pines and mangoes. Which branch should you mark?
This would be one change.

The seed has a similar pattern. Where does the change happen?

What is the total number of changes in character states required on this tree?
[Answer: 2]

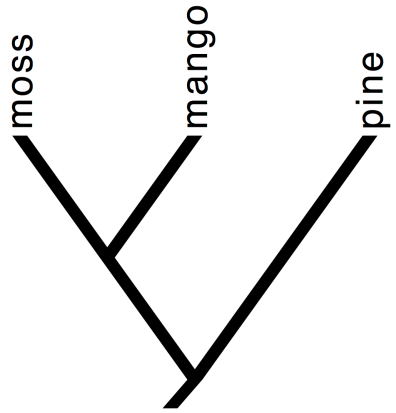
Do the same for the other two trees:



Where does vascular tissue arise?

Where do seeds arise?

How many changes on the tree?
[Answer: 4]



Where does vascular tissue arise?

Where do seeds arise?

How many changes on the tree?

Which tree requires the fewest number of changes?
This would be the shortest, or MP tree.

5 Further exercises

Larger data matrices (whether morphological or molecular) are commonly analysed using computer programmes that apply different methods of phylogenetic analysis. There are also methods to assess the robustness or statistical support for the results.

Our understanding of the origins of coronaviruses, Beta-coronaviruses and SARS-CoVs are based on our knowing evolutionary relationships among these groups of viruses. How do we find out the relationships among viruses?

Usually these phylogenetic trees are obtained by analysis of aligned sequences of particular genes (X, Y) or entire genomes usually using neighbour-joining, maximum likelihood, or Bayesian methods, and usually involve very large datasets. There are websites e.g., Nextstrain (`https://nextstrain.org/`) that act as a point of access for sequence data for coronaviruses, as well as analytical tools – anyone can used these to get an updated picture of the evolutionary history of SARS-CoV-2, SARS-CoV like betacoronaviruses, and Betacoronaviruses.