Video Tutorial 4.1: Creating a phylogenetic tree

A phylogenetic tree is a diagram showing inferred evolutionary relationships between a set of organisms. Phylogenetic trees have a branching pattern, like this figure from Darwin's notebook, where he has sketched out this branching pattern to reflect the process of descent with modification; a process central to Darwin's theories of evolution. This second tree shows the inferred evolutionary relationships among Darwin's finches on the Galapagos Islands.

But how are phylogenetic trees like this one built?

Well, any characteristic can be used to infer relationships and build phylogenetic trees, and DNA sequence data has proven incredibly helpful in clarifying and building many important phylogenetic trees.

Using large DNA data sets to build trees can get quite complicated but the underlying principle and logic is pretty straight forward: sequences separated by shorter evolutionary distance are expected to be more similar to one another than sequences separated over longer evolutionary distances.

We'll go through a highly simplified example to familiarize ourselves with the basic logic of phylogenetic tree building from DNA sequence data. We will do this in a series of steps so you can pause the video at any point to work through the steps yourself, returning to the video to check your progress.

In our simplified example we have a set of 6 sequences. These are the equivalent stretch of sequence taken from the equivalent gene, from each of the six species. And we're just going to refer to the species as A through F.

The first step in the process is to line the sequences up against one another.

Here we have the sequences aligned so that the equivalent nucleotide in each sequence lines up to form a column. This makes comparing the sequences much easier.

With the sequences aligned we can now compare each sequence with one another. We'll do this in a pair-wise fashion starting with sequences A and B. So our next step is to determine how many differences there are between sequence A and sequence B.

When we do this we see that in three positions the sequences are the same. [slight pause]. And in nine positions they are different. So we have 9 differences between sequences A and B.

We can start building a table of the differences between the sequences and our first entry is 9 differences between A [slight pause] and B.

The next step is to compare sequences A and C to determine how many differences there are between these two sequences. In this case we can see that the sequences are much more alike with ten positions that are the same between the sequences [slight pause] and only two that are different. So our next entry in the table is 2 differences between A and C.

And we can continue comparing the sequences in this way until we have completed the rest of the table. This is our next step.

And this is what the completed table looks like.

With the table completed we can move to the next step which is to use the table to identify the sequences with the fewest differences between them. We will infer that these are the sequences that are the most closely related to one another.

In our table we see that the sequences with the fewest differences between them are A and C with only 2 differences, as well as B and E that also have only 2 differences between them.

With this information we can draw the first groupings on our phylogenetic tree.

We'll group A and C together to reflect the fact that these two sequences show the closest relationship we have here to one another. And we'll group B and E together to reflect the fact that they also show an equivalently close relationship to one another.

With the first groupings made on our tree we now need to re-work our table with the grouped sequences combined together as a grouping rather than two individual entries in the table. We'll start by combining A and C, this group.

To do this we'll take the average difference that A and C show to each of the other sequences. Let's start with the differences they show to B. We see there are 9 differences between A and B, and nine differences between C and B so the average difference of A and C to B is 9. And we can make that entry on our new table.

Let's move to the next position, D.

We can see there are 4 differences between A and D and there are 5 differences between C and D. So the average difference of A and C to D is 4.5. And we can add that entry on our new table.

We can complete the rest of the table in this way. [slight pause]. We have an average of 9 differences between "A and C" to E, and 10 differences to F. The rest of the table can be copied down from the first table.

With the A-C grouping now added to our new table we can proceed to also add the B-E grouping [slight pause] and to complete the table with B and E grouped together using the same approach as before.

For B and E we have an average of 9 differences to the A-C group, [slight pause]. We have 4.5 for A-C to D, and 10 for A-C to F.

For B and E to D we have an average of 6 differences, and an average of 10 for B and E to F, and there are also 10 differences between D and F.

With this table completed, we can proceed to the next step.

This step is to identify the sequences with the fewest differences between them in our new table.

We can see that the A-C group has only 4.5 differences to D so this is the next close relationship in our tree.

And we can add this next grouping to our tree [slight pause] like this, with D as the next grouping out from A and C, reflecting the fact that D is more closely related to them than it is to the other sequences we have.

With this new grouping added to the tree we need to re-work our table again with the A-C-D grouping incorporated.

We have an average of 7.5 for A-C and D to the B-E group. And we have an average of 10 for A-C and D to F. And we also have 10 differences between the B-E group and F.

With the new table completed [slight pause] we can now determine the next relationship by identifying the sequences in the table with the fewest differences between them again.

We can see that this is the A-C-D group with the B-E group, with an average of 7.5 differences. So we can add the next grouping to the tree [slight pause], like this, grouping the A-C-D group with the B-E group.

This now leaves us with one more sequence, F, that is equally distantly related to all the other sequences with 10 differences to each of them.

So we can add this last grouping to the tree like this, reflecting the distant relationship between F [slight pause] and all the other sequences [slight pause].

And this completes our phylogenetic tree built from these 6 DNA sequences.

The method we used to build this tree is a Distance Method (specifically an approach called unweighted pair-group method with arithmetic mean or UPGMA). And the approach we took works well for a straight forward situation like this. But when we start to look at larger and more complex data sets of DNA sequences things can become more complicated and we need to start taking a number of other factors into account.

And it's often a good idea to compare the results from multiple different approaches to develop a sense of how reliable different parts of your tree are.

But despite these potential complications the underlying logic remains, that sequences separated by shorter evolutionary distance will be more similar to one another than sequences separated over longer evolutionary distances.

DNA sequences therefore provide a powerful source of information to help us infer evolutionary relationships that can be depicted in the form of phylogenetic trees.