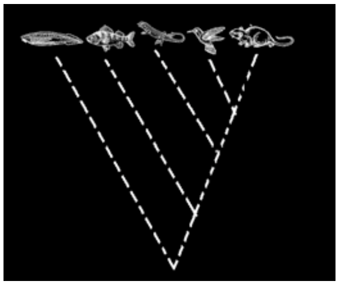
Molecular phylogeny textbooks & papers say:

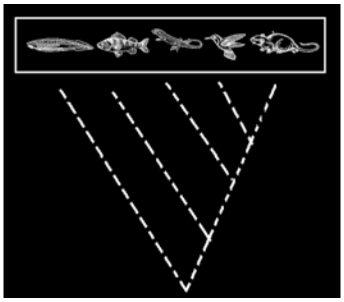
Which means, in other words:

1. Starting point: "Phylogenetic inferences are premised on the inheritance of ancestral characteristics, and on the existence of an evolutionary history defined by changes in these characteristics."[1]



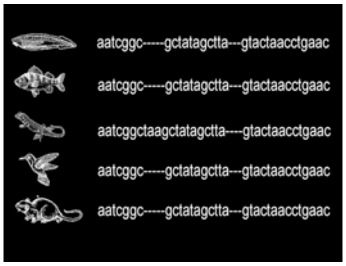
1. ***Assume*** that the organisms in question share a common ancestor. Some branching pattern (tree) links them all as relatives, and your task is to find that pattern. You're not asking "Do these organisms share common ancestry?" That's a given. Rather, you want to know how they are related (e.g., which group branched first).

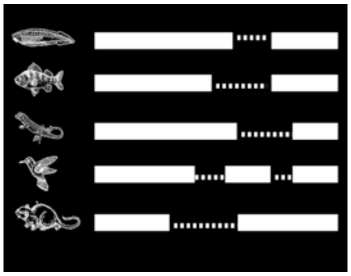
2. Choose the data: "The single most important component…of a phylogenetic analysis is the decision as to which method(s) or sequence(s) are appropriate to the phylogenetic question at hand. The method chosen must yield sufficient variation as to be phylogenetically informative, but not so much variation that convergences and parallelisms overwhelm informative changes."[2]



2. Since you have already assumed that the organisms share common ancestry (see step 1), select data and methods that are "informative," that is, which fit your theoretical expectations. Don't use gene sequences that are too different from each other, or that might be misleading ("convergences," "parallelisms"). Your phylogenetic tree should make good evolutionary sense.

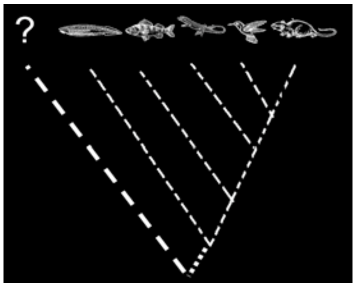
3. Align the sequences: "Alignment is often the most difficult and least understood component of a phylogenetic analysis."[3] "It is up to the user to ensure that the sequences in the dataset are actually homologous [related by common descent]. At this stage you need to examine the alignment to see if most of the gaps make sense. If many of the gaps seem to be arbitrary…then you will need to improve the alignment. Likewise, if they are large regions that are present in only one or two sequences (i.e., they appear as gaps in all other sequences), you may need to delete those regions in the sequence input file. Such regions do not share homology with the other sequences, and their presence will only contribute to artifacts when a tree is eventually generated."[4]


aatcggc-----gctatagctta---gtactaacctgaac
aatcggc-----gctatagctta---gtactaacctgaac
aatcggctaagctatagctta----gtactaacctgaac
aatcggc-----gctatagctta---gtactaacctgaac
aatcggc-----gctatagctta---gtactaacctgaac



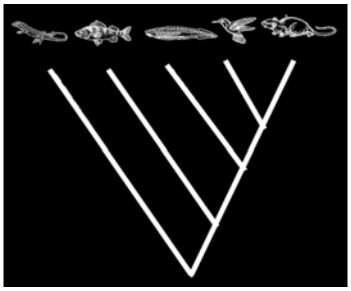
3. Your phylogenetic tree won't make any sense, however, if you compare sequences that are not genuinely homologous—that is, related by common descent. Sequences must therefore be aligned to locate their regions of homology. This decision process requires sound biological judgment. Allowing non-homologous regions to remain in an alignment will only create problems when the tree-generating algorithm or method is applied to the data. Thus, be sure to remove confounding regions from your set of sequences: "carefully and thoughtfully examine each alignment to see whether it makes good biological sense."[5]

4. ROOT THE TREE: "Obviously, this set of taxa [groups] had some common ancestor; the problem is where we should place the node that represents the ancestor –the root. The sequence alignment alone does not provide sufficient information…the choice of a root is often made on the basis of other information, which must be justified."[6] "The repercussions of outgroup choice are enormous…for they can determine which character states are interpreted as shared derived features and which are ancestral."[7]



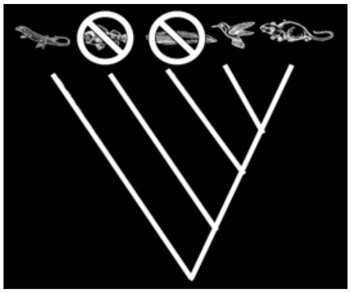
4. An unrooted tree doesn't give the branching order of groups, and thus, isn't really an evolutionary tree. So you need to pick a root, the point where your particular evolutionary tree joins the larger Tree of Life. Choose an "outgroup," an evolutionary relative lying outside, but close to, the group you're analyzing. Your choice should be reasonable in evolutionary terms—i.e, "be justified." The wrong outgroup will lead to an erroneous inference (pattern) of relationships.

5. GENERATE THE TREE: "[T]he field of phylogenetics is quite contentious with respect to which method is best. If you ask an evolutionary colleague which method to use, you are likely to get an answer such as 'You must use Parsimony (or Neighbor Joining or Maximum Likelihood, etc., depending on which colleague you ask). 'Other methods are just shoddy or worse.' Much of the opinion amounts to religious conviction, and you need not worry about it….In the end, it probably matters little which method you use."[8]



5. A wide range of different computational methods exist for generating evolutionary trees from molecular data. Use whatever method you think is best. Be aware, however, that other investigators may disagree with you, sometimes violently.

6. TROUBLESHOOT OR ELIMINATE UNRELIABLE DATA: "Sometimes a phylogenetic analysis will imply relationships that appear dead wrong in the context of previously published results or long-held view. A celebrated example in vertebrates is the placement of bony fishes within tetrapods…Of course, the possibility always exists that the widely held view is in fact wrong, and that a new phylogenetic paradigm has been uncovered….Troubleshooting such cases then involves demonstrating that a relationship is questionable and attempting to overcome the cause of the mistaken inferences."[9]



6. Sometimes, despite your best efforts, your chosen data and method will generate a tree that is just crazy ("dead wrong"). Perhaps you aligned the sequences incorrectly; maybe you picked a fast-evolving gene; you might have selected the wrong outgroup; or maybe you sampled too few species in your analysis. On the other hand, maybe your new tree is actually correct, and it's the established phylogeny that needs to be rejected. So now it's up to you! Troubleshoot your data so that they will fit in with the long-held view—or start fighting for the acceptance of your new phylogeny. Either option is open.

[1] David L. Swofford and Gary J. Olsen, "Phylogeny Reconstruction," in David M. Hillis and Craig Moritz, eds., *Molecular Systematics*. 1st ed. (Sunderland, MA: Sinauer, 1990), pp. 411-501; p. 413.

[2] Peter R. Baverstock and Craig Moritz, "Project Design," in David M. Hillis, Craig Moritz, and Barbara K. Mable, eds., *Molecular Systematics*, 2nd ed. (Sunderland, MA: Sinauer, 1996), pp. 17-27; p. 25.

[3] David L. Swofford, Gary J. Olsen, Peter J. Waddell, and David M. Hillis, "Phylogenetic Inference," in David M. Hillis, Craig Moritz, and Barbara K. Mable, eds., *Molecular Systematics*, 2nd ed. (Sunderland, MA: Sinauer, 1996), pp. 407-514; p. 412.

[4] Barry G. Hall, *Phylogenetic Trees Made Easy*, 1st ed. (Sunderland, MA: Sinauer, 2001), p. 30.

[5] Barry G. Hall, *Phylogenetic Trees Made Easy*, 2nd ed. (Sunderland, MA: Sinauer, 2004), p. 27.

[6] Barry G. Hall, *Phylogenetic Trees Made Easy*, 2nd ed. (Sunderland, MA: Sinauer, 2004), p. 55

[7] Michael J. Sanderson and H. Bradley Shaffer, "Troubleshooting Molecular Phylogenetic Analyses," *Annual Review of Ecology and Systematics* 33 (2002):49-72; p. 61.

[8] Barry G. Hall, *Phylogenetic Trees Made Easy*, 2nd ed. (Sunderland, MA: Sinauer, 2004), pp. 68, 76.

[9] Michael J. Sanderson and H. Bradley Shaffer, "Troubleshooting Molecular Phylogenetic Analyses," *Annual Review of Ecology and Systematics* 33 (2002):49-72; p. 50.