Introduction to Phylogeny

36-149 The Tree of Life

Christopher R. Genovese

Department of Statistics

132H Baker Hall x8-7836

http://www.stat.cmu.edu/~genovese/

.

Tracing the Tree of Life

The theory of common descent tells us that all organisms on Earth, in all their wondrous diversity, are descendants of a common ancestor.

We've seen how evolutionary processes can explain this diversity, and we've studied the mechanisms of change.

But what can we learn about the complex flow of evolution through the dim recesses of the past?

It turns out, we can learn quite a lot through the proper comparative study of organisms living and extinct.

To do this, we need a solid understanding of evolutionary processes, careful studies of organisms, and an interesting array of statistical tools.

Phylogenetics

Phylogenetics is the branch of systematics that involves understanding and reconstructing the evolutionary history of organisms.

The *basic* assumptions of a phylogenetic analysis are as follows:

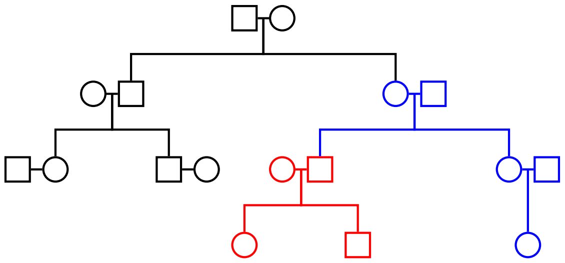
1. Organisms are related by descent from a common ancestor.
2. Characters change over time as organisms evolve.
3. New clades are created by binary splitting.

The first is just the theory of common descent. The second states simply that phenotypes evolve; the rate of evolution need not be constant. Neither of these assumptions is controversial.

The third assumption is more interesting. Why?

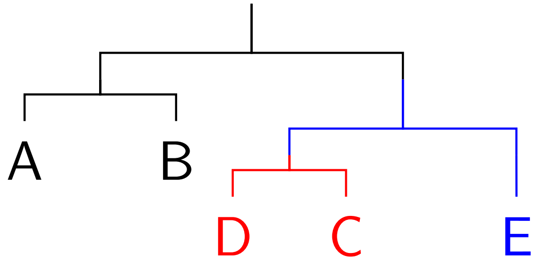
If these assumptions hold, then we can summarize the evolutionary history of taxa with what kind of diagram?

Analogy: Family Trees



- A family tree is a nested hierarchy defined by descent, as depicted by the tree.
- Branches in the tree represent ancestor-descendant relationships among individuals; in some cases, the lengths of the branches represent time.
- These relationships can often be observed from features of the family members.
- A family encompassing a group of individuals consists of the last common ancestors (a couple) of all those individuals and all the descendants of those ancestors.
- Having a family tree serves several useful purposes:
  - Delineate who is related to whom and classify the relationships.
  - Enables us to predict features of some members of the family from the features of others (e.g., propensity for certain diseases).
  - Offers insight into the family's history.

Phylogenetic Trees



- A phylogenetic tree is a nested hierarchy defined by descent, as depicted by the tree.
- Branches in the tree represent ancestor-descendant relationships among taxa; in some cases, the length of the branches represent time.
- These relationships can often be observed from features of organisms.
- A clade encompassing a group of taxa consists of the last common ancestor (a taxon) of those taxa and all taxa that descended from that last common ancestor.
- Having a phylogenetic tree serves several useful purposes:
  - Builds a classification of organisms.
  - Enables us to predict features of some members of a clade from the features of others.
  - Offers insight into how the taxa evolved.

Inferring Phylogenies: Inputs

The basic data of phylogenetic inference is a set of *characters* of various organisms.

Recall: A *character* is a heritable feature of an organism (e.g., number of appendages, feathers?, wishbone?, number of teeth).

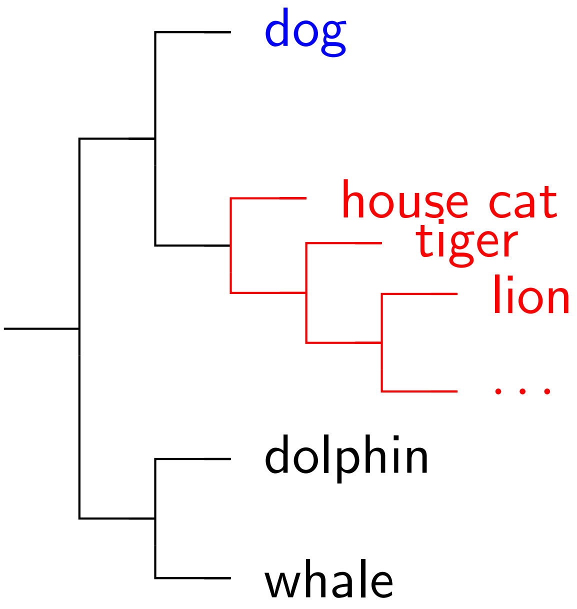
The values a character can take are called *character states* (e.g., 4, present, present, 0).

Two important types of characters used for reconstructing phylogenies are as follows:

1. molecular characters, including nucleotide sequences, proteins, and chemicals produced by the organism
2. morphological characters, including skeletal features preserved in fossils.

Inferring Phylogenies: Outputs

The output of a phylogenetic analysis is a “tree” that representes evolutionary history.



Terms: Root, node, branch, leaf, clade, basal, sister group

Types of trees: Rooted versus unrooted, scaled versus unscaled.

(The tree above is rooted and unscaled. An unrooted tree is called a *network*.)

Inferring Phylogenies: Outputs (cont'd)

Two kinds of trees produced by typical phylogenetic analyses

- phylogenetic (evolutionary) tree: a tree representing ancestor-descendant relationships among taxa, with the length of branches often indicating the time between speciation events.

- cladograms: diagrams that depict a hypothetical series of branchings. Taxa are grouped into a nested hierarchy defined by sharing a common ancestor.

Inferring Phylogenies: Computational Challenge

Number of possible trees to consider grows *quickly* with number of taxa:

Number of Taxa	Number of (Rooted) Trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575

Inferring Phylogenies: Computational Challenge (

How can we know if we’ve found the best tree?

How do we compare trees?

How do we express uncertainty about trees?

How do we summarize our analysis?

Inferring Phylogenies: The Logic

One school of systematic thought, called phenetics, uses overall physical similarity in as many characters as possible.

So, given a group of taxa, can we be confident that the two taxa that are most similar overall are the most closely related?

Inferring Phylogenies: The Logic (cont'd)

No. The observed similarities can come from two sources. The similar characters could have been inherited from a common ancestor or could have evolved independently (convergent evolution).

We have names for these two relationships between characters:

Homology is the similarity between structures in different organisms that is attributable to their inheritance from a common ancestor.

Homoplasy refers to similar traits in different organisms that do not share a common ancestor; convergent evolution.

Only homologous characters reflect an evolutionary relationship between taxa.

So then the two taxa that share the largest number of homologies are the closest relatives, right?

Inferring Phylogenies: The Logic (cont'd)

No. When it comes to inferring evolutionary history, all homologies are not created equal, so to speak.

Some homologous characters were recently derived, others are ancient features of a lineage.

(Only shared, derived features are informative about close relationships.)

Another school of systematic thought, called cladistics, rejects the use of overall similarity and focuses instead on shared derived homologies between taxa.

Cladistics

Cladistics was introduced by German entomologist Will Hennig in the 1950s (and possibly by some others independently as well).

Hennig put forward two key principles of systematics.

Principle #1: the only features that are informative about phylogenetic relationships are shared, derived homologies.

This rules out *both* homoplasies and ancient shared features of the taxa being studied.

Cladistics (cont'd)

For example, if we want to reconstruct a phylogenetic tree for human, horse, and iguana, we cannot use the character of having five digits on the feet.

Why?

How might this mislead?

Which is the ancient character and which the derived?

What role does the horse's single toe have on the analysis?

Cladistics (cont'd)

To avoid misleading, normative terms like primitive or advanced for ancient or recently derived characters. Systematists have devised an obscure terminology.

PLESIOMORPHY (adj. plesiomorphic): An ancestral character state, relative to another more recently derived state.

APOMORPHY (adj. apomorphic): A derived character state, relative to another more ancient state.

SYMPLESIOMORPHY: A plesiomorphy shared by two or more taxa.

SYNAPOMORPHY: An apomorphy shared by two or more taxa.

AUTAPOMORPHY: An apomorphy possed by only one taxon under consideration.

Which is which in the horse, human, lizard example?

Cladistics (cont'd)

Principle #2: Proper classification of organisms requires grouping taxa into clades, or monophyletic groups.

There are three different types of groups:

MONOPHYLETIC TAXON (aka clade): A group of organisms that includes the most recent common ancestor of all the organisms in the group and all the descendants of that most recent common ancestor.

PARAPHYLETIC TAXON: A group of organisms that includes the most recent common ancestor of all the organisms in the group, but unlike a monophyletic group, need not include *all* the descendants of the most recent common ancestor.

POLYPHYLETIC TAXON: A group of organisms that does not include the most recent common ancestor of all the organisms in the group, which is often excluded because it does not share features with the remainder of the group

Cladistics (cont'd)

What do these categories look like when represented by trees? What type are each of the following groups?

- mammals
- reptiles
- insects
- marine mammals
- dinosaurs (as traditionally defined)
- fish
- birds
- flightless birds
- invertebrates
- angiosperms
- trees

Cladistics (cont'd)

In practice – because we do not observe the organisms (or their remains) at most of the nodes in the tree – there is some flexibility in the definition of a taxon.

Three basic methods are used to define taxonomic groups in practice:

- Node-based taxon: as above, all the organisms descended from and including some basal node.

  Example: Aves (birds) are defined as *Archaeopteryx*, the Neornithes (modern birds), their common ancestor, and all its descendants.

- Stem-based taxon: all descendants from a particular splitting (cladogenesis) event.

  Example: *Ornithischia* (a dinosaur group) is defined as all dinosaurs more closely related to *Triceratops* than to *Tyranosaurus*.

Cladistics (cont'd)

- Apomorphy-based taxon: defined by the presence of one or more specified characters.

  Example: Aves (birds) consists of all archosaurs with feathered wings.

In addition, a clade defined in terms of living organisms is called a crown group.

Example: The clade defined by living birds (Neornithes) is a crown-group taxon that does not include *Archaeopteryx*. (Compare above.)