COSC 348:
Computing for Bioinformatics

Lecture 11:
Phylogenetic tree inference: introduction

Lubica Benuskova
Prepared according to the notes of Dr. Richard O'Keefe

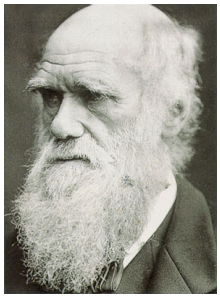
http://www.cs.otago.ac.nz/cosc348/

Evolution: inheritance and mutation

- Organisms (animals or plants) produce a number of offspring which are almost, but not entirely, like themselves:
  - Variation is due to sexual reproduction (offspring have some characteristics from each parent, alleles from 2 parents combine randomly)
  - In addition, variation is due to mutation (random changes) in the fertilised egg.

Evolution: natural selection

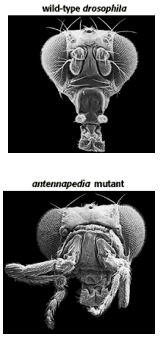
- Some of these offspring survive to produce offspring of their own—some won't:
  - The offspring with bigger fitness are more likely to survive and reproduce
  - Over time, later generations become better and better adapted to a given environment because only the fittest individuals have a higher chance to survive and reproduce.



Charles Darwin

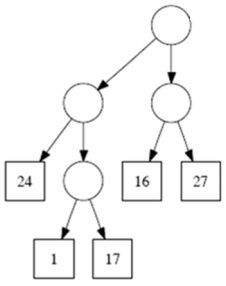
Mutations drive evolution

- If a mutation happens in the DNA of a fertilised egg, then all the cells of an offspring carry this mutation because this mutation is replicated by DNA replication during cell division.
- If that offspring survives, every offspring of his own will carry the same mutation. That's how mutations are preserved over evolution.
- Mutations that result in an improved trait, drive evolution.
  - E.g., bigger claws, better eyesight, opposing thumb, bigger brains, etc.
- Mutations that accumulate over time lead to appearance of new species through intermediate forms.



What is a phylogeny?

- A phylogeny is a tree showing ancestor/descendant relationships between taxonomic units (e.g., species, genera, etc).
- The square boxes in the tree represent the *observed taxonomic units* (living or fossils).
- The circles represent *hypothetical taxonomic units* (ancestral or intermediate species).
- Edges represent descent.



Basic terms from dictionary

- *Taxonomy*: The classification of items in an ordered system that indicates natural relationships.
- *Species:* A fundamental category of taxonomic classification, consisting of related organisms capable of interbreeding.
- *Subspecies*: A taxonomic subdivision of a species consisting of an interbreeding, usually geographically isolated population of organisms.
- *Genus (plural genera)*: A taxonomic category consisting of a **group of species** exhibiting similar characteristics.

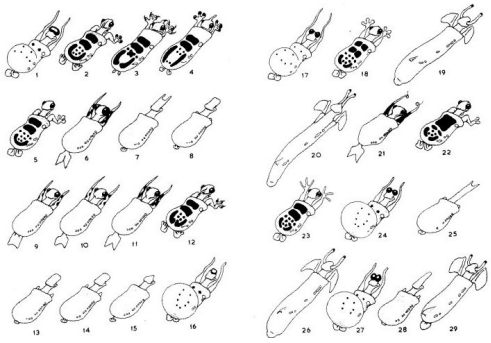
Taxonomy hierarchy

- Taxonomic organization of species is hierarchical. Each species belongs to a genus, each genus belongs to a family, and so on through order, class, phylum, and kingdom.
- Carl von Linne, an 18th-century Swedish botanist, founder of modern taxonomy.

Common name	Dog (Wolf)
Species	Canis familiaris (Canis lupus)
Genus	Canis
Family	Canidae
Order	Carnivora
Class	Mammalia
Phylum	Chordata
Kingdom	Animalia

How to construct phylogeny?

- Let's have an imaginary species called *Caminalcules* invented by Professor Joseph H. Camin (University of Kansas, USA) as a tool for modelling and understanding phylogenetics.



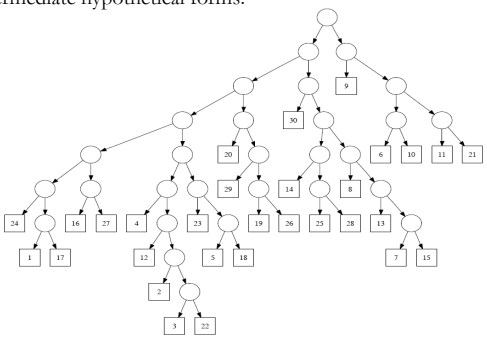
"Evolution" of *Caminalcules*

- The phylogenetic tree is supposed to reflect the evolution of *Caminalcules* from simple to more advanced forms.

[Source: http://hudsonvalleygeologist.blogspot.co.nz/2012/01/caminalcules.html]

Phylogenetic tree of *Caminalcules*

- This phylogenetic tree shows ancestor/descendent relationships between subspecies of *Caminalcules*. In order to explain properties of observed subspecies we have to assume the existence of intermediate hypothetical forms.



Observed and hypothetical taxonomical units

- The *observed taxonomic units* (OTUs) are the ones that we actually have access to measure them. They are placed at the leaves of the phylogeny. They are the only ones we can call real.
- We fill in the internal nodes of the phylogeny with *hypothetical taxonomic units* (HTUs); imaginary ancestors whose properties we are free to invent in any way that will make a good explanation for the observed properties of the OTUs.
- HTUs are explanatory entities, which we ***hope*** have some connection with reality, but it is only a hope. This hope is often unfounded. We have no guarantee that an inferred phylogenetic tree is correct.
- Taxonomic unit is also called **taxon** (plural taxa).

More on hypothetical taxonomical units

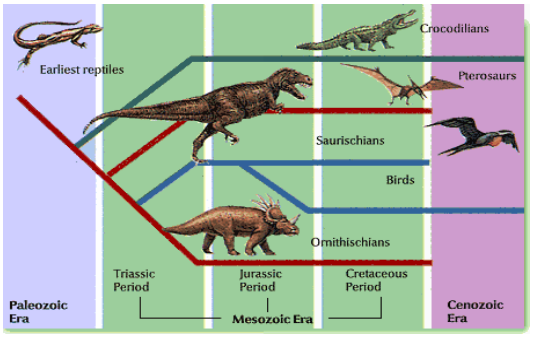
- How many branches should internal nodes have, and what should their order be?
- In the real world, a species can go extinct without leaving any descendants. If it leaves fossil evidence, it goes into our tree as a leaf, ***not*** an imaginary ancestor. In the real world, a species can split off one new species, and then go extinct.
- The only justification we have for inventing an imaginary ancestor is a *difference* between two (or more) groups of OTUs.
  - An HTU with no children would be one having no reason to exist.
  - An HTU with one child would have no evidence to distinguish it from that child.
- Assumption: every internal node must have at least two children.

Principles of phylogeny construction

- Biologists have been building classification systems for over two hundred years based on fossils using the principles of phenetics, patristics and cladistics.
- *Phenetic*: relating to a system of classification of organisms based on overall or observable morphological similarities rather than on genetic or evolutionary relationships.
- *Patristic*: related to fathers.
- *Cladistic*: A system of classification based on the evolutionary history of groups of organisms.
  - *Clade*: a group of organisms considered as having evolved from a common ancestor

Cladogram

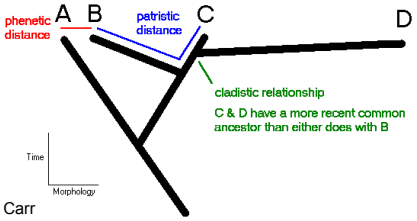
- is a diagram used in cladistics, which shows ancestral relations between organisms, to represent the evolutionary tree of life.



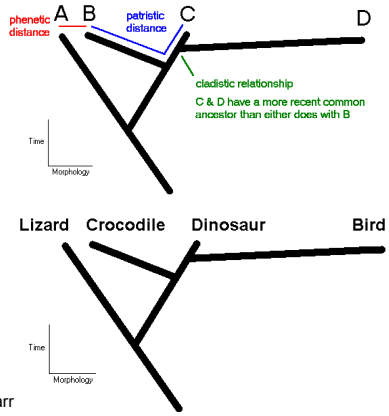
Source: http://rst.gsfc.nasa.gov/Sect20/A12d.html

Phenetic, patristic and cladistic

- The diagram shows the evolutionary and morphological relationships among four taxa A, B, C, & D:
  - A & B are most similar **phenetically**: i.e. the *morphological* difference between them is smallest.
  - B & C are most similar **patristically**: the *amount of change* that separate them is smallest.
  - C & D are most closely related **cladistically**: they have a more recent ***common ancestor*** than any other pair in the tree.

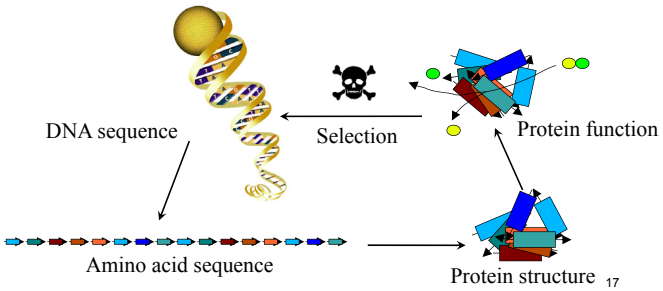


Phenetic, patristic and cladistic



Evolution "works" at a molecular level

- Traits of organisms depend on proteins.
- The function of a protein in the cell is determined by its AA sequence. Thus proteins select which genes will survive.



Phylogeny based on biomolecules

- The cladistic approach takes actual evolutionary descent as of primary importance.
- Molecular data seem to be better able to uncover the true evolutionary relationships than physical resemblance.
- If you have some idea of the actual evolutionary history of a group of organisms, you can pose all manner of interesting biological questions:
  - these organisms are not closely related but look alike, why?
  - these organisms are more closely related to ones from a distant land, why?
  - this lineage has changed a lot, this other lineage not much, why?
  - Etc.

Big and small approximation

- When we develop a computational or statistical model of the real world, we go through (at least) two approximations:
  - the **big** approximation is when we make a formal model of the aspects of the real world we are concerned with. For example, treating DNA as a string of letters is a **big** approximation.
  - the **small** approximation is when we simplify the formal model to something our statistical or computational techniques can handle.
- This is particularly relevant to inferring phylogenies.
- This distinction is attributed to the statistician George Box.

Big approximations in phylogeny

- We do not consider the horizontal gene transfer.
- Horizontal gene transfer occurs when an organism incorporates genetic material from another organism without being the offspring of that organism.
  - By contrast, *vertical* transfer occurs when an organism receives genetic material from its ancestor, e.g. its parent or a species from which it evolved.
- Horizontal gene transfer is a highly significant phenomenon, and amongst single-celled organisms the prominent form of genetic transfer, which complicates phylogeny construction for these organisms.

Big approximations in phylogeny

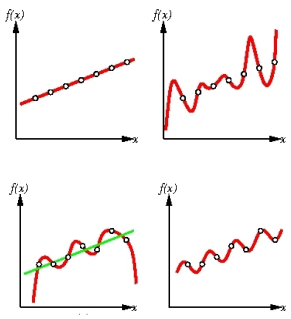
- A species is not a simple thing. It is a population of organisms, including individuals of many ages and other varieties (colour of skin, blood type, etc). So, variations within a single species are ignored.
- We do not consider hybridisation (i.e. crossbreeding).
  - From a taxonomic perspective, hybrid refers to offspring resulting from the interbreeding between two animals or plants of different taxa (e.g. between lions and tigers, sheep and goat, etc.)
- Our internal nodes (HTU) may correspond to a subset of the actual ancestors, we think there is only one.
- Thus, in our model, we are presuming that each taxon can be characterised simply, adequately, and consistently.

Three major approaches to phylogeny

- Parsimony methods.
  - Step 1: define a **cost** measure for phylogenies.
  - Step 2: announce that the best tree(s) is(are) the one(s) having the least cost.
- Distance-based methods.
  - Step 1: define a distance measure reflecting how different each OTU is from each other OTU.
  - Step 2: use a clustering algorithm to fit a tree to the mutual distances.
- Maximum likelihood.
  - Step 1: define a probability model for evolutionary change.
  - Step 2: announce that the best tree(s) is(are) the one(s) which ascribes the highest probability to the observed outcome.

Parsimony or Ockam's razor

- Ockham's razor: prefer the *simplest hypothesis* consistent with data
  - Ockham's razor or law of parsimony attributed to the 14th century Franciscan logician William of Ockham (Occam)
- In general, tradeoff between the complexity of function and degree to fit the data.



Conclusions

- Why do we care about phylogenies? Because biologists care and the problems are algorithmic, so we have to develop algorithmic answers to these problems.
- We have to start by understanding what the biological data and problems actually are.
- We have to understand the limitations of our answers, especially how well they scale with the problem size.
- We need to maintain a properly sceptical attitude to our own success. There is no such thing as a correct evolutionary tree.