

Phylogenetic Reconstruction

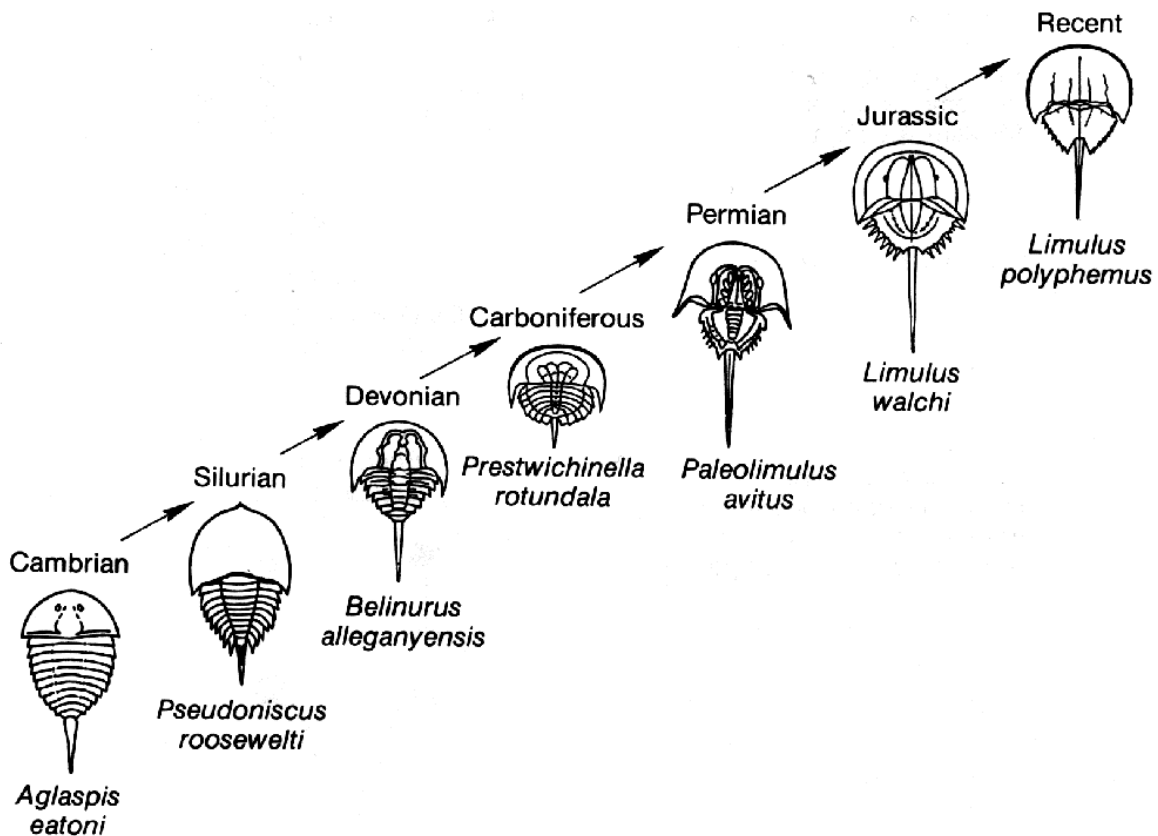
Phenetics; cladistics; practical reconstruction; homology; parallel evolution; morphology versus biochemistry.

Introduction

This lecture looks at the ways we try to reconstruct evolutionary history. This necessarily involves us in the thorny problems of classification, and introduces us to some of the worst jargon and confusing nomenclature around. I'll talk about some theory and then there'll be some practice.

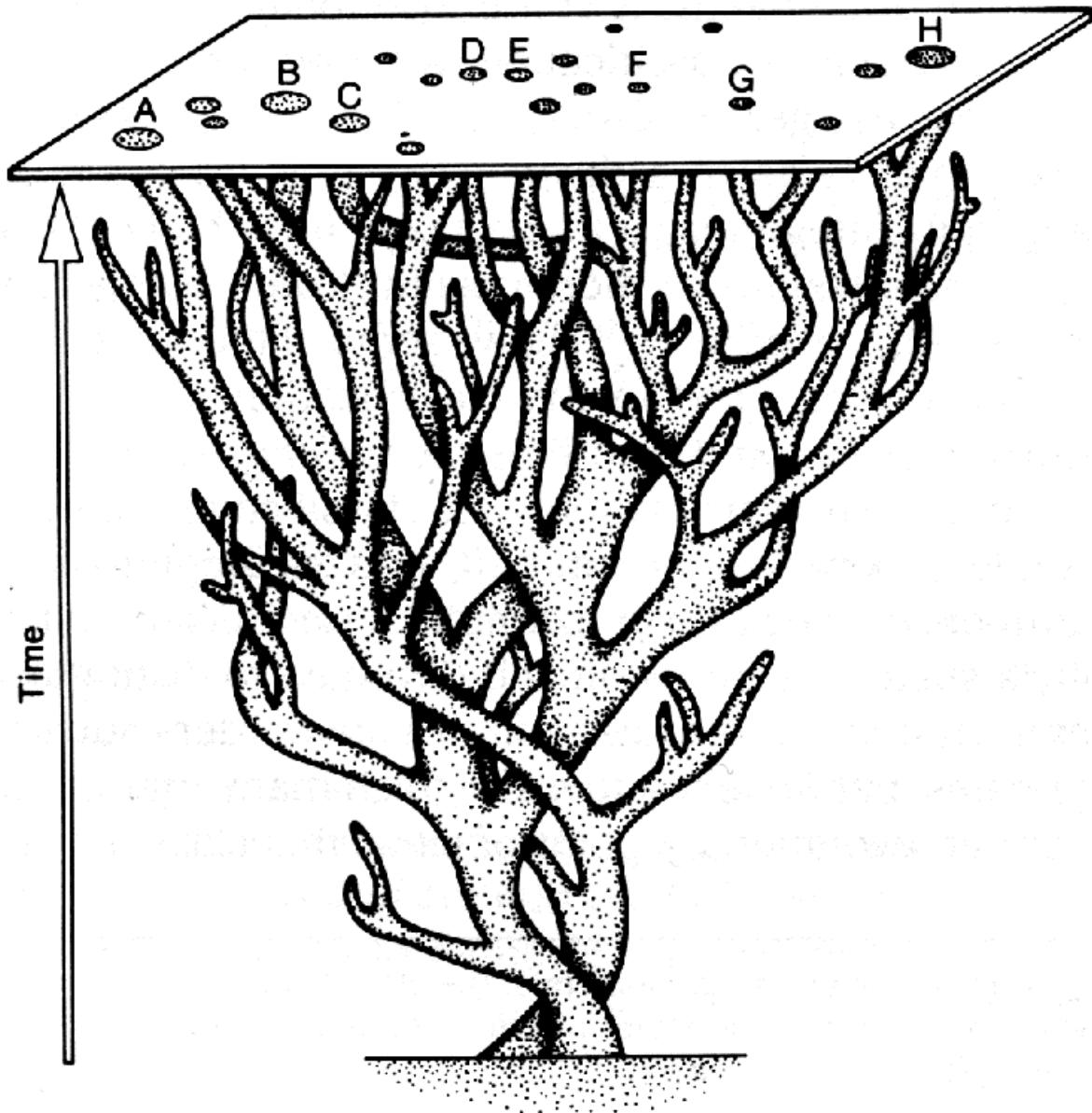
Principles

Diagram of horseshoe crabs illustrating Anagenesis



Firstly, evolution (change over time) doesn't have to lead to a branching pattern. A single group can change gradually without splitting into two distinct groups. This process is called Anagenesis. However, the much more interesting problem is the reconstruction of the branching pattern, where species split into two or more groups. This is called Cladogenesis and is what gives us our family tree.

Diagram of a generic phylogenetic tree



Determination of phylogenetic trees is difficult in practice because the common ancestors are usually long extinct and the fossil record is inadequate. However, the relationship can be inferred by looking at common, inherited characteristics: the more morphological, embryological, behavioural, physiological, biochemical, genetic and chromosomal inherited characteristics that organisms have in common, the more likely they are to have descended from a common ancestor.

Merely sharing common features is not enough since they may derive from different evolutionary causes:

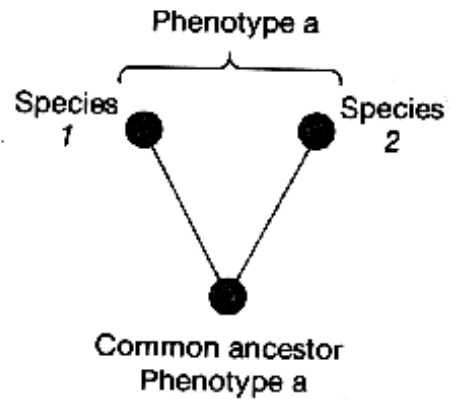
Homology

This is what we want. The feature is shared because it derives directly from a common ancestor. For example the bony features of the forelimbs in vertebrates.

Diagram illustrating homology and the two forms of analogy

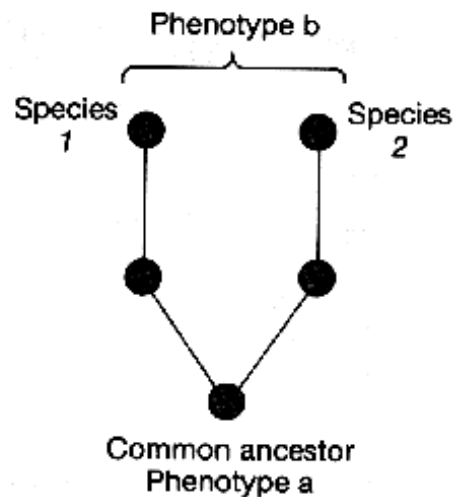
(a) Homology:

two species bearing the same phenotype caused by common ancestry for the same genotype



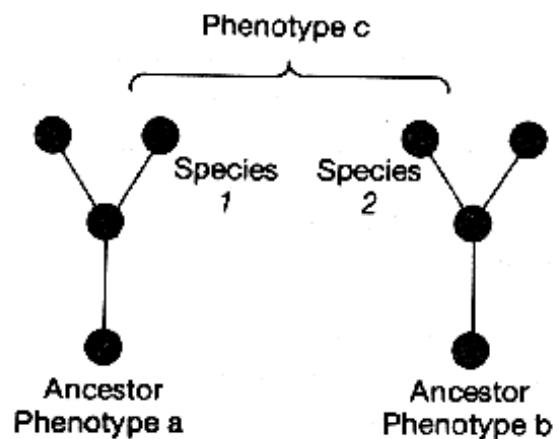
(b) Parallelism:

two species with the same phenotype descended from a common ancestor with a different phenotype



(c) Convergence:

two species with the same phenotype whose common ancestor is very far in the distant past



Parallelism

The similar feature occurs in different species, but it is not present in their immediate common ancestor. For example, anteater-like features in various different mammalian lineages. These shared features are very much functional adaptations.

Anteater parallelism

Prototheria (*Echidna*)



Metatheria (*Myrmecobius*)



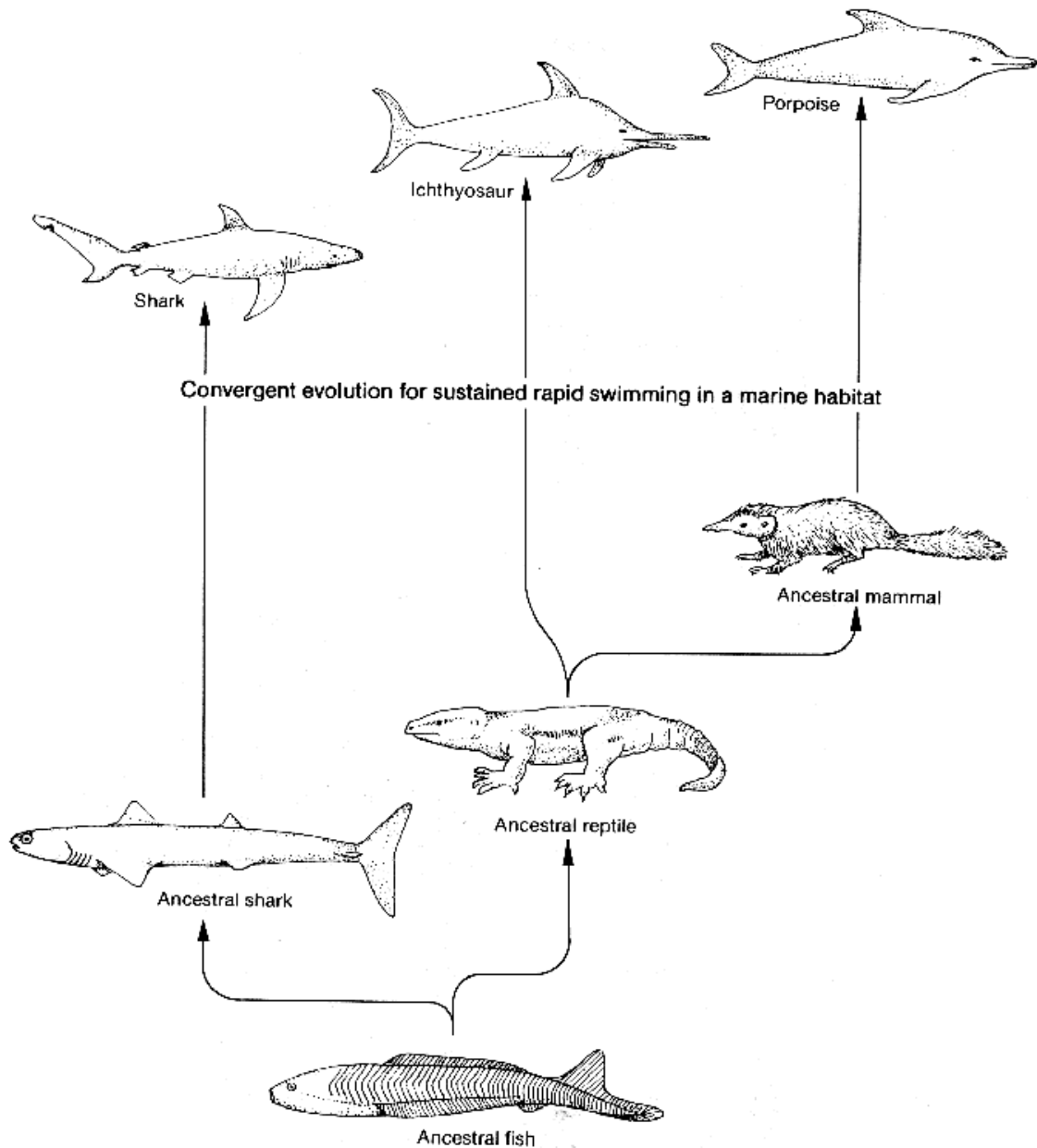
Eutheria (*Myrmecophaga*)



Convergence

Similar to parallelism, but the ancestral lineages differed for a considerable period of time. For example vertebrate and octopus eyes, or the hydrodynamic morphology of marine predators from the widely separated fish, reptile and mammalian classes.

Marine predator convergence



Practice

Obviously, homologies are what we need to consider to reconstruct phylogenies. However, they are not always easily separated from the other 2 (together often referred to as analogy). Consider the convergence example: the shape of the pectoral fins in these animals is very

similar due to convergence. However, there is a great deal of homology there two. Especially between the reptile and the mammal due to a common land vertebrate ancestor.

Classical taxonomy

The process of sorting into groups when applied to living or once living organisms is called taxonomy. Some workers consider taxonomy to be synonymous with classification:

'the ordering of organisms into groups (or sets) on the basis of their relationships', Simpson 1961

Others consider taxonomy to be synonymous with systematics:

'the scientific study of the kinds and diversity of organisms and of any or all of the relationships among them' Simpson 1961

Evolution has produced the natural groups of life; classification should seek to be natural rather than artificial. Classification should therefore represent evolution.

This approach, together with all other types of classification, runs into problems. Notably:

1. variation. Any group of animals, like any box of matches, contains individuals that vary in size and shape. The variation in a box of matches is small: the variation in a group of animals is larger. imagine a group of individuals such as this class.
2. branching. Evolutionary theory dictates that at some point a group will split in to two.
3. divergence. Once two groups have formed they will become less similar. How can we assess the rate of this divergence? Do dissimilar groups ever converge (because of similarity of behaviour, habitat, food?)

Simpson (a famous palaeontologist) attempted to describe how he classified animals. He described eight overlapping and intertwining steps:-

1. The selection of organisms to be studied
2. The observation and recording of data, the study of the geography, fossil record, and literature concerning the groups under analysis.
3. The sorting of organisms into taxonomic units and the study of variations and polymorphic forms.
4. Comparisons are made among the characteristics of the varying taxa, special attention is given to the kinds and degrees of resemblances.
5. The interpretation of the relationships revealed by the comparisons, especially in terms of homology, parallelism, convergence, primitiveness and specialisation.
6. On the basis of these interpretations, inferences are made as to the evolutionary pattern among the populations studied. A family tree may be drawn, and the conclusions will take account of factors such as degrees of divergence that may not be readily shown in a diagram.

7. The translation of conclusions on affinities, divergence etc. into hierarchical terms i.e. Order, Family, Genus, Species.
8. The naming of taxa.

Modern systematics

Modern systematicists try to create groups in a more logical way. Their groups are:-

- a. based only upon the branching pattern of evolution
- b. based only upon the degree of overall similarity
- c. some functional mix of the two

Relationships

Because of the different definitions of taxonomy the relationships between animals can be thought of in different ways. If we believe that taxonomy = classification then the relationships between what we classify are based on overall similarity. This is a phenetic relationship.

If we believe that taxonomy = systematics then we arrive at a set of relationships based on the true branching sequence of the evolutionary process. This is a cladistic relationship.

Phenetics

Pheneticists attempt to

- a. describe patterns of similarity/difference between organisms
- b. derive stable classifications based on these patterns

This involves: -

1. The description of the organisms (qualitative/quantitative)
2. The calculation of a similarity coefficient or difference coefficient (distance).
3. The identification of clusters of similar organisms which are then classified together.

The process of identifying clusters of similar organisms involves the plotting of graphs of their relationships and/or the use of clustering algorithms. (A clustering algorithm is a computer program which will identify groups in a sample of suitable data).

Similarity coefficients and distances.

Imagine we are studying two animals A and B, by identifying three characters a, b, c. which can be given a numerical value (For instance leg length; legs may be absent (=0) 0-1cm long (=1) 1-2cm long (=2) etc.):

Phylogenetic Reconstruction

	Character	a	b	c
Organism				
A		1	2	3
B		2	2	3

This table is called a character matrix

Similarity coefficient S

$$S = \frac{\text{No. of matches}}{\text{No. of matches} + \text{No. of mismatches}} = \frac{2}{2+1} = \frac{2}{3}$$

for the above matrix.

Distance coefficient E

May be measured, for example by the Euclidean distance

$$E = \sqrt{\sum_{i=1}^n (X_iA - X_iB)^2}$$

where n is the number of characters

$$E = \sqrt{(X_1A - X_1B)^2 + (X_2A - X_2B)^2 + (X_3A - X_3B)^2}$$

$$E = \sqrt{-1^2 + 0 + 0} = 1$$

for the above matrix.

Examples

Below are reproduced two character matrices based upon fruit. The fruit character matrix is worked to show a table of distances and a phenogram indicating the pattern of similarities between fruits. A phenogram is a table of similarities.

Classification of fruit

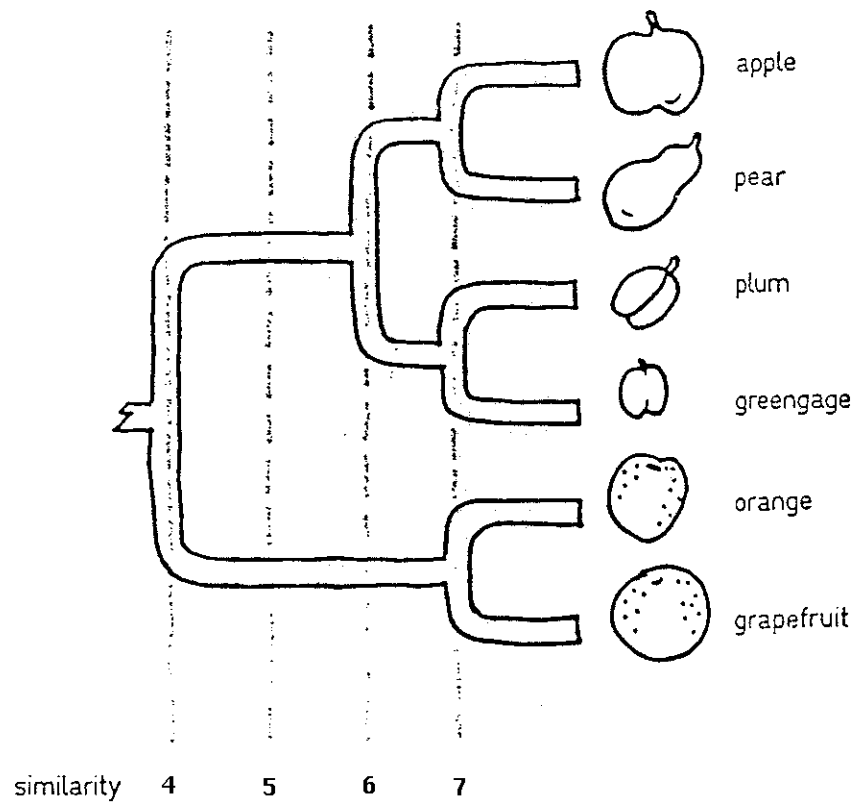
Features

	pear	orange	grapefruit	apple	plum	greengage
hard	+	+	+	+	-	-
round	-	+	+	+	-	+
stone	-	-	-	-	+	+
thin skin	+	-	-	+	+	+
smooth	+	-	-	+	+	+
sweet	+	+	-	+	+	+
stalk	+	-	-	+	+	+
segments	-	+	+	-	-	-

Similarities

pear					
orange	3				
grapefruit	2	7			
apple	7	4	3		
plum	6	1	0	4	
greengage	5	2	1	6	7
	pear	orange	grapefruit	apple	plum

Phenogram



Phenetics - summary

Phenetic classification can thus be summarised:

- a. Organisms are grouped by overall similarity
- b. The resultant pattern reflects patterns of between group variation
- c. There is no pretence that the resultant groupings reflect the branching pattern of evolution
- d. The biologist is free to postulate branching sequences which might result in the observed relations.

One important problem is that animals evolve and change all the time: some of these changes result in speciation - the formation of new groups - others do not, but simply modify the characteristics of the whole population (see diagram above). Phenetics fails to distinguish these two types of change. Another type of classification, cladistics, identifies only the events which result in speciation.

Cladistics

We have seen that phenetics groups organisms on the basis of similarities. Cladistics does this too, but tries to derive relationships on the basis of splitting events e.g. speciation.

Similarities between organisms exist for three reasons:

1. They are inherited from a common ancestor
2. They evolved in common from an ancestral state
3. They evolved in parallel from different ancestral states.

Gould 1983 has a rather straightforward approach to cladistics and defines two terms:

Sister Group	"Two lineages sharing a common ancestor that no other lineage has sprung from."
Shared Derived Characters	Shared features "present only in members of an immediate lineage. They are unique and newly evolved."

Cladists attempt to construct hierarchies of sister groups in order to specify temporal order of branching. For example (and here I deviate from Gould because he is rather out of date here), humans and chimps form a sister group because no other primate group branched from their common ancestor. We extend this process adding gorillas next and then orang-utans. From this it can be seen that in the mind of the cladist, the grouping "great ape" makes no sense as a genealogical unit.

Shared derived features are the only thing to be used for identifying sister groups. Primitive characters are characters present in a distant common ancestor and must be avoided. For example, humans and many salamanders have five toes and horses have one. We may not therefore state that humans are more closely related to salamanders and that the concept of "mammal" is a fiction. Rather, that five toes is an inadmissible primitive character. Many mammals have lost toes! Hair, though, is a unique and newly evolved character possessed by mammals and no other order.

Definitions for masochists

(Non masochists stick to Gould's sister group and shared derived character)

- a. plesiomorphous character state = primitive
- b. symplesiomorphous character state = shared primitive state
- c. apomorphous character = derived
- d. autapomorphous = uniquely defined
- e. synapomorphous = shared derived character

For instance:

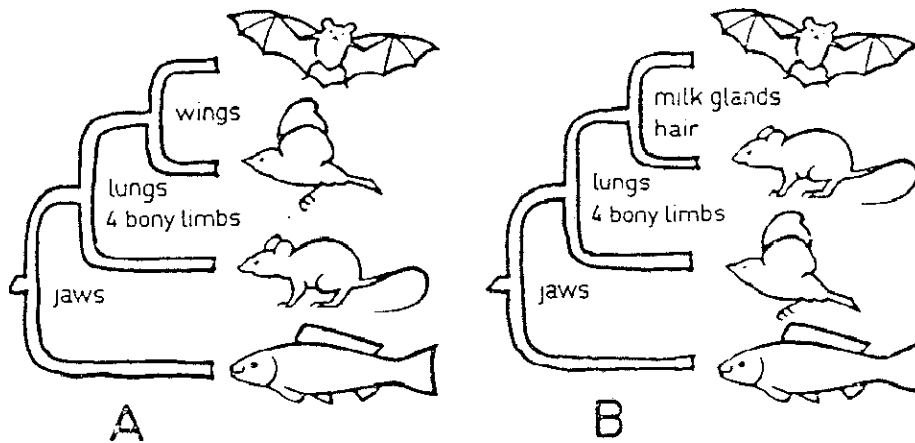
Phylogenetic Reconstruction

- bilateral symmetry, homoiothermy and a vertebral column are plesiomorphous for primates: they are primitive for all mammals and so imply no close relationship
- the presence of an opposable thumb is synapomorphous for primates .
- bilophodont molars are autapomorphous for new world monkeys.

Cladistics operates by finding synapomorphies and finding groupings based on them.

Below is a character matrix and two different cladograms based upon it.

	mouse	bat	bird	fish
lungs	+	+	+	-
4 bony limbs	+	+	+	-
milk glands	+	+	-	-
hair	+	+	-	-
wings	-	+	+	-
feathers	-	-	+	-
jaws	+	+	+	+



Problems with cladistics

It works very well at higher taxonomic levels. At lower levels, it often has to be interpreted statistically since we don't actually know what the primitive state of a character is. This means that a numerical process is used that examines a large number of possible cladograms and chooses the one that requires the smallest number of evolutionary events (mutations) to produce the observed diversity of features. In this respect, it becomes another sort of clustering algorithm. There are problems, though:

- Who said evolution should be parsimonious?
- How can you test all trees? 9 taxa can have over 2 million possible cladograms, and this number increases very rapidly - it rapidly becomes "computationally impossible". i.e. not unreasonable problems taking thousands of years of computer time, and each increase in numbers of taxa vastly increasing the computation time.
- Trees with very different topologies may be equally (or very nearly) parsimonious.

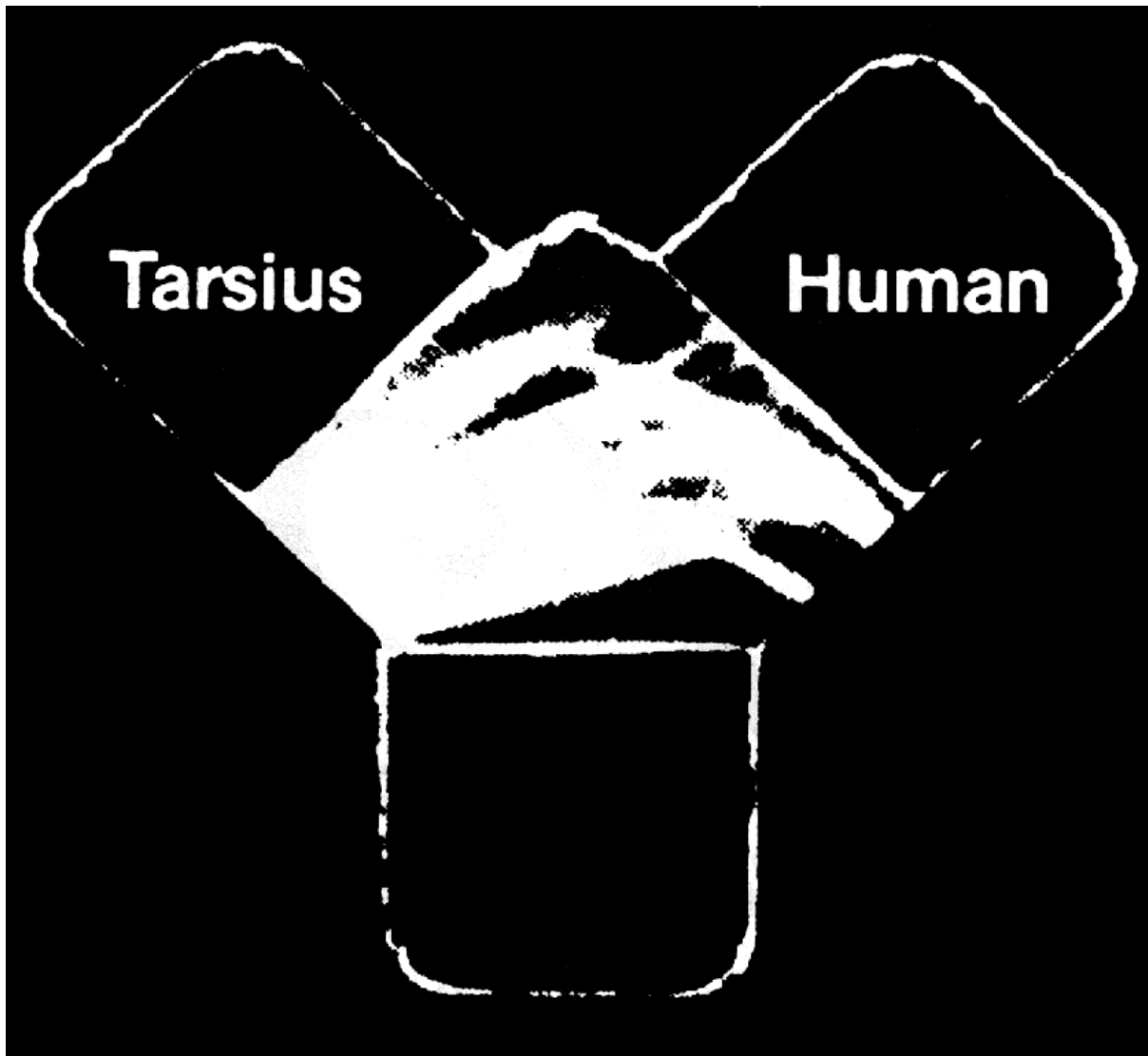
Molecular Approaches

In the 1960, with the understanding of the role of DNA in heredity and the manufacture of proteins, it was assumed that the molecular record of phylogeny would be a good deal easier to unravel than the one hidden in morphology.

Distance Measures

The first attempts to reconstruct phylogeny from molecular of techniques that can be described as "distance measures". These are a variety of methods that measure the degree of difference between two molecules; whether that is two proteins or two DNA strands. The information obtained can then be dealt with in exactly the same way as the various coefficients in phenetic analysis.

Immunodiffusion

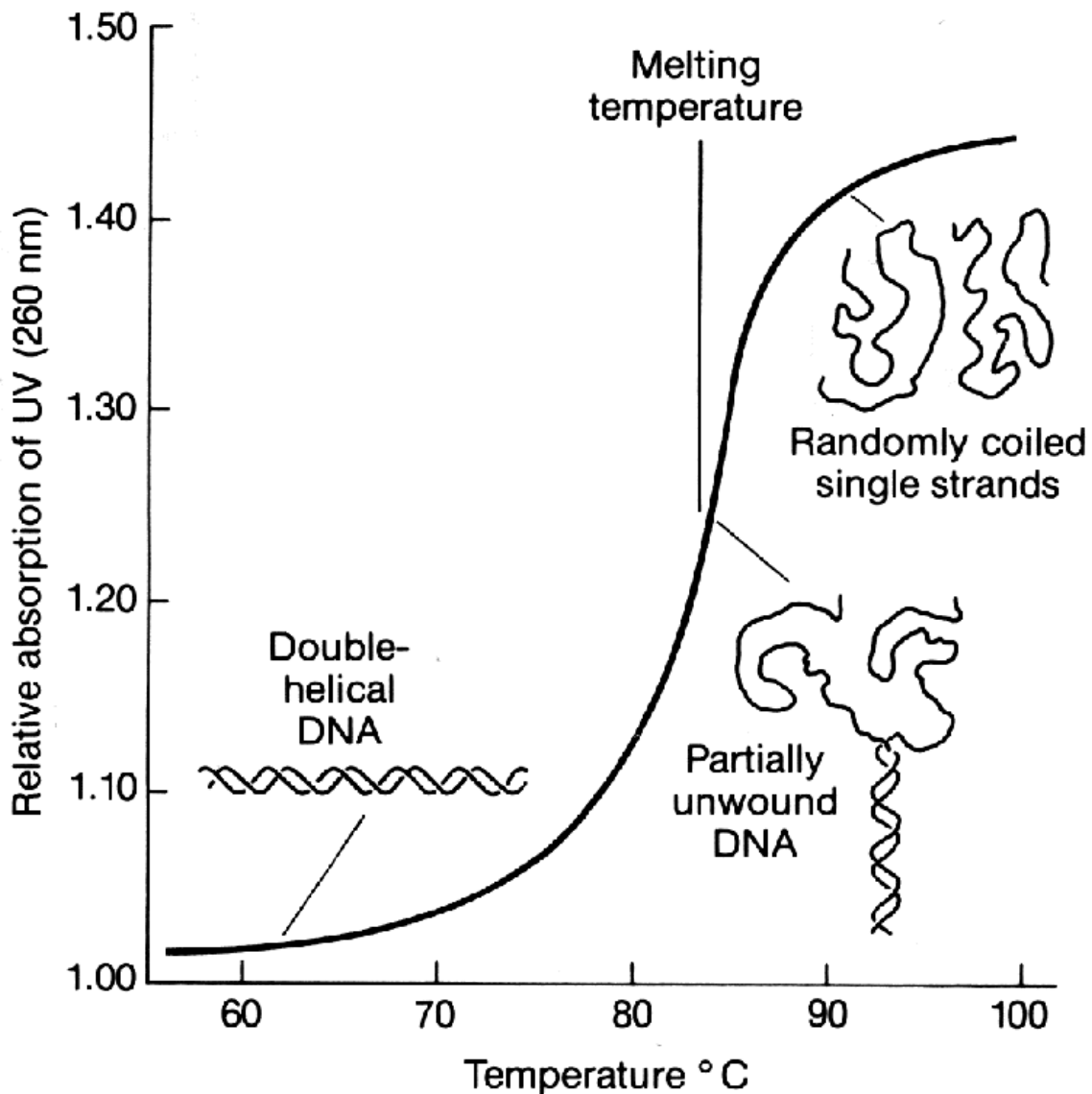


The simplest form is an immunological technique called "immunodiffusion" which produces "antigenic distances" dependent on the degree of dissimilarity. In this case, antibodies are raised by injecting an animal with blood serum of one (reference) species. These antibodies are placed in a centre well on an agar plate with the serum of the two (comparison) species to be compared in nearby wells forming a trefoil arrangement. The antibodies and antigens diffuse through the agar and produce precipitation bands where specific antibody/antigen reactions occur. Because the various proteins diffuse at different rates, a number of separate bands are produced. If the proteins constituents of the two comparison species are alike, then the bands will fuse. If they differ, then spurs will form indicating greater antibody reaction with one species than with the other. By analysing the number and lengths of the spurs, a figure for antigenic distance can be generated.

Another example uses DNA-DNA hybridisation to measure the overall similarity between DNA sample. This uses the change in thermal stability as a measure of dissimilarity. However, the early Sibley and Alquist papers need to be read with a great deal of care since it is

fairly clear that a certain amount of over-optimistic reading of the data was going on. In reality, these distance measures do not have a particularly high resolving power, and they are unable, for example, to separate out the human/chimp/gorilla trichotomy (no matter what Sibley and Alquist may have claimed).

DNA-DNA hybridisation



Sequencing

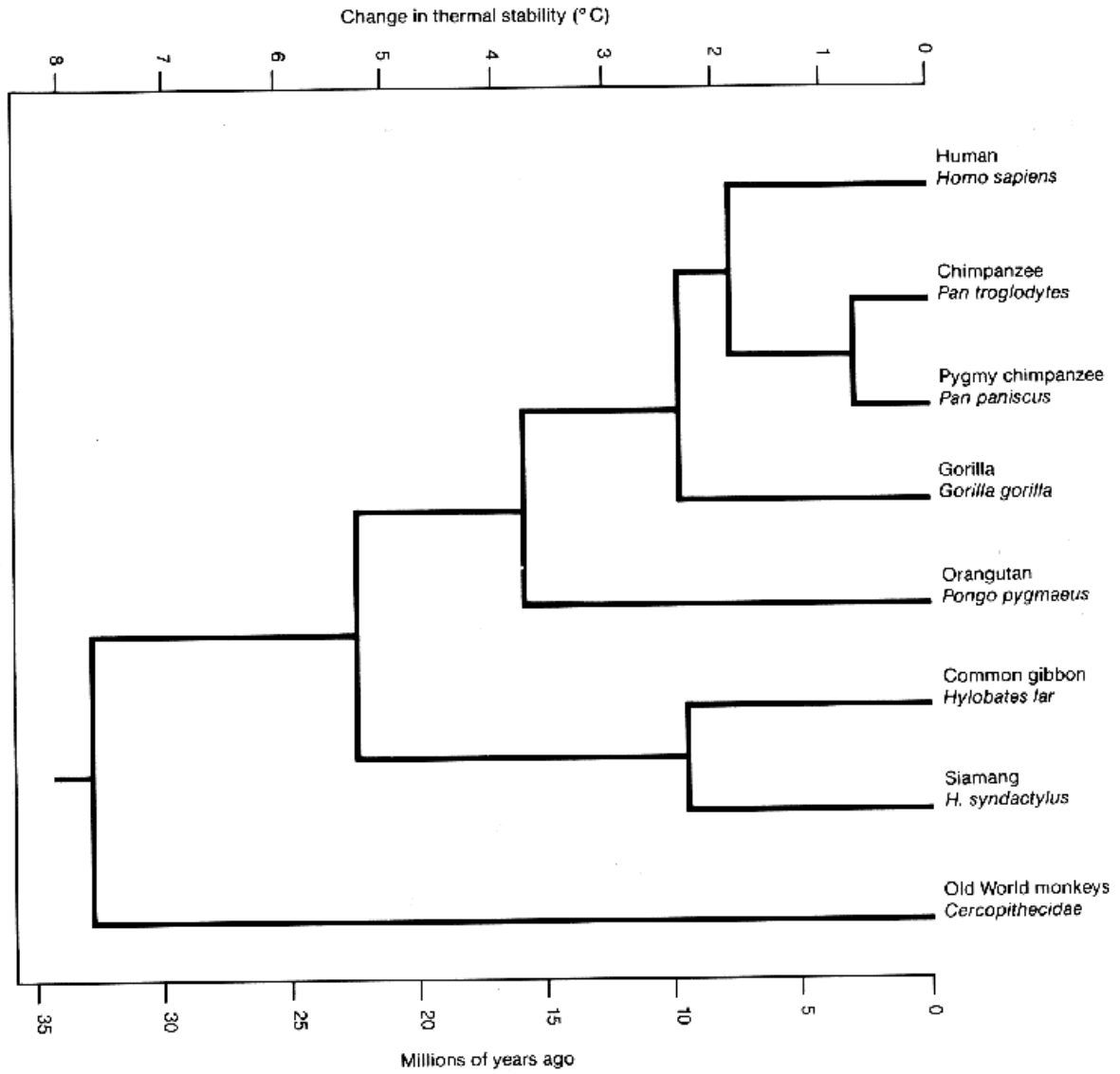
Newer approaches use the actual sequence data (amino acids in proteins and nucleotides in DNA) to obtain phylogenetic trees. These methods are more akin to cladistics since they try and identify homologous sequences and look for the differences. However, because there is no direction information (since there is no fossil record of DNA - currently), this information has all to be dealt with statistically, looking for the most parsimonious tree structure. Particular proteins and DNA sequences are particularly good depending on the timescale involved. Fundamental proteins such as Cytochrome-C are very well conserved and can be

used for comparisons over the last billion years or so. Others, such as mitochondrial DNA are much more changeable, and thus only good for the last few million years.

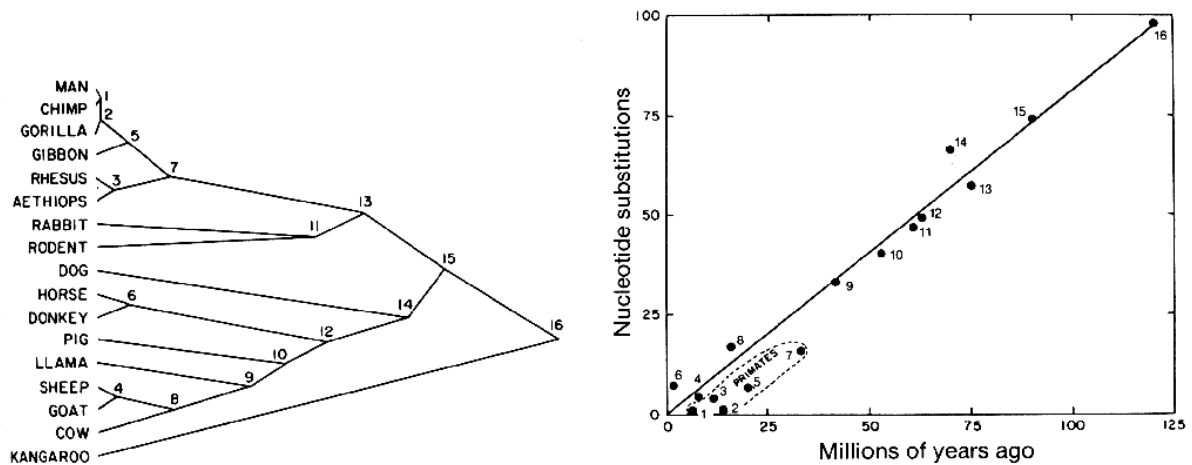
Molecular Clock

If the rates of change of molecules over time can be considered to be fairly constant, then, once calibrated by using the fossil record, molecular phylogenies can produce predicted divergence times. This seems to be quite likely for neutral, non-adaptive changes in sequences. This has been tested by comparing molecular phylogenies with commonly accepted divergence times. It turns out that there is considerable variation in the rates of molecular change between different proteins and between different DNA sequences, and even within a particular protein, the rate is not constant. However, on average, the correlation between molecular change and time is quite good. So, bearing these caveats in mind, the molecular clock can be extremely useful.

Molecular clock based on DNA-DNA hybridisation



Molecular clock based on sequence difference



Panacea?

Unfortunately, no. Molecular data is enormously useful, but it has its own specific set of problems. It once promised to answer all the questions, but different molecular methods give different answers. Nowadays, it is generally seen as yet another method. In the end, all methods get used and the accepted evolutionary tree is usually a compromise affair. See Colin Paterson's book for an excellent discussion.

- Are sequences homologous? (i.e. is there any point comparing them?) Or are they analogous?
- A homologue, anatomically is a single part or organ "anatomical singulars" (e.g. femur, pancreas).
- A homonym is a multiple homologue. e.g. leaves or red blood cells.
- All homologous molecules are obviously homonyms... And homologous molecules can be searched for statistically (since you can calculate the chance of random matching).
- Also, can look at tertiary structure for homology.
- Parsimony (Occam's Razor - minimum change - not to be confused with Occam's brom which is used to sweep unexplainable experimental data under the carpet). Can be calculated by a variety of computer programs although there are severe problems with the computing time required.

Different kinds of homology:

- Orthology - genes and gene products that can be used to infer common ancestry among organisms.

- Paralogy - genes and gene products that can be used to infer common ancestry among GENES. e.g Globin family of genes coding for myoglobin, alpha, beta and gamma haemoglobin
- Paralogy is like serial homology - the relationship between multiple derived structures such as limbs in crustacea or vertebrae in mammals.
- Can be very difficult to distinguish between paralogous and homologous sequences.

Other problems:

- Genes can mix - bits of one gene can combine with bits of another gene - partial homology.
- And pseudogenes - segment of DNA (often produced by accidental reverse transcription) that is like a gene but with no promoter so that it is not expressed.
- Xenology - foreign genes acquired by transfection.

Bibliography

Goodman, M., *Macromolecular sequences in systematic and evolutionary biology*. London: Plenum, 1982.

Gould, S. J., *Hen's teeth and horse's toes*. Harmondsworth: Penguin, 1983.

Hennig, W., *Phylogenetic systematics*. Urbana: University of Illinois Press, 1979.

Patterson, C., *Molecules and morphology in evolution : conflict or compromise?* Cambridge: Cambridge University Press, 1987.

Simpson, G. G., *Principles of animal taxonomy*. Columbia University Press, New York. 1961.

Sneath, P. H. A. and Sokal, R. R., *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman, 1973.

Strickberger, M. W., *Evolution*. Jones and Bartlett Publishers, Boston. 1990