Computational Molecular Biology and Bioinformatics

Phylogenetic Analysis

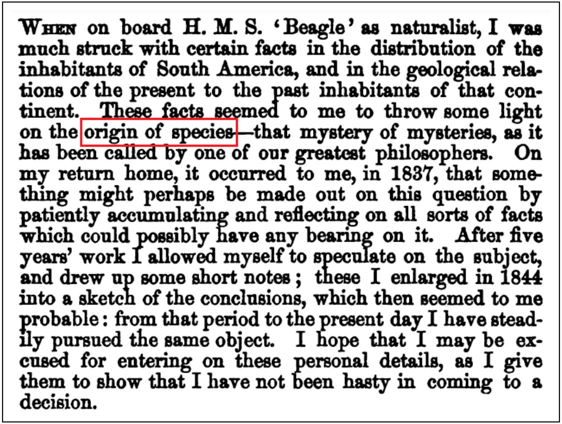
Malay Bhattacharyya

Assistant Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

December, 2021

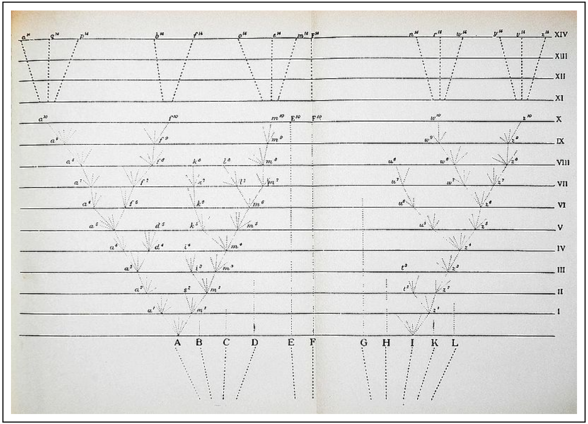
The theory of evolution



WHEN on board H. M. S. 'Beagle' as naturalist, I was much struck with certain facts in the distribution of the inhabitants of South America, and in the geological relations of the present to the past inhabitants of that continent. These facts seemed to me to throw some light on the origin of species—that mystery of mysteries, as it has been called by one of our greatest philosophers. On my return home, it occurred to me, in 1837, that something might perhaps be made out on this question by patiently accumulating and reflecting on all sorts of facts which could possibly have any bearing on it. After five years' work I allowed myself to speculate on the subject, and drew up some short notes; these I enlarged in 1844 into a sketch of the conclusions, which then seemed to me probable: from that period to the present day I have steadily pursued the same object. I hope that I may be excused for entering on these personal details, as I give them to show that I have not been hasty in coming to a decision.

Source: Charles Darwin, *On the Origin of Species by Means of Natural Selection*, 1859.

Evolution to phylogeny



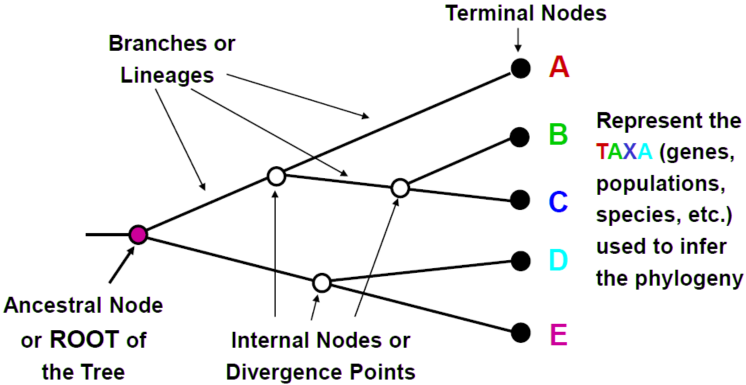
The illustration of species divergence as a tree by Charles Darwin

Phylogenetics

Phylogenetics is the study of relationships among different groups of organisms and their evolutionary development.

Phylogeny (also termed as phylogenetic tree) is a diagrammatic hypothesis of relationships that reflects the evolutionary history of a group of organisms.

Basic terminologies



Represent the TAXA (genes, populations, species, etc.) used to infer the phylogeny

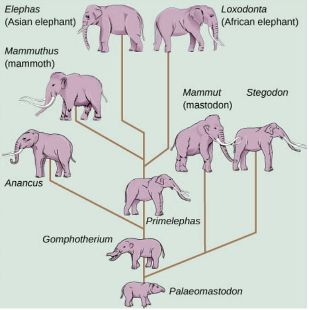
The generic problem

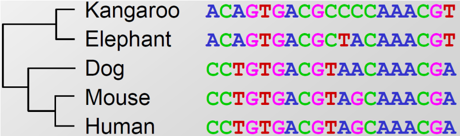
Phylogenetics

Inferring complete ancestry of a set of *objects* based on knowledge of their *traits*.

- Objects can be – species, genes, cell types, tissue types, diseases, etc.
- Traits can be – morphological, molecular, gene expression, TF binding, motifs, etc.

The progress in phylogenetics





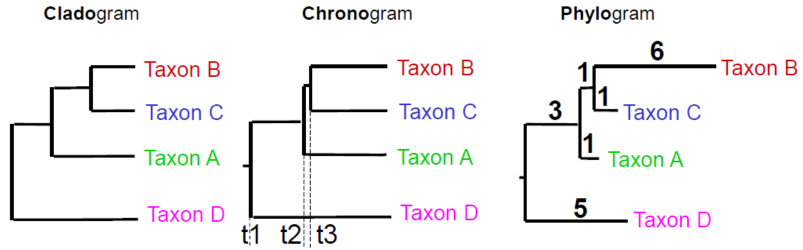
From morphology data
(traditional traits)

From molecular data
(modern traits)

Traditional vs. modern phylogenetics

Traditional phylogenetics	**Modern phylogenetics**
1. Building species trees	1. Building gene trees and species trees
2. Small number of traits (e.g., hoofs, nails, teeth, horns, etc.)	2. Very large number of traits (e.g., every DNA base, every protein residue, etc.)
3. Well-behaved traits, each arose once	3. Frequently ill-behaved traits

Types of phylogenetic trees

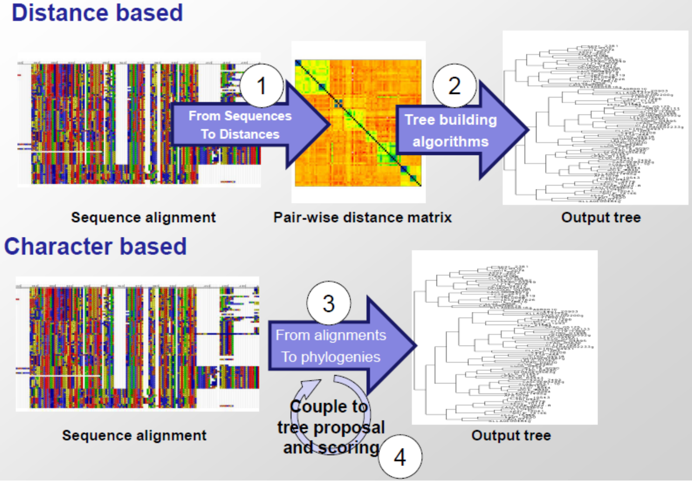


Topology only

Topology +
Divergence times

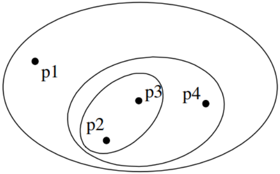
Topology +
Divergence times +
Divergence rates

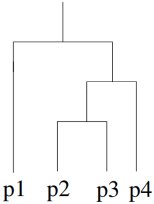
Phylogenetic inference from molecular data



Hierarchical clustering for phylogenetic analysis

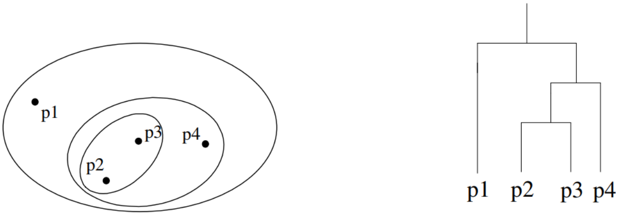
Hierarchical clustering seeks to build a hierarchy of clusters.





Hierarchical clustering for phylogenetic analysis

Hierarchical clustering seeks to build a hierarchy of clusters.



It adopts one of the following two strategies:

- **Agglomerative:** A bottom-up approach wherein each vector is assigned to a separate cluster, and pairs of clusters are recursively merged as one moves up the hierarchy.
- **Divisive:** A top-down approach wherein all vectors are assigned to a single cluster, and it is recursively split as one moves down the hierarchy.

Linkage methods

The linkage criterion is a function of the pairwise distances that determines the distance between two vectors (say D_1 and D_2) to be combined.

Linkage methods

The linkage criterion is a function of the pairwise distances that determines the distance between two vectors (say D_1 and D_2) to be combined.

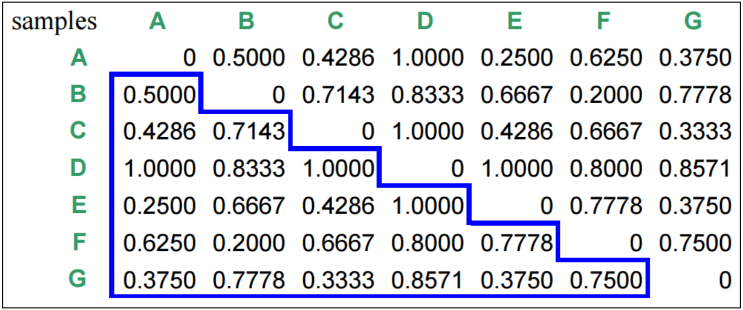
The different linkage methods used in hierarchical clustering are as follows:

- single linkage ($\min_{d_i \in D_1, d_j \in D_2} dist(d_i, d_j)$)
- complete linkage ($\max_{d_i \in D_1, d_j \in D_2} dist(d_i, d_j)$)
- average linkage ($\frac{1}{|D_1||D_2|} \sum_{d_i \in D_1} \sum_{b \in D_2} dist(d_i, d_j)$).

Hierarchical clustering in action

Let us illustrate this step by step with an example.

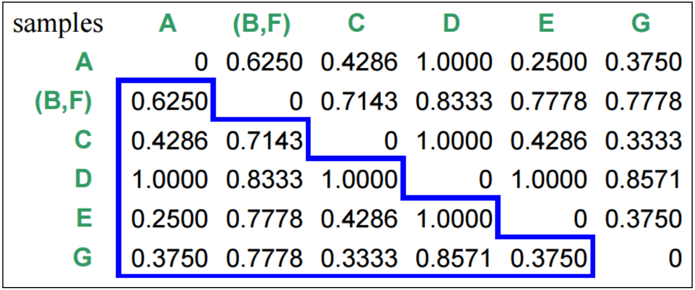
Step 1:



samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Hierarchical clustering in action

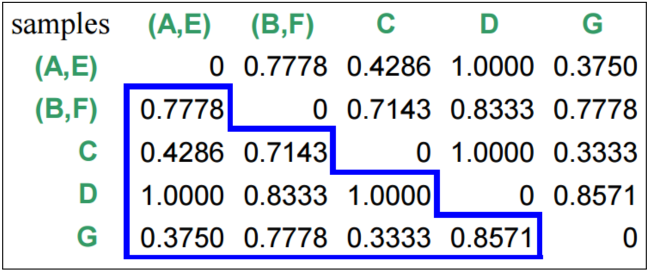
Step 2:



samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0

Hierarchical clustering in action

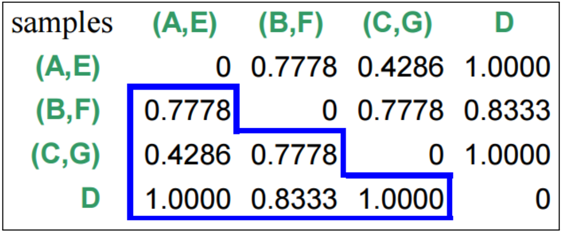
Step 3:



samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0

Hierarchical clustering in action

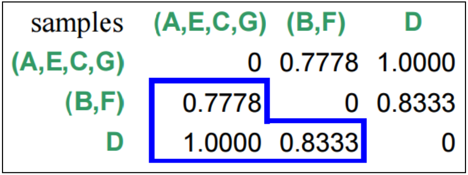
Step 4:



samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0

Hierarchical clustering in action

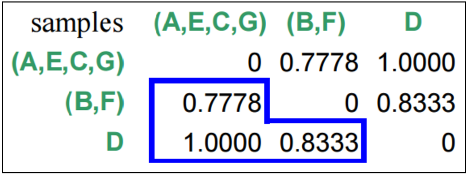
Step 5:



samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

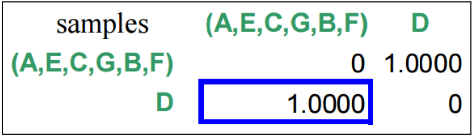
Hierarchical clustering in action

Step 5:



samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

Step 6:



samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0

Hands-on

1. Read the following paper and reproduce its results.
    - i) Palandacic, A., Naseka, A., Ramler, D. and Ahnelt, H., 2017. Contrasting morphology with molecular data: an approach to revision of species complexes based on the example of European Phoxinus (Cyprinidae). BMC Evolutionary Biology, 17(1), pp.1-17, 2017.
      Link: `https://bmcecolevol.biomedcentral.com/articles/10.1186/s12862-017-1032-x`.