Numerical Taxonomy

A **phylogeny** is a tree representation of the evolutionary history of the species we are interested in.

At each leaf of the tree is a species – we also call it a taxon in phylogenetics (plural form is taxa). They are all distinct.

Each internal node corresponds to a speciation event in the past.

When reconstructing the phylogeny we compare the characteristics of the taxa, such as their appearance, physiological features, or the composition of the genetic material.)

Data set used in phylogenetic analysis

- Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
- Molecular markers (e.g., SNPs, RFLPs, etc.)
- Morphology
- Gene order and content

Types of data used in phylogenetic inference:

Character-based methods: Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

Distance-based methods: Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Example 1: Uncorrected "p" distance (=observed percent sequence difference)

Example 2: Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

Slide by Caro-Beth Stewart

Numerical Taxonomy or Phenetics: Taxonomy is viewed and practiced as an empirical science.

Before proceeding, it is necessary that we clearly define our use of the term "numerical taxonomy." We mean by it *the grouping by numerical methods of taxonomic units into taxa on the basis of their character states.* The term includes the drawing of phylogenetic inferences from the data by statistical or other mathematical methods to the extent to which this is possible. These methods require the conversion of information about taxonomic entities into numerical quantities. We have preferred

1.4 THE ESTIMATION OF RESEMBLANCE

Estimation of resemblance is the most important and fundamental step in numerical taxonomy. It commences with the collection of information about characters in the

taxonomic group to be studied. This information may already exist and merely require extraction from the literature, or it may have to be discovered entirely or partly de novo. In most cases both of these procedures will need to be applied. For the method to be reliable, many characters are needed. All kinds of characters are equally desirable: morphological, physiological, ethological, and sometimes even distributional ones. One must guard only against introducing bias into the choice

Numerical Taxonomy Example:

	Character					
Taxon	1	2	3	4	5	6
A	0	0	0	0	0	0
B	0	1	1	1	0	1
C	0	1	0	1	1	1
D	0	1	0	1	1	1
E	0	1	1	1	0	1

0= character absent
1= character present

This would lead to the following similarity matrix:

(count up how many characters are shared)

	A	B	C	D	E
A		2	2	2	2
B			4	4	6
C				6	4
D					4

Or, equivalently, the following distance matrix:

	A	B	C	D	E
A		4	4	4	4
B			2	2	0
C				0	2
D					2

This leads to the following relationships ((B,E), (C,D), A)

Similarity or Distance Coefficients which express relationships of either similarity or difference between units defined by binary data are especially important in systematic biology because of the wide availability of data of this type, such as features present or absent in specimens or species present or absent in samples.

http://evolution.genetics.washington.edu/**phylip**/general.html

http://**paup**.phylosolutions.com

https://www.researchgate.net/publication/246982444_**NTSYS**-pc_-_Numerical_Taxonomy_and_Multivariate_Analysis_System

Taxa	Characters									
	1	2	3	4	5	6	7	8	9	10
A	1	1	1	1	1	1	1	1	1	0
B	1	1	1	0	0	1	1	1	0	0
C	1	1	1	1	0	1	1	1	0	1
D	1	1	0	0	0	1	0	0	0	0
E	1	1	0	0	0	0	1	0	0	0