

## Lecture 2 – Phylogenetic Characters

**I. Introduction** – Nearly all methods of phylogenetic analysis rely on characters as the source of data.

A. Character variation is coded into a **character-by-taxon matrix**.

	Taxon A	Taxon B	Taxon C	Taxon D
Character 1	0	0	1	1
Character 2	1	1	0	1
Character 3	0	0	0	1
Character 4	1	1	1	0
Character 5	1	1	0	0
Character 6	1	2	2	0
Character 7	G	A	C	C
Character 8	*	*	&	&
Character 9	15	10	10	25

Exceptions. Even in current studies that use genetic distances to estimate phylogeny, those distances most often summarize character data. The few exceptions to this are some methods that estimate genetic distances indirectly. These include Immunological Distance and genetic distances estimated by DNA-DNA hybridization. Neither of these methods currently sees much usage, so we won't worry too much about them. However, many comparative genomics studies are developing distance-based approaches.

We make the assumption that **characters are independent**.

**B. Definition** – A character is a trait, feature, or attribute of an organism.

1. In the **most common usage**, a **character is composed of a number of states**. Character states are the manifestation of the character in particular taxon.

For example **Eye Color** is a character and its states are Blue, Brown, & Green

2. Another usage of the term character, that many cladists have employed, is equivalent to character state.

In this usage, there is a single transformation series (TS), and the various manifestations of that TS are the characters. So Eye Color would be the TS and Blue, Brown & Green are the characters. We will not use the term in this manner, although we need to be aware of this usage.

3. **Selection of Characters** - In principle, any homologous character may be used in phylogeny analysis, but efforts are usually made to only focus on characters that are thought to exhibit levels of variation that are appropriate in the context of the study at hand. Our example, Eye Color wouldn't be considered an appropriate character for a study of mammalian relationships because it exhibits **far too much variation**. Conversely, the number of heart chambers wouldn't

be an appropriate character in the same study, because it is invariant across mammals and, therefore, uninformative.

Therefore, the variation shown by a character must match the **level of universality**. Presence of hair is primitive if we're working at the level of mammalian orders. It's derived and unites mammals if we're working at the level of Vertebrata.

The selection of characters on which to focus prior to collection of data is an extremely important consideration. It also represents an enormous source of subjectivity in phylogenetic analysis that is usually not acknowledged.

**II. Homology** – The concept of homology is central to phylogenetic inference. Only homologous characters can be compared (although there is a growing dissenting view).

**A. Original Definition** – It was first used in 1848 by Richard Owen. “The same organ in different organisms in all its varied forms”

Richard Owen was a very influential vertebrate morphologist of Darwin's day, and one of Darwin's biggest critics; in fact he denied evolution. Nevertheless, he recognized amazing skeletal similarities across a broad range of vertebrates. If we look at the forelimbs of a bat, a mole and a dugong (a relative of manatees), we see the same elements in spite of the fact that they perform very different functions. To Owen, these represented variations on an **archetype**.

**B.** To a systematist, a **more appropriate definition** of homology would be something like this.

“Possession by two or more species of a trait inherited from a common ancestor, either with or without modification.”

**C. Criteria** for establishing homology in choosing morphological characters.

Prior to analysis, **hypotheses of character homology must be established**. We'll talk about molecular characters later; for now I want to focus on morphological characters. I also want to exemplify the extreme care that the best morphological cladists exhibit.

Certainly one of the fundamental texts on the issue of morphological phylogenetics is Wiley's 1981 book *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. We have this in our library.

**1. Morphological Criteria** (used to erect hypotheses of homology)

- a. Similarity of position – This includes overall position (i.e., humerus occurs in the front limb)  
It also includes position relative to other structures (scapula, ulna)
- b. Special similarity – This is a rather vague criterion, but it refers to either similarity on a finer scale, or perhaps similarity in developmental pathways.

c. Continuance through intermediate forms (i.e., intermediate fossils).

**2. Phylogenetic Criteria** – Once we've conducted a phylogenetic analysis, including this character in our analysis, we have tested the hypothesis of homology further.

a. Are the putatively identical character states we've identified in generating our character by taxon matrix indeed synapomorphies that unite the groups that possess them?

b. Character congruence either supports or refutes these hypothesized homologies.

**III. Coding:** We've skipped a step here however. We need to transform our putative hypotheses about character-state variation into our character-by-taxon matrix.

This is of fundamental importance in morphological phylogenetics, because the matrix is actually what is analyzed.

We need to make decisions regarding whether to accept multi-state characters, and whether to order the character states or leave them unordered.

**Simplest are qualitative binary characters.** For example, is the first upper premolar present or absent?

0 <---> 1

Also simple are **linear, ordered multi-state characters.**

0 <----> 1 <----> 2 <----> 3

**Unordered multistate characters** are similarly easy to deal with. These are very common, e.g., molecular sequence data.

More difficult to deal with are **non-linear, ordered, multistate characters.** These become necessary when dealing with complex morphological characters. These require a complex character-state tree that must be reflected in the character coding.

Frequently, these are decomposed into a series of binary characters:

0 → 0 0 0 0

1 → 1 0 0 0  
2 → 1 1 0 0  
3 → 1 1 1 0  
4 → 1 1 0 1

These all represent decisions that have to be made in order to convert your understanding of character variation into a character-by-taxon matrix that can be used to infer phylogenies.

**IV. Polarity** – Recall that the Hennigian approach focuses only on synapomorphies, not on shared primitive characters (Sympleisiomorphies).

This led most early cladists to conduct an assessment of character polarity prior to erecting a matrix.

There was a period in which a number of criteria were proposed to establish polarity (i.e., determine the primitive state), including common=primitive (not useful), ontogeny (following von Baer's Law), and evidence from fossils (as we've already exemplified).

By far, the most important criterion is **outgroup analysis**.

An **outgroup** is a taxon (or taxa) included in the analysis that is thought, **based on independent data**, to be more distantly related to the focal taxa, which are called the **ingroup**.

Another way of saying this is that the ingroup must be monophyletic relative to the outgroup.

A character state that is **shared between the outgroup and at least one member of the ingroup is deemed to be the primitive state** for that character.

For several years, the only acceptable approach was to polarize characters *a priori*, that is, before a phylogenetic analysis.

It became apparent during the early 1980's though that simply including an outgroup(s) in an analysis would allow *a posteriori* polarity determination, and that these are equivalent (Maddison et al., 1984). We'll talk more about this when we discuss rooted vs. unrooted trees.

Some really good references for the material covered here are:

Wiley, E. O. 1981. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. John Wiley and Sons, Inc., NY.

Wiley, E. O., D. Siegel-Causey, D. R. Brooks, and V. Funk. 1991. *The Compleat Cladist: A Primer of Phylogenetic Procedures*. University of Kansas Museum of Natural History, Special Publication 19. This is out of print, but a pdf is available from <http://www.nhm.ukans.edu/cc.html>.

Maddison, W. P., M.J. Donoghue, and D. R. Maddison. 1984. Outgroup analysis and parsimony. *Systematic Zoology*, 33:83-103.

## V. Molecular characters

**A. Types of Molecular Data:** Certainly the lion's share of molecular data used in phylogeny estimation is DNA sequence data, but we should examine other types of molecular data, and some of the characteristics of those data that influence phylogeny estimation.

### 1. Inherently Distance-based Data

There are molecular phylogenetic approaches in which the nature of the data produced is a matrix of pair-wise distances. The units for these distances vary, but the matrix can then be subjected to a number of potential phylogenetic analyses.

	cwk1056	eaa292	cwk1025	eaa448	dsr5032	eaa028	fac1117	cwk1007
cwk1056	-----							
eaa292	0.05840708	-----						
cwk1025	0.01769911	0.05398230	-----					
eaa448	0.08672567	0.08141593	0.08230089	-----				
dsr5032	0.02566372	0.05929204	0.01946903	0.08495575	-----			
eaa028	0.06725664	0.07433628	0.06371681	0.07522124	0.07168142	-----		
fac1117	0.02123894	0.05575221	0.00530973	0.08053097	0.02123894	0.0637168	-----	
cwk1007	0.05221239	0.02920354	0.05132743	0.08230089	0.05486726	0.07610620	0.05132743	-----
eaa667	0.05840708	0.01238938	0.05221239	0.07787611	0.05752213	0.07433628	0.05398230	0.02743363

Again, historically inherently distance based methods have been things like **DNA-DNA hybridization** and **immunological distances**. We won't focus on those because they're never used anymore. However, much emerging information from comparative genomics may be presented as inherently distance data.

### 2. Discrete Characters

The most common type of molecular data is **sequence data**. These can obviously be from DNA or from proteins.

**a. Gene sequences** (DNA sequences) are the most common.

This is quite straightforward – nucleotide positions are the characters, and the four nucleotides are the character states (there are only four possible states).

We'll spend the majority of the semester with this type of data.

**b. Protein sequences** (amino-acid sequences) are also used, particularly to estimate old phylogenetic relationships.

So, amino-acid positions are the characters and the 20 amino acids are the character states. There's a much bigger character state space, and this is seen by some as an advantage.

However a disadvantage is that there may be more than one path of substitutions from one amino acid to another, and this has led some to view this as an inferior character type.

Typically, nucleotide sequences are collected and amino-acid sequences are inferred. This is due to the fact that the DNA sequencing technology is far easier.

**5. Higher order molecular characters.** There are several different types of higher order molecular characters. Some examples include:

**a. Genome structure** (Gene order or gene content) – This type of data certainly is being used more as genomics matures, and there has been a recent flurry of activity in using genome rearrangement data in phylogeny estimation.

These have often focused on organellar genomes (e.g., rearrangement of genes in mtDNA genome), but have recently been applied to the growing database of prokaryotic genomes.

Most of the new methods are distance based, whereas others attempt to use inferred rearrangement events as characters. The papers don't tend to show up in the phylogenetics literature.

**b. Protein Domains** – Presence/absence of protein domains (called fold superfamilies, or FSF's) in complete genomic sequences.

A paper was published recently (Yang et al., 2005. PNAS. 102:373) in which the authors show that NJ tree based on distances derived from presence/absence data produces bacterial trees (deepest level in the tree of life) consistent with other data. (174 genomes – 1294 FSF)

A recent MB&E paper (Wang & Caetano-Anolles 2006. Mol., Biol. Evol., 23:2444) provides a similar analysis that include metazoan complete genomes as well.

**B. Homology of Molecular Characters** –For molecular characters, there are two levels of homology that we have to deal with.

**1. Alignment** and homology of characters (e.g., nucleotide sites).

Because of its centrality to molecular phylogenetics, alignment will be addressed separately in Lecture 5. Alignment of sequences determines positional homology.

**2. Genic homology** - At a higher level, we have to worry about homology of the source of the sequence data (i.e, homology of the genes).

Remember that homology is similarity due to descent from a common ancestor.

Paralogy is homology due to gene duplication.

Orthology is homology due to speciation,

Gene duplication has been an incredibly important and widespread evolutionary process.

It's easy to demonstrate that if a mixture of orthologous and paralogous genes are compared, the true history of the genes (a.k.a. the gene tree) won't correspond to the species phylogeny (a.k.a. the species tree) that we're trying to estimate.

Because gene duplication is such a widespread phenomenon, there have been several methods devised to try to take advantage of them a character data. Unfortunately, we don't have time to get into them, but a good introduction is provided by: Slowinski & Page (1999. *Systematic Biology* 48:814-825).