Phylogenetic and genealogical homology

Phylogenies distinguish homology from similarity

Previously, we examined how rooted phylogenies provide a framework for distinguishing similarity due to common ancestry (HOMOLOGY) from non-phylogenetic similarity (ANALOGY). Here we extend the concept of phylogenetic homology by making a further distinction between a HOMOLOGOUS CHARACTER and a HOMOLOGOUS CHARACTER STATE. This distinction is important to molecular evolution, as we often deal with data comprised of homologous characters with non-homologous character states. The figure below shows three hypothetical protein-coding nucleotide sequences (for simplicity, only three codons long) that are related to each other according to a phylogenetic tree. In the figure the nucleotide sequences are aligned to each other; in so doing we are making the implicit assumption that the characters aligned vertically are homologous characters. In the specific case of nucleotide and amino acid alignments this assumption is called POSITIONAL HOMOLOGY. Under positional homology it is implicit that a given position, say the first position in the gene sequence, was the same in the gene sequence of the common ancestor. In the figure below it is clear that some positions do not have identical character states (see red characters in figure below). In such a case the involved position is considered to be a homologous character, while the state of that character will be non-homologous where there are differences.

<table>
<thead>
<tr>
<th>Phylogenetic perspective on homologous characters and homologous character states</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancestral character states</td>
</tr>
<tr>
<td>C ➩ T</td>
</tr>
<tr>
<td>Implicit in the above alignment is the assumption of positional homology for the “red” position above. At this position C and T are non-homologous character states. Note that the pair of T’s in the first example, and the pair of C’s in the second example represent homologous character states.</td>
</tr>
</tbody>
</table>

It is possible for character states to be the same yet not be homologous. Identical but non-homologous character states are called HOMOPLASY. Homoplasy can arise during the course of evolution by CONVERGENCE, PARALLELISM, or REVERSAL. Note that identity of character states due to shared ancestry represents true phylogenetic signal, whereas identity in character states due to homoplasy represents non-phylogenetic, or false, signal.
CONVERGENT EVOLUTION: Non-phylogenetic identity in character states of a homologous character that arose by
substitutions in independent evolutionary lineages that arrived at identical states by coincidence. Note that this
happens more frequently in nucleotide sequences as compared with amino acid sequences due to the much
more restricted “state space” (i.e., 4 possible states for a DNA based character as compared with 20 possible
states in amino acid character). An example is provided in the figure below.

<table>
<thead>
<tr>
<th>Convergent (non-phylogenetic) similarity of nucleotide character states; i.e., homoplasy</th>
</tr>
</thead>
<tbody>
<tr>
<td>T ⇒ C ⇒ A</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>G ⇒ A</td>
</tr>
</tbody>
</table>

CHARACTER REVERSAL: A non-phylogenetic identity in character states of a homologous character that arose by
one or more substitutions followed by a coincidental substitution back to the original character state (a
REVERSAL). Such identity is not necessarily phylogenetically misleading; however, it can lead to underestimated
branch lengths [this can lead to all sorts of problems]. An example is provided in the figure below.

<table>
<thead>
<tr>
<th>Reversal leads to non-homologous similarities in character state.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ⇒ C ⇒ A</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>T</td>
</tr>
</tbody>
</table>

PARALLEL EVOLUTION: A Non-phylogenetic identity in character states of a homologous character that arose by
an identical series of substitutions in independent evolutionary lineages. An example is provided in the figure
below.
The longer the separation of two genes (in two independent evolutionary lineages), the more likely that the above types of homoplasy will occur. You probably realized that differences between sequences accumulate after the evolutionary separation of two gene sequences; but now you know that a certain amount of non-phylogenetic similarity accumulates in such gene sequences as well. As long as the amount of phylogenetic signal accumulates at a faster rate than non-phylogenetic signal, we can recover a reasonably good estimate of the phylogeny from gene sequences. However, there are special cases where non-phylogenetic signal exceeds the true phylogenetic signal; we will return to this problem later.

We saw in the above examples the homoplasy arises when MULTIPLE SUBSTITUTIONS OCCUR AT A SINGLE SITE along a single lineage. This phenomenon is sometimes called SUPERIMPOSED SUBSTITUTIONS, MULTIPLE SUBSTITUTION, or MULTIPLE HITS. Whether or not superimposed substitutions lead to non-phylogenetic similarity between unrelated sequences, it leads to difficulty in accurately estimating a branch length.

**Evolutionary dissociation**

Mindell and Meyer (2001) define EVOLUTIONARY DISSOCIATION as a “change in linkage or effects that particular traits from different levels of biological organization have on each other over evolutionary time”. A classic case involves continuity in the patterning of morphological development among divergent organismal lineages while the underlying genetic mechanisms of the developmental process have changed. In *Drosophila* the transcription factor called even-skipped is responsible for important pattern formation in embryos, while in the distantly related wasp (*Aphidius ervi*) and locust (*Schistocerca americana*) the same structure are present but develop under the control of a different transcription factor (Patel et al. 1992; Gribic and Strand 1998) Although, the evolution of development provides a good example, dissociation could involve any complex network of gene expression associated with an organism’s phenotype.
A general model for evolutionary dissociation might involve **genetic co-option**; here existing genetic systems are modified, or co-opted, for new uses. In evolutionary terms this is much more efficient than having to construct a system from scratch to provide a new or modified function. The process of gene duplication is likely to play an important role, as it provides a source of material for modification by natural selection while still maintaining a functional copy of the ancestral system. Over time, with multiple rounds of duplication and genetic co-option the process of evolutionary displacement could take place.

The phenomenon of evolutionary dissociation illustrates an important point: **homologous characters can have non-homologous character states**. The example of the alignment of gene sequences provides a familiar example. The positions in the alignment are assumed to be homologous (assuming no alignment errors); this is called **positional homology**. Many sites in this alignment will be variable. If we relate the individual gene sequences to each other by a phylogeny, then we can distinguish between homologous states (i.e., identities in the states A, C, T, or G due to shared ancestry) among lineages at a position, from homoplasies (i.e., identities in the states A, C, T, or G due to convergence, reversals, or parallelisms).

**Alternatives to the phylogenetic concept of homology**

For a long period of time the criterion for defining homology was simple similarity. In fact, there are still disciplines that equate homology to similarity; such practice can lead to misunderstanding. For example, it is common practice for molecular biologists with no evolutionary training to refer to "percent homology" between two DNA sequences. In such cases they are referring to the percentage of nucleotide sites that are the same between the two sequences. There is considerable potential for misunderstanding because they are usually willing to assume that the entire set of DNA sequences have a single evolutionary origin, yet an evolutionary biologist might assume that such sequences have some percentage of sites that have a unique evolutionary history, perhaps by the process of lateral gene transfer.

**Serial homology** refers to repeated structural units of morphology of an organism. Examples of such serially arranged structures are the body segments of arthropods and the individual vertebrae of vertebrates. It seems reasonable that such units are derived from homologous genes and developmental processes; whether this assumption turns out to be generally applicable to morphological features exhibiting serial homology has not yet been demonstrated.

**Functional homology** is generally used when the functional attributes of organisms are similar due to shared ancestry. However, due to a long history of using this term to denote simple similarity, one should not assume this usage unless unequivocally stated.

**Biological homology** is defined by Wagner (1989) as morphological structures that share "developmental constraints, caused by locally acting self-regulating mechanisms or organ differentiation". Although phylogeny can be used to help diagnose such homology, the definition does not explicitly include the condition of shared ancestry.

As is traditional in the discipline of Molecular Evolution, **we will use common ancestry as the single criterion for homology**. We have seen that the concept of similarity leaves open too many alternatives to be useful in the study of the evolutionary history of characters such as genes and genomes. We will see shortly that the dynamic and complex process of the evolution of genes and genomes requires even more precise definitions of different forms of homology.
In molecular evolution we are interested in the evolutionary histories of genes. A species phylogeny is a diagram of the macroevolutionary relationships of the organism lineages we have been calling a species. As you know, any point in time along the branch of a phylogeny represents a population of individuals. Contained within these organism lineages are the evolutionary histories of the organism’s own genes and genome.

In the above figure, only the genealogies of the individual organisms are depicted. In a diploid population, each individual possesses two copies of each gene. Each one of these copies has its own genealogy, and the different copies within the same individual will have different genealogies. Due to largely random mating in natural populations, the connections between genes in individuals will be very "tangled" (see figure below). The "tangled" pattern can be greatly simplified by removing the circles that indicate individual organisms, and sorting the gene genealogies to untangle the connecting branches (see figure below). Note the structure of these gene-genealogies is dependent on micro-evolutionary forces such as selection, mutation, drift, etc. Thus the effective size of the population ($N_e$) of the involved organism is relevant to the structure of the gene genealogies, and the relationship between microevolution of the genes and the macroevolutionary pattern of the species!
In this less complicated format we can take a fresh look at the notions of mutation, polymorphism, and substitution.

Genealogies of genes in a population of diploid individuals. The tangle of branches is simplified by removing the circles that represent individual organisms, and swapping genes left and right to untangle the branches.

Individuals are represented by circles. Diploid genotypes indicated by pairs of dots. Relationships are tangled because of random mating.

Removing the circles shifts the focus away from individual organisms and towards individual copies of genes.

Now the genes can be sorted right and left to untangle the lineages.

NOTE: Above looks much better, but genes from the same individual are no longer next to each other!

Polymerism and substitution (highly simplified) along a branch of a phylogeny

Residence time: the time that a particular neutral polymorphism is present in a population.

Mean residence time is determined by effective population size ($N_e$)

Population substitution 1: $A \Rightarrow C$

Population substitution 2: $C \Rightarrow A$

Population substitution 3: $A \Rightarrow G$

Coalescent
**Forms of Homology**

The evolutionary history of any type of character, not just a gene, can be different from that of the organism in which it is measured. As we have seen, there are numerous evolutionary processes (e.g., LGT, recombination, etc.) that can lead to such a discrepancy. In order to accurately portray such complexity, we need a more refined set of definitions. Below is a set of definitions for various forms of evolutionary relationships; the involved evolutionary processes are defined in terms of the gene. It is important to remember that the characters need not be genes; the characters could be organismal traits such as a chromosome, developmental role, or morphology.

1. **Orthology.** Orthologous genes are derived from the divergence of an organismal lineage; i.e., a speciation event. Thus if we look at orthologs on a phylogeny we see that their most recent common ancestor represents the coalescence of two organismal lineages.

2. **Paralogy.** Paralogous genes are derived from the divergence event within a genomes; i.e., a gene duplication event. In this case if we look at paralogs on a phylogeny we see that they coalesce at a gene duplication event.

3. **Pro-orthology.** A gene is pro-orthologous to another gene if they coalesce at a speciation event that predates a gene duplication event. Thus a single-copy gene in organism A is pro-orthologous to a gene that is present in multiple copies in organism B due to gene duplication events that followed the divergence of organisms A and B.

4. **Semi-orthology.** This is a term that simply takes the reverse perspective of pro-orthology. Any one of the multi-copy genes in the genome of organism B is said to be semi-orthologous to a single copy gene in the genome of organism A, if the most recent common ancestors of those genes coalesce at a point in time that predates the gene duplication event.

5. **Partial homology.** This refers to the situation that arises when the evolutionary histories of different segments within the same gene coalesce at different ancestors. This can arise from evolutionary processes such as homologous recombination or exon shuffling.

6. **Gametology.** Gametologs coalesce at an event that isolated those genes on opposite sex chromosomes; i.e., they coalesce at the point when they became isolated from the process of recombination.

7. **Xenology.** Genes that coalesce at either a speciation or duplication event, but whose evolutionary histories do not fit with that of the organismal lineages which carry such genes due to one or more lateral gene transfer events.

8. **Sinology.** Homologous genes found within the same organism’s genome have different evolutionary histories due to the fusion of formerly evolutionarily independent genomes, such as in endosymbiosis.
The mammalian Ldh-A and Ldh-C gene family is used as an example to illustrate the various forms of homology (ORTHOLOGY, PAROLOGY, PRO-ORTHOLOGY, and SEMI-ORTHOLOGY) that are important when dealing with gene families.

Examples of different types of homology in gene families:

* Homo and Rattus (rat) Ldh-C are orthologous.

* Homo Ldh-C and Homo Ldh-A are paralogous.

* Homo Ldh-C and Rattus Ldh-A also are paralogous.

* Gallus (chicken) Ldh-A is pro-orthologous to both Homo Ldh-C and Homo Ldh-A.

* Homo Ldh-C is semi-orthologous to the Gallus Ldh-A.

All mammalian Ldh-A genes are semi-orthologous to the non-mammalian Ldh-A. Note that the gene duplication that gave rise to the Ldh-C gene is specific to an ancestor of all present-day mammals.

Mammalian Ldh-C and Ldh-A genes are paralogous.
**Organism history can differ from genic history**

The condition of homology is often important, and most unclear, when trying to relate processes across different levels of biological organization. For example, the relationship between homology at the genetic level and say, similarity in the embryological origin of a morphological structure might be important. At one level we are working with genetic material that functions within the context of a genome, and at another level we are considering a morphological structure in the context of the development of an organism. Unfortunately there are many evolutionary genetic processes that results in significant discrepancy between the genealogical relationships of organisms and the genealogical relationships of the genes carried by those organisms.

1. **Ancestral polymorphism**: When polymorphism goes to fixation between speciation events, the gene tree will track with the species tree (this is called **RECIPROCAL MONOPHYLY**). Unfortunately, when the time \( t \) between speciation events is very short a polymorphism may persist through the process of speciation. Remember that under genetic drift the direction of fixation is random. So those polymorphism whose fate is determined by drift and persist through speciation events, will be "sorted" (i.e., go to fixation) at random with respect to the phylogeny. The result is that some of these polymorphisms are "sorted" in a non-phylogenetic pattern among macroevolutionary lineages; this is called **NON-PHYLOGENETIC LINEAGE SORTING**, and in such cases the gene tree differs from the species tree. The figure below illustrates this. The probability that a polymorphism will persist long enough for non-phylogenetic sorting depends on two factors: (i) the residence time of the polymorphism, which is determined by \( N_e \); and (ii) the time between successive organism splitting events \( t \). If residence times are small (small \( N_e \)) and the time between splitting events \( t \) is large, the probability will be quite low.
2. Birth-death evolution of gene families: Gene families are sometimes found in linear (tandem) arrays along a chromosome. In such cases the rate of gene “birth” by a duplication event, and “death” by a deletion event, can be quite high. In cases with very high rates of birth and death the gain or loss of genes happened one or more times between speciation events, and the pattern of birth and death is often random. This means that the genealogies of the surviving genes will be independent of the underlying organism tree. The problem is that genes with high similarity and apparently the same position on a chromosome may be paralogs; i.e., they coalesce at a gene duplication event. Treating such genes as if they were orthologous can lead to incorrect interpretation of a gene tree as a species tree. The figure below illustrates this. Some gene families are quite well known for undergoing a quite fast rate of evolution by birth-death. Well known examples include histone gene families and immunoglobulin gene families.
Trans-species evolution: We discussed in an earlier lecture how overdominant selection leads to stable polymorphisms. When a polymorphism is stable it passes easily through one or more speciation events. In such cases different alleles at the same locus can be more closely related to alleles in another species! We have seen this in the MHC locus. Such trans-species polymorphisms are quite common at mammalian MHC loci.

3. Recombination: Recombination leads to partial homology. Imagine a case where there is a tandem array of two genes. As we have seen they can have different genealogies due to birth-death evolution. If there is a partial recombination event between them, the result is that one or both genes could contain a mosaic of phylogenetic signal (see figure below).

Note there are other sources of partial homology. Examples include co-infection of a single host cell by viruses followed by a partial recombination event, or recombination between alleles with very different genealogies within a single diploid individual.
4. **Horizontal or lateral gene transfer (HGT or LGT):** LGT is defined as the transfer of genetic material from one genome to another; such an event is generally assumed to be between different species. Homologous sequences that coalesce at an LGT event are an example of **xenology**. Note that transfers between organelle and nuclear genomes are well known and are sometimes considered LGTs even though they do not occur between different species. LGT is quite common in bacteria. The figure below illustrates how such events lead to conflicts between gene trees and organism trees.

![Gene transfer illustration](image-url)

Lateral gene transfer invalidates the simile of a great tree of life. For frequent LGT, a network might provide a better model of evolutionary relationships.