"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

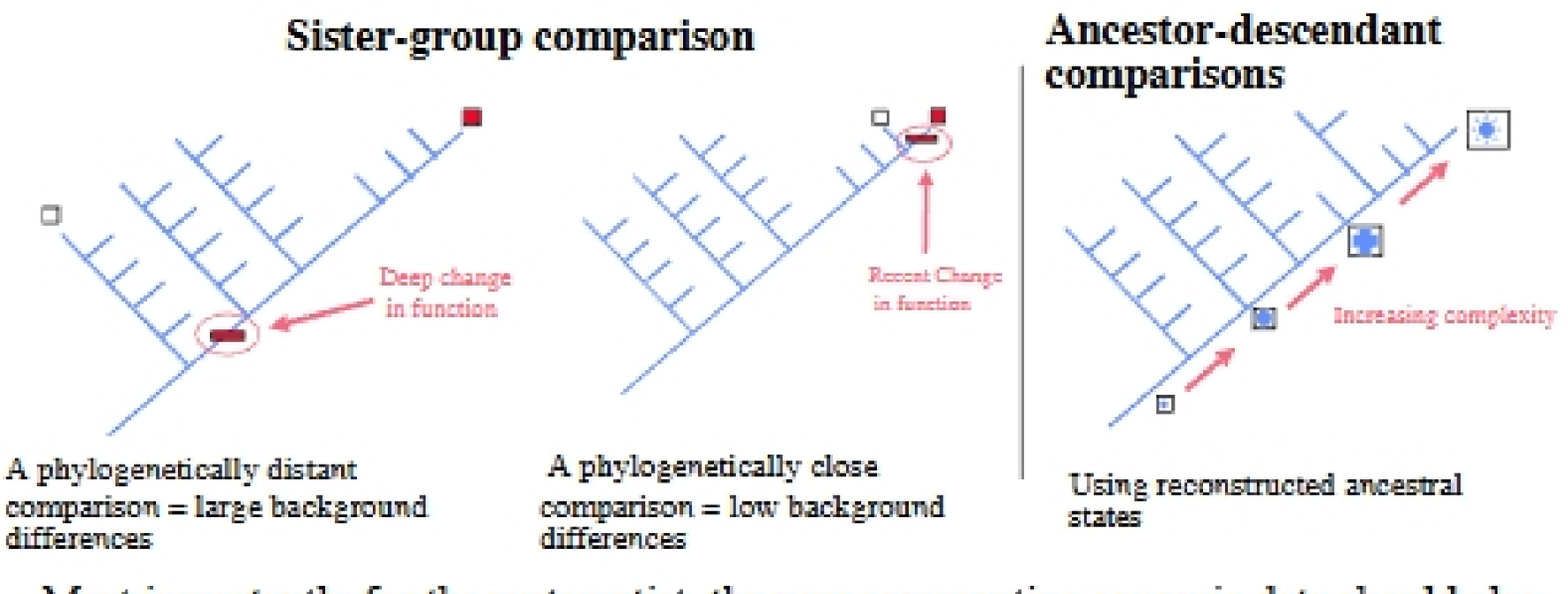
Integrative Biology 200B Spring 2011

University of California, Berkeley Nick Matzke, revised from B.D. Mishler

March 3, 2011. **Comparative genomics; Evolution and development**

This is the era of whole-genome sequencing; molecular data are becoming available at a rate unanticipated even a few years ago. Sequencing projects in a number of countries have produced a growing number of fully sequenced genomes, providing computational biologists with tremendous opportunities. However, comparative genomics has so far largely been restricted to pair-wise comparisons of genomes. The importance of taking a phylogenetic approach to systematically relating larger sets of genomes has only recently been realized.

A recent synthesis of phylogenetic systematics and molecular biology/genomics – two fields once estranged – is beginning to form a new field that could be called "phylogenomics" (Eisen 1998). Something can be learned about the function of genes by examining them in one organism. However, a much richer array of tools is available using a phylogenetic approach. Close sister-group comparisons between lineages differing in a critical phenotype (e.g., desiccation or freeze tolerance) can allow a quick narrowing of the search for genetic causes. Dissecting a complicated, evolutionarily advanced genotype/phenotype complex (e.g., development of the angiosperm flower), by tracing the components back through simpler ancestral reconstructions, can lead to quicker understanding. Hence, phylogenomics allows one to go beyond the use of pairwise sequence similarities, and use phylogenic comparative methods as discussed in this class to confirm and/or to establish gene function and interactions.



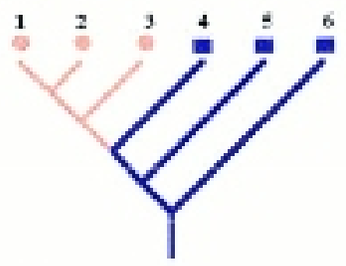
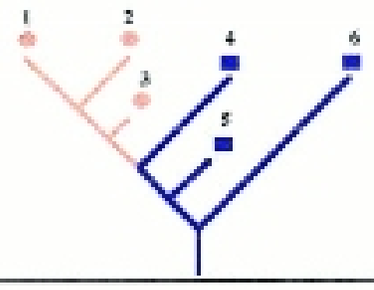
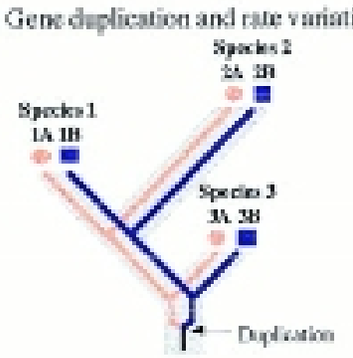
Most importantly for the systematist, the new comparative genomic data should also greatly increase the accuracy of reconstructions of the Tree of Life. Even though nucleotide sequence comparisons have become the workhorse of phylogenetic analysis at all levels, there are clearly phylogenetic problems for which nucleotide sequence data are poorly suited, because of their simple nature (having only four character states) and tendency to evolve in a regular, more-or-less clocklike fashion. In particular, "deep" branching questions (with relatively short internodes of interest mixed with long terminal branches) are notoriously difficult to resolve with DNA sequence data.

It is fortunate therefore, that fundamentally new kinds of structural genomic characters such as inversions, translocations, losses, duplications, and insertion/deletion of introns will be increasingly available in the future. These characters need to be evaluated using much the same

principles of character analysis that were originally developed for morphological characters. They must be looked at carefully to establish likely homology (e.g., examining the ends of breakpoints across genomes to see whether a single rearrangement event is likely to have occurred), independence, and discreteness of character states. Thus close collaboration between systematists and molecular biologists will be required to code these genomic characters properly, and to assemble them into matrices with other data types.

Next two figures from: Jonathan A. Eisen and Claire M. Fraser, Phylogenomics: Intersection of Evolution and Genomics , *Science*, Vol 300, Issue 5626, 1706-1707 , 13 June 2003

Table 4 Examples of Conditions in Which Similarity Methods Produce Inaccurate Predictions of Function

Evolutionary Pattern and Tree of Genes and Functions[1]	Gene With Unknown Function[2]	Highest Hit Method: Predicted Function[3]	Highest Hit Method: Accurate?	Phylogenomic Method: Predicted Function[4]	Phylogenomic Method: Accurate?	Comments
A. Functional change during evolution. 	1 ●	●	+	●	+	• Phylogenomic method cannot predict functions for all genes, but the predictions that are made are accurate. • Highest hit method is misleading because function changed among homologs but hierarchies of similarity do not correlate with the function (see Bolker and Raff 1996).
	2 ●	●	+	●	+	
	3 ●	●	+	●/■	±	
	4 ■	●	-	●/■	±	
	5 ■	●/■	±	■	+	
	6 ■	●/■	±	■	+	
B. Functional change & rate variation. 	1 ●	●	+	●	+	• Similarity based methods perform particularly poorly when evolutionary rates vary between taxa. • Molecular phylogenetic methods can allow for rate variation and reconstruct gene history reasonably accurately.
	2 ●	●	+	●	+	
	3 ●	■	-	●/■	±	
	4 ■	●	-	●/■	±	
	5 ■	●	-	■	+	
	6 ■	■	+	■	+	
C. Gene duplication and rate variation. 	1A ●	●	+	●	+	• Most-similarity based methods are not ideally set up to deal with cases of gene duplication since orthologous genes do not always have significantly more sequence similarity to each other than to paralogs (Fitch et al. 1995, Zardoya et al. 1996, Tatusov et al. 1997). • Similarity-based methods perform particularly poorly when rate variation and gene duplication are combined. This even applies to the COG method (see Table 1) since it works by classifying levels of similarity and not by inferring history. Nevertheless, the COG method is a significant improvement over other similarity based methods in classifying orthologs. • Phylogenetic reconstruction is the most reliably way to infer gene duplication events and thus determine orthology.
	2A ●	●	+	●	+	
	3A ●	■	-	●	+	
	1B ■	■	+	■	+	
	2B ■	■	+	■	+	
	3B ■	●	-	■	+	

[1] The true tree is shown but it is assumed that it is not known. Different colors and symbols represent different functions. Numbers correspond to different species.
[2] The function of all other genes is assumed to be known.
[3] The top hit can be determined from the tree by finding the gene in the shortest evolutionary distance away (as determined along the branches of the tree).
[4] It is assumed that the tree of the genes can be reproduced accurately by molecular phylogenetic methods (see Fig. 1).

Outline of a phylogenomic methodology (next page). In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. (*A*) A gene family has undergone a gene duplication that was accompanied by functional divergence. (*B*) Gene function has changed in one lineage. The true tree (which is assumed to be unknown) is shown at the *bottom*. The genes are referred to by numbers (which

represent the species from which these genes come) and letters (which in *A* represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in *A* (*bottom*) correspond to the phylogeny of the species in which the duplicate genes evolve in parallel (as paralogs). Different colors (and symbols) represent different gene functions; gray (with hatching) represents either unknown or unpredictable functions.