9. Gene trees in species trees

EEOB563

Spring 2021

"The gene phylogeny for a set of OTUs is defined as the ancestral order of branching of evolutionarily related genes leading to the expressed genes in the contemporary OTUs. The species phylogeny for a set of OTUs is defined as the evolutionary branching of species leading to the contemporary species or OTUs. Since a given gene may duplicate within its species lineage, and distinct sister genes may be expressed by different species, the gene lineage topology for a given set of OTUs may not coincide with the species lineage topology for that set of OTUs". Goodman *et al.*, 1979.

1 Preliminary considerations

The beginning of molecular phylogenetics can be traced to the influential paper by Emile Zuckerkandl and Linus Pauling "Molecules as documents of evolutionary history", which was written in 1963 and first published in a Russian translation in 1964. In this article the authors argued that the "macromolecules that carry the genetic information or a very extensive translation thereof" are the best fit for molecular phylogenetics. While development of molecular phylogenetics led to revolutionary changes in multiple areas of biology, it was realized relatively early that there always exist a disconnect between the gene or protein phylogeny we infer and the species phylogeny we often extrapolate from it.

The three main reasons (Fig. 1) why a correctly reconstructed gene tree may not correspond to the species tree are:

- horizontal gene transfer (HGT);
- gene duplication and loss (GDL); and
- incomplete lineage sorting (ILS) or late coalescence

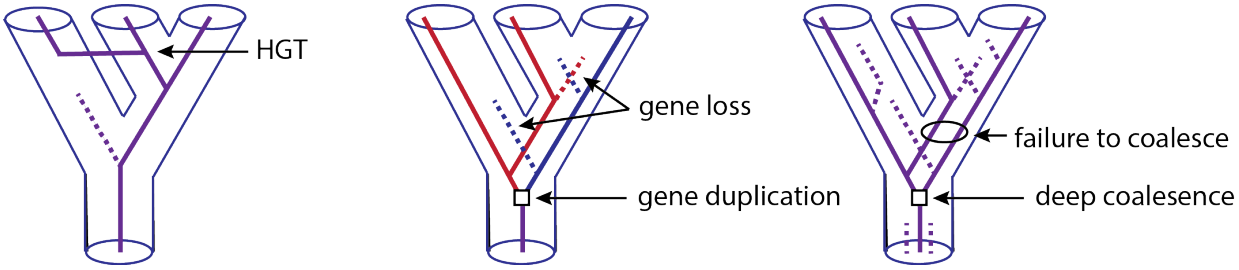

Figure 1: HGT, GDL, and ILS as three main reasons for gene tree/species tree incongruence.

Interestingly, while the reasons for the factors leading to gene tree/species tree discord have been elucidated relatively early (Maddison 1997), the transition in attention of broader biological community to each of them followed the development of sequencing technologies.

1.1 Horizontal Gene Transfer

In the early days of molecular phylogenetics, scientists were lucky to have one or a few molecular sequences from their favorite organisms and the question of gene tree/species tree discord was usually ignored. The development of automated sequencing and shotgun assembly methods opened the possibility of sequencing complete organismal genomes. Starting from *Haemophilus influenza* in 1995, the end of the XX century was the "gold rush" of bacterial genomics. It was quickly realized that horizontal gene transfer is a major player in the evolution of prokaryotic genome and the topic of HGT became one of the dominant topic in the field of molecular evolution. At the apex of that period an issue of New Scientist was published with a cover depicting a tree of life, along with the statement "Darwin was Wrong" and an editorial titled "Uprooting Darwin's tree" (2009, 201[2692]). While the headline was immediately seized upon by creationists, it simply referred to the fact that it might be impossible to build a single evolutionary tree given the rampant HGT among bacterial lineages. Hybridization and endosymbiosis can be seen as special types of gene transfer, affecting a large portion of the genome.

1.2 Gene duplication and loss

In 1996, the first eukaryotic genome (that of *Saccharomyces cerevisiae*) has been sequenced, and several animal and plant species followed soon, including human in 2001. One of the surprises that emerged from these sequences was the frequency of gene duplication and loss in eukaryotic genomes. For example, the rate of gene duplication in *C. elegans* is on the order of 10^{-7} duplications/gene/generation. Estimates of such rates in other eukaryotes range between 0.001 – 0.03 /gene/myr. Interestingly, humans and chimpanzees differ by at least 6% (1,418 of 22,000 genes) in their complement of genes, which stands in stark contrast to the oft-cited 1.5% difference between orthologous nucleotide sequences. These observations led to the "Revolving door" model of gene family evolution.

1.3 Incomplete lineage sorting

For a long time molecular phylogeneticists used a single gene (or genome) sequence to represent a species. The advent of population genomics brought the issue of the influence of population-level variation on phylogenetic reconstruction to the forefront of phylogenetic analysis. In particular, the issue of Incomplete Lineage Sorting (ILS) or Deep Coalescence drove most of the recent theoretical and computational development in the field. ILS can be defined as "the failure of two or more lineages in a population to coalesce, leading to the possibility that at least one of the lineages first coalesces with a lineage from a less closely related population" (Degnan and Rosenberg, 2009).

2 Inferring species trees from gene trees

There are two general approaches to deal with the issue of gene tree/species tree incongruence. First, one can try to filter out the data to eliminate "problematic genes". In fact, a lot of early effort

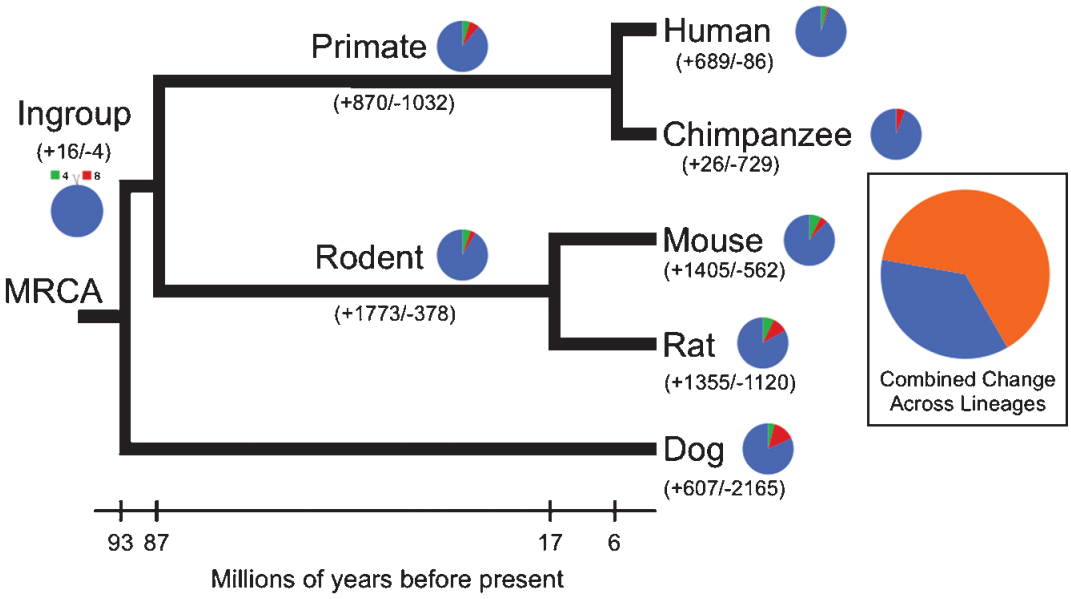

Figure 2: Distribution of gene gain and loss among mammalian lineages. From Demuth et al. 2006

in phylogenomics was devoted to recognition of genes derived through HGT and/or recognition of orthologous gene sequences. In case of HGT, one would identify unusual feature(s) of subsets of genes (such as unusual phyletic pattern, unusual genome location, anomalous DNA composition, etc.) that distinguishes them from the bulk of genes in the genome. It was eventually proposed the existence of "core genes" that are mostly inherited vertically and that can be used to estimate the organismal phylogeny. Similarly, several methods were developed to filter out paralogous gene copies. These efforts resulted in lists of primarily "single-copy" genes, databases of orthologous sequences (e.g., OrthoDB) and software (e.g., Orthograph).

An alternative approach to the gene tree/species tree problem is to incorporate various gene-level processes into the inference of species phylogeny. Thus models were created to reconstruct gene trees when species trees are known, and species trees when gene trees are known. However, because of the computational complexity only a few studies have attempted to jointly infer gene trees and species trees and such analyses are usually limited to a few genes and few taxa (see Szollosi et al. 2014).

While models incorporating gene duplication and horizontal gene transfer remain relatively uncommon in phylogenetic analysis, those dealing with ILS have been very popular in the last 10 years.

2.1 Incorporating ILS into phylogenetic inference

While GDL and HGT can be regarded as somewhat exceptions to the "normal" pattern of gene inheritance, ILS is clearly a part of this process. Our understanding of the behavior and ultimate fate of multiple gene lineages in a population is based on the **coalescent theory**, which models genealogies within population. This theory can be applied to phylogenetic analysis, by representing phylogenies as multiple populations connected by an evolutionary tree ("multispecies coalescent" model). The multispecies coalescent can be used to describe a probability distribution of random

gene trees that evolve along the branches of a species tree.

Based on the coalescence theory, if a branch is shorter than 5 coalescence time units (T/Ne, where T is the number of generations and Ne is the effective number of chromosomes), multiple gene lineages are able to persist between the two nodes it connects. Indeed, 11.4% of the genes in human genome are more closely related to those in gorilla, despite the fact that as a species we are more closely related to chimps (see this blog for a nice graphical explanation).

We can apply the multispecies coalescent to calculate the probability of a given gene tree within a species tree. We can also attempt to co-estimate species tree and gene tree. Unfortunately, the latter approach is very computationally intensive and so far has been limited to a relatively few species and few genes. By contrast, "shortcut" coalescence methods are readily available for a large number of genes and taxa. These methods first estimate individual gene- or even c-gene trees. Such estimates are highly problematic because they are based on very limited sequence data and are subject to multiple biases.