Molecular evolution and natural selection

All the events of biological evolution are played out somewhere along the branches of phylogenetic trees. As the result, phylogenetic trees preserve traces of the historical evolutionary events that gave rise to the diversity of contemporary species. The combination of phylogeny and information on species can be used to infer what the past was like and how the present came about. This is the field of **molecular evolution**.

The neutral theory of molecular evolution

The 2 most striking claims in molecular evolution:

1. DNA sequence change occurs with little influence from natural selection (the Neutral Theory of Molecular Evolution).
2. The rate of molecular evolution (DNA sequence change) is approximately constant (there is a molecular "clock").

The neutral theory of molecular evolution was proposed independently by M. Kimura and J. King & T. Jukes.

Claim: The large majority of observed molecular polymorphisms reflect neutral changes. Likewise, most substitutions observed between homologous genes are selectively neutral.
Implications: Gene (protein) families evolve through neutral mutations and purifying selection. Most genes (proteins) have not been improved during the period of metazoan evolution.

Motoo Kimura – my hero in biology. Before (and up to this day) people believed that changes in genes / proteins occur for adaptive reasons.

Neutral theory has been embraced, either explicitly or implicitly, by almost all molecular biologists.

It does not mean that there is no selection. Quite to the contrary it allows us to detect selection:
Selectively neutral changes should be fixed at the rate at which they occur due to mutation.
However, deleterious mutations will not be fixed most of the time, while advantageous mutations will be fixed rapidly.
If we consider a protein coding sequence, the synonymous positions are likely to be neutral, so they provide a "baseline" rate for comparison.

There are several interesting expectations based on the theory:

- the probability that a neutral mutation is lost after just one generation is closely approximated by $e^{-1} \approx 0.37$
- the probability of ultimate fixation of a neutral allele is 1/2N
- the long-term rate of neutral evolution is equal to the genic mutation rate μ
- the average time to fixation for a neutral mutation is 4Ne.
- the probability that a selectively advantageous mutation will be lost after one generation is $(1\text{-}s)\ e^{-1}$ or ≈ 0.33 for s=0.1.
- the probability of fixation of a beneficial allele with additive effects is less than twice its selective advantage in the heterozygous state.
- For a given population size, the time to fixation of a beneficial mutation is identical to the time to fixation of a deleterious allele with the same selection coefficient (of opposite sign). For |4Nes|>>1, the mean time to fixation is about (4/|s|)ln(2Ne) generations.
- Finally, the rate of fixation of beneficial mutations is $(2sNe/N)x2N\mu_b = 4Ne\mu_b s$

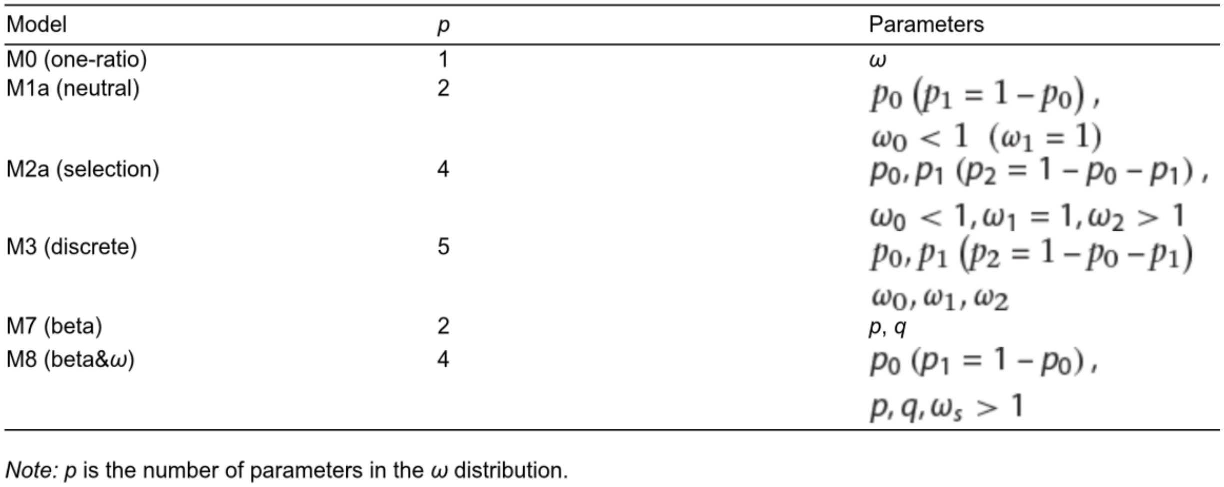
Tests for selection:

One way to get insights into the mechanism of molecular evolution is by analyzing synonymous and nonsynonymous substitutions (dN/dS). If the synonymous rate is used as a benchmark, one can infer whether the fixation of nonsynonymous mutations is aided or hindered by natural selection that acts on the protein. The nonsynonymous/synonymous rate ratio, $\omega = d_N/d_S$, measures selective pressure at the protein level.

Early studies using the d_N/d_S criterion conducted pairwise sequence comparison, averaging the substitution rates over all codons and over the whole time period separating the two sequences. Such approach rarely detects positive selection. Indeed, only 35/8079 genes with $\omega>1$ were found in a human - chimpanzee genome-wide comparison. Later efforts have focused on detecting positive selection affecting particular lineages on the phylogeny or individual sites in the protein.

Such analysis can be made more rigorous by taking a likelihood approach under a model of codon substitutions, analyzing all sequences jointly on a phylogenetic tree. The big idea is to use ω as a part of the model and allow for different levels of heterogeneity in ω among lineages. **Branch models** allow various ω for different branches. The simplest model would assume the same ω for all branches. The most general model assumes an independent ω for each branch in the phylogeny (too many parameters for large trees). These models can be compared using LRT to examine interesting hypotheses. **Note** that the lineages of interest should be identified a priory. Also, although variation in the ω ratio among lineages is a violation of the strictly neutral moel, it is not by itself a sufficient evidence of adaptive evolution. The criterion that the ω ratio, averaged over all sites in the sequence, should be greater than 1 is very stringent; as a result, the branch test often has little power to detect positive selection.

Site models allow various ω for different codons in the alignment. In principle, the sites can be pre-defined based on some previous knowledge one ω can be estimated for every site. However, a better strategy is to use a statistical prior distribution to describe the random variation of ω over sites. The null hypothesis of no positive selection can be tested using an LRT comparing two statistical distributions, one of which assumes no sites with ω > 1 while the other assumes the presence of such sites. When the LRT suggests the presence of sites with ω > 1, an EB approach is used to calculate the conditional probability distribution of ω for each site given the data at the site. Two pairs of models are commonly used (see table below).



Model	*p*	Parameters
M0 (one-ratio)	1	ω
M1a (neutral)	2	$p_0\ (p_1 = 1 - p_0)$, $\omega_0 < 1\ (\omega_1 = 1)$
M2a (selection)	4	$p_0, p_1\ (p_2 = 1 - p_0 - p_1)$, $\omega_0 < 1, \omega_1 = 1, \omega_2 > 1$
M3 (discrete)	5	$p_0, p_1\ (p_2 = 1 - p_0 - p_1)$ $\omega_0, \omega_1, \omega_2$
M7 (beta)	2	p, q
M8 (beta&*ω*)	4	$p_0\ (p_1 = 1 - p_0)$, $p, q, \omega_s > 1$

Note: p is the number of parameters in the *ω* distribution.

The first pair involves the null model M1a (neutral), which assumes two site classes with frequencies p0 and p1 and 0<ω0<1 and ω1=1. The alternative model M2a adds another class of sites with ω2 > 1. The second pair of models consists of the null model M7(beta), which assumes a beta distribution for ω (0 < ω <1). The alternative model, M8 (beta&ω), adds another class of sites with ωs > 1 for positive selection. Both alternative models add two free parameters and can be tested using χ^2 distribution with 2df. The tests appear conservative in both cases. The models discussed above assume that the synonymous rate is constant across all sites while non-synonymous rate varies.