8. Molecular evolution and natural selection

EEOB563

Spring 2025

> "Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones". Kimura, 1968.

1 Preliminary considerations

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.

Four areas of molecular evolution have particularly benefited from advances in phylogenetics and will be considered in more details in this class:

- predicting adaptive evolution;
- reconstruction of ancestral characters states;
- estimation of timing of evolutionary events;
- comparative studies

2 The Neutral Theory of Molecular Evolution

The two most striking claims in molecular evolution:

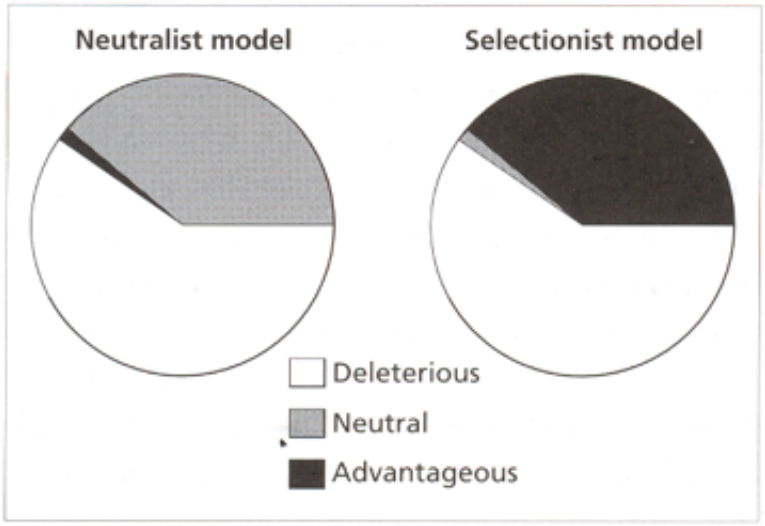
1. DNA sequence change occurs with little influence from natural selection (the Neutral Theory of Molecular Evolution).
2. The rate of molecular evolution (DNA sequence change) is approximately constant (molecular "clock").

The neutral theory of molecular evolution was proposed independently by M. Kimura in 1968 and by J. King & T. Jukes in 1969, but is usually associated with M. Kimura because of his tireless championing of the theory and his many subsequent contributions. The chief tenet of the theory is that most genetic differences between species and polymorphisms within species are selectively neutral. This view contrasted sharply with previously held notion that most of these variations were adaptive and led to a lengthy and generally productive neutralist – selectionist debate. Oxford Bibliographies provide an excellent summary of the topic.

Figure 1: Motoo Kimura

Motoo Kimura – my hero in biology. Before coming to ISU, I was told by one of my professors that Motoo Kimura studied there. I was very excited, but could not find any reference about his time at ISU. Eventually, I found the following quote: "After nine unsatisfying months at Iowa State, he [Motoo Kimura] joined James Crow's laboratory at the U. of Wisconsin, from which he received his Ph.D. in 1956". Here is an interesting piece of history from Daniel Hartl, who also did his Ph.D. with Dr. Crow: "In the 1960s Crow's laboratory was a crossroads of evolutionary genetics with many visitors for weeks, months, or years. Motoo Kimura visited one summer. Slightly built, softspoken, and formal, he would enter the lab each morning at 8 AM dressed in a three-piece suit, sharpen a dozen #2 yellow wooden pencils with a mechanical sharpener affixed to the wall, and disappear into his office until precisely noon when he would emerge to meet Professor Crow for lunch, return at around 1 PM, resharpen the pencils, and disappear again until precisely 5 PM. His work habits denoted discipline, not unfriendliness. He was always glad to answer questions or chat if you caught him between the pencil sharpener and his office."

Neutral theory has been embraced, either explicitly or implicitly, by almost all molecular biologists in a sense that we equate faster evolving regions of genes and proteins with relaxed functional constrains. Note, that the Neutral Theory does not imply the absence of selection! Quite to the contrary, it allows one 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 more frequently. 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 $\frac{1}{2N}$;
- the long-term rate of neutral evolution is equal to the genic mutation rate;
- 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 $(1-s)e^{-1}$ 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).
- Finally, the rate of fixation of beneficial mutations is $4N_e\mu_b s$

3 Tests for selection

The theory of natural selection is based on the simple observation that fitness-enhancing traits increase in prevalence in the population over time. Most of the organisms seem to be well "adapted" to their environment. However, finding genes responsible for these adaptations has been surprisingly difficult. In many cases correlations between genomic and phenotypic traits have been presented as causations, leading to what is known as "just-so stories". Neutral theory of molecular evolution, while being continuously challenged, provides a null hypothesis for rigorous testing of evolutionary forces. Many tests of natural selection have been developed. They are generally classified into three groups, depending on whether the genetic variation considered is *divergence*, *polymorphism*, or *both*. Because this is a phylogenetics class, we will focus on the methods that use divergence data (*e.g.*, differences among species).

Selection, but what selection? There are many specific modes of selection that have been described, some of which share conceptual overlap, and some of which are referred to by multiple names. In molecular evolution we usually talk about **positive** and **negative** selection, meaning that selection acts in a directional manner either favoring or disfavoring an allele. Negative selection is also known as **purifying** selection. The removal of novel (deleterious?) alleles from the gene pool before they can achieve detectable frequency within the population is known as **background** selection. When the strength and direction of selection depends on the presence and interaction of multiple alleles at the same locus, organisms can experience **balancing** or **diversifying** selection. Most of the genetic and genomic methods have focused on detecting positive selection, which leaves a more conspicuous footprint on the genome and is also believed to be the primary mechanism of adaptation.

3.1 Tests for selection using divergence data

The standard method of testing the neutral hypothesis by using divergence data compares the number of synonymous substitutions per synonymous site (d_N) with the corresponding number of nonsynonymous substitutions per nonsynonymous site (d_S). Under the null hypothesis of neutrality, $d_N = d_S$. Using the synonymous rate as a benchmark, one can infer whether the fixation of nonsynonymous mutations is aided or hindered by natural selection that acts on the protein by calculating the nonsynonymous/synonymous rate ratio, $\omega = d_N/d_S$ (also referred to as Ka/Ks).

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 that includes ω, analyzing several sequences jointly on a phylogenetic tree (Fig. 2).

$$\mathbf{Q} = \{q_{ij}\} = \begin{cases} \omega\kappa\pi_j & : \text{ nonsynonymous transition} \\ \omega\pi_j & : \text{ nonsynonymous transversion} \\ \kappa\pi_j & : \text{ synonymous transition} \\ \pi_j & : \text{ synonymous transversion} \\ 0 & : \ i \text{ and } j \text{ differ at more than one position,} \end{cases} \quad \textbf{[1]}$$

Figure 2: The Goldman and Yang (1994) model

Furthermore, models can include more than one ω value.

Branch models allow various ω for sets of 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. It is important to note that although variation in ω values among lineages is a violation of the strictly neutral model, it is not a sufficient evidence of adaptive evolution. The criterion that the 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 for different codons in the alignment. In principle, the sites can be predefined based on some previous knowledge one can be estimated for every site. However, a better strategy is to use a statistical prior distribution to describe the random variation of 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 $\omega > 1$ while the other assumes the presence of such sites. When the LRT suggests the presence of sites with $\omega > 1$, an EB approach is used to calculate the conditional probability distribution of 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 p_0 and p_1 and $0 < \omega_0 < 1$ and $\omega_1 = 1$. The alternative model M2a adds another class of sites with $\omega_2 > 1$. The second pair of models consists of the null model M7(beta), which assumes a beta distribution for ω $(0 < \omega < 1)$. The alternative model, M8 (beta&ω), adds another class of sites with $\omega_s > 1$ 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.

3.1.1 Other rate-based tests

Some studies have used comparative genomic data to identify positive selection by a significantly accelerated rate of substitution in a particular species of lineage. These relative-rate approaches are applicable to coding as well as non-coding sequences and have been used to identify genomic regions "that make us humans". Because high rate of sequence evolution can reflect both increase in positive selection and decrease in negative selection, population-level tests should be utilized to test such inferences. In particular, the Hudson-Kreitman-Aguadé (HKA) test can be used to examine the ratios of fixed interspecific differences i.e., substitutions (D) to within-species polymorphisms (P) across loci. For a neutral site, both D and P should be functions of the site's mutation rate and multilocus comparisons can be used to derive the expected neutral D/P ratio for a lineage.