Testing the Neutral Theory

1. The Neutral Theory of Molecular Evolution

The neutral theory was first proposed by Kimura in the late 1960s and was inspired by the observation of the protein molecular clock and the abundance of protein variation within natural populations. The theory was further developed and explained in Kimura's1983 book.

What is the neutral theory? The theory states that the molecular polymorphism observed in natural populations is due to neutral mutations under mutation-drift equilibrium. Molecular divergence between species is an extension of this process, caused by neutral mutations going to fixation in one or the other species.

What is it not? The theory does not state that all *possible* mutations are neutral, only those that are observed. There may be negative (purifying) selection against mutations that have a deleterious effect on an organism's fitness. This can explain why different proteins evolve at different rates and why silent substitutions occur more frequently than amino acid replacements.

The theory states that *positive* selection need not be invoked to explain molecular polymorphism and divergence. However, Kimura spent an entire chapter of his book discussing the discordance between morphological and molecular evolution. Morphological evolution provides ample evidence for adaptation. How can this be reconciled? Perhaps adaptive mutations do occur, but they are rare and so greatly outnumbered by neutral mutations that they have little or no detectable effect on polymorphism or divergence.

Two key predictions of the neutral theory:

- Divergence between species, $K = \mu$, where μ is the neutral mutation rate
- Heterozygosity within species, $\theta = 4Ne\mu$, where *Ne* is effective population size.

2. Tests of the Neutral Theory

a) Tajima's *D* - tests the frequency spectrum of segregating sites. Requires only intraspecific polymorphism data. Two estimates of θ; θ_w and π should be equal under the neutral theory.

$\theta_w = S / a_1$, where *S* is the number of segregating sites, $a_1 = (1 + 1/2 + 1/3 + \ldots + 1/ n\text{-}1)$, and *n* = number of sequences

$\pi = \sum_{ij} x_i x_j \pi_{ij}$, where x_i and x_j are the frequencies of the *i*th and jth type of DNA sequence and π_{ij} is the proportion of pairwise differences between sequence *i* and *j*.

The key difference is that θ_w depends only on the number of segregating sites, while π also depends on their frequencies. Tajima's *D* is calculated as

$$D = (\pi - \theta_w) / \mathrm{SD}(\pi - \theta_w)$$

where the demoninator is the standard deviation of the difference between π and θ_w. and is always a positive number. Thus, the sign of D is determined by the numerator and there are three possible outcomes:

$D = 0$: neutral expectation
$D < 0$: excess of low frequency variants. This could have a number of causes, for example purifying selection keeping deleterious mutations at low frequency, expanding population (*e.g.* after a bottleneck), recovery of variation after a selective sweep.
$D > 0$: excess of intermediate frequency variants. Possible causes included balancing selection and population admixture.

b) HKA test (Hudson, Kreitman, and Aguade) – tests for differences in the ratio of polymorphism to divergence between two different loci. Requires both polymorphism and divergence data for the two loci.

Under neutrality, $K = \mu$ and $\theta = 4Ne\mu$, thus $\theta = 4NeK$ and θ and K should be directly proportional. So for two loci, $\theta_1 / K_1 = \theta_2 / K_2$.

Example: Comparison of two loci in *D. melanogaster* and *D. sechellia*.
Locus 1 (*Adh* 5' flanking region), $\theta_1 = 0.0066$, $K_1 = 0.052$, $\theta_1 / K_1 = 0.127$
Locus 2 (*Adh*, silent only), $\theta_2 = 0.0090$, $K_2 = 0.056$, $\theta_2 / K_2 = 0.161$

The significance of the test can be determined from the chi-squared distribution with 1 degree of freedom. In the above case, $X^2 = 6.07$, $P = 0.016$.

Because you are testing ratios from two loci, it is often difficult to determine the cause of a departure from neutrality. One possibility is that selection has acted recently to reduce polymorphism at one of the loci (see selective sweeps below).

c) MK test (McDonald and Kreitman) - tests for differences in the ratio of polymorphism to divergence at synonymous and nonsynonymous sites. Requires both polymorphism and divergence data for a protein-encoding sequence.

This can be tested using a 2 x 2 contingency table (sometimes called a DPRS table):

Example: *G6pd* (Glucose-6-Phosphate Dehydrogenase); *D. melanogaster* vs. *D. simulans*

	Divergence	Polymorphism	
Synonymous	26	21	The table contains the counts of div. and poly.
Replacement	36	2	sites of each type.

$X^2 = 16.5$, $P = 0.00003$

The assumption is that synonymous sites evolve neutrally. Departures from neutrality could be caused by positive selection for amino acid changes (when there is a relative excess of nonsynonymous divergence) or balancing selection (when there is a relative excess of nonsynonymous polymorphism). The latter could also be a result of weak purifying selection, where slightly-deleterious nonsynonymous mutations are present as polymorphisms in a population, but do not contribute to the fixed differences between species.

d) Haplotype tests – test the distribution of segregating sites among alleles. Requires only intraspecific polymorphism data.

- Haplotype number test - tests the total number of haplotypes in a sample
- Hudson's haplotype test - tests for subsamples with unusually low polymorphism

The word "haplotype" refers to the combination of nucleotide sites present on the same chromosome. If all sequences are identical, there is only one haplotype. If each sequence differs from every other sequence, then the number of haplotypes equals the sample size. For example, consider the following alignment of eight sequences. The asterisks (*) at the bottom indicate sites where there is polymorphism:

```
Seq1  AACTGTGCACTGCATGATGA
Seq2  AACTGTGCACTGCATGATGA
Seq3  AAGTGTGCACTGCCTGATGA
Seq4  AAGTGTGCACTGCCTGATGA
Seq5  AACTGTGCACTGCATGATGA
Seq6  AACTGTGCACTGCATGATGA
Seq7  AACTGTGCACTGCATGCTGA
Seq8  AACTGTGCACTGCATGATGA
         *           *   *
```

Here there are a total of three haplotypes: Sequences 1, 2, 5, 6, and 8 are identical and represent one haplotype. Sequences 3 and 4 are identical and represent a second haplotype. Sequence 7 is unique and represents a third haplotype.

For both of the above tests, the probability of the observed data is determined by computer simulations.

Departures from the neutral expectation could be caused by balancing selection, ongoing positive selection, recent population expansion, or population admixture.

3. Genetic Hitchhiking and Selective Sweeps

It is important to note that the neutral theory predictions for polymorphism and divergence depend only on the effective population size and the mutation rate. The recombination rate does not enter into the equations. However, experimental results indicate that there is a correlation between polymorphism and rate of recombination. In particular, regions of the genome experiencing very low (or no) recombination, show very low levels of polymorphism relative to regions experiencing normal or high levels of recombination. Divergence between species is not affected by recombination. This decoupling of polymorphism and divergence often leads to rejection of the neutral theory in regions of low recombination by the HKA test.

This deviation from neutrality can be explained by "Genetic Hitchhiking". Under this model, a positively selected mutation goes to fixation, bringing all linked neutral mutations to fixation with it. Since linkage is stronger in regions of reduced recombination, the effect of reduced polymorphism is strongest in these regions. This is often referred to as a "selective sweep".

4. Background Selection

An alternative explanation for reduced polymorphism in regions of low recombination is "Background Selection". This model also relies on the "hitchhiking" of neutral polymorphisms, except here they are linked to deleterious mutations that are removed from the population by negative (purifying) selection. This has the effect of reducing the effective population size in regions of low recombination relative to those with normal recombination. Since $\theta = 4Ne\mu$, a reduction in *Ne* leads to a reduction in polymorphism. Thus, background selection is compatible with the neutral theory. It is typically very difficult to distinguish between selective sweeps and background selection experimentally and this is currently a topic of much research and debate in molecular evolution and population genetics. Note that the two mechanisms are not mutually exclusive, so both may be acting simultaneously.

5. More about Haplotypes and Selective Sweeps

Because selective sweeps remove variation in a region of a chromosome, they lead to a reduction in the number of haplotypes. If selection is strong, there is little chance for recombination or new mutations to occur on the selected chromosome and a single haplotype will extend over a large region of the chromosome. Furthermore, the selected haplotype should be in high frequency. These features of selective sweeps have inspired approaches for scanning entire genomes to find regions that have experienced recent positive selection.