Lecture 11: Divergence-based tests: I. HKA and MK tests, codon models

UNE course:

The search for selection

3 -- 7 Feb 2020

Bruce Walsh (University of Arizona)

jbwalsh@email.arizona.edu

Divergence-based tests

- **Population-based divergence tests**. Contrast the levels of polymorphism within a reference population with the level of divergence between populations or species
  - different classes of sites within the same gene (the **McDonald-Kreitman**, or **MK**, **test**)
  - different genes (the **Hudson-Kreitman-Aguadé**, or **HKA**, test)
- **Phylogeny-based divergence tests**. Contrasts the rates of evolution at different sites within a gene over a number of species in a phylogenetic context.
  - K_A to K_S ratios
  - Codon models

A History of Selection Alters the Ratio of Polymorphic to Divergent Sites

Population-based tests contrast the patterns of within-species polymorphism and between-species divergence to see if they are in concordance with their neutral expectations. Under the equilibrium neutral model, two standard measures of polymorphism under the infinite-sites model are functions of $4N_e\mu$ (where μ is the per-site mutation rate): the nucleotide diversity, π, and the number of segregating sites, S. These have expected values of $E[\pi] = 4N_e\mu$ and $E[S] = 4N_e\mu\, a_n$, where a_n is a constant that depends only on the sample size, n (Equation 9.21a). Under the assumptions of the equilibrium neutral model, the relationship between polymorphism (measured by nucleotide diversity, π) and the between-population divergence (D) for the ith gene being considered is

$$\pi_i = 4N_e\mu_i, \qquad D_i = 2t\mu_i \tag{10.1a}$$

where N_e is the effective population size, and t is the divergence time in generations. Hence,

$$\frac{\pi_i}{D_i} = \frac{4N_e\mu_i}{2t\mu_i} = \frac{2N_e}{t} \tag{10.1b}$$

Because the gene-specific mutation rates cancel, under the equilibrium neutral model, the π/D ratio at all loci should be roughly the same, namely $2N_e/t$ (subject to random sampling). When polymorphism is instead scored as the number of segregating sites, S, then

$$\frac{S_i}{D_i} = \frac{2N_e a_n}{t} \tag{10.1c}$$

Example 10.1. McDonald and Kreitman (1991a) examined the *Adh* (alcohol dehydrogenase) locus in the sibling species *Drosophila melanogaster* and *D. simulans*. Within this gene, they contrasted **replacement (nonsynonymous)** and **silent (synonymous)** sites. At the DNA level, a replacement-site mutation results in an amino acid change, while a silent-site mutation still codes for the ancestral amino acid. Equation 10.1c indicates that, under neutrality, the ratio of the number of segregating sites to the number of fixed differences should be the same for both categories of sites. This results in a simple association test, and significance can be assessed using either a χ^2 approximation or the (much better) Fisher's exact test, which accommodates small numbers in the observed table entries. Of the 24 fixed differences between the two species seen by McDonald and Kreitman, 7 were replacement-site mutations and 17 were silent-site mutations. The total number of polymorphic sites segregating in either species was 44, 2 of which were replacement and 42 of which were silent. The resulting association table becomes

	Fixed	Polymorphic
Silent	17	42
Replacement	7	2

Fisher's exact test gives a p value of 0.0073, indicating a highly significant lack of fit to the neutral equilibrium model. Based on the ratio of 42:2 silent/replacement polymorphisms, the expected number, x, of replacement fixations is $17/x = 42/2$, or $x = 0.81$, i.e., $\sim$1 replacement polymorphism is expected under neutrality. Because 7 were seen, this suggests roughly 6 adaptive substitutions, or that 86% (6/7) of the *Adh* amino acid substitutions between these species are adaptive.

A History of Positive Selection Alters the Ratio of Silent- to Replacement-site Substitution Rates

Phylogeny-based divergence tests do not require polymorphism data, but rather simply contrast the divergence rates at silent versus replacement sites. Silent sites are treated as proxies for neutral sites, although we have seen that they may be under (at least) weak selection (Chapter 8). Mutations at replacement sites are generally viewed as being under much stronger selection, most of it purifying. The primary evidence that such *negative* selection (removal of new deleterious mutations) is widespread is the observation that silent-site substitution rates are almost always much higher than those for replacement sites, when averaged over an entire gene. This pattern is expected if a higher fraction of mutations in replacement sites is deleterious relative to that in silent sites. However, there are cases where, for a limited region within a gene, the replacement-site substitution rate exceeds that for silent sites, suggesting the presence of adaptive fixation (i.e., positive selection).

While there are several variant notations in the literature, we use K_s to denote the per-site silent substitution rate and K_a to denote the per-site replacement rate between taxa (the subscript *a* indicating a change in an amino acid); K_{ns} and K_n are also used in the literature to denote replacement-site (i.e., nonsynonymous) substitution rates. A value of $K_a/K_s > 1$ indicates a long-term pattern of positive selection at replacement sites. As Example 10.2 illustrates, even if this is occurring at *specific regions* within a gene, when averaged over an *entire* gene, K_a/K_s is usually < 1. Thus, while an observation of $K_a/K_s > 1$ is almost universally accepted as a signature of a long-term pattern of multiple episodes of positive selection, such inflation is almost never seen if the entire gene is taken as the unit of analysis. Phylogeny-based methods (examined below) accommodate this concern by taking the codon as the unit of analysis, first placing genes within a phylogeny and then using codon-evolution models to test whether $K_a/K_s > 1$ for some subset of codons.

Austin Hughes

Masatoshi Nei

Example 10.2. One of the classic examples of using sequence data to detect signatures of positive selection is the work of Hughes and Nei (1988, 1989). They examined the major histocompatibility complex (MHC) Class I and Class II loci of mice and humans, highly polymorphic genes involved in antigen recognition. A large number of prior studies on other genes had found that an excess of silent substitutions is almost always the norm, implying that most replacement changes are selected against. Indeed, when one looks over an entire Class I (or II) MHC gene, this pattern is also seen. The insight of Hughes and Nei was to use data on protein structure to specifically focus on the putative antigen-binding site and to compare this region with the rest of the gene as an internal control.

Hughes and Nei compared the ratio of silent- to replacement-site nucleotide substitution rates in the putative antigen recognition sites versus the rest of the gene. For both Class I and Class II loci, they found a significant excess of replacement substitutions in the recognition sites and a significant deficiency of such substitutions elsewhere. If both types of substitutions were neutral, the per-site rates should be roughly equal. If negative selection is acting, the expectation is that the silent-site substitution rate would be significantly higher (reflecting the removal of deleterious replacement mutations). However, if positive selection is sufficiently common among new mutations, one expects to find an excess of replacement substitutions. The observed patterns for both Class I and II loci were consistent with positive selection within the part of the gene coding for the antigen recognition site and purifying selection on the rest of the gene.

Divergence-based Tests are Biased Toward Conservative Sites

A major (but subtle) distinction between most methods in this chapter and those in Chapter 9 are that the latter usually have very little restrictions on the kinds of sequences being scanned for selection. In contrast, most divergence-based tests were built (at least initially) around analyses of protein-coding sequences (HKA is an exception), such as contrasts between silent and replacement sites or the substitution patterns at a codon (or set of codons) over a phylogeny. In such settings, these methods focus almost exclusively on detecting *structural* adaptations, namely, adaptive changes in the amino acid sequence. As we saw in Chapter 9, regulatory changes are thought to be at least as important as structural changes for short-term adaptation.

One reason for the focus on protein-coding regions in divergence-based tests is that one must be able to ***align homologous sequences***. Because they accept relatively few insertion or deletion mutations, long open-reading frames allow one to align homologous coding sequences, even over fairly substantial periods of evolutionary time. By contrast, this is often *not* the case for regulatory sequences, especially when considering that we still have a limited (albeit improving) ability to detect the full universe of such sequences. As shown in several examples below, divergence-based approaches have been applied to *highly conserved* regulatory regions, which offer a better opportunity for comparing homologous sequences over evolutionary time. However, this also biases these tests toward regions under strong functional constraints. Thus, the very interesting question of whether structural changes may be more important than regulatory changes for long-term adaptation cannot be fully addressed by divergence-based data alone, as these have a bias toward detection in highly conserved regions, whether structural or regulatory. Extensive regulatory changes in less-conserved regions may be entirely missed by most divergence-based tests. Despite these isses, there are hints starting to emerge of at least as many adaptive substitutions in noncoding regions as there are in coding regions (as we detail below).

What Fraction of the Genome is Under Functional Constraints?

The amount of metazoan DNA that codes for proteins and structural RNAs (the so-called **coding DNA**) is usually just a fraction of their total genome. The role of the remaining (and usually majority) component of the genome, the **noncoding DNA**, has been the subject of numerous debates as to its evolutionary role and function. This raises a central question of just what fraction of the genome is under some sort of functional constraint (and therefore, selection). Chiaromonte et al. (2003) denoted this fraction by α_{sel}, which is somewhat unfortunate notation given the widespread use of α for the fraction of *adaptive* substitutions (to be covered in detail shortly). One obvious approach for estimating α_{sel} is from the amount shared conserved sequences between two divergent taxa. For example, early studies searched for regions first shared between mice, humans, and dogs, and later over a wider range of mammals, arriving at the result that around 6% of the human genome is conserved over such time scales (Lindblad-Toh et al. 2005, 2011). This is six-fold more than the 1% of the human genome that codes for proteins (~33 MB out of a total of ~3100 MB; Church et al. 2009). Andolfatto (2005) estimated a much higher value of α_{sel}, between 40% and 70%, for *Drosophila melanogaster*, with about twice as many constrained sites in noncoding, as opposed to coding, regions. Such comparisons, especially when based on widely-divergent taxa, are simply *lower bounds*, as sequences under functional constraints can still turnover through time, escaping detection (Dermitzakis and Clark 2002). Indeed, Pheasant and Mattick (2007) suggested that the functional portion of the human genome may exceed 20%, basing their argument on the fact that rapidly evolving regions will not be detected through sequence conservation studies.

Further insight into α_{sel} can be gained by examining how the amount of conserved sequences shared between species pairs changes with their divergence times. This approach was used by Meader et al. (2010), who found that the fraction of shared conserved sequences among mammals decreased over time, and used the rate of this decrease to estimate that between 200 and 300 MB (6.5% to 10%) of the human genome is under functional constraints. A more refined estimate arrived at a value of around 8% (Rands et al. 2014). Hence, roughly 88% (7/8) of human constrained sites are found in noncoding regions. Meader et al. also used their approach on *Drosophila melanogaster*, finding an α_{sel} value of between 47% and 55%. Given around 22 MB for coding DNA and their estimate of 35–45 MB of constrained noncoding DNA, roughly two-thirds of the constrained sites are in noncoding regions.

These estimates of the amount of constrained noncoding DNA raise a number of important evolutionary questions (beyond the obvious one of their functional role). How strong is selection in noncoding regions? How often do adaptive mutations arise from these noncoding regions? What fraction of segregating deleterious mutations are attributable to these regions? While unbiased answer to these questions remain elusive, preliminary estimates based on conserved noncoding regions and on transcription factor binding sites suggest that noncoding DNA is likely a rich source of adaptive substitutions.

HKA test

- Hudson, Kreitman, and Aguadé (1987) proposed the first approach to jointly use polymorphism and divergence data.
  - Their HKA test
- Unlike many of the other divergence-based tests, HKA can be applied to any type of sequence data (not just a contrast between replacement and silent sites).

Dick Hudson

Marty Kreitman

Montserrat **Aguadé**

The Hudson-Kreitman-Aguadé (HKA) Test

Hudson, Kreitman, and Aguadé (1987) proposed the first approach to jointly use polymorphism and divergence data. Unlike many of the other divergence-based tests, their's can be applied to any type of sequence data (not just a contrast between replacement and silent sites). Their **HKA test** is formulated as follows. Consider two species (or very distantly related populations) A and B, which are both at mutation-drift equilibrium with effective population sizes of $N_A = N_e$ and $N_B = \delta N_e$. Further assume that they separated $\tau = t/(2N_e)$ generations ago from a common population of size $N_e^* = (N_A + N_B)/2 = N_e(1+\delta)/2$, the average of the two current population sizes. Suppose $i = 1, \cdots, L$ unlinked loci are examined in both species. We allow the neutral mutation rate, μ_i, to vary over loci, but assume (for a given locus) that it has been the same in both species, and hence unchanged during divergence. The expected number of neutral segregating sites at locus i is a function of $\theta_i = 4N_e\mu_i$ in species A, and $4N_B\mu_i = 4(\delta N_e)\mu_i = \delta\theta_i$ in species B. The expected divergence between A and B is $2t\mu_i$, which we can express as

$$2t\mu_i = 2\,\frac{t}{2N_e}\,2N_e\mu_i = \tau\theta_i$$

$$\widehat{E}[S_i^A] = \widehat{\theta}_i a_{n_A}, \quad \widehat{E}[S_i^B] = \widehat{\delta}\,\widehat{\theta}_i a_{n_B}, \quad \text{where} \quad a_{n_x} = \sum_{i=1}^{n_x-1} \frac{1}{i} \tag{10.3a}$$

$$\mathrm{Var}(S_i^A) = \widehat{\theta}_i a_{n_A} + \widehat{\theta}_i^{\,2} b_{n_A}, \quad \mathrm{Var}(S_i^B) = \widehat{\delta}\,\widehat{\theta}_i a_{n_A} + \widehat{\delta}^2\,\widehat{\theta}_i^{\,2} b_{n_B}, \quad b_{n_x} = \sum_{i=1}^{n_x-1} \frac{1}{i^2} \tag{10.3b}$$

$$\widehat{E}[D_i] = \widehat{\theta}_i \left(\widehat{\tau} + \frac{1+\widehat{\delta}}{2} \right) \tag{10.3c}$$

$$\mathrm{Var}(D_i) = \widehat{\theta}_i \left(\widehat{\tau} + \frac{1+\widehat{\delta}}{2} \right) + \left(\frac{\widehat{\theta}_i(1+\widehat{\delta})}{2} \right)^2 \tag{10.3d}$$

Equations 10.3a and 10.3b follow from the infinite-sites model (Equations 4.3a and 4.4a, respectively). Equation 10.3c follows if we rewrite

$$\theta_i \left(\tau + \frac{1+\delta}{2} \right) = 4N_e\mu_i \left(\frac{t}{2N_e} + \frac{1+\delta}{2} \right) = 2\mu_i t + 4\mu_i \frac{N_e(1+\delta)}{2} = 2\mu_i t + 4N_e^*\mu_i$$

More formally, the HKA test statistic, X^2, is given by

$$X^2 = \sum_{i=1}^{L} X_i^2 \tag{10.2a}$$

where

$$X_i^2 = \frac{\left(S_i^A - \widehat{E}[S_i^A]\right)^2}{\mathrm{Var}(S_i^A)} + \frac{\left(S_i^B - \widehat{E}[S_i^B]\right)^2}{\mathrm{Var}(S_i^B)} + \frac{\left(D_i - \widehat{E}[D_i]\right)^2}{\mathrm{Var}(D_i)} \tag{10.2b}$$

The HKA test statistic is approximately chi-square-distributed with $3L - (L + 2) = 2L - 2$ degrees of freedom, given the $3L$ observations and $L+2$ parameters ($\theta_1 \ldots \theta_L$, τ, δ) to estimate

Example 10.3. Hudson et al. (1987) partitioned the *Adh* gene into two regions, silent sites and 4-kb of the 5′ flanking region, corresponding to a test using $L = 2$ loci. (The careful reader might be concerned that these loci are linked, while the HKA test assumes independence across loci. The high recombination rates in *Drosophila* result in LD generally being over only very small distances.) A sample of 81 *Drosophila melanogaster* alleles was examined, along with a single allele from its sibling species *D. sechellia*. Based on sequencing data, the divergence was 210 differences in the 4052-bp flanking region and 18 differences in the 324 silent sites, amounting to roughly equal levels of divergence per base pair between the two "loci." Based on restriction-enzyme data, within *melanogaster*, 9 of the 414 5′ flanking sites were variable, while 8 of 79 *Adh* silent sites were variable. Thus, while the per-site divergence was roughly equal, there was a four-fold greater polymorphism level at silent sites.

The test static value as 6.09. Because $\Pr(\chi_1^2 > 6.09) = 0.014$, the test indicates a significant departure from the equilibrium neutral model.

Example 10.4. Ingvarsson (2004) examined chloroplast (cpDNA) diversity in two plants in the genus *Silene* (family Caryophyllaceae). A standard HKA test contrasting four noncoding regions of the chloroplast (treated as a single locus) and two unlinked autosomal genes between *S. vulgaris* and *S. latifolira* gave a highly significant value, with most of the signal (using Equation 10.2b) coming from the cpDNA region. However, the estimated F_{ST} value (Chapter 2) for cpDNA was 0.546 versus 0.056 for nuclear genes, showing strong population structure at the organelle-gene level but only modest structure for nuclear genes. Ingvarsson attempted to correct for these between-gene differences in the amount of structure as follows. Under an island model of migration (Chapter 2), to a first approximation, population structure increases the amount of segregating sites and decreases the divergence, both by a factor of $1 - F_{ST}$. Ingvarsson thus corrected the observed number, S, of segregating sites by using $S_c = (1 - F_{ST})S$ and the divergence by $D_c = D/(1 - F_{ST})$. Applying these corrections to both the cpDNA and nuclear genes and using the S_c and D_c values in the HKA test yielded a nonsignificant result. Thus, the apparently strong signal of selection appears to simply be an artifact generated by nuclear and organelle genes having different population structures.

Note: The entire cpDNA or mtDNA genome is considered as a single locus under the HKA test

The McDonald-Kreitman (MK) Test

- One of the most widely-used tests
- Requires a polymorphism sample from one species and a divergence sample between species
- Robust to demography, *provided* that the effectively neutral mutation rate is the same during the polymorphism and divergence phases
- Contrasts the amounts of polymorphism and divergence between two categories of sites within a single gene
  - Typically silent vs. replacement sites

The McDonald-Kreitman (MK) Test: Basics

One of the most straightforward, and widely used, tests of selection was proposed by McDonald and Kreitman (1991a), who contrasted the amounts of polymorphism and divergence between two categories of sites within a single gene (Example 10.1). Typically, these categories are silent versus replacement sites, but the basic logic can be extended to other comparisons. Under the neutral theory, deleterious mutations are assumed to occur, but to then be quickly removed by selection, thus not contributing to either polymorphism or divergence (Figure 7.1). In the standard neutral-theory expressions for the amount of polymorphism ($4N_e\mu$) and divergence ($2t\mu$), μ is the *effectively neutral* mutation rate, which is the rate at which effectively neutral ($4N_e|s| \ll 1$) mutations arise. While most mutations at silent sites may often be effectively neutral, a much smaller fraction, f, of new mutations at replacement sites are neutral, resulting in a lower effectively neutral mutation rate, $f\mu$. Given that f is the fraction of replacement mutations that is effectively neutral, $1 - f$ is a measure of *functional constraints*, with values of $1 - f$ near one ($f \simeq 0$) implying that most new mutations are not effectively neutral (i.e., they are deleterious). A minor bookkeeping detail is that the silent and replacement mutation rates in the MK test refer to the sum over all sites, so that $\mu_s = \mu n_s$ and $\mu_a = \mu f n_a$ are the total neutral mutation rates over the collection of n_s silent and n_a replacement sites in the gene of interest (generally $n_a > 2n_s$, as all second-base and many third-base positions within codons are replacement sites).

As before, under the equilibrium neutral model, the expected number of substitutions (D_i) in site class i is $2t\mu_i$, while the expected number of segregating sites (S_i) in a sample of n sequences is $a_n\theta_i$ (Equation 9.21a). Because S_i is a measure of the amount of polymorphism, we denote it by P_i to conform to the standard notation for MK tests. Thus, under neutrality,

$$\frac{D_a}{D_s} = \frac{2t\mu_a}{2t\mu_s} = \frac{2t\mu f n_a}{2t\mu n_s} = f\frac{n_a}{n_s}, \qquad \frac{P_a}{P_s} = \frac{S_a}{S_s} = \frac{a_n\theta_a}{a_n\theta_s} = \frac{4N_e\mu f n_a}{4N_e\mu n_s} = f\frac{n_a}{n_s} \tag{10.5a}$$

where the subscript a denotes replacement (amino-acid changing) sites, and s denotes silent sites. Hence, under the equilibrium neutral model, we expect that, on average,

$$D_a/D_s = P_a/P_s \tag{10.5b}$$

If some replacement sites are under positive selection, because of their rapid sojourn times relative to drift, these will generally contribute very little to the within-species polymorphism (Kimura 1969; Smith and Eyre-Walker 2002; Figure 7.1), but they will result in an excess of replacement substitutions, so that $D_a/D_s > P_a/P_s$. Similarly, note that

$$\frac{P_a}{D_a} = \frac{a_n\theta_a}{2t\mu_a} = \frac{a_n 4N_e\mu f n_a}{2t\mu f n_a} = \frac{a_n 2N_e}{t}, \qquad \frac{P_s}{D_s} = \frac{a_n\theta_s}{2t\mu_s} = \frac{a_n 2N_e}{t} \tag{10.5c}$$

and thus, under neutrality, we also have

$$P_a/D_a = P_s/D_s \tag{10.5d}$$

McDonald and Kreitman provided a more general derivation of the polymorphism ratio in Equation 10.5a, replacing $4N_e$ (the equilibrium value) by T_{tot}, the total time on all of the within-species coalescent branches (Chapter 2). By considering the ratio of the number of polymorphic sites in the two categories, the common term T_{tot} cancels, so that any effects of demography also cancel. Hence, ***provided the effectively neutral mutation rates remain unchanged***, the MK test is ***unaffected by population demography*** (Hudson 1993; Nielsen 2001). Because the coalescent structure that determines the amount of polymorphism is explicitly removed by using the P_a/P_s ratio, there is no assumption that the allele frequencies are in mutation-drift equilibrium nor any assumption about constant population size. This is a very robust feature not shared by most other tests of selection.

Thus, while Zhai et al. (2008) found that the HKA test was more powerful than the MK test when the equilibrium assumptions hold, the robustness of the MK test (and lack of robustness of the HKA test) when demographic issues are present favors the use of the former. However, as we will see shortly, the MK test is by no means foolproof, as changes in the effective population size can influence the *effectively neutral* mutation rates (the rate at which alleles with $4N_e|s| < 1$ arise), which can bias some of the comparisons used by the test. Another complication is that mildly deleterious alleles can contribute to within-species polymorphisms, but not to between-species divergence, and thus their presence inflates the polymorphism ratio over the divergence ratio, reducing the power to detect positive selection.

The MK test is performed by contrasting polymorphism and divergence data at silent and replacement sites for the gene in question. Given that these two ratios are expected to be equal under neutrality, the test uses a simple 2×2 contingency table (Example 10.1). The presentation of the data required for the MK test is often referred to as either an **MK table** or a **DPRS table**, the latter based on the (clockwise order) of the table's four categories: **D**ivergence (number of substitutions), **P**olymorphism (number of segregating sites), **R**eplacement, and **S**ilent (or **S**ynonymous):

	Divergence	Polymorphism
Silent	D_s	P_s
Replacement	D_a	P_a

Example 10.1 presented the original data used by McDonald and Kreitman, while Example 10.5 shows how their test can be modified to examine different regions within the same gene.

Example 10.5. Le Corre et al. (2002) examined the *FRIGIDA* (*FRI*) gene in *Arabidopsis thaliana*, a key regulator of flowering time. European populations show significant variation in flowering time, with potentially strong selection for earlier flowering having arisen following the end of the last ice age. For the data below, fixed differences (divergence) were obtained by comparing *A. thaliana* with *A. lyrata*, while data on numbers of segregating sites are based on *A. thaliana* populations.

Entire coding region	Fixed	Polymorphic	
Silent	59	7	
Replacement	68	21	Fisher test $p = 0.056$
Exon 1	Fixed	Polymorphic	
Silent	30	2	
Replacement	38	16	Fisher test $p = 0.013$
Exons 2 and 3	Fixed	Polymorphic	
Silent	29	5	
Replacement	30	5	Fisher test $p = 1.000$

The *FRI* locus clearly shows heterogeneity in patterns of selection when contrasting exon 1 with the remaining exons, and detecting such within-gene heterogeneity may provide important clues for a putative region under functional selection.

These data could be interpreted simply as a reduction on functional constraints in exon 1, resulting in a smaller fraction of segregating replacement mutations being deleterious. In principle, this could occur because of a shift in the selection pressures or for purely demographic reasons, such as a recent reduction in the effective population size increasing the effectively neutral mutation rate. However, there is a nice internal control in that exons 2 and 3 do not display a decrease in the ratio of fixed to polymorphic replacement sites relative to silent sites, which appears to rule out a reduction in effective population size in *thaliana* accounting for the reduction in constraints. The authors noted that roughly half of the replacement polymorphisms in exon 1 are loss-of-function mutations, which result in early flowering. Hence, it appears that the excess number of replacement polymorphisms in exon 1 likely results from selection for early flowering in some populations. Further, because a nonfunctional copy of *FRI* results in early flowering, there are a large number of mutational targets to achieve this phenotype (and hence a high effective mutation rate), which likely explains the large number of replacement polymorphisms. In effect, these data appear to show an ongoing multiple-origins soft sweep (Chapter 8).

A McDonald-Kreitman test will be significant when P_a/D_a is significantly different from P_s/D_s (Equation 10.5d). Because it is assumed that the silent-site ratio is unchanged by selection, a significant MK test can occur either through an excess of replacement polymorphisms (P_a too large relative to D_a and P_s/D_s) or through an excess of replacement substitutions (D_a too large relative to P_a and P_s/D_s). The **neutrality index** of Rand and Kann (1996),

$$NI = \frac{P_a/D_a}{P_s/D_s} = \frac{P_a\, D_s}{P_s\, D_a} \tag{10.6a}$$

indicates which of these two scenarios occurs. Note that *NI* is simply the odds ratio for the MK contingency table (Jewell 1986). A value greater than one indicates more polymorphic replacement sites than expected, while a value less than one indicates an excess of replacement substitutions. Values less than one suggest that some of the substitutions are adaptive, while values greater than one are suggestive of weakly deleterious segregating alleles.

Example 10.7. Consider Le Corre et al.'s data on the *FRI* gene (Example 10.5). For exon 1, the neutrality index is

$$NI = \frac{P_a/D_a}{P_s/D_s} = \frac{16/38}{2/30} = 6.42$$

showing that the significant result is due to an excess of segregating replacement sites. Conversely, for exons 2 and 3

$$NI = \frac{5/30}{5/29} = 0.97$$

suggesting a good fit to the neutral model, with neither an excess of polymorphic site nor of fixed replacement sites.

Our interpretation of the signal in exon 1 was as a sign of ongoing selection of alleles for earlier flowering (Example 10.5). However, the *NI* value is also consistent with an excess of slightly deleterious alleles in this region, thus inflating the levels of replacement polymorphisms. The lack of such a signal in exons 2 and 3 argues against this, but it remains a formal possibility that slightly weaker selection in exon 1 (relative to exons 2 and 3), coupled with a genomewide reduction in N_e, could account for the excess polymorphism in exon 1. However, evidence for a recent population expansion argues against this.

Peter Andolfatto

Carlos Bustamante

Example 10.6. Andolfatto (2005) examined 35 coding and 153 noncoding fragments from a Zimbabwe sample of 12 *D. melanogaster* X chromosomes, with a single *D. simulans* X as an outgroup. The numbers of observed polymorphic and divergent sites were then lumped into various classes as follows:

		Polymorphisms		Fisher Test p value	
Mutational Class	Fixed	All sites	Minus singletons	All sites	Minus singleton
Silent	604	502	323	——	——
Replacement	260	115	52	$4.7 \cdot 10^{-7}$	$4.3 \cdot 10^{-10}$
Noncoding	3168	2386	1295	$1.4 \cdot 10^{-2}$	$5.2 \cdot 10^{-3}$
5′ UTRs	328	160	71	$2.7 \cdot 10^{-6}$	$1.7 \cdot 10^{-10}$
3′ UTRs	143	86	36	$3.3 \cdot 10^{-2}$	$8.2 \cdot 10^{-5}$

Given the small sample size ($n = 12$ chromosomes), polymorphism data are reported both as the total number of segregating sites (all sites) and the total number of segregating sites minus the singletons. The logic for removing singletons is the concern that slightly deleterious alleles can contribute to segregating sites (although they will be rare) but are unlikely to become fixed, and if retained in the analysis, will result in the polymorphism ratio overpredicting the number of fixed sites. Using the silent class as the neutral reference, McDonald-Krietman tests were performed against each of the four remaining categories (replacement, noncoding, 5′ UTR, and 3′ UTR), and computed separately using either all polymorphisms or only polymorphisms that were not singletons. The exclusion of singletons ("Minus singletons" column above) decreases the p values (increasing significance) in all cases. Even after correcting for multiple tests, all of the comparisons based on polymorphisms minus singletons were highly significant.

Example 10.8. Bustamante et al. (2005) sequenced roughly 11,600 genes in 39 humans and contrasted the results with human-chimp divergence at these same loci. Summing over all sites, the resulting DPRS table (where SNPs denote polymorphic sites) was

	Divergence	SNPs
Silent	34,099	15,750
Replacement	20,467	14,311

As in Example 10.6, this analysis differs from a standard MK test, as the values for a large number of loci are aggregated into a single table. The resulting p value, $< 10^{-16}$, was highly significant, meaning that the neutral model is rejected.

What is the source of the discrepancy? Equation 10.6a gives the neutrality index as

$$NI = \frac{P_a/D_a}{P_s/D_s} = \frac{14,311/20,467}{15,750/34,099} = 1.514$$

showing that the lack-of-fit to the neutral model is driven by an excess of replacement polymorphisms (SNPs). The authors suggest that these polymorphisms are mainly deleterious, a view echoed by Hughes et al. (2003). Consistent with this conclusion, in an analysis of ~47,500 replacement SNPs in a sample of 35 humans, Boyko et al. (2008) used the site-frequency spectrum to estimate that 27–29% of these SNPs were effectively neutral, 30–42% were moderately deleterious, and nearly all of the rest were highly deleterious (we will discuss how such values are obtained shortly). This large fraction of segregating deleterious alleles significantly lowers the power of MK tests. Indeed, Charlesworth and Eyre-Walker (2008) noted that because of excessive replacement polymorphisms, MK tests in humans are very underpowered.

One potentially significant advantage of the MK test is that it does not assume constant population size or that mutation-drift equilibrium has been reached, and hence is rather robust against many of the demographic concerns that plague other tests. Balancing this advantage are two subtle (but serious) problems, both relating to how the distribution of fitness values for new alleles impacts the observed data (polymorphisms and substitutions).

First, the MK framework assumes that *deleterious mutations are strongly deleterious and make essentially no contribution to either the number of segregating or fixed sites.* In fact, however, weakly deleterious mutations (i.e., $-10 < 4N_e s < -1$) can contribute to segregating polymorphisms (especially because the MK test uses the number of polymorphic sites, not their frequencies), but they are highly unlikely to become fixed (Figure 7.1). Such mutations are overrepresented in polymorphic sites relative to fixed sites, which reduces the power of the MK test to detect an excess of replacement substitutions (and hence a signature of positive selection). We assume that the impact from any overrepresentation of selected polymorphisms at silent sites (our neutral proxy) is small, as these are either neutral or under very weak purifying selection. Conversely, overrepresentation is potentially a significant problem at polymorphic replacement sites. One proposed correction for this problem is to drop "rare" polymorphisms, but this is a rather subjective endeavor. Dropping singletons (Templeton 1996) as in Example 10.5 provides one simple correction, while other authors (e.g., Fay et al. 2002; Smith and Eyre-Walker 2002; Gojobori et al. 2007) have suggested including only "common" polymorphisms in the analysis, such as those with minor-allele frequencies above 0.10. We return to this issue shortly.

The second concern is even more problematic. At the heart of the MK test is Equation 10.5a. Under the neutral hypothesis, the ratio of polymorphic sites and the ratio of substitutions both estimate the same quantity, f (scaled by the sample-size correction factor n_a/n_s), the ratio of effectively neutral mutation rates for the two categories. Recalling (Chapter 7) that any mutation for which $4N_e|s| \ll 1$ behaves as if it were effectively neutral, the caveat is that the *effectively neutral mutation rate, $f\mu$, changes with N_e*. It is important to stress that the *total* mutation rate, μ, remains unchanged, but the fraction, f, of these mutations that are effectively neutral can decline with increasing N_e, resulting in a decline in $f\mu$. Figure 10.1 shows that estimates of f do indeed decrease as the effective population size, N_e, increases, as the amount of constraint, $1 - f$, increases with N_e. For the same distribution of selection coefficients, one can raise (or lower) f (and hence the effectively neutral substitution rate) by decreasing (or increasing) the effective population size. If the effective population size is significantly different during the divergence phase (when substitutions were fixed) than in the current phase (which generates the observed number of polymorphisms), then these two phases could have different fractions of mutations that are effectively neutral. Because the ratios D_a/D_s and P_a/P_s estimate the f values for these two different phases, they can have different expected values.

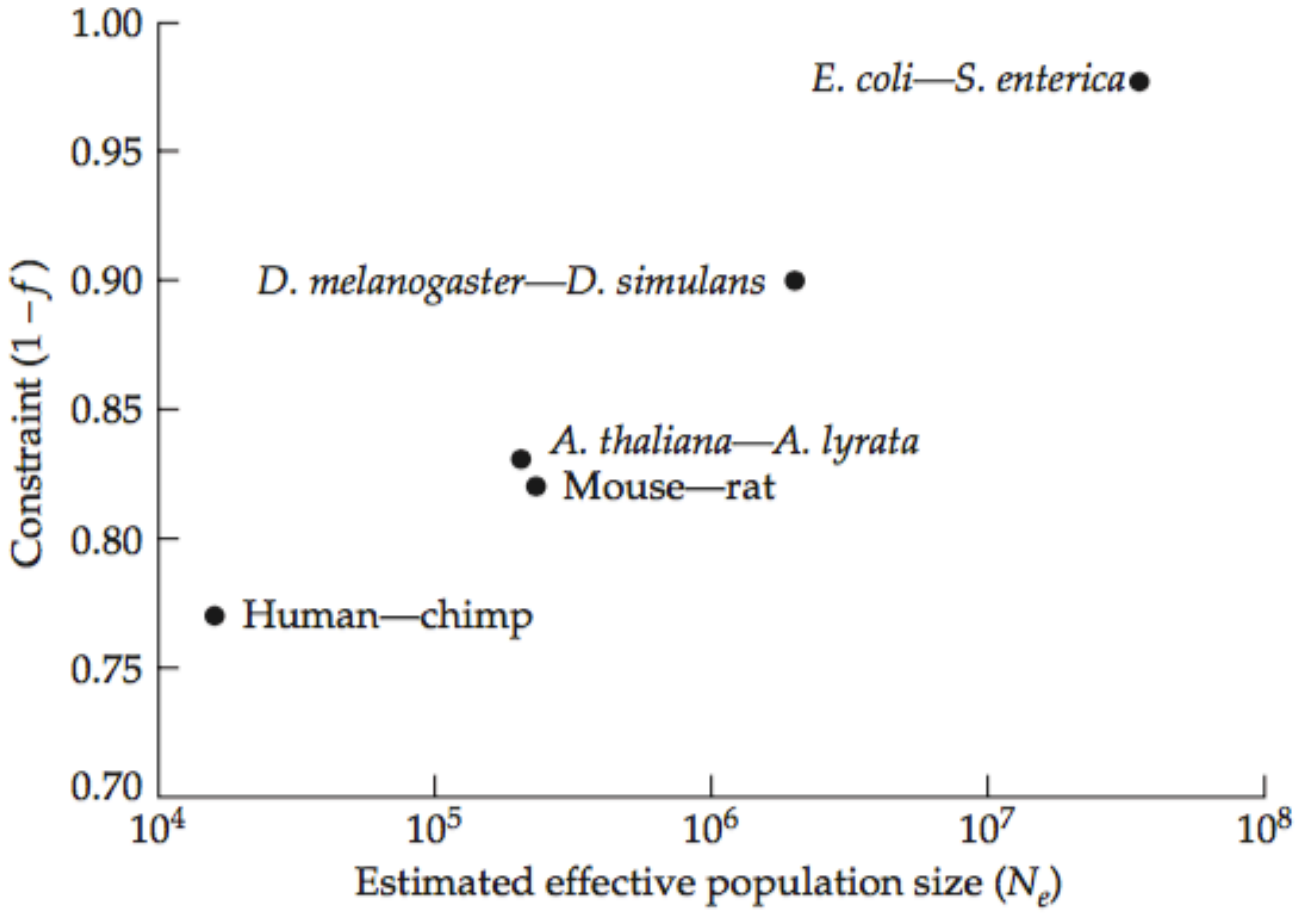


Figure 10.1 The estimated constraint, $1 - f$, on replacement sites as a function of effective population size, where f is the ratio of effectively neutral mutation rates (the fraction of new mutations that efficiently behave as neutral alleles) at replacement versus silent sites. As N_e increases, more deleterious mutations move from the effectively neutral class into the strongly deleterious class (f decreases), reducing the effectively neutral mutation rate and increasing the amount of constraint on a gene. (After Wright and Andolfatto 2008.)

Example 10.9. An example of some of the potential difficulties in interpreting the results of a McDonald-Kreitman test was seen in a study of the human melanocortin 1 receptor (*MC1R*), a key regulatory gene in pigmentation (Harding et al. 2000). In comparing the canonical *MC1R* haplotype in humans with a sequence from chimpanzees, these authors found 10 replacement and 6 silent substitutions. An African population sample revealed no replacement and 4 silent polymorphisms, giving the MK table as

	Fixed (Human-Chimp)	Polymorphic (African)
Silent	6	4
Replacement	10	0

Fisher's exact test gives a p value of 0.087, close to significance. Taken at face value, one might assume that these data imply that the majority of the replacement substitutions between human and chimp were selectively driven. However, the authors also had data from populations in Europe and East Asia, which showed 10 replacement and 3 silent polymorphisms, resulting in a new MK table:

	Fixed (Human-Chimp)	Polymorphic (Europe / East Asia)
Silent	6	3
Replacement	10	10

with a corresponding p value of 0.453. The authors suggested that the correct interpretation of these data is as very stringent purifying selection due to increased functional constraints in African populations (due to selection for protection against high levels of UV exposure), with a release of constraints in Europe and East Asia. Asians in Papua New Guinea and India (populations living in high-UV environments) also showed very strong functional constraints (few replacement polymorphisms), consistent with a model of selection for UV protection.

Phylogeny-based tests

- Need divergence data from a number of species, placed within a known phylogeny
- K_a to K_s ratio, ω
- Likelihood-based codon models

PHYLOGENY-BASED DIVERGENCE TESTS

Finally, we briefly consider divergence tests that examine the pattern of substitutions over a known phylogeny. These tests are designed to detect a rather different pattern of selection than was assumed in Chapter 9 (single events) or earlier in this chapter (multiple substitutions over an *entire gene* between two populations or species). While multiple substitutions are also required for a signal in phylogeny-based divergence tests, these must be at the same *site* (typically a codon) within a gene. Single substitutions over a number of different codons across a gene may leave very little signal for these tests (unless very few silent substitutions have occurred). As such, phylogenetic tests are biased toward detecting sites that undergo repeated evolution, and are likely to miss many, indeed perhaps *most*, adaptive substitutions (Hughes 2007). Given this restriction, these methods may work well in so-called "**arms race**" scenarios, in which trait values between two interacting species escalate to increasingly extreme values (Bergelson et al. 2001), such as the interactions between hosts and parasites.

The required input for phylogeny-based tests is a set of aligned DNA sequences and a predetermined phylogenetic tree for the sampled species. The assumption is that all sequence differences are the result of fixation events. Thus, if a site is segregating in one (or more) of the taxa from which a single sequence is drawn, one may incorrectly infer that it is a substitution event. The taxa must also have the correct amount of divergence, as either too little or too much, will result in very low power. With too little divergence, there are not enough substitutions, and hence there is little power to detect small percentage differences in silent versus replacement changes at particular sites. Further, if little true divergence has occurred, even a few segregating sites incorrectly called as substitutions can significantly inflate the divergence. Conversely, with too much divergence, multiple substitutions at single sites may occur between lineages, and adjustments for such multiple hits can introduce substantial bias if an incorrect statistical model is used to account for these.

The K_a to K_s ratio, ω

The basis for divergence-based tests is $\omega = K_a/K_s$, the (per-site) ratio of replacement (nonsynonymous) to silent (synonymous) substitution rates, which Miyata and Yasunaga (1980) referred to as the **acceptance rate** and which also appears in the literature as the **width of the selective sieve**. For sites under the standard neutral model (deleterious mutations can arise, but are quickly removed), the expected value of ω at a site (or gene) is $\omega = \mu f/\mu = f \leq 1$, where f is the ratio of the effectively neutral mutation rates. Thus, in the absence of positive selection, we expect $\omega < 1$. Moreover, if adaptive mutations are absent (or very rare), then $1-\omega$ is a direct measure of the amount of constraint $(1-f)$ on a site. Conversely, $\omega > 1$ is usually taken as an unmistakable signature of selection (Kimura 1983). Even if a demographic change results in a lowering of the effective population size (increasing the effectively neutral mutation rate at replacement sites), such a change (in the absence of positive selection) only brings K_a/K_s closer to, but still likely leaves it smaller than, 1.0.

There are cases where $\omega > 1$ is *not* a signal for positive selection. Ratnakumar et al. (2010) noted that a resolution of heteroduplex DNA during gene-conversion events often results in a bias toward G and C bases (also see Galtier et al. 2001; Webster and Smith 2004; Lassalle et al. 2015). Given that replacement-codon positions often have lower GC content than synonymous sites, there can be more opportunities for A/T at these sites to be changed to G/C, resulting in replacement substitutions and potentially inflating the K_a/K_s ratio (Berglund et al. 2009; Galtier et al. 2009). Ratnakumar et al. analyzed a dataset of roughly 18,000 human genes compared against their orthologs in at least two other mammalian genomes. They found that genes giving divergence-based signals of selection had a significant tendency to also display genomic signals of GC conversion bias. They estimated that >20% of elevated ω values in this dataset could be the result of biased gene conversion. A second factor is mutational bias. McVean and Charlesworth (1998) and Lawrie et al. (2011) found the counterintuitive result that weak selective constraints that oppose a mutational bias can actually *accelerate* the rate of evolution over that of a neutral site. In the words of Lawrie et al., this occurs because

> Common mutations drive substitutions away from the fitter states despite purifying selection, whereas selection favors fixation of uncommon mutations resulting in faster back substitutions to the fitter states. This allows for greater overall flux between states and thus a higher rate of substitution at the constrained sites compared with the neutrally evolving sites.

A final factor that can upwardly bias estimates of ω is the presence of strong selective constraints on *silent* sites, was found in *Drosophila* by Lawrie et al. (2013). Chamary et al. (2006) reviewed some of the evidence that silent sites may still be subjected to constraints (beyond any weak ones from codon usage bias; Chapter 8) because they affect mRNA stability, splicing, or microRNA binding. A cautionary tale is offered by the work of Hurst and Pál (2001), who examined constraints on the breast cancer *BRCA1* gene. A sliding window of roughly 300 nucleotides, allowing for average regional estimates of K_a and K_s, was used to scan across this gene in two pairs of comparisons, human-dog and mouse-rat. The window around position 200–300 showed a relatively normal level of K_a (relative to the rest of the gene), while K_s plummeted dramatically, especially in the human-dog comparison. The result was an ω value significantly greater than one, not due to an elevation in the replacement-substitution rate, but rather to a decrease in the silent-substitution rate. Wolf et al. (2009) found that an upward bias in ω from reduced K_s values can be especially problematic when using closely related taxa, as a small value of D_s causes excessive stochastic variation in the denominator of a K_a/K_s ratio. Pond and Muse (2005) noted that if variation in K_s occurs over the gene, failure to include this heterogeneity in the model can easily result in false positives (estimated $\omega > 1$ for particular codons). Thus, while $\omega > 1$ is usually taken as a gold standard for positive selection, a little more humility in its use may be in order.

While conceptually straightforward, the operational problem in using ω is that while one or a few sites may be under *repeatedly* strong directional selection ($\omega > 1$ at these residues), most sites in a protein are expected to be under some selective constraints ($\omega < 1$), so that the average over all sites yields $\omega < 1$. Indeed, a meta-analysis by Endo et al. (1996) found that only 17 out of 3595 proteins (from a wide range of species comparisons) showed $\omega > 1$. There were, however, a few early success stories. Example 10.2 discussed the work of Hughes and Nei (1988), who used the three-dimensional (3-D) protein structure of the major histocompatibility complex to suggest potential sites to examine (those amino acids on the surface in critical positions). Within this set of residues, $\omega > 1$, while $\omega < 1$ when averaged over the entire gene. Unfortunately, most proteins lack this amount of detailed biological knowledge for an investigator to draw upon. Because amino acid residues in close proximity on the 3-D structure of a protein can be scattered all over the primary (i.e., linear) sequence, grouping sites for analysis by their position in the primary sequence can be very ineffective, and even misleading. The key is to base tests of ω values on a *codon-by-codon* basis, so that codons, rather than genes, become the unit of analysis. The limitation for this approach is the need for sequences from a sufficiently dense and well-supported phylogeny.

Two general approaches have been suggested to estimate ω. Both require a phylogeny, and issues such as the correct multiple-sequence alignment as well as errors in the assumed tree potentially loom in the background. **Parsimony-based** approaches reconstruct the ancestral sequence at each node in the tree, and then use these to count up the number of silent and replacement substitutions for each codon. **Likelihood approaches** (LW Appendix 4) are on a much firmer statistical footing, but they are computationally intense and can be rather model-specific. Both approaches allow for tests of whether a protein is under positive selection and, more specifically, tests of positive selection at specific *sites* in that protein. As with extensions to PRF models, more recent tests are being built around **Bayesian approaches** that extend the ML models (Appendix 2), which allow for the management of uncertainty in very complex statistical models.

First, it is well known that **transitions** (A ↔ G, C ↔ T) can occur at different rates than **transversions** (e.g., A ↔ T, etc.), and (at third-base positions) transitions are more likely to give synonymous changes. Failure to incorporate these rate differences can result in an overestimation of the number of replacement substitutions (Yang and Nielsen 2002). Second, any codon usage bias (Chapter 8) must be accommodated. Third, when divergence times are modest to large, to avoid undercounting the number of the actual substitution events one must correct for the possibility of multiple substitutions between lineages at a site. All of these issues can have a highly significant effect on estimates of ω (Yang and Bielawski 2000). Finally, given that the ancestral states are likely estimated with error, parsimony analysis has no formal procedure to take this uncertainty into account. Bayesian posterior distributions can account for these errors, but this requires moving from a parsimony to a likelihood framework. For these reasons, most analyses use likelihood-based approaches (and their Bayesian extensions), wherein one explicitly allows the model to account for transitions vs. transversions, codon usage bias, and multiple substitutions.

Joe Felsenstein

Ufeus felsensteini

ML methods require a specific probability model for the movement among the 64 different codons. They start with a vector representing the 61 different nonstop codon states (stop codons are assumed lethal). At any point in time, a codon can mutate to one of nine other codons following a single base change (Figure 10.4). The model given by Goldman and Yang (1994) defines the following relative rates for movement between codons i and j,

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a silent transversion} \\ \kappa\pi_j & \text{for a silent transition} \\ \omega\pi_j & \text{for a replacement transversion} \\ \omega\kappa\pi_j & \text{for a replacement transition} \end{cases} \quad \text{for } 1 \leq i, j \leq 61 \qquad (10.18)$$

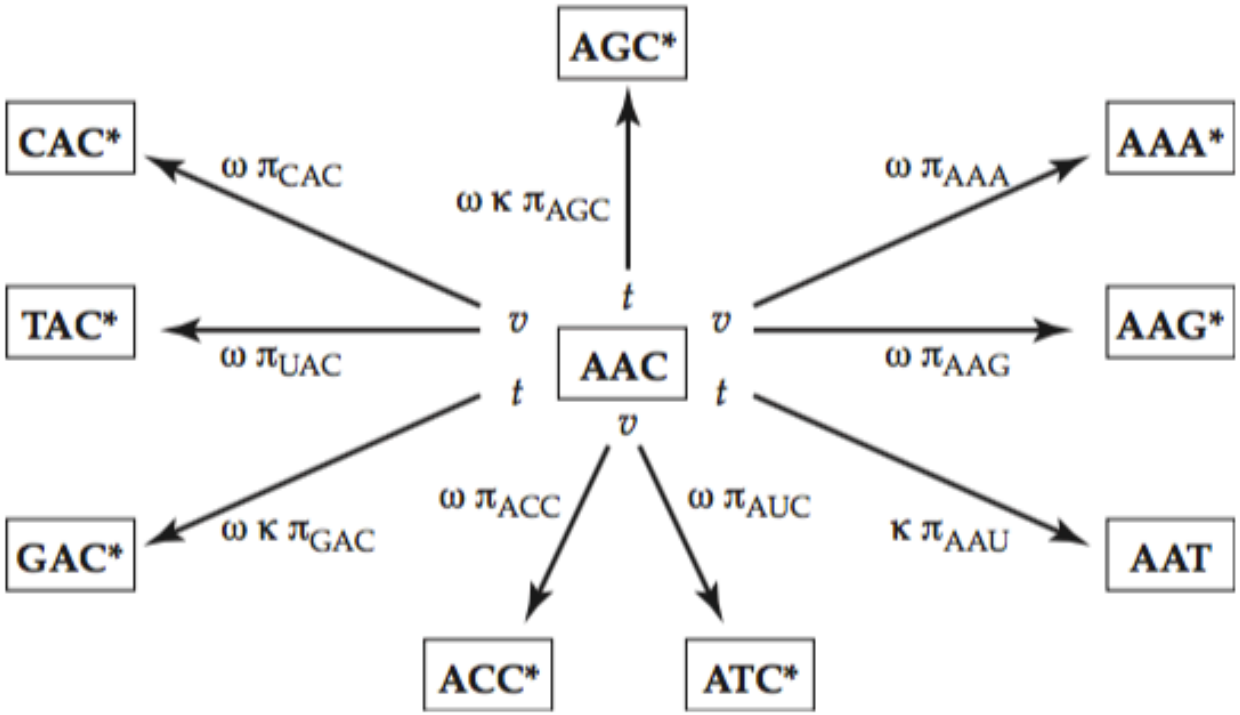


Figure 10.4 The various possible state changes and their rates under the codon evolution model (Equation 10.18) for the nine new codons that are within a single nucleotide change from the target codon (here AAC). Asterisks denote a replacement change, where the rate is a function of selection, and hence ω. Because transitions (denoted in the figure by t) and transversions (v) may occur at different rates, setting the transversion rate as the baseline, κ denotes any transition rate correction (with $\kappa = 1$ if the two rates are equal). All changes are a function of π_j, the equilibrium frequency of the mutant codon, j. Performing these same calculations over all 60 other nonstop codons generates the full transition matrix , $\mathbf{Q}$.

Tests for directional selection on a gene are accomplished by using this codon model superimposed on the phylogenetic tree to run the likelihood calculation (over all codons) to find the ML solutions for the $\mathbf{Q}$ matrix parameters. This allows for a direct test that $\omega > 1$ using the likelihood-ratio approach (LW Appendix 4). The key to these likelihood calculations is that $\mathbf{P}(t)$, the codon state matrix at time t, is related to the instantaneous rate matrix, $\mathbf{Q}$, by

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) \tag{10.19}$$

The corresponding elements of the 61×61 matrix $\mathbf{P}$ are

$$P_{ij}(t) = \Pr(\text{codon } = i \text{ at time } t \,|\, \text{codon is } j \text{ at time } t = 0) \tag{10.20}$$

The matrix exponential, $\exp(\mathbf{Q}t)$, is computed by diagonalizing the $\mathbf{Q}$ matrix by writing $\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix, whose ith diagonal element is the eigenvalue λ_i of $\mathbf{Q}$ (Equation A5.10a), and $\mathbf{U}$ is a matrix of the eigenvectors of $\mathbf{Q}$ (Equation A5.10c). With this transformation, Equation 10.19 now becomes

$$\exp(\mathbf{Q}t) = \mathbf{U}\exp(t\boldsymbol{\Lambda})\mathbf{U}^T$$

where

$$\exp(t\boldsymbol{\Lambda}) = \begin{pmatrix} e^{t\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{t\lambda_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & e^{t\lambda_n} \end{pmatrix} \tag{10.21}$$

The base model (Equation 10.18) assumes that all codons within a given gene have the same ω value, which is not only unreasonable but also destroys most of the power of this approach, as it returns an estimate of ω based on a gene-wide average. Given that $\omega < 1$ for most codons, the signal from the majority of codons then masks the signal from any small fraction of codons where indeed $\omega > 1$ (e.g., Example 10.2). Nielsen and Yang (1998) and Yang et al. (2000) extended the base model by assuming a mixture-model (LW Chapter 13), with the codons in a sequence being drawn from one of several selection categories, each with different ω values. For codons from selection category k, Equation 10.18 becomes

$$Q_{ij}^{(k)} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a silent transversion} \\ \kappa\pi_j & \text{for a silent transition} \\ \omega^{(k)}\pi_j & \text{for a replacement transversion} \\ \omega^{(k)}\kappa\pi_j & \text{for a replacement transition} \end{cases} \tag{10.22a}$$

The simplest version of biological interest has three selection classes, with codons either being neutral (with probability p_0), deleterious (with probability p_d), or advantageous (with probability $p_b = 1 - p_n - p_d$). Within each class there is a fixed selective value, with

$$\omega^{(k)} = \begin{cases} 0 & \text{deleterious class} \\ 1 & \text{neutral class} \\ \omega > 1 & \text{positively selected class} \end{cases} \tag{10.22b}$$

The parameters p_0, p_d, and ω are estimated from the data by maximum likelihood (LW Chapter 13 examines ML on mixture models). The idea is that one fits a base model (allowing only neutral and deleterious classes), and then fits the full model (Equation 10.22b or other extensions), using a likelihood-ratio test to see if the fit is significantly improved. If so, this is taken as support for a history of repeated positive selection on a subset of codons in the gene of interest.

While Equation 10.22b is clearly an improvement over models assuming a single value of ω for *all* replacement mutations, assigning all codons in the deleterious class an ω value of 0 (i.e., no substitutions) is clearly too restrictive, as is assigning all codons in the advantageous class the same ω value. Nielsen and Yang (1998) and Yang et al. (2000) further expanded Equation 10.22b by taking

$$\omega^{(k)} = \begin{cases} w^{(d)} \sim (0,1) & \text{deleterious class} \\ 1 & \text{neutral class} \\ w^{(a)} \sim (1,\infty) & \text{positively selected class} \end{cases} \tag{10.23}$$

where now the fitness values, $\omega^{(k)}$, for any particular codon in class k are *random draws* from some specified distribution (as opposed to Equation 10.22b, which assumed they are unknown constants) whose parameters are again estimated by maximum likelihood. This is exactly the approach used previously to allow γ to vary over genes in the PRF model (e.g., Equations 10.17a and 10.17b). A number of candidate distributions for ω are possible, depending on whether we wish to restrict values to between (0, 1) or to (1, ∞), for codons in the deleterious and positively selected classes (respectively). For example, Nielsen and Yang (1998) and Yang et al. (2000) used either a beta or truncated gamma distribution (restricted to returning values of $0 < \omega < 1$) for the deleterious class and a truncated gamma (restricted to returning values of $\omega > 1$) for the positively selected class (Appendix 2 reviews the beta and gamma distributions). Again, a model-fitting approach is used where one first fits a

Bayesian Estimators of Sites Under Positive Selection

Suppose there are k classes, with each class having a different associated ω. The posterior probability that a specific codon is in fitness class i is

$$\Pr(\text{class } i \mid D) = \frac{\Pr(D \mid \text{class } i)\Pr(\text{class } i)}{\Pr(D)} = \frac{\Pr(D \mid \omega_i)\Pr(\omega_i)}{\sum_{i=1}^{k}\Pr(D \mid \omega_i)\Pr(\omega_i)} \quad (10.24a)$$

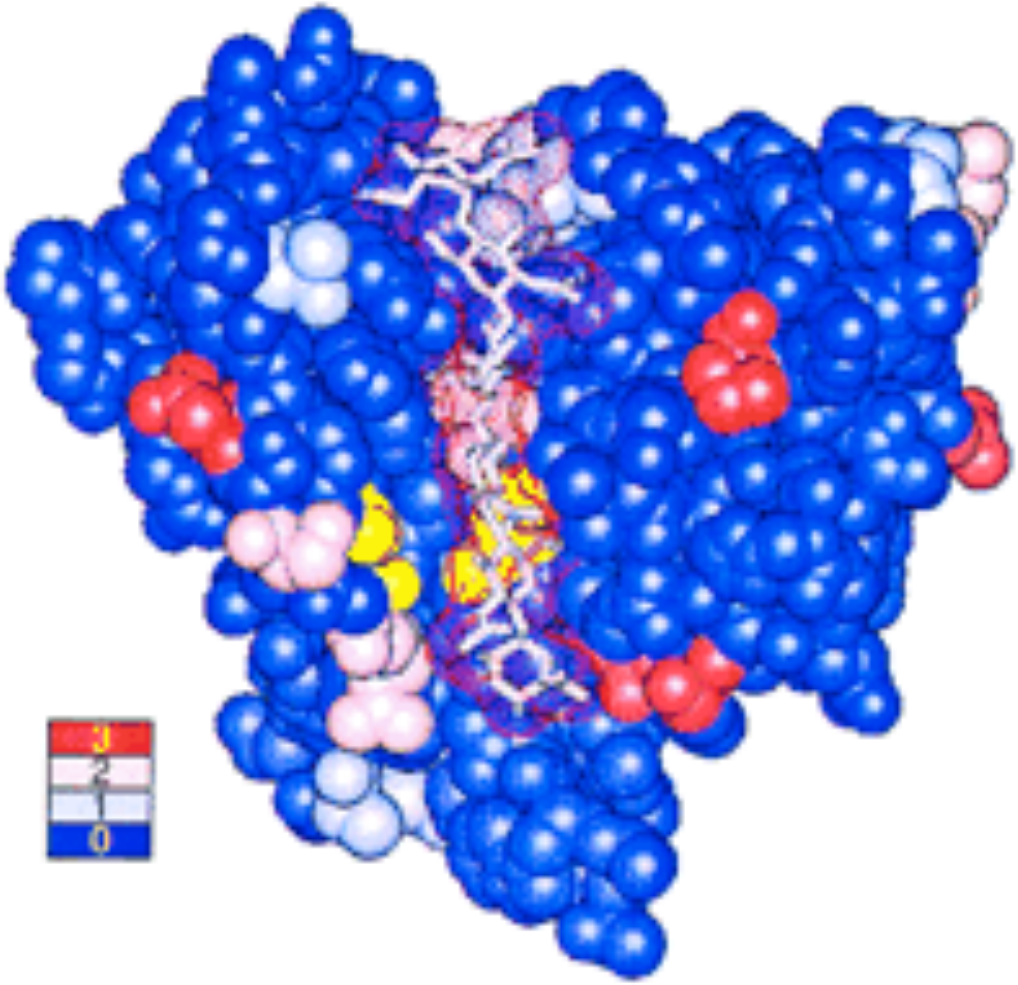
where D is the pattern of codons for that site in the tree, and the prior $\Pr(\text{class } i)$—the values for p_0, p_b, and p_d —is estimated by maximum likelihood. The case of interest is whether the codon belongs to the class of advantageous sites, $\Pr(\omega > 1 \mid D)$,

$$\Pr(\text{advantageous} \mid D) = \frac{\Pr(D \mid \omega > 1)p_b}{\Pr(D \mid \omega < 1)p_d + \Pr(D \mid \omega = 1)p_0 + \Pr(D \mid \omega > 1)p_b} \quad (10.24b)$$

This approach allows us to directly assign probabilities of selective classes to any particular site. Anisimova et al. (2002) found that large ω values and a modest to large number of sequences are required for this approach to have reasonable power. A number of technical issues that arise when applying Equation 10.24a were examined by Huelsenbeck and Dyer (2004), Newton et al. (2004), Scheffler and Seoighe (2005), Yang et al. (2005), Aris-Brosou (2006), Guindon et al. (2006), and Anisimova and Liberles (2007).

Example 10.15. Bishop et al. (2000) examined the class I chitinase genes from 13 species of mainly North American *Arabis* (tower mustards), crucifers closely related to *Arabidopsis*. Chitinase genes are thought to be involved in pathogen defense, as they destroy the chitin in cell walls of fungi. Many fungi have evolved resistance to certain chitinases, so these genes are excellent candidates for repeated cycles of selection (i.e., an "arms race" scenario). Codon-evolution models estimated that between 64 and 77% of replacement substitutions are deleterious, with 5–14% being advantageous (analyses using phylogenies estimated by different methods all yielded similar results). These favored sites had an estimated value of $\omega = 6.8$. Using the criterion of a posterior probability of membership in the advantageous class in excess of 0.95 (i.e., $\Pr(\text{advantageous class} \mid D) > 0.95$), 15 putative sites were located (using Equation 10.24b). Seven of these sites involved only one substitution type, which evolved multiple times over the phylogeny. The authors had access to the 3-D structure of chitinase, which shows a distinctive cleft thought to be the active site. Mapping putative sites of positive selection showed a significant excess of these sites clustered at the cleft.

Balancing this apparently successful application of these methods to detect selected sites is the work of Yokoyama et al. (2008). These authors examined the evolution of dim-light vision in vertebrates, which is determined by the wavelength of maximal absorption of rhodopsin. This can be directly measured in the lab, allowing the authors to experimentally determine the role of particular substitutions in dim-light adaptation using 11 engineered ancestral rhodopsin sequences. They found that most of the change in maximal absorption can be accounted for by 12 sites. In contrast, Bayesian methods predicted a total of 8 positively selected sites, none of which corresponded to sites shown by mutagenesis to have adaptive roles.



Class I Chitinase *(Arabis)*