Synonymous Nucleotide Divergence: What Is "Saturation"?

J. Maynard Smith and N. H. Smith

School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, United Kingdom

Manuscript received July 30, 1995
Accepted for publication November 10, 1995

ABSTRACT

The nucleotide divergence at synonymous third sites between two lineages will increase with time since the latest common ancestor, up to some saturation level. The "null-hypothesis divergence" is defined as the percentage of difference predicted at synonymous third sites, allowing for amino acid composition and codon bias, but assuming that codon bias is the same at all sites occupied by a given amino acid, when equilibrium has been reached between forward and backward substitutions. For two highly expressed genes, *gapA* and *ompA*, in the enterobacteria, the estimated values of the null-hypothesis divergence are 39.3 and 38.15%, respectively, compared to estimated values of saturation divergence of 19.0 and 25.4%. A possible explanation for this discrepancy is that different codons for a given amino acid are favored at different sites in the same gene.

TO interpret the percentage of nucleotide divergence between two taxa at synonymous third sites, we must know how different they would be if equilibrium had been reached between forward and backward substitutions. The question arose when trying to decide whether codon bias is the same at all sites specifying a given amino acid, or whether it varies between sites (MAYNARD SMITH and SMITH 1996), but it arises also in other contexts.

The nucleotide divergence at equilibrium depends on what assumptions are made. It is helpful to distinguish between three sets of assumptions:

1. Random divergence. There is no codon bias; that is, all codons for a given amino acid are used equally. Divergence will still depend on amino acid composition.

2. Null-hypothesis divergence. There is codon bias, but it is the same in direction and intensity for all taxa and for all sites occupied by a given amino acid. In practice, codon bias is usually greater for some genes than others (SHARP and LI 1987). However, it is still reasonable to calculate the null-hypothesis divergence for a given gene or class of genes with a similar codon bias.

3. Saturation divergence. There is codon bias, but it may vary from site to site either in direction or intensity, even for the same amino acid in the same gene. However, site-specific preferences, whether caused by mutation or selection, are the same in both taxa.

In general, null-hypothesis divergence will be less than random divergence, and saturation divergence will be less than null-hypothesis divergence. If codon bias varies in intensity but not direction, the difference between saturation and null-hypothesis divergence will be small, but if different codons are favored at different sites, the difference can be large. Saturation itself, however, is only a temporary equilibrium. Ultimately, changes in codon bias, caused, for example, by changes in G + C ratio, will result in increased divergence.

This work describes these models in greater detail and shows how a difference between saturation and null-hypothesis divergence can be used to detect site-specific codon bias.

Random divergence: If there is no codon bias, the equilibrium divergence at synonymous third sites is 0% for tryptophan and methionine, 50% for twofold redundant amino acids, 75% for fourfold redundant amino acids, and $(13/18) \times 100 = 72.2\%$ for sixfold redundant amino acids. The value for any gene will depend on amino acid composition. For example, for the *gapA* gene of *Escherichia coli,* discussed below, it is 63%.

Null-hypothesis divergence: If there is codon bias, the divergence will be lower. In the limit, if only one codon is used for each amino acid, the divergence will be 0%. On the null-hypothesis that bias is the same at all sites specifying a given amino acid, the equilibrium divergence is $1 - \Sigma p_i^2$, where p_i is the frequency of the i^{th} nucleotide at third sites. For twofold redundant amino acids, the value is $2p(1 - p)$. The degree of bias is greater for some genes than others, and within a gene, for some amino acids.

If the difference between two taxa is lower than the null-hypothesis divergence, then either there has not been time for forward and backward substitutions to reach equilibrium, or there is site-specific bias. Consider, for example, a twofold redundant amino acid with the two codons used in the proportion 0.8:0.2. The null-hypothesis divergence is then 32%. If the observed difference is, say, only 15%, then either equilibrium has not been reached, or one of the two codons is preferred at some sites and the other codon at other sites.

Corresponding author: J. Maynard Smith, School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, United Kingdom.

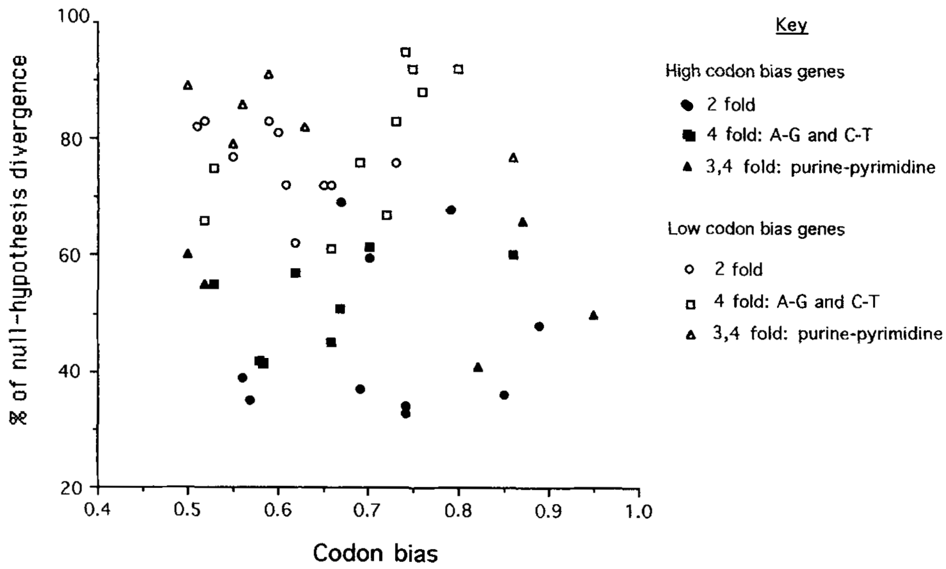

FIGURE 1.—Null-hypothesis divergence and codon bias. The divergence at synonymous sites between homologous genes of *E. coli* and *S. typhimurium* is compared, as a percentage of the null-hypothesis divergence, to the codon bias of those sites. Genes have been partitioned into high and low codon bias groups, and sites are partitioned into two-, fourfold and purine-pyramidine redundant sites.

This possibility is illustrated in Figure 1, which analyzes a number of homologous genes of *E. coli* and *Salmonella typhimurium*. The figure shows the divergence for two- and fourfold redundant amino acids in a set of highly and lowly expressed genes (sixfold redundant amino acids were omitted for simplicity). For the fourfold redundant sites, three values have been calculated: one for sites coded for by C or T in both species, one for sites coded for by A or G, and one for sites coded for by a purine in one species and a pyramidine in the other. In all cases the observed difference has been compared to the null-hypothesis divergence, $2p(1 - p)$, where p is the frequency of the favored nucleotide, in the highly expressed or lowly expressed genes, respectively.

The observed differences are close to the null-hypothesis divergence for the lowly expressed genes, but well below the null-hypothesis divergence for highly expressed genes. The difference is highly significant (mean values, 0.79 ± 0.018 and 0.50 ± 0.025; $p < 0.001$). Hence, either there is site-specific bias in the highly expressed genes, or for some other reason, the approach to saturation is slower in these genes. EYRE-WALKER and BULMER (1995) have suggested that the mutation rate may be lower in highly expressed genes. This would explain why highly expressed genes are further from the null-hypothesis divergence. However, we prefer site-specific bias as an explanation. A reduced mutation rate would not explain why, for two highly expressed genes, *gapA* and *ompA*, different codons are fixed at different sites in 10 genera of enterobacteria, although, as we show below, these genera are at, or close to, saturation. These observations are explained by site-specific bias.

Furthermore, the divergence, relative to the null-hypothesis, is the same for purine *vs.* pyridine as it is for C *vs.* T and A *vs.* G. If saturation has not been reached in the highly expressed genes, we would expect the purine-pyrimidine values to be lower, because the rate of divergence depends on mutation rates and transversions are rarer than transitions (NEI 1987; see MAYNARD SMITH 1994 for evidence that similar differences exist in bacteria). (The full equations for the divergence at a fourfold redundant site with bias and different transition and transversion rates are hard to solve. However, numerical simulation of the equations confirms the obvious expectation that C/T and A/G divergence approach saturation faster than purine-pyrimidine divergence). This observation makes it unlikely that highly expressed genes are further from saturation because of a lower mutation rate. However, we recognize that it would be desirable to analyze the enterobacterial data to see whether transitions are indeed commoner than transversions, as they are (by a factor of 4.5) in Neisseria (MAYNARD SMITH 1994).

Hence the divergence in highly expressed genes is further from the null-hypothesis divergence, and this cannot be explained by a slower rate of approach to saturation. The lower divergence is predicted by site-specific bias and is discussed in more detail in the accompanying manuscript (MAYNARD SMITH and SMITH 1996).

Rate of approach to saturation: Theory also predicts that the approach to saturation should be as fast, or faster, in highly biased genes. It is shown in the appendix that, at a site with two alternatives (C and T say) and no bias

$$d = d_{\max}(1 - e^{-4mt}),$$

where d is the actual divergence at time t, $d_{\max}$ the null-hypothesis divergence, and m the probability per unit time that a C will be substituted by a T, or T by C. If there is bias, with probabilities of substitution a for C → T and b for T → C, the corresponding equation is

$$d = d_{\max}(1 - e^{-2(a+b)t}).$$

It is also shown in the appendix that if $2N_e s < 1$ (where s is the selective advantage of the favored allele), then $a + b \simeq 2m$; if $2N_e s = 1$, then $a + b = 2.16m$; and if $2N_e s \gg 1$, then $a + b \gg 2m$. Hence the rate of approach to saturation is as high or higher in the presence of bias.

Saturation divergence: If one knew the absolute times of divergence between a set of taxa, one could plot the synonymous differences against time and deduce whether they were close to saturation. Usually we do not know the absolute times. However, one can use the numbers of amino acid substitutions between pairs of taxa as estimates of the relative times since divergence from a common ancestor.

Figure 2 shows such an analysis for the *gapA* and *ompA* genes in 10 genera of enterobacteria (MAYNARD SMITH and SMITH 1996). The percentage nucleotide difference at synonymous third sites against numbers of amino acid differences, for the 45 pairwise comparisons of all strains, is shown. For both genes, we have fitted a curve of the form d = $d_{\max}(1 - e^{-kt})$, where $d_{\max}$ and k have been estimated so as to minimize the sum of the squared departures of the observed differences from the curve. The saturation values estimated in this way are 19% for *gapA* and 25.4% for *ompA*.

These values can be compared to values of 39.3 and 38.1%, respectively, predicted by the null-hypothesis divergence. These values were calculated for third sites only, for all codons, assuming the amino acid composition of the two genes in *E. coli*. We assume the codon usage estimated for very highly expressed genes [*i.e.*, the VH category in BULMER (1988)] because single genes, *gapA* and *ompA*, contain too few amino acids to provide reliable estimates.

Although the estimates from the observed data are not precise, they are substantially below the null-hypothesis values. The data for *gapA* are particularly convincing. They show that most of the pairwise divergences have reached saturation at a level about half the null-hypothesis value (Figure 2A). This is most easily explained by site-specific bias. Part of the deficiency between the null-hypothesis divergence and the saturation divergence may be explained by site-specific differences in the intensity of selection for the most favored codon. However, differences in the intensity of selection alone will not explain the fixation of unfavored codons in the *gapA* and *ompA* genes from taxa that are at, or close to, saturation. It seems, therefore, that different codons are favored at different sites within the gene, a conclusion that agrees with the fact that different codons are in fact fixed at different sites across genera.

The estimated values of k are 0.45 for *gapA* and 0.122 for *ompA*. The difference between the estimates may reflect a real difference. As shown in the appendix, k, which is equal to $2(a + b)$, increases sharply with $2N_e s$, provided that $2N_e s > 1$. There is no reason why s, and therefore k, should be the same for different genes.

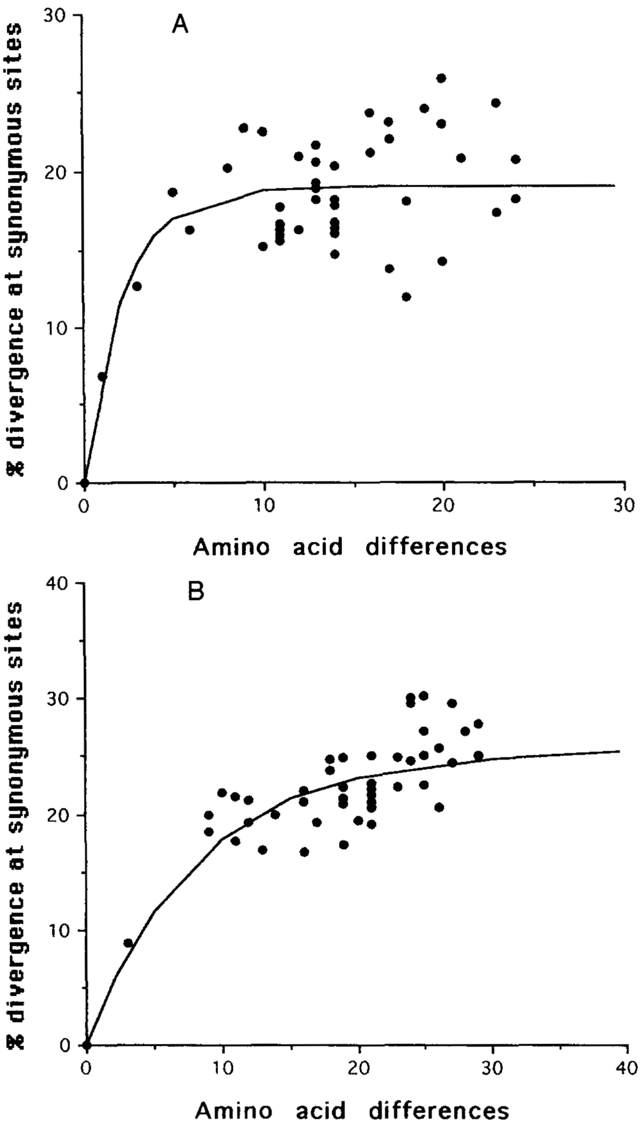

FIGURE 2.—Synonymous divergence of *gapA* and *ompA*. The percentage nucleotide divergence at synonymous third sites compared with the number of amino acid differences for the 45 pairwise comparisons of all strains for *gapA* (A) and *ompA* (B). For both genes, we have fitted a curve of the form $d = d_{\max}(1 - e^{-kt})$, where $d_{\max}$ and k have been estimated so as to minimize the sum of the squared departures of the observed differences from the curve.

DISCUSSION

The predicted nucleotide divergence at synonymous third sites varies greatly according to the assumptions made. For example, for the *gapA* gene in enterobacteria, it is 63.0% if codon bias is ignored, 39.3% if codon bias is allowed for (but assumed to be the same at all sites coding for a given amino acid), and 19.0% from the observed data that presumably include site-specific bias (MAYNARD SMITH and SMITH 1996).

These are three estimates, of increasing plausibility, of the percentage divergence when equilibrium has been reached between forward and backward substitution. Such equilibria, however, are themselves temporary. In time, differences in codon bias or G + C content between diversifying lineages will result in further increases in the percentage difference between synonymous sites.

N.H.S. was supported by a Welcome Trust grant to Professor B. G. Spratt.

LITERATURE CITED

Bulmer, M., 1988 Are codon usage patterns in unicellular organisms determined by selection-mutation balance? J. Evol. Biol. **1:** 15–26.

Maynard Smith, J., 1994 Estimating selection by comparing synonymous and substitutional changes. J. Mol. Evol. **39:** 123–128.

Maynard Smith, J., and N. H. Smith, 1996 Site-specific codon bias in bacteria. Genetics **142:** 000–000.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Sharp, P. M., and W. H. Li, 1987 The codon adaptation index—a measure of directional synonymous codon bias, and its potential applications. Nucleic Acids Res. **15:** 1281–1295.

Communicating editor: W.-H. Li

APPENDIX

Consider a site with two synonymous alternatives, say C and T. There are two independently evolving lineages. Let p, q, and r be the probability, respectively, that both lineages have C, that one is C and one T, and that both have T.

In unit time, the probability that a C lineage will change to T is a, and that a T lineage will change to C is b. Changes are supposed to be instantaneous. If we consider a time interval, dt, short enough that we can ignore two changes in one interval:

$$\mathrm{d}p = (-2ap + bq)\mathrm{d}t$$

$$\mathrm{d}q = (2ap - aq - bq + 2br)\mathrm{d}t$$

$$\mathrm{d}r = (aq - 2br)\mathrm{d}t.$$

Eliminating q and r from the resultant differential equation gives

$$\mathrm{d}^2p/\mathrm{d}t^2 + 3ab\mathrm{d}p/\mathrm{d}r + 2(a + b)^2p - 2b^2 = 0.$$

The solution of this equation is

$$p = [b/(a + b)]^2 + Ae^{-2(a+b)t} + Be^{-(a+b)t}.$$

If we suppose that the two lineages are descended from a single ancestor at time $t = 0$, then $p_o = b/(a + b)$; $r_o = a/(a + b)$; $\mathrm{q}_o = 0$,

and $p = [b/(a + b)]^2 + ab/(a + b)^2e^{-2(a+b)t}$,

and $q = 2ab/(a + b)^2(1 - e^{-2(a+b)t})$.

As $t \rightarrow \propto$, $q \rightarrow 2ab/(a + b)^2$. This is the difference between the two sequences at saturation, $q_{\max}$. Hence

$$q = q_{\max}(1 - e^{-2(a+b)t}).$$

If there is no bias, $a = b = m$, and $q = q_{\max}(1 - e^{-4}mt)$. If we want to know whether bias accelerates or slows down the approach to saturation, we need to know whether $a + b > 2m$.

The probability, in a haploid, that a single mutant will be fixed, is

$$u = S/[N(1 - e^{-S})],$$

where $S = 2N_es$ and s is the selective advantage. If the mutation rates C → T and T → C are both equal to m, then the substitution rates are

$$a = Nmu_1 = mS/(1 - e^{-S}),$$

$$b = Nmu_2 = -mS/(1 - e^{S}), \quad \text{and}$$

$$a + b = mS(e^S - e^{-S})/(e^S + e^{-S} - 2) = \mathrm{Rm}.$$

When $S \rightarrow 0$, $R \rightarrow 2$, and

for	$S = 0.1$	1	10	100
then	$R = 2.004$	2.16	10	100

Hence, if bias is caused by selection, it either hardly alters the rate of approach to saturation ($S < 1$) or accelerates it ($S > 1$).