Citation:
Filipski A, Prohaska S & Kumar S (*2007*) Molecular signatures of adaptive evolution. Evolutionary Genomics and Proteomics edited by Pagel M & Pomiankowski A. *Sinauer Associates, Inc.*, Massachussets, 241-254.

CHAPTER 11

Molecular Signatures of Adaptive Evolution

Alan Filipski, Sonja Prohaska, Sudhir Kumar

Introduction

DARWIN'S GREAT DISCOVERY OF NATURAL SELECTION was rightfully considered to be the fundamental principle of adaptive evolutionary change at both the phenotypic (Fisher 1930) and, later, the DNA sequence level. At the same time, suggestions were made that evolutionary changes could be random as well as adaptive. This was given a mathematical basis by Sewall Wright's description of random evolutionary change in populations in the form of what is now called *random genetic drift* (Wright 1931). The relative importance of genetic drift and natural selection on the evolution of protein sequences was much debated at that time. Because of the belief that most protein change had selective impact and that effective population sizes were so large that random effects would be smoothed out, the effect of genetic drift tended to be underappreciated.

In the 1960s, the genetic code was deciphered and more protein sequences became available; hence, it finally became possible to assess in a quantitative way the amount of genetic variation within and among species. In 1968, the tide began to turn as Motoo Kimura argued, on the basis of the limited sequence data then available, that most protein differences within and among species could be attributed to random genetic drift, and that the differences were not the result of natural selection only (Kimura 1968, 1983). This argument was based on calculations that demonstrated that the observed quantity of protein polymorphism in populations and the inferred rate of evolution between species were far greater than would be expected if selective mechanisms were the primary determinants (Kimura 1983).

Kimura made an early and compelling argument about the relative importance of natural selection and genetic drift; this argument exploited the degeneracy of the genetic code (Kimura 1977). Many DNA sequence changes in certain codon positions do not result in amino acid changes (i.e., they are synonymous changes), which allows them to escape natural selection and accumulate *neutrally* by random genetic drift. This stands in contrast to other codon positions where the vast majority of changes will alter the amino acid encoded (nonsynonymous changes). These changes will often be subject to positive selection, which speeds up the rate at which they proceed to fixation in a population, or to what is often called purifying selection, which removes them from the population. If the mutations are fixed by positive selection, then the rate of fixation of nonsynonymous changes will be greater than the rate of fixation (substitution) of synonymous changes in any codon.

In contrast, Kimura's neutral theory predicts that the evolutionary rate of synonymous DNA substitutions will be greater than that of nonsynonymous substitutions, because purifying selection is operating to remove the vast majority of nonsynonymous mutations. This prediction was confirmed (Kimura 1977, 1991). Over time, these observations were extensively supported by growing sequence data sets, which gradually led to the acceptance of the neutral theory.

In comparing the rate of synonymous and nonsynonymous substitutions in the same protein, Kimura implicitly assumed that the observed substitution rate of synonymous codon sites may be used as an estimate of mutation rate for the entire gene (Kimura 1977, 1983). Some form of this estimation method has been virtually universal since then in studies relating to adaptation at the DNA level (Miyata and Yasunaga 1980; Nei and Kumar 2000; Bustamante et al. 2005; Nielsen et al. 2005a). The common modus operandi in these tests of selection is to estimate synonymous divergence per synonymous site (denoted d_S or K_s), and nonsynonymous divergence per nonsynonymous site (d_N or K_a) for any protein-coding gene, and then to derive the selection ratio $\omega_S = d_N/d_S$. Many methods have been developed for estimating d_S and d_N based on somewhat different assumptions, but the principle remains the same (see overviews in Nei and Kumar 2000; Yang 2006). If ω_S is significantly greater than 1.0, then we attribute this excess to positive selection on the gene. If ω_S is significantly less than 1.0, then we attribute the difference to negative (purifying) selection.

For a vast majority of proteins, ω_S is found to be less than 1.0 when it is computed as an average over all the codons in a protein. This is because most of the nonsynonymous mutations will have negative fitness effects and will be eliminated by selection. Even in the presence of positive selection on some codons, ω_S will generally be less than 1.0 because of the purifying selection against most of the nonsynonymous mutations. For this reason, estimating ω_S for each codon separately is advocated (e.g., Nielsen and Yang 1998; Suzuki et al. 2001; Suzuki and Nei 2001; Zhang et al. 2005). A recent account by Nei (2005) provides an excellent overview of many historical and recent studies, where scientists have looked for genes and codons that have undergone positive selection (Darwinian evolution) at the molecular level.

Even though d_N/d_S and K_a/K_s have been used historically for referring to the extent of natural selection, it is most accurate to define the selection ratio (ω) as the ratio of the rate of nonsynonymous *substitutions* per site to the rate of nonsynonymous *mutations* per site. That is, $\omega = (d_N/2t)/\mu_N$, where μ_N is the rate of nonsynonymous mutations per nonsynonymous site, and t is the time of species divergence.

Of course, the rate of nonsynonymous mutation is difficult to determine directly, and synonymous divergence is often used as a proxy for the denominator. In this case, ω becomes identical to ω_S. A number of authors have also used sequence divergence in introns (d_I) and in pseudogenes (d_φ) in the denominator, which makes ω_I and ω_φ, respectively, equal to ω (Li and Tanimura 1987; Wolfe et al. 1989; Li and Graur 1991; Bergström et al. 1999; Keightley and Eyre-Walker 2000; Nachman and Crowell 2000b; Chen et al. 2001).

In the following, we discuss various factors that often invalidate the procedure of equating the rate of synonymous, intron, and pseudogene divergence with the rate of nonsynonymous mutations to estimate ω_S, ω_I, and ω_φ, respectively.

Codon Usage Bias and the Estimates of Selection Ratio (ω)

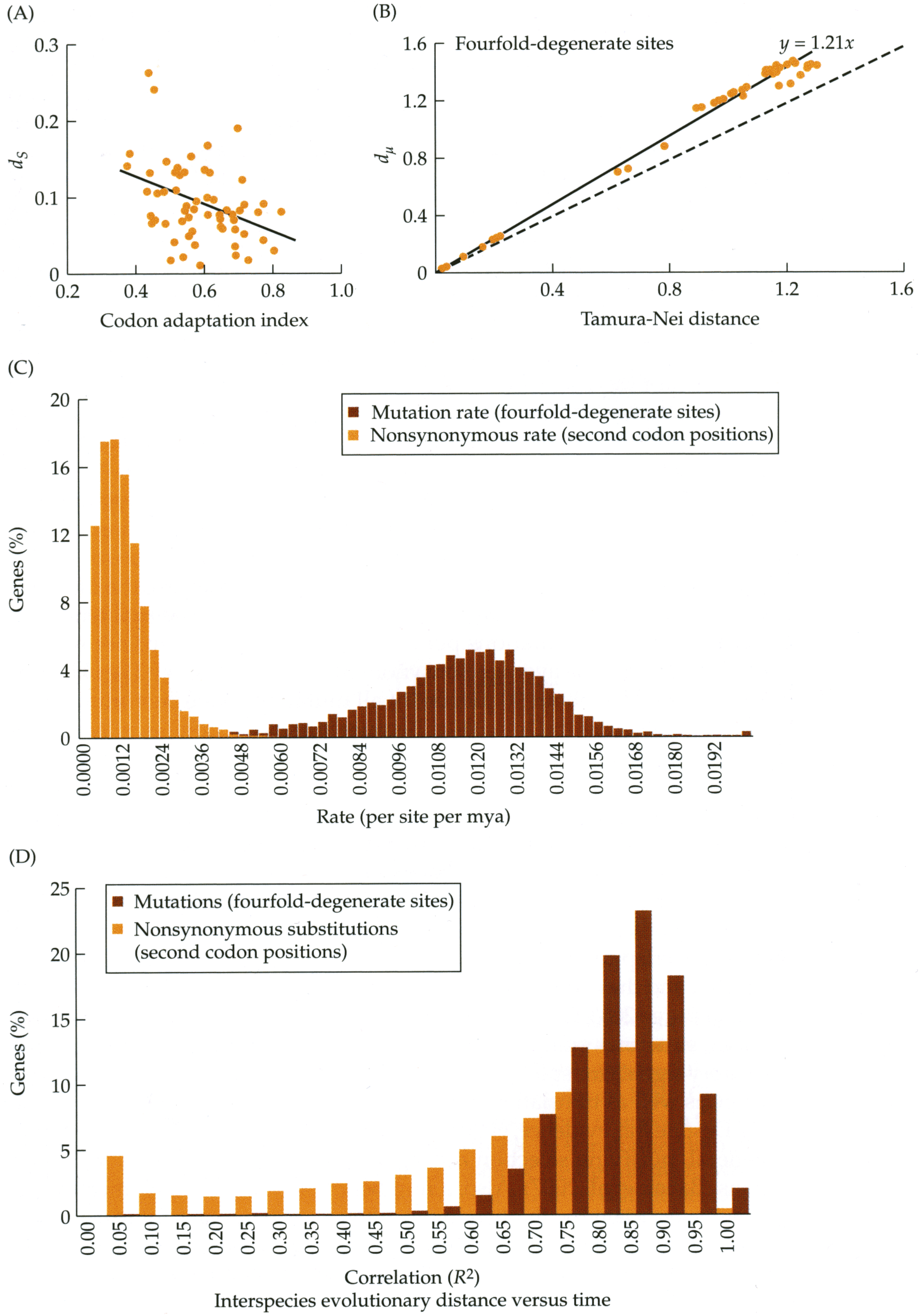
We begin with the assumption that the rate of synonymous mutations (μ_S) is identical to the rate of nonsynonymous mutations (μ_N). In this case, if all synonymous mutations are "strictly neutral," then we can directly use the synonymous divergence per synonymous site (d_S) in estimating the selection ratio ($\omega = \omega_S$). In the 1970s, it was thought that the synonymous mutations were all indeed strictly neutral, and, therefore, directly useable for estimating ω (Kimura 1977). However, recent comparisons of the human and chimpanzee genomes have established that the synonymous codons for a given amino acid appear with frequencies more biased than expected based on base composition differences (Chamary and Hurst 2005; Lu and Wu 2005; Parmley et al. 2006). This property has been seen in many species (Grantham et al. 1980; Bennetzen and Hall 1982; Blake and Hinds 1984; Eyre-Walker 1991; Akashi 2001; Mikkelsen et al. 2005).

Why certain codons are preferentially used is not completely understood, but reasons may include mutational bias, local GC content, and translational efficiency due to varying availability of tRNA for different codons. While mutational and compositional bias can explain some of the observed nonrandom synonymous codon usage, it is now abundantly clear that many synonymous mutations are eliminated by natural selection for, among other reasons, optimizing translational efficiency (e.g., Shields et al. 1988; Comeron et al. 1999; Carpen et al. 2006; Kimchi-Sarfaty et al. 2007).

Any existence of purifying selection on synonymous mutations compromises the assumption that the substitution rate at synonymous positions can be used directly as an estimate of mutation rate (Shields et al. 1988; Ochman et al. 1999; Tamura et al. 2004). A well-studied example of codon usage bias is from the genus *Drosophila*. In *D. melanogaster*, for example, the codon CTG for leucine is more than eight times as frequent as the codon TTA for the same amino acid. In *Drosophila*, synonymous distance (d_S) between orthologous genes correlates with the codon adaptation index (Tamura et al. 2004; Figure 11.1A). This is evidence that the codon usage bias constraint acts as negative selection. The relative effect of codon usage bias on the estimate of mutation rate from synonymous sequence divergence can be seen by comparing codon bias–corrected estimates of synonymous dis-

Figure 11.1 (A) Negative relationship of codon adaptation index and the number of substitutions per fourfold-degenerate site (d_S) between *Drosophila melanogaster* and *Drosophila simulans*, based on a study of 62 genes. ($R^2 = 0.148$). As codon constraint increases, synonymous rate goes down because more mutations become eliminated by purifying selection. (B) Effect of codon usage bias correction at the genomic level in *Drosophila*. Each plotted point represents a pair of species in the genus *Drosophila*. Distance between species was computed in two different ways: Tamura-Nei distance between orthologous genes from the species pair, averaged over all genes; and d_μ (codon usage bias–corrected Tamura-Nei distance between orthologous genes from the species pair, averaged over all genes). The dotted line shows identity and the solid line shows a linear best fit regression line. This is based on 9850 coding genes that have orthologs in all of the following twelve species: *D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis*, and *D. grimshawi*. See http://rana.lbl.gov/drosophila/wiki/index.php/Datasets for more information about the gene set. (C) Histogram showing the distributions of mutation rates and nonsynonymous substitution rates in 9850 genes. Mutational distances were estimated by correcting the evolutionary distance at the fourfold-degenerate sites for the effect of codon usage bias and differences in base composition biases between sequences under a sophisticated model of nucleotide substitution (Tamura and Kumar 2002; Tamura et al. 2004). Nonsynonymous distances at the second codon positions were estimated using the Tamura and Kumar method for each gene (Tamura and Kumar 2002). For each gene, the mutational and nonsynonymous distances were regressed against the evolutionary time for each pair of species. Divergence times were estimated using the genomic mutation distance and a clock calibration of 0.011 mutations per site per million years (Tamura et al. 2004). Per-gene rates were computed as follows: Using the above interspecies times as the independent variable (x), we calculated regressions against evolutionary distances (either for fourfold-degenerate or second position) as the dependent variable obtained from individual genes. The best fit slope (evolutionary rate) for that gene was estimated via a regression forcing the y-intercept to zero. (D) Histogram showing the extent of linearity of rate with time for both second codon position and fourfold-degenerate differences, expressed as the linear coefficient of determination, R^2. This was done by taking, for each gene, all species pairs available for that gene, and computing a linear regression of the interspecies distance (second codon position or fourfold-degenerate) against interspecies times as the independent variable. Because of known problems in calculating and interpreting the R^2 coefficient of determination when the y-intercept is forced to zero, we calculated the R^2 for each regression without constraining the y-intercept to be zero (Eisenhauer 2003). (A, after Tamura et al. 2004.)

▶


(A)
0.3
0.2
0.1
0
d_S
0.2
0.4
0.6
0.8
1.0
Codon adaptation index
(B)
1.6
Fourfold-degenerate sites
$y = 1.21x$
1.2
0.8
0.4
0
d_μ
0.4
0.8
1.2
1.6
Tamura-Nei distance
(C)
20
16
12
8
4
0
Genes (%)
🟫 Mutation rate (fourfold-degenerate sites)
🟧 Nonsynonymous rate (second codon positions)
0.0000
0.0012
0.0024
0.0036
0.0048
0.0060
0.0072
0.0084
0.0096
0.0108
0.0120
0.0132
0.0144
0.0156
0.0168
0.0180
0.0192
Rate (per site per mya)
(D)
25
20
15
10
5
0
Genes (%)
🟫 Mutations (fourfold-degenerate sites)
🟧 Nonsynonymous substitutions (second codon positions)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Correlation (R^2)
Interspecies evolutionary distance versus time

tances (corrected-d_S, which is referred to as d_μ as well) with those that only correct for multiple substitutions under sophisticated models of nucleotide substitutions (e.g., Keightley and Eyre-Walker 2000; Tamura et al. 2004).

A comparison of the estimates of selection intensities with and without correcting for codon usage bias for 9850 protein-coding genes from 12 *Drosophila* species provides a glimpse into the overall effect of codon usage bias on estimates of selection. Figure 11.1B shows that over all genes the bias correction increases the estimate of d_S by slightly more than 20 percent. Because d_S occurs in the denominator in the selection ratio, the failure to correct for codon usage bias will cause false positives when looking for genes and codons with positive selection. In the present data set, the 20 percent difference represents an average over all genes. However, much more dramatic effects are seen when we examine individual genes. For example, application of the correction to the *Adh* gene in the subgenus *Sophophora* (of genus *Drosophila*) approximately doubles the estimate of the mutation rate, which translates into much higher purifying selection on *Adh* than estimated based on synonymous divergences (Tamura et al. 2004).

Figure 11.1C shows the distribution of the rates of nonsynonymous and corrected-d_S for 9850 fruit fly genes. As expected, the distribution of nonsynonymous rates is highly skewed, with a large number of proteins evolving slowly, and only a few evolving with fast rates. In contrast, the distribution of mutation rates is symmetrical and shows a strong central tendency (average rate = 11.3 mutations per thousand base pairs [kbp] per million years [my]). Using the average nonsynonymous divergence (1.2 per kbp per my), it is clear that about 90 percent of all mutations have been eliminated by natural selection in *Drosophila*. Because different proteins may evolve with different relative evolutionary rates in the same set of species, we also present the correlation of nonsynonymous divergence with estimates of divergence times for the 12 *Drosophila* species (Figure 11.1D). Many proteins show rather low correlation between the nonsynonymous divergences with time, confirming a nonclocklike behavior.

In addition to the codon usage bias, synonymous mutations may be under selection for other reasons. For example, conserved RNA structures have been shown to overlap coding regions, and the functional relevance of both the structural RNA and the encoded protein has been determined (Konecny et al. 2000; Katz and Burge 2003; Chooniedass-Kothari et al. 2004; Meyer and Miklos 2005). The local structures on the mRNA in these cases serve as post-transcriptional regulatory signals for splicing, transport, and mRNA stabilization or destabilization, as well as translation efficiency (Cartegni et al. 2002; Chamary et al. 2006). Despite this evidence, genomic screens for structural RNAs suggest that, in general, purifying selection acts against secondary structures in coding regions (Babak et al. 2007). Also, a number of synonymous polymorphisms have been associated with phenotypic differences in humans for reasons that are not always clear (Oeffner et al. 2000; Carpen et al. 2006; Kimchi-Sarfaty et al. 2007). It is currently unclear how the effects of these (potentially minor) factors can be handled efficiently in correcting estimates of d_S.

Hypermutability of CpG Dinucleotides and the Estimates of Selection Ratio

In the above discussion, we assumed that the rate of synonymous mutations (μ_S) is identical to the rate of nonsynonymous mutations (μ_N). If this is not true, then synonymous divergence, even though estimated perfectly, will not be directly applicable for estimating μ_N. In vertebrates, and in nonanimal species, the existence of hypermutable CpG dinucleotides (a cytosine followed by a guanine on the same DNA strand; Figure 11.2) has been shown to cause a disparity between μ_S and μ_N (Subramanian and Kumar 2006).

Codon positions involved in CpG dinucleotides are expected to mutate at between five to twenty times the rate of other positions in the codon (Krawczak et al. 1998; Bird 1999; Subramanian and Kumar 2003), because CpG dinucleotides are often methylated in coding regions and are known to mutate rapidly to TpG and CpA. In this case, assuming that all CpG positions are likely to be methylated with equal probability, the rate of synonymous and nonsynonymous mutations in a given protein will only be equal,

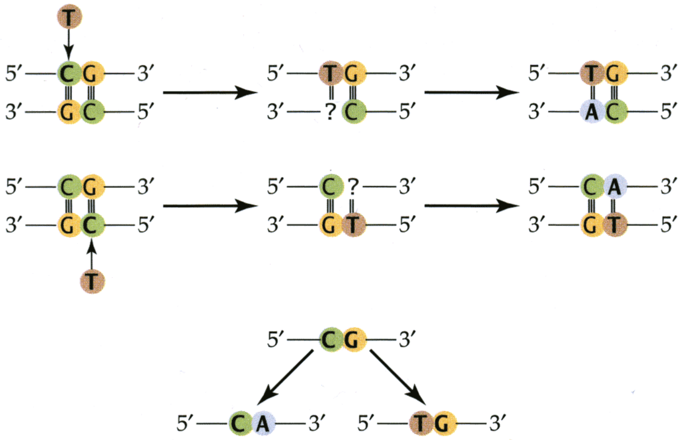


Figure 11.2 Hypermutability of CpG dinucleotides. The nucleotide sequence CpG is subject to mutation at a rate approximately ten times faster than other point mutations in mammals (Bird 1980; Krawczak et al. 1998; Anagnostopoulos et al. 1999). The methylation and subsequent mutation of the cytosine in the CpG configuration to thymine via spontaneous deamination can lead to either a C→T or G→A transitional mutation and thus cause the CpG to be transformed into either a CpA or a TpG. In the upper sequence, the cytosine of the CpG in the forward strand mutates to a thymine. The resulting mismatch with its complementary nucleotide is resolved by exchanging that complementary nucleotide for an adenine. In the lower sequence, the cytosine on the complementary DNA strand mutates to a thymine, and the resulting mismatch is again repaired by replacing its counterpart in the forward strand with an adenine.

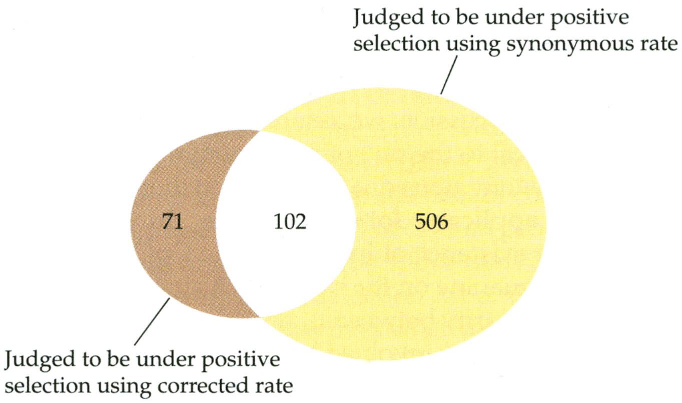

Judged to be under positive selection using synonymous rate
71
102
506
Judged to be under positive selection using corrected rate

Figure 11.3 Venn diagram showing the numbers of human genes under adaptive evolution as predicted by two different methods (silent divergence and estimated replacement mutation rate) to estimate the coefficient of selection. The numbers in the overlapping area indicate the genes that were identified by both the methods, and the numbers to the left and the right represent the genes that were predicted exclusively by one of the methods. (After Subramanian and Kumar 2006.)

on average, if the number of CpG sites per synonymous site is equal to the number of CpG sites per nonsynonymous site. This is unlikely to be true because the fraction of second codon positions involved in CpG dinucleotides will be greatly influenced by the amino acid frequencies in specific proteins. In contrast, the involvement of fourfold-degenerate positions in CpG dinucleotides is dependent on the mutational pattern of the genomic region, which dictates the G + C content of these sites, and the nucleotides in the flanking positions.

Incorporating the difference in the proportion of synonymous and nonsynonymous sites participating in CpG dinucleotides drastically changes the inferences about the proportion of genes predicted to be under positive selection (Subramanian and Kumar 2006). Figure 11.3 shows that, in a sample of human genes, the number judged to be under positive selection decreases to less than a third of its previous value when the CpG correction is applied. While the affects of CpG may be correctable when substitution rates are estimated over an entire polypeptide, the future looks unpromising when we begin to apply these concepts to the identification of individual codons under selection. We now need to know which CpG-involved codon positions are methylated in the germ line, and we need to ascertain whether the homologous codon positions in different species have the same methylation patterns. In the absence of any such information, we will not be able to reliably estimate nonsynonymous mutation rates for use in finding codons that have undergone positive selection.

Using Pseudogenes to Estimate Nonsynonymous Mutation Rates

Less than 25 percent of codon positions can experience synonymous mutations, and codons make up only about 2 percent of the genomes of higher organisms. Therefore, scientists have looked beyond the codons to generate estimates of nonsynonymous mutation rates, as mutations in the noncoding regions are considered to be not selected against (although this is now known to be an oversimplification; e.g., Andolfatto 2005). One possibility is to consider "dead" genes (pseudogenes), because these genes are expected to evolve without any selective pressures. For example, comparison of human and chimpanzee pseudogenes was used to estimate the mutation rate for these primates (Nachman and Crowell 2000b). Because pseudogenes are found to show higher divergence, they are sometimes considered to evolve with strict neutrality more readily than synonymous sites (e.g., Bustamante et al. 2002).

However, the hypermutable CpG positions also cause problems when using pseudogenes. In coding DNA, the CpG content is a function of the amino acid frequencies and codon usage; in noncoding DNA, this proportion is instead determined by GC content and the balance between mutations that establish CpG and those that disrupt it (Hwang and Green 2004; Fryxell and Moon 2005). This can vary considerably by organism and by location in the genome. For most vertebrates, between two and seven percent of the total cytosine nucleotides in the genome are methylated and subject to the acceleration of mutation rate caused by CpG hypermutability (Razin and Riggs 1980; Colot and Rossignol 1999). Furthermore, DNA methylation patterns are not distributed evenly over the genome. They tend to be targeted to mobile elements for host genome defense and to other regions for long-term silencing.

Moreover, pseudogenes have now been shown to evolve with very fast rates soon after they lose function, and with decreasing rates as further time passes (Figure 11.4). This nonlinear rate of evolution occurs because the CpG

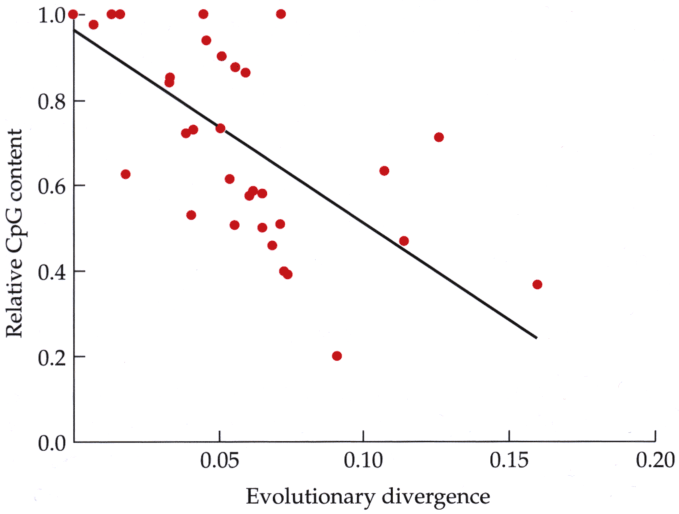


Figure 11.4 Relationship between evolutionary divergence and the ratio of CpG contents in pseudogenes and their functional counterparts, based on an analysis of 39 human pseudogenes. CpG content is seen to drop linearly as evolutionary divergence increases ($R^2 = 0.52$; $P < 0.01$). (After Subramanian and Kumar 2003.)

content at first and second codon positions of the pseudogene decays due to CpG→TpG and CpG→CpA mutations over time (Subramanian and Kumar 2003). Also, the rapid divergence rate makes the alignment of pseudogenes and functional genes difficult. Another drawback of using pseudogenes is that they may not be good indicators of mutation rate for a given protein, because they may reside in another part of the genome that has a different local mutation rate and pattern (Matassi et al. 1999).

Rate of Intron Divergence as a Proxy for Nonsynonymous Mutation Rates

In eukaryotes, comparison of homologous intron sequences is used as another candidate for estimating nonsynonymous mutation rates, because they are putatively neutral regions in proximity to the coding portion of a gene and not subject to cryptic selective pressures, especially if they are not located near exon boundaries (Castresana 2002). In fact, the mutation rate inferred from introns is less than the rate obtained from synonymous substitutions in mammals (Hoffman and Birney 2007), which is expected given that a smaller fraction of positions are involved in CpG dinucleotides in introns (Subramanian and Kumar 2003).

The mean observed (extrinsic) substitution rate within introns of a gene is a good estimate of the prevailing (intrinsic) mutation rate for the coding sequence of the gene if we assume that all intronic sites considered are strictly neutral. While restricting extrapolation of mutation rates from the introns to the exons of the same gene is more conservative than extrapolating over greater distances in the genome (Wolfe et al. 1989), it is clear that the GC content of third codon positions is significantly higher than that of introns (Bernardi 1986; Eyre-Walker 1991; Hughes and Yeager 1998). This would invalidate the assumption of equality of mutation rates between introns and exons.

Most recently, sequence divergence in introns and intergenic regions has been the preferred method to establish mutation rates for replacement positions in codons in human and chimpanzee lineages (Hellmann et al. 2003; Lu and Wu 2005; Mikkelsen et al. 2005; Parmley et al. 2006). However, the evolutionary divergence at introns (d_I) cannot always be directly equated with the rate at which mutations occur at codon positions that can experience nonsynonymous mutations. This is because nonsynonymous positions are found to be involved in CpG configurations twice as often as intronic sites. Furthermore, the per-gene frequency distribution of CpG content of introns is strikingly different from the distribution of CpG content of amino acid replacement positions, with the latter having a much higher CpG density (Figure 11.5A; Subramanian and Kumar 2006). Using the proportions of CpG contents in sites that experience nonsynonymous CpG mutations, rates can be corrected to account for the differences in CpG content. We previously have taken such an approach to compare two closely related species, human and chimpanzee, and found that the number of genes predicted to

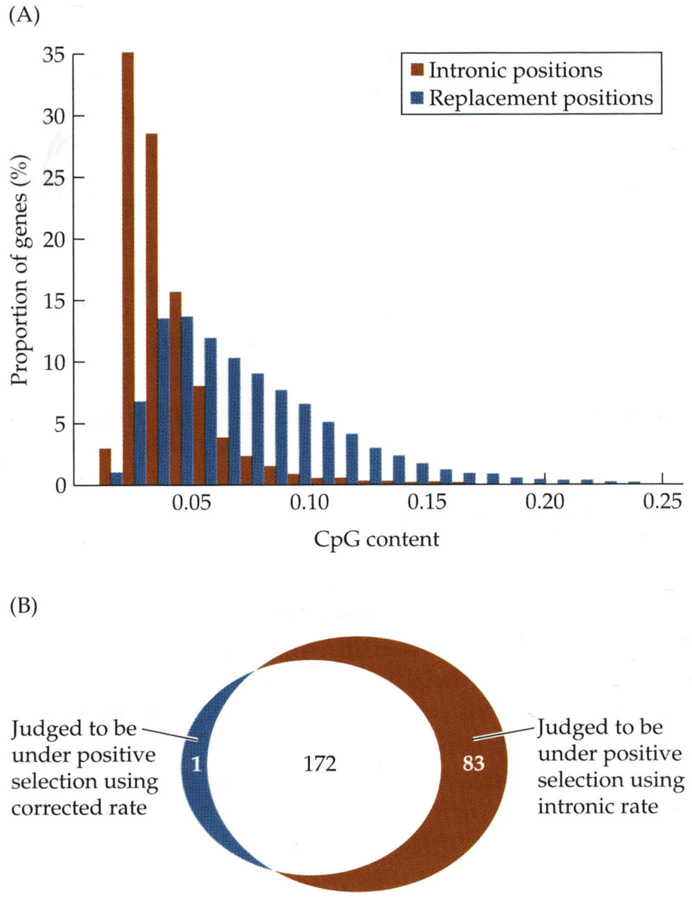


Figure 11.5 (A) The differential distribution of the fraction of intronic and replacement positions involved in CpG dinucleotide configurations from 10,196 functional human genes containing at least one intron. On average, the replacement positions involved in CpG configurations are two times higher than those of intronic positions (6.2% and 2.9%, respectively). The dispersion indices (the ratio of variance to mean) of the distributions of intronic and replacement positions are 0.015 and 0.024, respectively. (B) Venn diagram showing the numbers of common human genes inferred to be under adaptive evolution (ω>1), where ω estimated using intron distances is compared with the estimates derived using intrinsic mutational distances at nonsynonymous sites. The numbers in the overlapping area indicate the genes that were identified by both methods, and the numbers in the individual circles represent the genes that were predicted exclusively by one of the methods. (Data from Subramanian and Kumar 2006.)

be evolving with purifying selection changes considerably. This is shown in Figure 11.5B, which indicates that testing genes for selection based on CpG-adjusted versus unadjusted estimates of nonsynonymous mutation rates leads to both false positives and false negatives, primarily the former.

There are other potential problems with using introns as neutral regions. For one, introns may contain unknown conserved regulatory sites or RNA genes under purifying selection. For example, snoRNAs (small nucleolar RNAs) seem to be located exclusively in introns (Huang et al. 2005). Significant intron sequence conservation has also been detected in regions near exons (Hare and Palumbi 2003; Louie et al. 2003). Also, constraints from RNA structures (e.g., self-splicing introns), may also be sources of purifying selection. Similar to pseudogenes, the alignment of introns is more difficult than for exons, especially for more divergent species, and misalignment can have a significant effect on distance estimates. It has also been found that intronic rates differ more from species to species than synonymous rates; hence, the intron rates cannot be counted upon to be constant (Hoffman and Birney 2007).

How much difference is there between estimates of mutation rate derived from synonymous differences and those derived from presumably neutral intronic sites? This would depend on the extent of codon usage bias and the CpG density in different regions (among other factors). For 15,176 human–dog gene pairs, the median value of d_S (uncorrected for codon usage bias) was found to be 0.370, while median distance between intronic sites (d_I) was 0.305. For 16,183 mouse–rat orthologs, the distances are 0.212 and 0.158, respectively. Although correlations between the estimates of neutral distance are substantial (Spearman rank-order correlation coefficient of 0.57 for human–dog and 0.46 for mouse–rat), the two methods identify substantially different sets of genes as being under selection. In both the mouse–rat and human–dog cases, the sets of genes identified as being in the top 5 percent in terms of selection ratio do not overlap more than 65 percent between the two species (Hoffman and Birney 2007).

Direct Estimates of Mutation Rates in the Laboratory

Given the problems with comparative genomics and confounding factors, it is natural to ask why we don't just estimate mutation rates directly in the laboratory. Despite the inherent difficulties of large genomes and long generation times, some progress has been made in laboratory estimation of mutation rates for higher organisms. The first accurate experimental determination of a genome-wide mutation rate for a eukaryote (*Caenorhabditis elegans*) was only recently accomplished, which yielded a point mutation rate of 8.6×10^{-7} mutations per site per year. For reasons not yet completely clear, this estimate is considerably higher (by a factor of ten) than previous estimates using indirect methods (Denver et al. 2004). A similar method was used in 2007 for Drosophila, producing an annual rate of 5.8×10^{-8} (95% confidence interval 2.1×10^{-8} to 13.1×10^{-8}) under the assumption of 10 generations per year (Haag-Liautard et al. 2007). This rate is about 5 times higher than indirect estimates (Tamura et al. 2004).

Statistically significant variation was found among specific lineages within Drosophilids, however, with the slowest rate being 2.7×10^{-8}

(95% confidence interval 1.2×10^{-8} to 5.4×10^{-8}), and the fastest 11.7×10^{-8} (5.9×10^{-8} to 20.6×10^{-8}; Haag-Liautard et al. 2007). There are a number of additional difficulties with direct estimation: the cost and amount of time and data required for each species, the infeasibility of even maintaining mutation accumulation lines for many species, wide confidence intervals, uncertainty as to whether different generation times in the laboratory and in the wild affect the estimate, and the question of whether the sampled data are representative of the rest of the genome. Also, even if direct observation of mutation rate were easy to perform, it would give us only a snapshot of the rate for a species at the present time in the laboratory. The amount of error incurred by applying such rates to natural populations in the phylogenetic past is unknown.

Conclusion

Knowledge of mutation rates is the key not only to finding genes under selection, but to many other analyses, including detection of regulatory regions in genomic sequence data, identification of particular codons responsible for adaptive changes, and estimation of the number of deleterious mutations (Nei and Kumar 2000; Yang and Bielawski 2000; Nekrutenko et al. 2003; Yampolsky et al. 2005). For example, we may compute the number of deleterious mutations per diploid genome per generation (*U*) from the mutation rate, in order to examine whether the purging of synergistically interacting deleterious mutations can explain the maintenance of sex in a species. If *U* exceeds one, then the beneficial effect of sex in removing them can outweigh the cost to each organism of diluting its genetic material with that of a sex partner (Keightley and Eyre-Walker 2000). Understanding the strength of selection is also important, for example, in molecular clock analyses, since neutral substitution rates are expected to be less variable over lineages (over time) than nonsynonymous substitution rates (see Figure 11.1D).

How can we know the rate at which mutations occur in DNA? Mutation is a complex biochemical process that can be caused by inexact replication, chemical mutagens, or ionizing radiation (Brown 2002; Griffiths 2002). Mutation rates are further modulated by various repair mechanisms of differing efficacy (Brown 2002; Friedberg and Friedberg 2006). As a result, mutation rates may vary significantly among major taxa and within the same genome. They are not easy to determine experimentally in eukaryotes, because the rates are often so low that it takes a large amount of data and time to observe them in living systems.

Interspecific DNA comparison is a powerful tool for the estimation of the amount and type of selection on individual genes. Methods for doing this rely on accurate estimates of underlying mutation rates for the genes. This rate can vary from one lineage to another, from time to time, and over different regions of a single genome. Indirect methods for estimating mutation rate are based on the neutral theory and equate mutation rate with observed neutral substitution rate. The problem then becomes the identification of, and correction for, the effect of selective factors affecting putatively neutral substitutions. These factors include codon usage bias (in the case of synonymous changes in coding DNA) and CpG content (in the case of DNA in introns or pseudogenes). There are no doubt other factors of unknown magnitude to consider, but inroads have been made in correcting for these first two, and the resulting estimates of genes under selection are significantly affected.

Acknowledgments

We thank Kristi Garboushian for editorial support, Sven Findeiss for assistance with data preparation, and Wayne Parkhurst for drawing figures. This work was supported by a research grant from the National Institutes of Health (S. K.).