The Positive Correlation between $dN/dS$ and $dS$ in Mammals Is Due to Runs of Adjacent Substitutions

Nina Stoletzki*\(^1\) and Adam Eyre-Walker\(^1\)

\(^1\)Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom

*Corresponding author: E-mail: nstoletzki@googlemail.com.
Associate editor: John H McDonald

Abstract

A positive correlation between $\omega$, the ratio of the nonsynonymous and synonymous substitution rates, and $dS$, the synonymous substitution rate has recently been reported. This correlation is unexpected under simple evolutionary models. Here, we investigate two explanations for this correlation: first, whether it is a consequence of a statistical bias in the estimation of $\omega$ and second, whether it is due to substitutions at adjacent sites. Using simulations, we show that estimates of $\omega$ are biased when levels of divergence are low. This is true using the methods of Yang and Nielsen, Nei and Gojobori, and Muse and Gaut. Although the bias could generate a positive correlation between $\omega$ and $dS$, we show that it is unlikely to be the main determinant. Instead we show that the correlation is reduced when genes that are high quality in sequence, annotation, and alignment are used. The remaining—likely genuine—positive correlation appears to be due to adjacent tandem substitutions; single substitutions, though far more numerous, do not contribute to the correlation. Genuine adjacent substitutions may be due to mutation or selection.

Key words: positive correlation, $dN/dS$, $dS$, adjacent substitutions, bias $dN/dS$.

Introduction

In protein-coding sequences, we can differentiate between synonymous and nonsynonymous substitutions. If we assume that nonsynonymous mutations are either neutral or deleterious, then the rate of nonsynonymous substitutions, $dN$, is $fu$, where $u$ is the nucleotide mutation rate and $f$ is the proportion of nonsynonymous mutations that are neutral. Although it is known that selection can operate on synonymous sites, even in mammals (Chamary et al. 2006; Resch et al. 2007), their evolution is often assumed to be neutral and their rate, $dS$, would then become a measure of the mutation rate ($dS = u$). Under this assumption, the ratio of the nonsynonymous over the synonymous substitution rate ($\omega$) is a measure of the proportion of mutations that are neutral, $f$. Because both, $dN$ and $dS$, depend upon the mutation rate, they are expected to be positively correlated.

However, it has been known for some time that the correlation between $dN$ and $dS$ is stronger than one would expect under this simple evolutionary model (Ohta and Ina 1995; Smith and Hurst 1999; Smith et al. 2003). Furthermore, Wyckoff et al. (2005) and Vallender and Lahn (2007) have recently shown that their ratio, $\omega$, is positively correlated to $dS$ between several species pairs. Although the correlation is not always very strong between some species pairs (Liao and Zhang 2006), may be negative (Studer et al. 2008; Li et al. 2009), and can depend upon the method used to estimate $\omega$ and $dS$ (Li et al. 2009), the cause of the correlation remains unclear. The positive correlation between $\omega$ and $dS$ in mammals is quite unexpected under simple evolutionary models because it would appear to imply that the strength of natural selection acting upon a protein, which is measured by $\omega$, is correlated to the mutation rate, as measured by $dS$.

Wyckoff et al. (2005) suggest a number of explanations by which such a correlation between the strength of selection upon a protein and the mutation rate could arise. First, an increase in the mutation rate might decrease the effective population size of a genomic region through genetic hitchhiking and background selection. This reduction in the effective population size would allow more slightly deleterious mutations to fix, increasing $\omega$. Second, the correlation could arise through compensatory evolution because under such a model, the rate of intragenic compensatory substitutions would depend on the square of the gene’s mutation rate (Kimura 1985). Third, and most intriguingly, the mutation rate of a gene might be adapted to the characteristics of that gene; for example, if a gene is very important, then there may be selection to reduce the mutation rate of that gene; although it should be noted that such selection would be very weak and may not be effective in the face of genetic drift.

We consider two other potential explanations for the correlation between $\omega$ and $dS$. First, there is a yet unconsidered statistical concern: when DNA sequences are very similar, the rates of nonsynonymous and synonymous substitutions will be small and noisy and their ratio, $\omega$, will hence be difficult to estimate without bias. Such a bias may induce an artificial positive correlation between $\omega$ and $dS$. Second, we note that if the correlation is genuine and not a statistical artifact, it does not necessarily imply that the rate of mutation is correlated to the selective constraint of the protein as suggested by Wyckoff et al. (2005). The positive correlation may be due to any connection between the processes affecting synonymous and nonsynonymous sites. Adjacent substitutions may represent the most obvious connection. As adjacent substitutions will
affect proportionally higher numbers of nonsynonymous sites than single substitutions (unless evolution is neutral), sufficient variation of their relative numbers across the genome may cause a positive correlation between $\omega$ and $dS$. In particular, there is evidence that adjacent nucleotides can mutate simultaneously (Averof et al. 2000). This will generate a correlation between $\omega$ and $dS$ because some synonymous mutations will be effectively selected by their linkage to a nonsynonymous mutation.

We show in this study, first that $\omega$ estimates are generally biased when sequence divergence is low. This bias can cause a positive correlation between $\omega$ and $dS$. However, by restricting our analyses to genes in which the statistical bias is minimal, we show that this bias is not the primary reason for the correlation. Second, we show that the positive correlation between $\omega$ and $dS$ is entirely dependent upon runs of two and more adjacent substitutions, whereas single substitutions do not contribute. This indicates that the correlation is not between the mutation rate of single point mutations and selective constraint as suggested by Wyckoff et al. (2005). Adjacent substitutions may be due to errors in sequencing, alignment, and exon assignment but also due to mutational processes (i.e., doublet mutations) or selection.

**Material and Methods**

**Data**

We used human–chimp–macaque–mouse–rat–dog alignments kindly provided by Schneider et al. (2009). To investigate the potential effect of errors in sequencing, alignment, and annotation, we investigated the correlation between $\omega$ and $dS$ in the whole data set of Schneider et al. (2009; 9,942 genes), and in two restricted data sets, they provide: first, the 2,890 genes that are longer than 200 codons and with less than 5% length difference between homologues) and second, genes within the first group that additionally had high trace sequencing coverage (excluding humans for which this information was not available), known annotation status, and 100% alignment HoT scores; this yielded 1,514, 137, and 49 genes for mouse–rat, human–macaque, and human–chimp, respectively.

**Substitution Rates**

In general, rates of synonymous and nonsynonymous substitutions ($dS$ and $dN$) are estimated by dividing the observed number of synonymous and nonsynonymous substitutions per gene ($DS$ and $DN$) by the number of synonymous and nonsynonymous sites ($S$ and $N$), that is, $dS = DS/S$ and $dN = DN/N$. Many different methods exist to estimate substitution rates. They can be divided into heuristic counting methods, such as Nei and Gojobori (1986; NG) and maximum likelihood (ML) methods, such as the methods of Goldman and Yang (1994, GY) and Muse and Gaut (1994, MG); they can also be divided into those that define the number of synonymous sites as “mutational opportunities” and those that define sites “physically” (for discussion on the definition of sites, see Bierne and Eyre-Walker 2003 and Yang 2006).

The effect of the method used to estimate $dS$, particularly with respect to the correlation between $dS$ and GC content, has been extensively discussed before (Smith and Hurst 1999; Bielawski et al. 2000; Williams and Hurst 2002; Bierne and Eyre-Walker 2003; Tzeng et al. 2004; Yang 2006). Furthermore, Li et al. (2009) highlight a large effect of the method used to estimate substitution rates on the correlation between $\omega$ and $dS$; they show that the correlation is generally positive for most methods but less positive for GY94 and even negative for those methods based upon the method of Yang and Nielsen (2000). Li et al. (2009) suggest that the negative correlation may be affected by the numerical computation of the later method. Because Yang and Nielsen recommend the use of the GY method, that is what we adopt here. We estimate $dS$ using three different methods. Using the PAML software (Yang 1997), we estimated pairwise substitution rates using the NG and Yang and Nielsen (1998) methods, the latter being a variant of GY method; we refer to the method as the GY method. Both methods are mutational opportunity methods. Using the HyPhy software (Pond et al. 2004), we estimated pairwise substitution rates using the MG method, a physical site method. We performed PAML ML estimates using the following parameters: pairwise comparison (i.e., runmode = −1), $\omega$, and the transition–transversion ratio, $\kappa$, estimated from sequence data using codon frequencies estimated from the nucleotide frequencies at the three codon positions (F3x4 model). The parameterization for MG, using the MG94custom.mdl file, is similar to that of GY without considering transition–transversion bias or varying nucleotide frequencies between positions in a codon. Additionally, we used PAML to estimate the synonymous distance at 4-fold degenerate sites ($d4$) using the physical definition of a site.

In heuristic counting methods, $\omega$ is undefined if genes do not have any synonymous substitutions (i.e., $DS = 0$). We adopted two strategies to overcome this problem. In the first, we did what is commonly done and simply removed genes with $DS = 0$; in the second, we calculated $\omega = dN/(DS + 1)/S$ for all genes. To add one to the denominator is a standard correction for ratios that has been applied to odds ratios (e.g., see Jewell 1986). Note that the GY ML method estimates $\omega$ even if there is no apparent synonymous substitution as it first estimates $\omega$ and other parameters and then calculates $dN$ and $dS$ from those parameters according to their definitions (Yang 2006). Similarly, the MG ML method can estimate $\omega$ without any apparent synonymous substitution.

**Simulations**

To investigate the bias in the estimation $\omega$ when sequence divergence is low, we simulated evolving sequences under a simple evolutionary model using the Evolver software from PAML (Yang 1997). For simplicity, we generated pairs of sequences (900 bp and 1,800 bp) under the codon substitution model, using MCDatat, with equal codon frequencies, and $\kappa = 1$. This is most similar to the Jukes–Cantor model which assumes equal base frequencies
and that all substitutions are equally likely (Jukes and Cantor 1969). We used the model option 6, also known as PAML Model M0 and we fixed \( \omega \) to 0.2 or 0.5 for all sites in the gene. We varied the divergence from 0.0005 to 0.0500 in steps of 0.001. For each divergence level, we simulated 1,000 sequences and estimated substitution rates as described above.

Adjacent Substitutions

To investigate the effect of adjacent substitutions, we classified substitutions into those affecting a single site and those affecting one or more adjacent sites. We further classified lineage-specific tandem substitutions affecting two adjacent sites using parsimony and outgroups (mouse for human–macaque and human for mouse–rat comparisons). For the analysis of the best quality mouse–rat genes (i.e., those Schneider et al. (2009) selected because of trace quality, annotation, and alignment), we used dog as the outgroup as this species is of higher quality. We then excluded single, adjacent, and tandem substitutions to investigate their effect on the positive correlation between \( \omega \) and \( dS \). To test for correlations, we used two-tailed Spearman rank correlations.

Results

We began our analysis by confirming that a correlation exists between \( \omega \) and \( dS \) for three pairs of mammalian species: mouse–rat, human–macaque, and human–chimp. We first estimated substitution rates using the popular method of Goldman and Yang (1994) and which we hence refer to as the GY method. As Wyckoff et al. (2005) observed, there is a positive correlation for mouse–rat, and as Vallender and Lahn (2007) found, there is a negative correlation in human–chimp (table 1). In contrast to Vallender and Lahn (2007), we find a positive correlation for human–macaque using the data of Schneider et al. (2009; table 1); this is due to differences in the data sets because we observe a negative correlation using the data of Vallender and Lahn (2007; results not shown).

It has been shown previously that the direction and strength of the correlation between \( dN \) and \( dS \) and between \( dS \) and the third-position GC content can depend upon the method used to estimate the synonymous substitution rate (e.g., see Smith and Hurst 1999; Bielawski et al. 2000; Williams and Hurst 2002; Bierne and Eyre-Walker 2003; Tzeng et al. 2004; Li et al. 2009). In particular, the mutation rate is not always very well estimated by methods that employ a mutational opportunity definition of a site (Bierne and Eyre-Walker 2003; Yang 2006). We therefore considered the rate of synonymous substitution at 4-fold degenerate sites (\( d4 \)). Finally, we investigated whether the correlation between \( \omega \) and \( dS \) was restricted to the GY method by estimating \( dN \) and \( dS \) using the NG and MG methods. Under the NG method, there is a problem in estimating \( \omega \) when \( dS = 0 \); we adopted two strategies: in the first, we removed genes with \( dS = 0 \) and in the second, we calculated \( \omega = dN/(dS + 1)/S \) for all genes. The different methods give qualitatively similar results to those obtained with the GY method suggesting that the correlations are not a product of the methodology used (table 1).

However, \( \omega \) and \( dS \) are not independent of each other and one might expect them to be negatively correlated through sampling error in \( dS \) (Wyckoff et al. 2005; Vallender and Lahn 2007). The nonindependence between \( \omega \) and \( dS \) arises because \( \omega = dN/dS \); hence, sampling error in \( dS \) will tend to generate a negative correlation between \( \omega \) and \( dS \). It seems possible this is why Li et al. (2009) observed negative correlations between \( \omega \) and \( dS \) in some of their simulated data. To remove this nonindependence, we divided our genes into even and odd codons (Smith and Eyre-Walker 2003) and considered the correlation between \( \omega_{\text{ODD}} \) and \( dS_{\text{EVEN}} \) and the converse, \( \omega_{\text{EVEN}} \) and \( dS_{\text{ODD}} \). When we do this, the correlation between \( \omega \) and \( dS \) is positive for all three species pairs when using \( d4 \) and NG methods (table 2). This is similar to results obtained by Vallender and Lahn (2007) when they combined genes to reduce sampling error. Independent of the method used to estimate substitution rate, \( \omega \) is significantly and positively correlated to \( dS \) (and \( d4 \)) in mouse–rat and human–macaque (table 2). For human–chimp, all the correlations are positive and most are significant; the exceptions are the correlations using the MG method, in which one correlation is negative and the other positive but neither is significant.

The nonindependence of \( \omega \) and \( dS \) may explain the lack of a strong positive correlation between human and mouse observed by Liao and Zhang (2006); if we analyze the human–mouse data of Schneider et al. (2009), we observe, as Liao and Zhang (2006) did, a very weak positive correlation using the GY method (\( r = 0.048, P < 0.0001 \)); if we use odd and even codons, the correlation becomes substantially stronger (\( r = 0.086 \) and 0.087, \( P < 0.0001 \) in both...
cases), but it is, nevertheless, weaker than the correlation we observe in most other species pairs; this may be due to the large divergence between human and mouse. Unless otherwise specified, all subsequent analyses were performed splitting the data into odd and even codons.

**Simulation Results**

Estimates of ratios are typically biased because when the denominator is small, the ratio becomes disproportionately large. To investigate how the three methods perform in estimating \( \omega \) at low levels of divergence, we ran a series of simulations; for each level of divergence, which varied between 0.0005 and 0.05, we simulated 1,000 sequences 900 bp in length. For each set of simulations (e.g., 1,000 replicates at a divergence of 0.0005), we calculated the mean DS and the mean of \( \omega \); the results are shown in figures 1–3 when \( \omega \) and dS are estimated using the GY, NG, and MG methods, respectively. Our results show that all methods are statistically biased when sequence divergence is low. Although the true \( \omega \) value should be 0.2, the mean values for GY, NG, and MG have maximum mean values of approximately 17, 0.26, and 1.1, respectively. The extreme overestimation of \( \omega \) by the GY method appears to be largely due to genes with very low DS values (i.e., DS < 0.5) for which PAML often assigns a \( \omega \) value of +99. If we exclude those simulated sequence pairs that have a DS < 0.5, we find that the bias in the GY \( \omega \) value is very reduced and more similar to the NG method (figs. 1b and 2a).

Note that the mean value of \( \omega \) is greater than the expected value even when DS is quite large (e.g., DS = 8) for all methods (figs. 1a, 2a, and 3a; expected \( \omega \) = 0.2). This is due to the fact that ratios tend to be overestimated. If we modify \( \omega \) by adding one to DS, that is, \( \omega_{\text{mod}} = \frac{\ln((DS + 1)/S)}{S} \), we find that the estimate of \( \omega \) is less biased for all methods (figs. 1c, 2b, and 3b), and it also means that \( \omega \) can be estimated for all genes using the NG method.

The estimates of \( \omega \) and \( \omega_{\text{mod}} \) are likely to depend principally upon the absolute numbers of synonymous substitutions (DS) that are estimated to have occurred. To investigate this, we reran our simulations on genes that were twice as long; as expected the behavior of all methods followed that for the shorter genes (supplementary fig. S1, Supplementary Material online). We also see similar behavior if we rerun the simulations but with \( \omega = 0.5 \) (supplementary fig. S2, Supplementary Material online). The bias seems to depend upon the expected value of DS; when DS \( \geq 5 \), there seems to be relatively little bias in the GY, NG, or MG methods (figs. 1–3, supplementary figs. S1 and S2, Supplementary Material online).

**Application to Empirical Data**

The value of \( \omega \) tends to be overestimated by the GY and MG methods and underestimated by the NG method when divergences are low. Depending on the divergences of genes under study, this may cause artificial positive and negative correlations. To check whether the positive correlation between \( \omega \) and dS is due to statistical problems, we reran the analysis on the different mammalian species pairs, excluding genes below a certain sequence length \( x \), such that all genes longer than \( x \) had DS\textsubscript{ODD} (if using \( \omega \text{ODD} \)) \( \geq 5 \), and, respectively, DS\textsubscript{EVEN} (if using \( \omega \text{EVEN} \)) \( \geq 5 \). Unfortunately, this restriction leaves almost no genes within the human–chimpanzee data set, so we do not consider this data set further. For the mouse–rat and human–macaque data sets, the correlation remains positive for all three methods, that is, GY, NG, and MG (table 3). Hence, it seems that the correlation, in these species pairs, is not due to statistical bias in the estimation of \( \omega \) because when we remove most of this bias, the correlation remains. All subsequent analyses were done excluding genes below a length such that for all genes DS\textsubscript{ODD} (if using \( \omega \text{ODD} \)) \( \geq 5 \), and, respectively, DS\textsubscript{EVEN} (if using \( \omega \text{EVEN} \)) \( \geq 5 \).

**Adjacent Substitutions**

The correlation between \( \omega \) and dS could be due to a direct link between nonsynonymous and synonymous substitutions, for example, by substitutions at adjacent sites. To investigate the effect of adjacent substitutions, we removed all adjacent substitutions from the human–macaque and mouse–rat data sets. This turns the positive correlation into a negative correlation, often significant, using all methods (table 4). This suggests that the positive correlation is due to adjacent substitutions. The negative correlation probably arises because some adjacent mutations occur by chance alone and removing these induce a negative correlation.

A large number of the adjacent substitutions are tandem substitutions, which may be due to the simultaneous mutation of adjacent sites. Doublet mutations could induce a positive correlation between \( \omega \) and dS because dS would become partially dependent upon the constraint at nonsynonymous sites, if the rate of recombination between adjacent sites is low. As such, highly constrained genes have

---

**Table 2.** Table Shows the Correlation between \( \omega \) and dS When Genes Are Divided into Odd and Even Codons.

<table>
<thead>
<tr>
<th>Method Used to Estimate Substitution</th>
<th>Species Pairs</th>
<th>Species Pairs</th>
<th>Species Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GY\textsubscript{r}–GY\textsubscript{S}</td>
<td>+0.1101***, +0.1103***</td>
<td>+0.1216***, +0.1179***</td>
<td>+0.0198 NS, +0.0306 **</td>
</tr>
<tr>
<td>GY\textsubscript{r}–GY\textsubscript{d}</td>
<td>+0.2062***, +0.2001***</td>
<td>+0.2021***, +0.1897***</td>
<td>+0.0497***, +0.0322***</td>
</tr>
<tr>
<td>NG–NG\textsubscript{d}</td>
<td>+0.1805***, +0.1841***</td>
<td>+0.2748***, +0.2641***</td>
<td>+0.0339***, +0.0377***</td>
</tr>
<tr>
<td>NG\textsubscript{mod}–NG\textsubscript{d}</td>
<td>+0.1880***, +0.1919***</td>
<td>+0.2793***, +0.2889***</td>
<td>+0.1054***, +0.1126***</td>
</tr>
<tr>
<td>MG–MG\textsubscript{d}</td>
<td>+0.0707***, +0.0755***</td>
<td>+0.1144***, +0.1181***</td>
<td>-0.0019 NS, +0.0044 NS</td>
</tr>
</tbody>
</table>

**Note:** All 9,942 genes from Schneider et al. (2009) were considered.

*P < 0.05, **P < 0.001, ***P < 0.0005, NS, not significant.
The bias in the GY estimate of $\omega$. The figure shows the mean and standard error of $\omega$ plotted against the mean value of DS for simulated sequences of 900 bp in length. The true $\omega$ value is 0.2. Each point represents the mean of 1,000 simulated sequences for a particular expected level of divergence. (a) GY $\omega$, (b) GY $\omega$ for genes with DS $> 0.5$, (c) GY $\omega_{mod} = dN/((DS + 1)/S)$. 

**FIG. 1.** Positive Correlation between $dN/dS$ and $dS$ in Mammals · doi:10.1093/molbev/msq320

MBE
low $\omega$ values and also low values of $dS$. To investigate the contribution of potential doublet mutations, we removed all lineage-specific tandem substitutions; but the positive correlation remains in both species pairs (table 4).

We also reran the analysis excluding not the adjacent substitutions from genes but the whole genes that had adjacent substitutions. The positive correlation disappears for the remaining 338 mouse–rat genes that had sufficient synonymous substitutions (GYd4: $-0.1314^{*}$ and $-0.0641$ NS).

Adjacent substitutions may be genuine but they may also be artifacts of errors in sequencing, alignment, or exon assignment. In order to investigate this further, we took advantage of two subsets of the data analyzed by Schneider et al. (2009), which they had subjected to two levels of quality control. In the first set, genes were retained if they were longer than 200 bp and all orthologs in the alignment were less than 5% different in length; in the second set, good genes had to have additionally high trace coverage, known annotation status, and 100% alignment HoT scores. We find that there is a significant positive correlation for both species pairs in the first subset using all methods of estimating $\omega$ and $dS$ (table 5). Unfortunately, the second subset of data is sufficiently small that the analysis could only be performed on mouse–rat. However, in this case, all correlations are positive and most are significant; the exceptions are using the GY and MG methods (table 5).

We note that the positive correlation disappears when we exclude lineage-specific tandem substitutions from the second high-quality subset of mouse–rat genes (table 6). In contrast, when excluding only tandem substitution, in

**Fig. 2.** The bias in the NG estimate of $\omega$. The figure shows the mean and standard error of $\omega$ plotted against the mean value of $dS$ for simulated sequences of 900 bp in length. The true $\omega$ value is 0.2. Each point represents the mean of 1,000 simulated sequences for a particular expected level of divergence. (a) NG $\omega$, (b) NG $\omega_{mod} = dN/((dS + 1)/S)$. 

**Stoletzki and Eyre-Walker · doi:10.1093/molbev/msq320 MBE**
which one substitution occurs in each lineage, the positive correlation remains (table 6). This difference is not a consequence of numbers because there are similar numbers of lineage-specific and nonspecific substitutions—5,499 and 4,249, respectively. We further find that the ratio of double to single substitutions per gene correlates strongly with \( \omega \) (\( r = -0.6987^{***} \)) supporting their contribution to the correlation between \( \omega \) and \( dS \).

**Discussion**

It has been shown, and we have confirmed, that \( \omega \) is significantly correlated to \( dS \) in mouse–rat and human–macaque comparisons. We have explored two possible explanations for this surprising correlation: first, a statistical bias in the estimation of \( \omega \) and second, substitutions at adjacent sites. We show that the correlation between \( \omega \) and \( dS \) remains, not surprisingly, if we control for nonindependence using odd and even codons because the nonindependence should generate a negative correlation. We also show that the estimation of \( \omega \) is biased when sequence divergence is low, and this could potentially cause the correlation. However, we find that this statistical bias is not the primary determinant of the correlation. Instead the relationship between \( \omega \) and \( dS \) appears to be a consequence of substitutions at adjacent sites.

The value of \( \omega \) is a simple and widely used measure of selection pressure acting on protein-coding genes. Being a ratio, however, there are potential difficulties with its estimation. We have shown that it is particularly difficult to estimate \( \omega \) using the GY, NG, or MG methods when the absolute number of synonymous substitutions is less than five. There is a bias in the estimation of \( \omega \) when comparing closely related species or genes; this could affect correlations between \( \omega \) and \( dS \), and it may also have implications for comparing \( \omega \) values between species that differ in their level of divergence. Because the methods, which we used, represent the extremes in terms of their statistical sophistication, it is likely that this behavior will be present in most methods for estimating \( \omega \). One may exclude genes below a certain length such that \( dS \geq 5 \); unfortunately, however, in closely related species such as humans and chimps, very few genes have a sufficient number of synonymous substitutions to make a reliable estimation of \( \omega \) per gene. Another problem with estimating a ratio is that the denominator must not be zero. Under heuristic methods for estimating rates of synonymous and nonsynonymous substitution, like NG, \( \omega \) is undefined when the gene does not experience a single synonymous substitution. To get around the exclusion of these genes, we suggest adding one to the observed numbers of synonymous substitutions for all genes (\( \omega_{MOD} \) modified = \( dN/((dS + 1)/S) \)). This simple correction has the added bonus of making the estimate of \( \omega \) less biased for all methods, although we have not shown that this is generally true, and there may be no estimate of \( \omega \) that is unbiased when there is little data. To eliminate this bias when there is too little data, we suggest summing the numbers of substitutions across genes before calculating an average \( \omega \) for a group of genes.

**Table 3.** Table Gives the Correlation between \( \omega \) and \( dS \).

<table>
<thead>
<tr>
<th>Method Used to Estimate Substitution Rates for ( \omega ) and ( dS )</th>
<th>Species Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mouse–Rat</td>
</tr>
<tr>
<td>GY(\omega)–GY(dS)</td>
<td>+0.1019***, +0.1089***</td>
</tr>
<tr>
<td>GY(\omega)–GY(d4)</td>
<td>+0.2011***, +0.2049***</td>
</tr>
<tr>
<td>NG(\omega)–NG(dS)</td>
<td>+0.1709***, +0.1865***</td>
</tr>
<tr>
<td>NG(\omega Mod)–NG(dS)</td>
<td>+0.1746***, +0.1901***</td>
</tr>
<tr>
<td>MGo(\omega)–MGdS</td>
<td>+0.0851***, +0.0834***</td>
</tr>
</tbody>
</table>

**Note.** To control for statistical bias in the estimate of \( \omega \), we exclude genes below a length such that for all genes \( dS \geq 5 \). For mouse–rat, we analyzed 4,706 genes with length >1,218 bp, and for human–macaque, we analyzed 821 genes with length >1,000 bp.

*P < 0.05, **P < 0.001, ***P < 0.0005.
Other statistical issues may be insufficient correction of multiple hits or an effect of how $d_S$ is estimated. Saturation in $d_S$ for some genes can cause an underestimate of $d_S$ and an overestimate of $\omega$, thereby leading to a positive correlation between the two statistics. However, because the average synonymous divergence is fairly low in mouse–rat (0.185) and human–macaque (0.079), this seems an unlikely explanation. Although correlations of $\omega$ with other variables can be sensitive to the method used to estimate $d_S$, the results are generally independent of the method used to estimate $\omega$ and $d_S$. In some cases, however, the estimates from the $G$ and $M$ methods lead to nonsignificant results when the other methods are significant (table 2 and table 4 mouse–rat 2 genes). The $G$ method accounts for unequal codon usage when estimating the number of synonymous sites, implicitly assuming that codon bias is mutational in origin. If codon bias is due to selection it may provide a less accurate estimate of the mutation rate (for detailed discussion, see Birne and Eyre-Walker 2003 and Yang 2006). It is therefore probably more appropriate to consider a measure of the synonymous substitution rate such as $d_S$ that is an estimate of the number of substitutions per physical site, rather than per mutational opportunity.

It seems that although statistical issues of estimating $\omega$ might cause a positive correlation between $\omega$ and $d_S$, this is not the principle reason for the observed correlation. The positive correlation between $\omega$ and $d_S$ appears to be due to adjacent substitutions. Such an effect of adjacent substitutions is not unexpected: compared with single substitutions, adjacent substitutions affect higher proportions of nonsynonymous sites and with increasing numbers of adjacent substitutions, $d_M$ increases more strongly than $d_S$, which causes $\omega$ to increase with $d_S$, leading to a positive correlation between $\omega$ and $d_S$. It has been shown before that models that allow multiple adjacent changes fit data better than models which do not (Whelan and Goldman 2004; Higgs et al. 2007; Kosiol et al. 2007) and that positive results of the branch-site test of selection for human and chimpanzee lineages are often caused by codons with multiple nonsynonymous substitutions (Suzuki 2008).

Adjacent substitutions may have various origins. First, adjacent substitutions may be a consequence of errors in sequencing, exon assignment, annotation, or alignment. Such errors may result in clustered substitutions and if the amount of errors varies sufficiently between genes, it may cause a positive correlation between $\omega$ and $d_S$. This is because errors will not differentiate between nonsynonymous
and synonymous sites, whereas selection does. Therefore, $dN$ will increase proportionally more rapidly with error rate than $dS$, so a positive correlation between $\omega$ and $dS$ will be produced. We show that restricting our data to genes with high-quality sequences, known annotation, and unambiguous alignments decreases the correlation between $\omega$ and $dS$, indicating a contribution of such errors to the correlation. However, albeit weaker, the correlation between $\omega$ and $dS$ in mouse–rat is also present if we restrict our analysis to those genes that are expected to bear few such errors.

Second, adjacent substitutions may result from mutational nonindependence—the correlation might, for example, be a consequence of simultaneous mutation of adjacent nucleotides or of sites or regions of the genome with higher rates of mutation. It is known that two adjacent nucleotides mutate more together than by chance alone (Wolfe and Sharp 1993; Averof et al. 2000), their rate, however, appears much smaller than initially estimated (Kondrashov 2002; Silva and Kondrashov 2002; Smith et al. 2003; Hodgkinson and Eyre-Walker 2010) and the observed numbers of doublet substitutions may be partly due to selection (Smith and Hurst 1999). Processes involving longer runs of simultaneous mutation are not known; for example, in human polymorphism data two adjacent single nucleotide polymorphisms (SNPs) are overrepresented but runs of three or four SNPs are not (Hodgkinson and Eyre-Walker 2010). Also, there is no evidence that sequence- or template-directed mutagenesis, a mutational mechanism that can simultaneously change several nucleotides, is operating in mammals (Ladoukakis and Eyre-Walker 2007).

Finally, adjacent substitutions may be due to correlated selection on adjacent sites. Such correlated selection may arise if synonymous and nonsynonymous sites are part of the same structure, for example, if a regulatory element exists within a gene and is under selection, or through the action of selection on synonymous codons. Wyckoff et al. (2005) performed simulations involving selection on synonymous codons and inferred that such selection could not induce a correlation between $\omega$ and $dS$. However, they assumed that the strengths of selection at synonymous and nonsynonymous sites were uncorrelated, and it is clear that if this assumption is broken—as for selection on motifs spanning synonymous and nonsynonymous sites or on translational accuracy—then $\omega$ and $dS$ are likely to be correlated. The precise form of this relationship depends upon how the strengths of selection on synonymous and nonsynonymous mutations are related. Negative selection could indirectly cause clusters of adjacent substitutions if only certain regions under less constraint are left in a gene where mutations accumulate. But probably more likely, we need to consider selective forces that might induce positive selection across a number of adjacent sites. First, adaptive evolution of regulatory elements within a gene could lead to multiple adjacent substitutions. Another obvious idea by Lipman and Wilbur (1985) relates to optimal codon use. They noted that if the last letter of the optimal or preferred codon differed between amino acids, then a nonsynonymous substitution would often change an optimal codon to a suboptimal codon and this could then lead to an adaptive synonymous substitution. There are generally not thought to be optimal codons in mammals; however, a recent study by Drummond and Wilke (2008) shows higher optimal codon use at conserved—potentially functionally important—amino acid sites in humans and mice. If such amino acids evolve adaptively, adaptive synonymous changes could follow. However, the adjacent changes we consider may span multiple codons, which cannot be explained by correlated evolution of $dN$ and $dS$ within codon.

We have confirmed that there is a correlation between $\omega$ and $dS$ in mouse–rat and human–macaque comparison. Although errors in sequencing, annotation, and alignments contribute, we have shown that the positive correlation is at least in part due to genuine adjacent substitutions. The remaining positive correlation in high-quality mouse–rat gene alignments disappears when excluding lineage-specific tandem substitutions. Tandem substitutions and hence the origin of the positive correlation may be of mutational or selective origin.

### Supplementary Material

Supplementary fig. S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

### Acknowledgments

We thank Paul Higgs and two anonymous referees for many helpful and insightful comments on the manuscript, Sergei Pond for help with HyPhy, and Bruce Lahn for helpful

---

**Table 6.** Table Shows the Effect of Tandem Substitutions on the Correlation between $\omega$ and $dS$ in the Second More Stringent Subset of High-Quality Mouse–Rat Genes.

<table>
<thead>
<tr>
<th>Method Used to Estimate</th>
<th>Exclude Lineage-Specific Tandem Substitutions</th>
<th>Exclude Mosaics of Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyu–GydS</td>
<td>$-0.0642^*, -0.0520 NS$</td>
<td>$+0.0218 NS, +0.0215 NS$</td>
</tr>
<tr>
<td>Gyu–Gyd4</td>
<td>$+0.0164 NS, +0.0259 NS$</td>
<td>$+0.1138^{<em><strong>}, +0.0966^{</strong></em>}$</td>
</tr>
<tr>
<td>Ngou–Ngds</td>
<td>$-0.0025 NS, +0.0111 NS$</td>
<td>$+0.0889^{*<strong>}, +0.0776^{</strong>}$</td>
</tr>
<tr>
<td>Ngoumod–NgodS</td>
<td>$+0.0018 NS, +0.0147 NS$</td>
<td>$+0.0772^{<strong>}, +0.0995^{</strong>*}$</td>
</tr>
<tr>
<td>MG</td>
<td>$+0.0456 NS, +0.0274 NS$</td>
<td>$+0.1813^{<em><strong>}, +0.1266^{</strong></em>}$</td>
</tr>
</tbody>
</table>

Note.—Two types of tandem substitutions are considered: lineage-specific tandems and nonlineage-specific tandems, in which one substitution occurs in each lineage. To control for statistical bias in $\omega$ estimates, we analyzed for 1,191 genes with length greater than 899 bp and 1,216 genes with length greater than 909 bp.

* $P < 0.05$, ** $P < 0.001$, *** $P < 0.0005$, NS, not significant.
discussion. N.S. was supported by the International Union of Biochemistry and Molecular Biology, the Munich Graduate School for Ecology, Evolution and Systematics, and the Biotechnology and Biological Sciences Research Council (BBSRC). A.E.W. was supported by the BBSRC.

References


