Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 378 (2006) 58–64
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation

Stéphane Aris-Brosou [a,*], Joseph P. Bielawski [b]

[a] *Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON, Canada K1N 6N5*
[b] *Department of Biology and Department of Mathematics and Statistics, Dalhousie University, Halifax, NS Canada*

Received 11 February 2006; received in revised form 20 April 2006; accepted 26 April 2006
Available online 22 May 2006

Abstract

A popular approach to examine the roles of mutation and selection in the evolution of genomes has been to consider the relationship between codon bias and synonymous rates of molecular evolution. A significant relationship between these two quantities is taken to indicate the action of weak selection on substitutions among synonymous codons. The neutral theory predicts that the rate of evolution is inversely related to the level of functional constraint. Therefore, selection against the use of non-preferred codons among those coding for the same amino acid should result in lower rates of synonymous substitution as compared with sites not subject to such selection pressures. However, reliably measuring the extent of such a relationship is problematic, as estimates of synonymous rates are sensitive to our assumptions about the process of molecular evolution. Previous studies showed the importance of accounting for unequal codon frequencies, in particular when synonymous codon usage is highly biased. Yet, unequal codon frequencies can be modeled in different ways, making different assumptions about the mutation process. Here we conduct a simulation study to evaluate two different ways of modeling uneven codon frequencies and show that both model parameterizations can have a dramatic impact on rate estimates and affect biological conclusions about genome evolution. We reanalyze three large data sets to demonstrate the relevance of our results to empirical data analysis.

Keywords: Codon model; Synonymous rate; Natural selection; *Drosophila*; Rodents; Bacteria

1. Introduction

Evolution is expected to favor efficient protein synthesis systems, with reduced energetic costs and enhanced translation efficiency. One means retained by selection appears to be biased codon usage, where anticodons corresponding to tRNA genes with the largest copy number are favored. This type of selection is called translational selection. Under this selection scheme all synonymous codons are not equal, so that some synonymous mutations are disfavored and synonymous substitution rates are reduced. Consequently, a negative correlation between codon bias and synonymous substitution rates is taken as evidence for translational selection (e.g., Akashi and Eyre-Walker, 1998). The inference of such a correlation has been a common component of many studies of genome evolution (e.g., Wall et al., 2005; Drummond et al., 2006).

Codon bias can be measured by the Codon Adaptation Index or CAI (Sharp and Li, 1987), which relies on a prior specification of preferred codons, or by the Effective Number of Codons (ENC) that does not rely on such specifications (Wright, 1990). Less agreement exists, however, as to how synonymous substitution rates should be estimated. Some authors use estimates based on counting methods such as that by Nei and Gojobori (NG) (Nei and Gojobori, 1986); others use model-based approaches and in particular that by Goldman and Yang (1994). One issue is that these two approaches can give different results, and thereby occasionally lead to opposing conclusions when it comes to finding evidence for translational selection (Bielawski et al., 2000; Dunn et al., 2001).

Several studies have suggested that model assumptions matter more than methods, with conflicting results mainly caused by the presence or absence of a correction for codon usage bias (Yang and Nielsen, 1998; Bielawski et al., 2000; Dunn et al.,

Abbreviations: CAI, Codon Adaptation Index; ENC, Effective Number of Codons; GY, Goldman and Yang; NG, Nei and Gojobori; MG, Muse and Gaut.

* Corresponding author. Tel.: +1 613 562 5800x6354; fax: +1 613 562 5486.
E-mail address: sarisbro@uottawa.ca (S. Aris-Brosou).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.04.024

2001). In the likelihood framework, specific parameterizations of codon frequencies can be chosen to emphasize a process where rates either depend on the underlying mutational process (Muse and Gaut, 1994; Whelan and Goldman, 2004; Wong and Nielsen, 2004) or depend on the codon towards which the mutations occur (e.g., Goldman and Yang, 1994). These two Markov processes, respectively denoted MG and GY hereafter, both have similar instantaneous rate matrices that distinguish nonsynonymous from synonymous substitutions by means of their rate ratio, ω. However, they differ by their state spaces: nucleotide for MG, and codon for GY. The possibility that biological conclusions might also be sensitive to these assumptions has not been investigated.

We illustrate with a simple example how alternative parameterizations of codon frequencies, as in GY and MG, can potentially lead to distinct biological interpretations. Let us consider a single mutation from nucleotide A to C, occurring in codon AAA (Lys). Depending on which codon position k is affected, mutation rates are proportional to:

	$k=1$	$k=2$	$k=3$
GY	π_{Gln}	π_{Thr}	π_{Asn}
MG	$\pi_{\mathrm{C}}^{(1)}$	$\pi_{\mathrm{C}}^{(2)}$	$\pi_{\mathrm{C}}^{(3)}$

Therefore, under GY, the mutation rate from A to C depends on the equilibrium frequency π_j of the target codon (j), while under MG, this mutation rate depends only on the equilibrium frequency $\pi_j^{(k)}$ of the target nucleotide (j) at a given codon position k. Because of this fundamental difference, these two models are expected to have different properties when codon frequencies are uneven. In this study we focus on the impact of these differences on the estimation of synonymous substitution rates, and investigate the potential for estimation errors to impact conclusions derived from genome-scale data analysis. We use simulations to investigate statistical properties such as systematic bias. We also analyze three real data sets, and show that biological conclusions can indeed be sensitive to the parameterization, that is, to our understanding of the underlying evolutionary process.

2. Methods

2.1. Parameterizations of codon models

The codon model of Goldman and Yang (1994) is specified by the instantaneous rate matrix $\boldsymbol{Q}=\{q_{ij}\}$ where non-diagonal elements represent the instantaneous rate of change from codon i to codon j. This rate of change takes into account the transition to transversion bias (κ), the selective pressure (ω) and uneven codon equilibrium frequencies (π_j). Diagonal elements are specified by the requirement that lines of the $\boldsymbol{Q}$ matrix sum to zero. Only one-step changes to sense codons are permitted, so that $\boldsymbol{Q}$ is a 61×61 matrix. The transition probability matrix $\boldsymbol{P}(t)=\{p_{ij}(t)\}$ that gives the probability of a substitution from codon i to codon j during a time period of length t is obtained by taking $\exp(\boldsymbol{Q}\cdot t)$. Equilibrium codon frequencies π_j are either estimated as free parameters, or calculated from the observed nucleotide frequencies. In this latter case, two schemes are possible: either estimate π_j from the average nucleotide frequencies (F1×4; see Yang, 2001) or from the average nucleotide frequencies at the three codon positions (F3×4). This parameterization of the codon model is denoted GY hereafter. The synonymous substitution rate is then calculated as in Yang (2001) by the ratio of the estimated proportion of synonymous substitutions per codon and of the estimated proportion of synonymous sites. It measures the mutational opportunities before the action of selection on the protein.

Another parameterization of the instantaneous rate matrix $\boldsymbol{Q}$, proposed by Muse and Gaut (1994), is based not on codon frequencies but on nucleotide frequencies. If we denote codon i by the triplet $i_1i_2i_3$, where i_k represents the nucleotide at codon position k, the instantaneous substitution rate from codon i to codon j is now proportional to the frequency of the target nucleotide j at position k: $\pi_j^{(k)}$. Unlike the original model description by Muse and Gaut (1994), the exact specification of the instantaneous rate matrix can be based on the HKY model (Hasegawa et al., 1985) so that it differs from the above GY $\boldsymbol{Q}$ matrix only by which equilibrium frequencies are considered: those of codons for GY, or those of nucleotides. This latter parameterization is denoted MG hereafter. It should not be confused with the model developed by Muse and Gaut (1994) or with MG94×HKY85 (Kosakovsky Pond and Frost, 2005a,b; Kosakovsky Pond and Muse, 2005) since these later models, unlike GY, describe synonymous and nonsynonymous rates using distinct parameters rather than their rate ratio. The MG parameterization used here is however similar to the codon substitution model used by Wong and Nielsen (2004) to describe the evolution of coding regions. Synonymous substitution rates are calculated as above; hence both GY- and MG-derived estimates are based on mutational opportunities, and provide comparable measures of the average substitution rate over the three codon positions.

Bierne and Eyre-Walker (2003) suggest that an alternative definition of the synonymous substitution rate, where the rate is measured per codon rather than synonymous site, might be more appropriate for some comparisons. It is not the purpose of this study to investigate the relative utility of physical-site and mutational-opportunity measures of the synonymous rate. The mutational opportunity approach has remained the method of choice in large-scale comparisons among genes (e.g., Wall et al., 2005; Drummond et al., 2006). Hence, continued assessment of the sensitivity of this approach to assumptions about the process of evolution is warranted.

2.2. Data sets analyzed

We reanalyzed three data sets, consisting of 128 genes from *Escherichia coli* and *Salmonella typhimurium* (Eyre-Walker and Bulmer, 1995; Smith and Eyre-Walker, 2001), 35 nuclear genes from *Drosophila melanogaster* and *D. pseudoobscura* (Dunn et al., 2001) and 356 nuclear genes from mouse and rat (Wolfe and Sharp, 1993). These data sets were chosen because their results previously exhibited diverse levels of agreement, ranging from total consensus across the different methods (Smith and Eyre-Walker, 2001) to total disagreement (Dunn et al., 2001). Synonymous rates were estimated with codeml

(Yang, 1997). Codon bias was estimated by the effective number of codons (Wright, 1990), denoted by ENC, and by the codon adaptation index, denoted by CAI (Sharp and Li, 1987), both implemented in codonw written by John Peden. Linear models were fitted using robust MM regression (Yohai, 1987; see also Aris-Brosou, 2005).

2.3. Computer simulations

Two different types of biases were simulated: (i) uneven G and C nucleotide frequencies over the three codon positions, denoted GC bias hereafter and (ii) uneven G and C nucleotide frequencies only at third codon positions, denoted GC3 bias. The computer program evolver (Yang, 1997) was used to simulate the data under these two types of biases. This implies that all simulations were ultimately based on the codon model developed by Goldman and Yang (1994).

Simulated codon frequencies were generated as in Fig. 1. Depending on the type of codon frequency bias, biased codons were those containing at least a G or a C (GC bias, Fig. 1a) or those for which 3rd positions were either a G or a C (GC3 bias, Fig. 1b). The equilibrium frequency of biased codons was described by the simulation parameter η scaled by Σ so that the 64 codon frequencies sum to 1. Note that the unbiased case, where the category of biased codons is as likely as the other category of codons, is at $\eta = 1/2$. For each two-species simulated data set, the transition to transversion rate ratio κ was set either to one or to five, and branch lengths were set to .2, .5, 1.0 and 2.0. The simulation conditions were determined by varying η from 0 to 1 by increments of 1/20, while the nonsynonymous to synonymous rate ratio ω was varied from .05 to 5 by increments of 1/20. This added up to a total of 8000 simulation conditions. In each case, 100 replicates were generated for two sequences with 100,000 codons. No relationship between codon bias and substitution rate was assumed in the simulations. Replicates were analyzed with codeml (Yang, 1997).

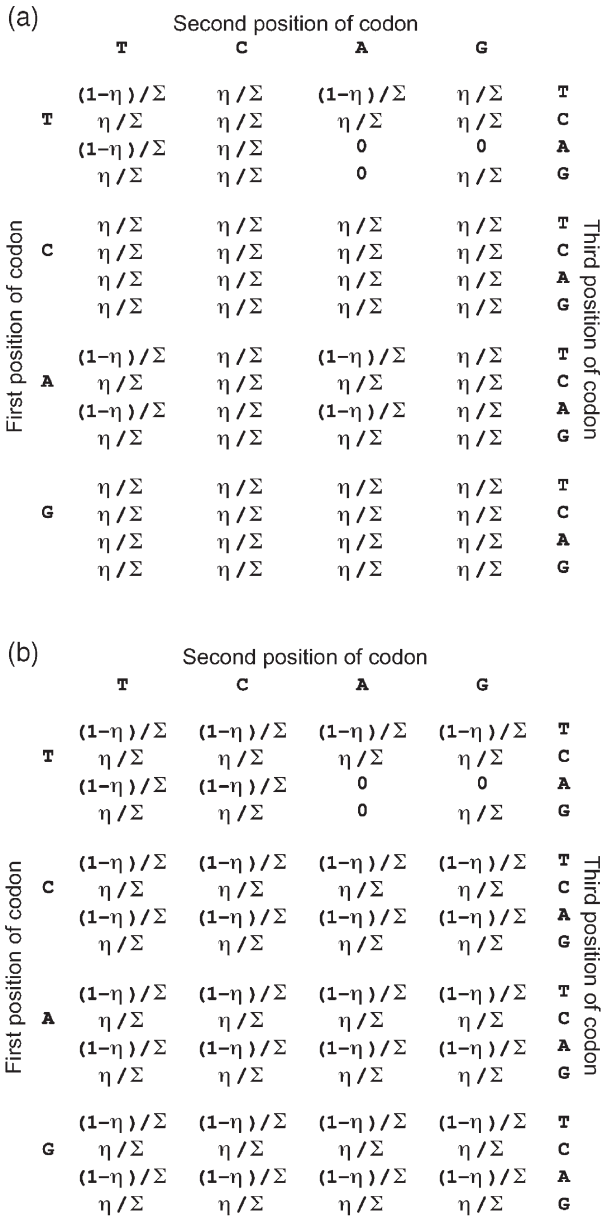

Fig. 1. Vectors of codon frequencies used to simulate (a) GC bias and (b) GC3 bias.

3. Results and discussion

3.1. Real data analyses

We first reanalyzed three published data sets to evaluate the impact of the parameterizations of codon models on detection of translational selection. The first data set consists of 128 genes from *E. coli* and *S. typhimurium* (Eyre-Walker and Bulmer, 1995; Smith and Eyre-Walker, 2001), two species belonging to the Enterobacteriales order. Unlike the original authors, we analyzed all the genes, and estimated synonymous substitution rates using the NG, GY and MG methods. Fig. 2a shows that irrespective of the method used, a highly significant ($p < .0001$) and positive linear relationship exists between synonymous substitution rates and ENC. Similarly significant (but negative, as expected) correlations were obtained with CAI for all methods (not shown). These results support the idea that translational selection might be acting in these enterobacteria species, which is consistent with the results obtained by the original authors.

The second example considered here is a comparison between two species of flies, *D. melanogaster* and *D. pseudoobscura* originally analyzed by Dunn et al. (2001). Fig. 2b shows that while using NG leads to a highly significant positive correlation between synonymous substitution rates and ENC, no correlation exists when rates are estimated using GY. These results are consistent with those found by the original authors, and would tend to support the idea that some features not considered by NG, such as codon bias, can affect rate estimates. However, the correlation between rates estimated by MG and codon bias is significant (ENC: 5% level; Fig. 2b), which suggests some evidence for translational selection. Note that under MG a large component (22%) of the variation in synonymous substitution rate is explainable by variation in codon usage bias, as measured by ENC.

The third example consists of 357 genes in a mouse–rat comparison (Wolfe and Sharp, 1993). A first analysis of the entire set of gene pairs revealed that when synonymous substitution

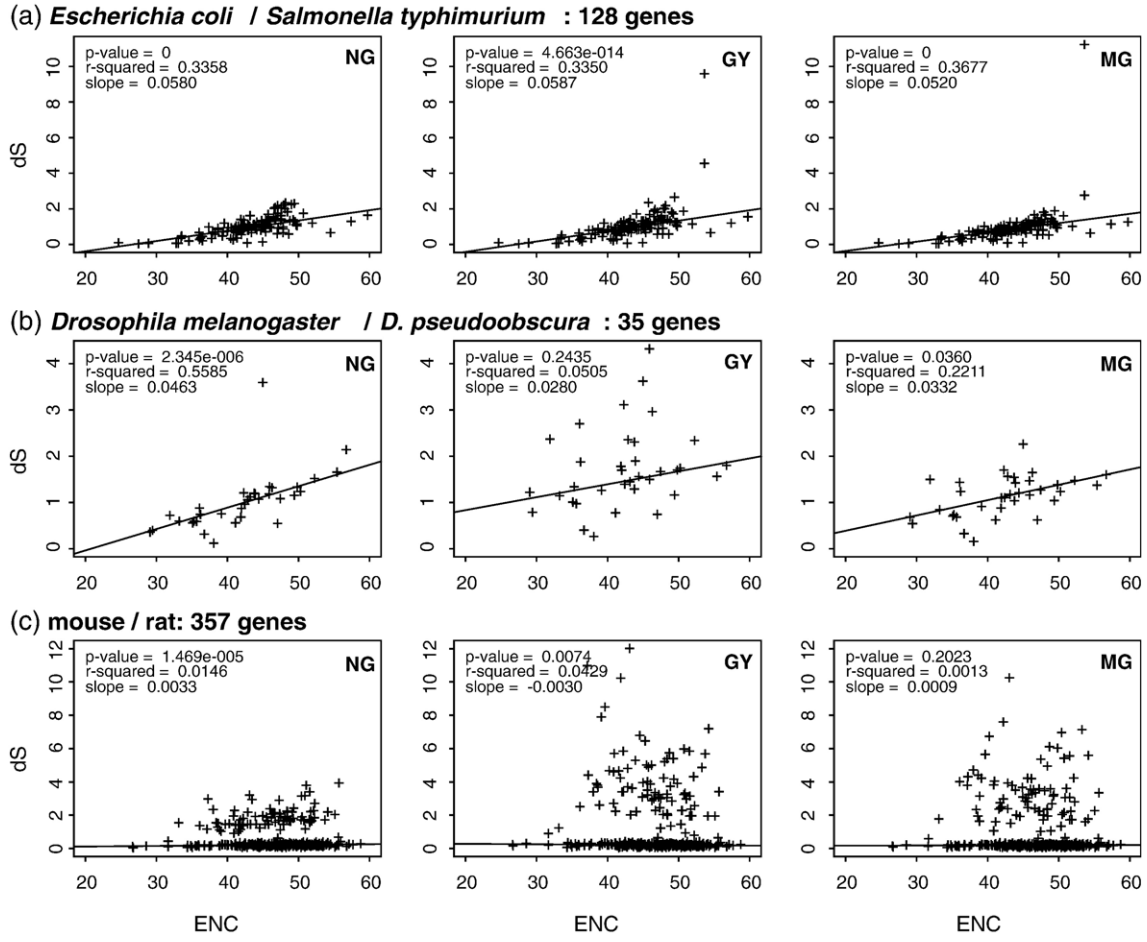

Fig. 2. Influence of the method used to estimate substitution rates (d*S*) on the detection of evidence for translational selection in several real data sets: (a) bacteria, (b) *Drosophila*, and (c) rodents. Three different approaches were used: Nei and Gojobori (NG), Goldman and Yang (GY) and our Muse and Gaut (MG — see Methods).

rates are estimated with NG or GY, a highly significant relationship is found with ENC. Note that outliers are accommodated by using the robust MM regression. However, while the slope obtained with NG is positive, the slope with GY is negative. On the other hand, no such relationship is detected when rates are estimated with MG (Fig. 2c), a model which differs from GY only in how codon bias is described. Although the proportion of the variance explained by the linear model is extremely small (r-squares $<.05$), such discrepancies hint at the existence of a problem.

These last two examples show that the parameterization of codon model can have a dramatic effect. Because our specification of MG and GY only differs in how equilibrium frequencies are defined, it suggests that our limited understanding of the relationship between mutational processes acting on protein coding nucleotide sequences and the precise parameterization of codon frequencies can impair our inference about synonymous rates.

3.2. Computer simulations

In order to understand how codon bias can affect our estimates of synonymous substitution rates, we carried out some simulations. One of the shortcomings common to all similar simulation studies is the need to adopt a tractable, and hence simplified, model to generate the data. We note that comparison among several competing models will be biased towards the generating model if it is among the comparators. However, our purpose is simply to assess if estimates of synonymous substitution rates obtained under a model different from the generating model have a bias sufficient to negatively impact biological conclusions. To do so, we simulated two types of composition biases. The first one, denoted GC, has uneven G and C nucleotide frequencies over the three codon positions (Fig. 1a). The second, denoted GC3, has uneven G and C nucleotide frequencies only at third codon positions (Fig. 1b). No relationship between codon bias and substitution rate was assumed in the simulations. Our results indicate that for a given type of bias (Fig. 1), estimates of synonymous rates under another model can be up to 30% different, and that this difference depends on codon bias and also, interestingly, on the strength of selection.

Fig. 3 shows that estimation errors of synonymous rates are more important at low GC or GC3 contents. More specifically, for GC biases (Fig. 3a–b), MG can underestimate synonymous rates for GC poor sequences (bias parameter $\eta<1/2$) by up to 30%, and underestimation is all the more important when κ, the transition to transversion rate ratio, and ω are small ($\kappa=1$ and $\omega<1.0$). Recall that the unbiased situation occurs for $\eta=1/2$.

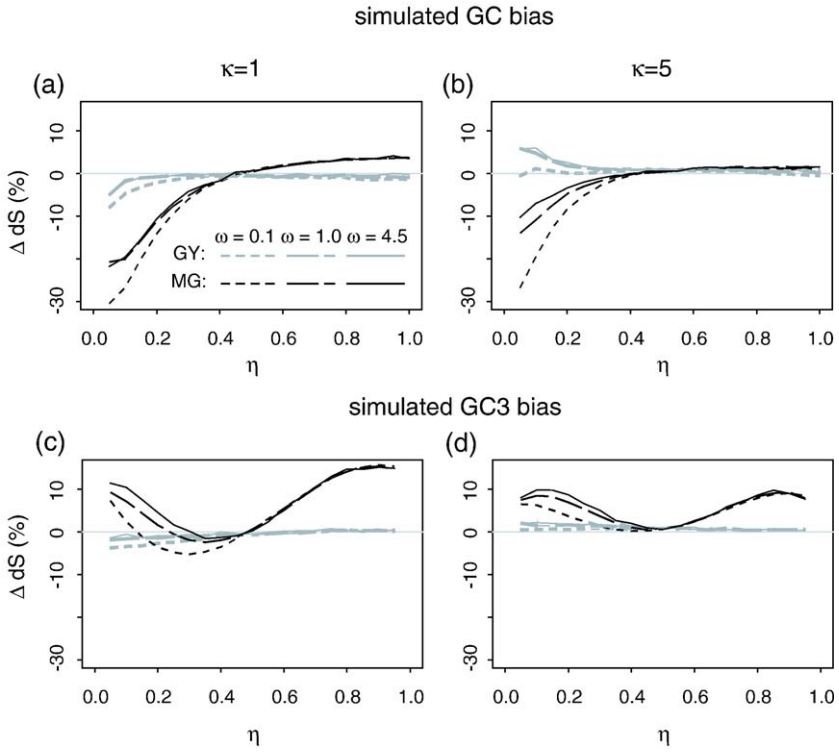

Fig. 3. Relative differences ($\Delta = 100\,(x-y)/y$) between parameter estimates (x) and their expected values (y) of substitution rates (dS) as a function of the simulation bias parameter (η) for a divergence (t) of 0.2. The unbiased situation is at $\eta = 1/2$. Parameter estimates were obtained either under Goldman and Yang (GY) or under our Muse and Gaut (MG) model. Expected values were obtained directly from the analyses using the simulation model (F61) and long sequences (100,000 codon sites).

As a result, such data sets will tend to show positive correlations between codon bias and synonymous rates (e.g., steep positive slope, leftmost part of Fig. 3a). With increasing GC content, the correlation decreases under MG, and is almost non-existent under GY. Note that for larger κ values, a small negative correlation is observed at low GC contents under GY. Remember that since the data were simulated by using a model where rates depend on the frequency of the codons towards which the mutation occurred (Goldman and Yang, 1994), MG was expected to exhibit a higher estimation bias than GY, as it is observed here (Fig. 3). These results should not be taken as an indictment of MG; we expect that GY would not have performed well had we simulated under a model where rates depended on the frequency of the target nucleotide at a particular codon position.

For GC3 biases the situation is slightly more complicated. GY gives a small positive correlation at low GC3 content and low κ (Fig. 3c), and a small negative correlation at low GC3 content and higher κ (Fig. 3d). On the other hand, MG tends to give a negative correlation at low GC3 content and a positive correlation at high GC3 content. So for GC3-poor genes, the simulations show that it is possible to obtain a weak positive correlation under one parameterization (GY in Fig. 3c) and a negative correlation under the other parameterization (MG in Fig. 3d). These results show that spurious correlations can be obtained between synonymous rates and codon bias when codon frequencies are uneven, and that both parameterizations, MG and GY, are sensitive to this effect.

To assess the impact of sequence divergence we simulated over a range of branch lengths. These specific values were chosen to cover the breadth of the median branch lengths estimated from our empirical data sets. We found that the estimation bias we observed increased with sequence divergence (Supplementary Fig. S1). Similarly, we found that relative numbers of changes measured on a physical-site basis were also affected by the specification of the model (Supplementary Fig. S2). We believe this reflects the increasing importance of an adequate correction for multiple substitutions at a site. Interestingly, both Muse (1996) and Dunn et al. (2001) pointed out a similar effect associated with counting methods with no correction for codon usage bias. Here the Markov processes are formulated such that multiple

Table 1
Distribution of the GC and GC3 contents for the real data sets analyzed

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Escherichia coli/Salmonella typhimurium						
GC	0.464	0.519	0.536	0.534	0.554	0.595
GC3	0.269	0.535	0.580	0.566	0.605	0.674
Drosophila melanogaster/D. pseudoobscura						
GC	0.458	0.535	0.560	0.561	0.576	0.672
GC3	0.397	0.671	0.731	0.708	0.778	0.881
Mouse/Rat						
GC	0.402	0.496	0.535	0.531	0.567	0.733
GC3	0.327	0.539	0.627	0.617	0.694	0.957

substitutions are not permitted at the same time, and probability theory takes care of correcting for superimposed substitutions over time. As the probabilities are dependent on the model, the discrepancy between the model and the data becomes more important with increasing sequence divergence. We note that for real data sets, a Markov process accommodating multiple-nucleotide changes (Whelan and Goldman, 2004) might improve the situation.

We can now try to better understand the results obtained with the real data. Table 1 shows that all the data sets analyzed here have little GC bias, but tend to be GC3 rich. Our simulations show that in this case, if the data had been generated under a process where codon frequencies were determined by more than a site independent mutational process (e.g., neighbor effects on mutation rates or weak selection) such as GY, a spurious positive correlation could be obtained when analyzing these data under a model of mutation formulated at individual nucleotides such as MG. The *Drosophila* data set was the most highly biased towards high GC3 content (mean = 71% — Table 1), and was one for which no correlation was found when analyzed under GY, but a correlation existed under MG. The mouse/rat data show intermediate distribution of GC3 values with a mean around 62% (Table 1), but the results are opposite to those found for the *Drosophila* data. The mouse/rat results would be those expected if the data had been generated under a mutational model such as MG, in the absence of translational selection.

3.3. Conclusions

Our results show that evidence for translational selection can be difficult to find when codon frequencies are uneven and suggest that more effort should be devoted to understanding and carefully modeling the relationship between the mutation process acting on protein coding genes and the precise parameterization of equilibrium frequencies in codon substitution models.

The results from the real data analyses confirm that different likelihood models can be consistent and detect some evidence for translational selection (e.g., the enterobacteria data set). The *Drosophila* data are much more problematic. Previous analysis suggested that a discrepancy between results for *Drosophila* was due to the combined effects of an inappropriate correction for multiple substitutions at a site and ignoring the effect of biased codon frequencies (Dunn et al., 2001; but see Bierne and Eyre-Walker, 2003). In this study, we illustrate that the results for *Drosophila* are even more sensitive to the precise model parameterization than previously recognized. We showed that two likelihood approaches, GY and MG, which both use a probabilistic model to correct for multiple substitutions and include a correction for biased codon usage, can lead to different results. The only difference between the two parameterizations is about how equilibrium frequencies are defined. This illustrates that biological conclusions derived from genome-scale analyses involving estimation of synonymous substitution rates can be sensitive to our understanding of the underlying mutation process. We suggest that in some cases it might be necessary to model the mutational process at the nucleotide level, and assume that their rates do not depend on the codon towards which the mutation occurs (possibly mammals), whereas in other cases it might be necessary to specify a model that depends on the equilibrium frequency of the target codon (possibly *Drosophila*). We note that all such models are idealizations; additional work, including diagnostics for model applicability, is warranted.

One alternative to modeling the equilibrium frequency of the target codon is to directly model a substitution process that puts the emphasis on mutations at the nucleotide level and that takes neighboring effects into account, either locally at the scale of a few surrounding sites (Siepel and Haussler, 2004), or globally at the scale of the entire protein (Robinson et al., 2003). When substitutions are known to be context-dependent in the genes under scrutiny, these latter approaches are expected to perform better, but this comes at a non-negligible computational cost. In genomic-scale studies, the concern might become that of a trade-off between computational throughput, and correctness of the estimates to reach robust biological conclusions.

Acknowledgments

This work was funded by an NSERC grant (DG 311625) to SAB. JPB was partially supported by a start-up grant from the Genome Atlantic Centre of Genome Canada, and by an NSERC grant (DG 298394). We thank Dr Giorgio Bernardi and an anonymous referee for comments that helped improve the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2006.04.024.

References

Akashi, H., Eyre-Walker, A., 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8, 688–693.

Aris-Brosou, S., 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol. Biol. Evol. 22, 200–209.

Bielawski, J.P., Dunn, K.A., Yang, Z., 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156, 1299–1308.

Bierne, N., Eyre-Walker, A., 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates. Implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics 165, 1587–1597.

Drummond, D.A., Raval, A., Wilke, C.O., 2006. A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23, 327–337.

Dunn, K.A., Bielawski, J.P., Yang, Z., 2001. Substitution rates in *Drosophila* nuclear genes. Implications for translational selection. Genetics 157, 295–305.

Eyre-Walker, A., Bulmer, M., 1995. Synonymous substitution rates in enterobacteria. Genetics 140, 1407–1412.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

Kosakovsky Pond, S.L., Frost, S.D., 2005a. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol. Biol. Evol. 22, 478–485.

Kosakovsky Pond, S.L., Frost, S.D., 2005b. A simple hierarchical approach to modeling distributions of substitution rates. Mol. Biol. Evol. 22, 223–234.

Kosakovsky Pond, S.L., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. In: Nielsen, R. (Ed.), Statistical Methods in Molecular Evolution. Springer, New York, NY, pp. 125–181.

Muse, S.V., 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. 13, 105–114.

Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N., Thorne, J.L., 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20, 1692–1704.

Sharp, P.M., Li, W.H., 1987. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Siepel, A., Haussler, D., 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. 21, 468–488.

Smith, N.G., Eyre-Walker, A., 2001. Nucleotide substitution rate estimation in enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. Mol. Biol. Evol. 18, 2124–2126.

Wall, D.P., et al., 2005. Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. U. S. A. 102, 5483–5488.

Whelan, S., Goldman, N., 2004. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 167, 2027–2043.

Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37, 441–456.

Wong, W.S., Nielsen, R., 2004. Detecting selection in noncoding regions of nucleotide sequences. Genetics 167, 949–958.

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87, 23–29.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., 2001. Adaptive molecular evolution. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, London, pp. 327–350.

Yang, Z., Nielsen, R., 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46, 409–418.

Yohai, V.J., 1987. High breakdown-point and high-efficiency robust estimates for regression. Ann. Stat. 15, 642–656.