

Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

Gene

journal homepage: [www.elsevier.com/locate/gene](http://www.elsevier.com/locate/gene)

## Genetic robustness at the codon level as a measure of selection

Marco Archetti\*

Department of Zoology, University of Oxford, UK  
St John's College, University of Oxford, UK

### ARTICLE INFO

#### Article history:

Received 9 July 2008

Received in revised form 15 May 2009

Accepted 19 May 2009

Available online 27 May 2009

Received by M. Di Giulio

#### Keywords:

Genetic robustness

Error minimization

Volatility

Codon usage

Natural selection

### ABSTRACT

Selection at the DNA level is usually detected by analysing substitution rates from multiple-species comparisons. It has been suggested that measures of genetic robustness at the codon level, which can be measured by analysing a single coding sequence, can be used to estimate selection, but the validity of these measures has been questioned. Here I test the efficiency of different measures of genetic robustness at the codon level to estimate the level of selection acting on a gene. I find that volatility and other measures of robustness are correlated with  $dN/dS$ , and that this is not simply the effect of a preference for translationally optimal codons. I discuss the possible implications and the possible problems of these methods based on single-sequence codon usage analysis.

© 2009 Elsevier B.V. All rights reserved.

### 1. Introduction

Natural selection at the DNA sequence level is usually assessed by comparing the number of synonymous substitutions per synonymous site ( $dS$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $dN$ ) from an alignment of at least two homologous protein-coding DNA sequences (Li, 1997; Nei and Kumar, 2000).

Two methods have been proposed, independently (“volatility” by Plotkin and Dushoff, 2003; Plotkin et al., 2004 and “degree of error minimization” by Archetti, 2004a,b), that measure genetic robustness of a coding sequence, and could be used to detect selection by analysing a single sequence, with no need of comparative analysis. These methods rely on the fact that synonymous codons, though neutral at the protein level, have different mutant codons with different fitness, and therefore different rates of back mutations (Archetti, 2004b; Plotkin et al., 2004) or because they affect differently the fitness of competing individuals (Archetti, 2006). The main difference between the two methods is that volatility (Plotkin et al., 2004) is a relative measure (relative to the other genes of the genome), whereas the

degree of error minimization (Archetti, 2004a,b) is an absolute measure (depending only on the genetic code used).

Both methods rely on codon usage analysis of single coding sequences, with no need of comparative analysis. Since homologous sequences are not always available, a method to detect selection based on direct single-sequence analysis would be a useful tool in the study of molecular evolution. Both Archetti (2004b) and Plotkin et al. (2004) found a correlation between genetic robustness and substitution rates in a limited set of genes. Therefore the importance of these methods as a complement to  $dN/dS$  is worth investigating.

Volatility (Plotkin et al., 2004) has been criticised on different grounds, and virtually all the subsequent analyses (Hahn et al., 2005; Nielsen and Hubisz, 2005; Sharp, 2005; Chen et al., 2005; Dagan and Graur, 2005; Friedman and Hughes, 2005; Stoletzki et al., 2005; Zhang, 2005; Pillai et al., 2005) seem to suggest the rejection of volatility as a method to detect selection, whereas the degree of error minimization (Archetti, 2004b) has received less attention, primarily in studies concerning the evolution of the genetic code (Goodarzi et al., 2005; Marquez et al., 2005).

The critiques to volatility belong to three broad categories: (i) theoretical or conceptual uncertainties on the method, (ii) existence of confounding factors, and (iii) doubts on the validity of the results (correlation with  $dN/dS$ ). Clearly, as Stoletzki et al. (2005) point out, critiques of the first and second kind are rather unimportant if the method does provide an alternative valid measure to  $dN/dS$ .

The critiques put forward by Hahn et al. (2005), Nielsen and Hubisz (2005), Sharp (2005), Chen et al. (2005), Zhang (2005) and Pillai et al. (2005) and also Dagan and Graur (2005), belong to the

Abbreviations: CAI, codon adaptation index; CRI, codon robustness index;  $dS$ , number of synonymous substitutions per synonymous site;  $dN$ , number of nonsynonymous substitutions per nonsynonymous site; ENC, effective number of codons; PCA, pretermination codon avoidance;  $R_N$ , degree of error minimization, unweighted;  $wR_N$ , degree of error minimization, weighted.

\* Current address: Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138-2902, USA.

E-mail address: [archetti@fas.harvard.edu](mailto:archetti@fas.harvard.edu).

first and second category. Stoletzki et al. (2005), and Friedman and Hughes (2005), and in part also Dagan and Graur (2005), belong to the third. Stoletzki et al. (2005) show that a correlation between  $dN/dS$  and volatility actually exists but suggest that it is a byproduct of a correlation with preference for codons that optimize translation efficiency, while Dagan and Graur (2005) and Friedman and Hughes (2005) question the very existence of a correlation.

I develop new measures of genetic robustness and I compare these and previous measures with  $dN/dS$  values derived from classical comparative analysis. I also re-analyse the data for which volatility has been questioned, to test whether the other measures described here can provide a valid alternative to  $dN/dS$  or whether they have the same problems as volatility. I then discuss whether the theoretical critiques that have been put forward about volatility can be applied to other measures as well.

## 2. Methods

### 2.1. Volatility

Volatility has been described and used by Plotkin and Dushoff (2003) and by Plotkin et al. (2004). The volatility  $v(i)$  of codon  $i$  is defined as

$$v(i) = \frac{1}{n} \sum_n D[A(i), A(i^*)]$$

where the sum is calculated over the  $n$  non-stop codons  $i^*$  that can mutate into  $i$  by a single point mutation and  $D$  is the distance (dissimilarity) between the amino acid  $A(i)$  coded by codon  $i$  and its mutant  $A(i^*)$  coded by codon  $i^*$ . Plotkin et al. (2004) use the simplest possible measure for  $D$ , the Hamming metric, which equals zero if two amino acids are identical, and one otherwise, although other similarity matrices can be used. Volatility as described in Plotkin et al. (2004) does not take into account mutations to and from stop codons, although termination codons are included in a previous version (Plotkin and Dushoff, 2003). The volatility of a gene is the summed volatility of the codons in the coding sequence. To calculate the volatility  $P$  value of the gene, the observed volatility is compared with a bootstrap distribution of  $10^6$  synonymous versions of the gene in which nucleotide sequences are produced with the same translation as the original but whose codons are drawn randomly according to the relative frequencies of codons in the genome as a whole. The  $P$  value for the gene is given by the proportion of the randomization trials in which volatility exceeds or equals the volatility of the original gene.

### 2.2. Degree of error minimization

This measure has been described by Archetti (2004a,b) and further used in Archetti (2006). The mean dissimilarity (MD) between the amino acid coded by each codon and its possible mutants is calculated by using one of the many amino acid similarity matrices available. Here I use either the Hamming metric (used previously for volatility by Plotkin et al. 2004) or McLachlan's chemical similarity matrix (used previously by Archetti, 2004a,b and 2006). Different matrices do not change the result drastically (Archetti, 2004a,b). For each pair of amino acids, the measure  $D_{AA/AA^*} = \omega_{AA/AA} - \omega_{AA/AA^*}$  can be derived from the matrix, where  $\omega_{AA/AA}$  is the similarity of amino acid AA with itself (this value is usually the same for all amino acids, but not in all matrices: in McLachlan's it is either 8 or 9) and  $\omega_{AA/AA^*}$  is the similarity of AA with the mutant amino acid AA\* obtained after an error at one of the three positions of the original codon. Hence,  $D_{AA/AA^*}$  is the distance (dissimilarity) between the original (AA) and the mutant (AA\*) amino acid.

Similarities between amino acids and termination signals ( $\omega_{AA/STOP}$ ) are not tabulated in similarity scoring matrices; however a measure of the damage produced by mutations to termination codons is considered in the calculation by introducing a score for  $\omega_{AA/STOP}$  that is less or equal to the lowest similarity score of the matrix (0 in McLachlan's matrix); here I will use  $\omega_{AA/STOP} = 0$  or  $-10$  (more extreme values do not seem to make much difference – Archetti, 2004a,b). Since  $\omega_{AA/AA} \geq \omega_{AA/AA^*}$  for every amino acid,  $D_{AA/AA^*}$  is always positive, and since there are three possible mutants for each position, there are nine measures of  $D_{AA/AA^*}$  for each codon, corresponding to the nine possible mutant codons. Their mean value is taken as a measure of distance (dissimilarity) between the original codon and its possible mutants. I call this measure  $MD$  (Mean Distance). Optimal codons are predicted to have small  $MD$  values.  $MD$  values can be calculated on one or more mutation events (this do not seem to make much difference – Archetti, 2004b). A complete explanation of the method to calculate the  $MD$  values on more than one mutation can be found in Archetti (2004b).

To calculate the degree of error minimization of a coding sequence, the correlation between the  $MD$  values and the corresponding codon frequencies is calculated for each synonymous codon family. The number of degenerate synonymous codon families on which the correlation is calculated ( $N$ ) depends on the genetic code and on the assumptions about mutation rates; for the standard genetic code, usually  $N = 18$ , but  $N < 18$  if for some amino acid there is no variance in the  $MD$  values or in the frequency of its synonymous codons (this depends on the similarity matrix used: the values in some matrices are equal for different amino acids, resulting in no correlation and therefore a lower value of  $N$  – Archetti, 2004b) or for the frequencies of their synonymous codons (this depends on the sequence).

The unweighted degree of error minimization ( $R_N$ ) is obtained by dividing the mean correlation among the amino acids ( $R$ ) by  $N$ ; ( $R_N = R/N$ , therefore, ranges between  $-1$  and  $1$ ) and measures genetic robustness with the assumption that all amino acids are weighted equally, irrespective of their frequency on the protein. If the value of each correlation is weighted (multiplied) by the frequency of the corresponding amino acid, and the mean correlation divided by  $N$ , then the degree of error minimization is denoted by  $wR_N$ . Since  $MD$  is a measure of dissimilarity, for both  $R_N$  and  $wR_N$  the lower the value the higher the degree of error minimization. A “robust” codon usage will be one with a very low  $wR_N$  (or  $R_N$ ), while an “anti-robust” (hypersensitive) codon usage will have a very high  $wR_N$  (or  $R_N$ ).

### 2.3. Codon robustness index

This measure is developed and used here for the first time. For an amino acid encoded by  $n$  synonymous codons  $i$ , let  $f(i)$  be the frequency of codon  $i$  among its synonymous codons in the sequence,  $MD(i)$  the  $MD$  value (calculated exactly as for the degree of error minimization – see above) of codon  $i$ ,  $MD_{\min}$  and  $MD_{\max}$  the highest and lowest, respectively,  $MD$  values among the synonymous codons coding for that amino acid. The codon robustness  $C$  of this amino acid in the sequence is defined by:

$$C = \sum_{i=1}^n f(i) \frac{MD(i) - MD_{\min}}{MD_{\max} - MD_{\min}}$$

The mean value of  $C$  over all the degenerate amino acids is the unweighted codon robustness index of the sequence. If the values of  $C$  are weighted according to the relative frequency of the corresponding amino acid in the sequence we obtain a weighted measure of the codon robustness index ( $CRI$ ) which is possibly more indicative of the properties of the sequence because each amino acid is assumed to have an importance corresponding to its frequency. In what follows I

will always refer to the weighted version of *CRI* unless otherwise specified. If for all the degenerate amino acids only the codon with the highest *MD* value (the less robust one) is used then  $CRI = 1$ , whereas if only the codon with the lowest *MD* value is used then  $CRI = 0$ . Intermediate frequencies produce intermediate values accordingly: high values (towards 1) indicate sensitivity to mutations (anti-robustness) while low values (towards 0) indicate robustness.

#### 2.4. Disposable pretermination codon avoidance

This measure is also developed and used here for the first time. Seven codons in the standard nuclear genetic code (GGA, AGA, CGA, TTA, TTG, TCA, TCG) are disposable “pretermination” codons (Modiano et al., 1981), that is one-mutation neighbour to termination codons that have alternative synonymous codons that are not pretermination codons. These disposable pretermination codons can be therefore avoided in favour of other synonymous codons that are not prone to mutate to termination codons. Note that there are other codons that are one-mutation away from termination codons but that are not disposable (that is all their synonymous codons are also one-mutation away from termination codons).

A simple way to measure robustness may be by measuring the frequency of disposable pretermination codons. More specifically I divide the observed (in the sequence) relative frequency (among their synonymous codons) of disposable pretermination codons by their expected frequency (1/4 for GGA and 1/6 for the others), double the values for TTA and TCA (because these codons have two one-mutation neighbour termination codons) and take the mean of these nine values. This mean value is called *PCA* (pretermination codon avoidance). If the observed frequency of disposable pretermination codons is not different from the expected frequency, then  $PCA = 1$ ; if  $PCA > 1$  then there are more disposable pretermination codons than expected in the sequence, whereas if  $PCA < 1$  the sequence tends to avoid disposable pretermination codons.

#### 2.5. *dN/dS*, codon usage and robustness

For *Saccharomyces cerevisiae* (4133 genes), *Plasmodium falciparum* (1827 genes) and *Mus musculus* (2021 genes), *dN/dS* measures were obtained from Friedman and Hughes (2005). The data used for *S. cerevisiae* by Stoletzki et al. (2005) are a subset of those used by Friedman and Hughes (2005). I managed to identify 90% of their genes in the gene list of Friedman and Hughes (2005). In order to avoid confounding effects due to short length, only genes of at least 100 codons were used.

Codon usage bias (as a pure measure of codon frequencies, with no regard for the capacity to reduce errors) is measured in two

ways: the effective number of codons (*ENC*, Wright, 1990) and the Codon Adaptation Index (*CAI*, Sharp and Li, 1987), calculated using the program CodonW (written by J. Peden and available at <http://www.molbiol.ox.ac.uk/cu>) based on the values in Sharp and Cowe (1991). The software used to develop these tests and the full set of data analysed are available on request or can be downloaded at <http://users.ox.ac.uk/~zool0643>. Volatility can be calculated at <http://volatility.cgr.harvard.edu/cgi-bin/volatility.pl> as described by Plotkin et al. (2004).

### 3. Results

#### 3.1. Genetic robustness is correlated with *dN/dS*

I analysed the correlation between *dN/dS* and measures of genetic robustness for species that had been analysed previously and for which no evidence of a correlation between *dN/dS* and volatility had been found. The results are in Tables 1 and 2. In *S. cerevisiae* I found a correlation between *dN/dS* and all the measures of genetic robustness except *PCA* and except the modified versions of volatility. Plotkin et al. (2004), mention that the correlation between *dN/dS* and volatility is higher ( $r = -0.38$ ) and according to Stoletzki et al. (2005)  $r = -0.224$ . The data used for *S. cerevisiae* by Stoletzki et al. (2005) are a subset of those used by Friedman and Hughes (2005) (90% of the genes of Stoletzki et al. in the gene list used by Friedman and Hughes). I also found a good correlation (Table 1) between the original measure of volatility and *dN/dS*. Because this is consistent with the results of Plotkin et al. (2004) and of Stoletzki et al. (2005), and because I used the same genes and *dN/dS* used by Friedman and Hughes (2005) – but measured volatility independently with the program of Plotkin et al. (2004) – it is surprising that Friedman and Hughes (2005) report no significant correlation between *dN/dS* and volatility for yeast.

#### 3.2. This correlation is not simply a byproduct of a correlation with codon bias for translation efficiency

Stoletzki et al. (2005), however, show that the correlation is stronger between codon bias (*CAI*) and *dN/dS*. Therefore they suggest that the correlation between volatility and *dN/dS* is only a byproduct of a correlation between codon bias (due to selection for translation efficiency) and *dN/dS*. Although in my results the correlation between *dN/dS* and volatility seems slightly stronger than the one reported by Stoletzki et al. (2005) – this may be due to the fact that my analysis includes 4133 genes, while Stoletzki et al. (2005) used only 1077 genes – I confirm that *dN/dS* seems better correlated with *CAI* than with volatility (Tables 1 and 2).

**Table 1**  
Correlations with *dN/dS* in *Saccharomyces cerevisiae* (4133 genes), *Plasmodium falciparum* (1827 genes) and *Mus musculus* (2021 genes).

	Correlation with <i>dN/dS</i>						
	<i>S. cerevisiae</i>		<i>M. musculus</i>		<i>P. falciparum</i>		
	[ENC]	[CAI]	[ENC]	[ENC]	[ENC]		
$R_N$	0.300**	0.214	0.189	0.156**	0.101	0.041	0.040
$wR_N$	0.317**	0.218	0.198	0.137**	0.086	0.239**	0.239
<i>CRI</i> (unweighted)	0.245**	0.174	0.133	0.144**	0.063	0.041	0.041
<i>CRI</i>	0.289**	0.185	0.161	0.061*	-0.025	0.300**	0.299
<i>PCA</i>	-0.001	0.055	-0.003	0.125**	0.047	-0.157**	-0.157
Volatility ( <i>H</i> )	0.035	-0.030	0.031	-0.108**	-0.020	0.167**	0.166
Volatility ( <i>M</i> )	0.028	-0.044	0.019	-0.112**	-0.026	0.152**	0.151
Volatility	-0.263**	-0.176	-0.171	-0.060*	-0.013	-0.038	-0.038

Volatility is calculated either with the original method or including mutations to termination codons ( $\omega_{AA/STOP} = -10$ ), using the Hamming metric (*H*) or McLachlan's matrix (*M*);  $R_N$ ,  $wR_N$ , and *CRI* are calculated with McLachlan's matrix ( $\omega_{AA/STOP} = -10$ , 10 generations). Volatility for *S. cerevisiae* is calculated with a transition–transversion ratio = 4.1 as indicated by Plotkin et al. (2004). In the other cases no transition–transversion ratio has been assumed. Significant correlations are marked (except partial correlations): \*\* $P < 0.00001$ ; \* $P < 0.01$ . Correlations that are opposite to the expected correlations are in italics. Note that for volatility *p*-values (as opposed to raw volatility) are compared to *dN/dS*, therefore a negative correlation coefficient is expected. Partial correlations are in the right columns for each species and the measure controlled for (*CAI* or *ENC*) is indicated in square parentheses.

**Table 2**  
Correlations of measures of codon bias with  $dN/dS$  for the same genes of Table 1.

	Correlation with $dN/dS$		
	<i>S. cerevisiae</i>	<i>M. musculus</i>	<i>P. falciparum</i>
CAI	−0.289*		
CAI [ $wR_N$ ]	−0.146		
CAI [CRI]	−0.162		
ENC	0.313*	0.167*	−0.026
ENC [ $wR_N$ ]	0.212	0.129	−0.026
ENC [CRI]	0.222	0.158	−0.029

Significant correlations are marked (except partial correlations): \* $P < 0.00001$ . Partial correlations for each species are in the rows indicated by the measure controlled for ( $wR_N$  or CRI) in square parentheses. for *M. musculus* and *P. falciparum* were not available.

The results for other measures of genetic robustness, however, are different. Correlations between  $dN/dS$  and most measures of genetic robustness are stronger, or at least equivalent, than the correlation between  $dN/dS$  and CAI in *S. cerevisiae* (Tables 1 and 2). Partial correlations with  $dN/dS$  (controlling for CAI) are also stronger, or at least equivalent, for measures of genetic robustness than for CAI. Therefore measures of genetic robustness (with the exception of PCA) seem a good estimate of selection ( $dN/dS$ ) in *S. cerevisiae*, and not just a byproduct of selection for translation efficiency based on tRNA abundance.

In *M. musculus* I also found significant, though weaker, correlations (Table 2); the correlation between  $dN/dS$  and CRI is the weakest. I also found correlations between  $dN/dS$  and alternative measures of volatility, though weaker than correlations with other measures. The correlation with the original measure of volatility is very weak, as reported by Friedman and Hughes (2005).

In *P. falciparum*, Plotkin et al. (2004) report a correlation between  $dN/dS$  and volatility, though Friedman and Hughes (2005) question their results and show that there is no correlation. I found that while some measures ( $wR_N$  and CRI) yield a strong correlation, volatility and other measures (the unweighted versions of CRI and  $R_N$ ) yield only a weak correlation or no correlation at all (Table 1) or with a sign opposite to the expectations (PCA and alternative measures of volatility).

To summarize, measures of genetic robustness seem capable to provide an estimate of  $dN/dS$ ,  $wR_N$  being the most reliable measure among species (CRI seems a better estimate but it fails to provide a strong correlation in *Mus musculus*). Volatility has also a good correlation with  $dN/dS$  in many cases.

#### 4. Discussion

I have shown that alternative measures based on the analysis of codon usage from a single sequence are correlated with  $dN/dS$  values derived from comparative analysis, at least in the species analysed here. Because the validity of one of these measures (volatility) has received many critiques, I think it is worth pointing out the differences between the idea on which volatility is based and the logic of other measures of genetic robustness.

The first main difference is that volatility is a relative measure of genetic robustness: it calculates the expected volatility from the mean codon usage of the genome and tests whether the gene under examination has a higher or lower volatility. It is therefore a relative measure – relative to the other genes of the genome – whereas the other measures presented here are absolute measures depending only on the structure of the genetic code. This means, among other things, that in order to calculate volatility it is necessary to know the mean codon usage of the genome, whereas in the other methods this is not necessary.

The second main difference is that Plotkin et al. (2004) do not take into account mutations to and from stop codons in the

calculation of volatility, because they consider volatility primarily the result of different rates of back mutations (which excludes stop codons as codons of origin – but see Plotkin and Dushoff, 2003). The degree of error minimization (Archetti, 2004b) and the other measures presented here, instead, were developed as measures of genetic robustness (phenotypic sensitivity to mutations) and they do take into account mutations to stop codons, because mutations to stop codons do have an impact on robustness. Mutations to stop codons will influence in complex ways the capacity to detect positive selection and cannot be ignored. If a sequence is under positive selection, codons that after mutations produce different amino acids (anti-robust) may have an advantage, yet these codons will still be under negative (purifying) selection against mutations to stop codons. It seems, however, that alternative measures of volatility in which mutations to stop codons are taken into account do not always change the results (see Table 1), therefore the difference in the results probably depends also on other causes.

Measures of genetic robustness at the codon level seem to provide a reliable way to detect selection in all the cases analysed here (*S. cerevisiae*, *M. musculus*, *P. falciparum*), and it is not just a spurious effect of preference driven by tRNA abundance. It might be argued with the same confidence, reverting the logic, that it is tRNA abundance that is a byproduct of selection for genetic robustness.

Because volatility has been criticised on different grounds, I will try to discuss these critiques and whether they also apply to the other measures and results presented here. Critiques to volatility have been put forward by Hahn et al. (2005), Nielsen and Hubisz (2005), Sharp (2005), Chen et al. (2005), Dagan and Graur (2005), Friedman and Hughes (2005), Stoletzki et al. (2005), Zhang (2005), and Pillai et al. (2005). They can be divided in three broad categories: (1) that there is no theoretical justification for the method; (2) that there is a confounding factor that may in principle impair the method; (3) that the method produces measures that are actually not correlated with  $dN/dS$ . Clearly the third critique is the main one: if the method does provide a reliable alternative to  $dN/dS$ , then the first and second critiques become unimportant. On the other hand, if the method is not reliable, then it is not necessary to analyse the details further. I will also discuss the first and second critique, though, because I believe it is important to clarify certain aspects of the idea and the methods.

##### 4.1. Theoretical basis

4.1.1. *There is no formal theoretical model describing the effects of selection at the protein level on the frequency of synonymous codons (Hahn et al., 2005; Chen et al., 2005)*

There are now formal mathematical models (Archetti, 2006; Plotkin et al., 2006) showing that with realistic mutation rates and selection coefficients, synonymous codons can actually change in frequency because of selection on protein conformation (with robust codons increasing under negative selection).

4.1.2. *The theoretical basis is trivial, because assigning an allele lower fitness will deterministically lower its frequency (Hahn et al., 2005). High-volatility still have the same fitness of their other low-volatility synonymous codons and therefore cannot increase in frequency (Zhang, 2005)*

These concerns are based on a misunderstanding (also noted by Plotkin et al., 2005). In the discussion of Plotkin et al. (2004), certain codons are supposed to increase in frequency not because they have a higher fitness, but because they have different mutants and therefore different rates of back mutations. Different rates of back mutations produce synonymous codon usage in a simple model of mutation-selection balance (Archetti, 2006; Plotkin et al., 2006). However there is one important circumstance, soft selection

(Archetti, 2006, 2009), in which the theoretical basis is not trivial and in which anti-robust codons can increase in frequency because of the effect on local competition.

*4.1.3. Positive selection rarely increases volatility in a theoretical model (Zhang, 2005); volatility does not provide a reliable way to identify positive selection in a theoretical model (Nielsen and Hubisz, 2005)*

This critique is centred on positive selection. Zhang (2005) shows that frequency dependent selection in a theoretical model can, however, affect volatility, and Nielsen and Hubisz (2005) also say that in principle models of stabilizing (purifying) selection can be constructed that have effects on volatility. Indeed this is what I have shown elsewhere (Archetti, 2006). Plotkin et al. (2005) do not agree with Nielsen and Hubisz because, they say, Nielsen and Hubisz (2005) ignore the effect of population variability (basically they ignore back mutations – the same misunderstanding mentioned in point 1.2). I tend to agree with Plotkin et al. (2005) on this point though a formal model of positive selection must be developed. Yet I agree that positive selection will be difficult to detect in species with soft selection (Archetti, 2006, 2009) because, in short, negative soft selection (that is purifying selection at certain stages of the life cycle) can produce codon frequencies that may not be distinguished from codon frequencies produced by positive selection (both positive selection and negative soft selection allow the increase of anti-robust codons).

*4.1.4. In a theoretical simulation, volatility can change through different selection regimes but can also increase without positive selection and decrease without negative selection (through neutral evolution) (Dagan and Graur, 2005)*

This does not necessarily mean that codon frequencies evolve only by drift. It is a possibility. Moreover, such simulations ignore population variability and therefore they cannot be expected to capture the relationship between selection on proteins and codon usage because back mutations are ignored (Plotkin et al., 2005).

*4.1.5. Population size is probably too small for a signal of negative selection (Zhang, 2005)*

This is a speculation, not demonstrated by data in Zhang (2005). As Plotkin et al. (2005) notice, the effective population size of microbes are debatable.

*4.1.6. The selective pressure is probably too small (the order of the mutation rate) to be effective (Sharp, 2005)*

This was also the argument of Kimura (1983) against the suggestion of Modiano et al. (1981) that disposable pretermination codons are avoided in globin genes to avoid mutations with drastic effect. If the selection coefficients are small, Kimura's argument clearly applies. However, for stronger selection, realistic mutation rates are effective in producing synonymous codon bias (Archetti, 2006; Plotkin et al., 2006).

## 4.2. Possible confounding factors

*4.2.1. Volatility is only another measure of codon bias, because there is a correlation between codon bias and volatility (Hahn et al., 2005; Nielsen and Hubisz, 2005; Chen et al., 2005)*

Volatility and other measures of genetic robustness are not only measures of codon usage bias, they incorporate information about the impact of mutations, which measures of pure codon bias (for example ENC, CAI) do not. Furthermore this statement, even if true, clearly (as also Plotkin et al., 2005 noted) would not invalidate the method per se. Moreover I have shown (Archetti, 2004b) that while an extreme genetic robustness or anti-robustness produces an extreme codon usage bias, the reverse is not necessarily true, that is an extreme codon usage bias does not necessarily lead to an extreme degree of genetic

robustness. It depends on the frequency of the robust codons: only if the most robust codons are preferred does a high codon bias correspond to a high robustness, and it is easy to see why this would also apply to volatility. Finally, as also Plotkin et al. (2005) remark, a correlation between codon usage bias and  $dN/dS$  is also often observed. This, however, does not mean that codon bias impairs our capacity to use  $dN/dS$  to measure selection.

*4.2.2. Random volatilities produce U-shape distributions like the one obtained by Plotkin et al. (2004) with the same genes at the extremes. (Hahn et al., 2005)*

This critique clearly applies only to volatility and not to other measures and it depends on the method (compare for example the distribution in Plotkin et al., 2004 and in Archetti, 2004b, 2006, in which random degrees of robustness do not produce U-shape distributions).

*4.2.3. Volatility depends only on 4 amino acids: this is insufficient to detect selection (Chen et al., 2005; Sharp, 2005)*

This critique only applies to volatility and not to other measures, which depend on up to 18 amino acids, according to the assumptions used in the method. It is true that volatility depends only on 4 amino acids, but neither Chen et al. (2005) nor Sharp (2005) provide any evidence for their statement that this is insufficient to detect selection. Plotkin et al. (2005) reply that comparative analysis often rely on fewer than 5% of sites.

*4.2.4. Codon bias is influenced by CG content (Sharp, 2005)*

Genetic robustness at the codon level, as measured here, however, is not simply equivalent to codon bias. I have shown that it is difficult to support the idea that genetic robustness (not simply codon bias) depends on mutation bias (Archetti, 2004b, 2006). However, it should certainly be kept in mind that isochore structure, if not properly controlled for, could bias the results of codon-based methods to detect selection (Plotkin et al., 2004).

## 4.3. Empirical data

*4.3.1. Certain genes identified as under positive selection by Plotkin et al. (2004) are probably not under positive selection (Chen et al., 2005); the genes detected by Plotkin et al. (2004) as most volatile genes have peculiar repetitive internal repetitive structures (Sharp, 2005) or peculiar amino acid usage (Dagan and Graur, 2005)*

It is not clear whether this argument applies to other measures too, though I also showed for *Drosophila melanogaster* (Archetti, 2006) that certain genes that are detected as hypersensitive are probably not under positive selection. I believe, unlike Plotkin et al. (2004) that hypersensitive genes (volatile in their terminology) can be under soft negative selection (Archetti, 2006, 2009). This means that it will be difficult to detect these genes as under positive or negative selection on the basis of codon usage alone, but it does not mean that the idea that selection for protein conformation affects codon usage is wrong. In any case, this is not an argument against the method in general. Presumably also  $dN/dS$  have similar problems in some cases. Does the method provide a reliable way to detect selection, in general? If it does, then we cannot discard it simply because of few false positives due to peculiar repetitive structures or peculiar amino acid composition.

*4.3.2. There is no correlation between volatility and  $dN/dS$  in *Mycobacterium tuberculosis* (Dagan and Graur, 2005), in *Plasmodium falciparum*, *Saccharomyces cerevisiae* and *Mus musculus* (Friedman and Hughes, 2005); Pillai et al. (2005) suggest that volatility does not correspond to what is expected for six genes of the HIV virus. However Stoletzki et al. (2005) confirm the correlation between  $dN/dS$  and volatility reported by Plotkin et al. (2004) for *S. cerevisiae*, though they*

find that a slightly different measure of volatility (not the one used by Plotkin et al., 2004) fails to provide a significant partial correlation

This is the crucial argument, but of course, as it is suggested by Stoletzki et al. (2005), it depends on the method used. What I have shown here is that methods other than volatility do provide a valid alternative to  $dN/dS$ .

4.3.3. It may be possible that, even if a correlation between  $dN/dS$  is observed, this depends only on the fact that (Zhang, 2005) codon bias is related to the level of gene expression and may reflect selection for translation efficiency through preferential use of synonymous codons whose cognate tRNA has higher frequencies. In other words that the correlations are the byproduct of a correlation between bias for translationally optimal codons and  $dN/dS$  (Stoletzki et al., 2005)

Clearly, however, another possibility is that codon bias depend primarily on genetic robustness (selection on the conformation of the protein) rather than on selection for matching tRNA abundance, and that tRNA abundance evolved after codon usage evolved due to genetic robustness. This hypothesis had simply not been tested before (Archetti, 2004b).

However, while Zhang (2005) does not provide any evidence for his statement, Stoletzki et al. (2005) show that the correlation between  $dN/dS$  and a measure of codon bias ( $CAI$ ) is stronger than the correlation between  $dN/dS$  and volatility. This seems to suggest that codon bias is the primary cause, not volatility, but again this depends on the method used. Here I have shown that the correlation (even partial) of  $dN/dS$  with other measures of genetic robustness is stronger, or at least equal, than the correlation with codon bias, including  $CAI$  in *S. cerevisiae*. A correlation with levels of gene expression is predicted not only by the hypothesis of selection for translation efficiency, but also by the hypothesis of selection for the minimization of errors: it may well be that genes with higher expression levels are also more robust against mutations and *mis-translation* (Archetti, 2004b).

Finally, translation efficiency and mutational robustness may have coevolved. As I have discussed above, from a purely theoretical point of view (see Archetti, 2006) tRNA abundance bias does not necessarily produces high robustness or low robustness. It would do so only if tRNA frequencies were biased for all synonymous families in the same direction (all towards the robust codons or all towards the anti-robust codons), which is of course difficult to obtain by chance. It seems rather more likely, or at least equally likely, that tRNA frequencies may adapt to a codon bias that depends on genetic robustness (which depends only on the structure – not modifiable – of the genetic code).

#### 4.4. Conclusion

All the theoretical and methodological concerns put forward against volatility must be proved empirically; some of the critiques are only relevant to volatility but not to the other measures of genetic robustness. The crucial point is whether the empirical data show a correlation between  $dN/dS$  and measures based on codon analysis. I have shown that measures developed to study genetic robustness have a correlation with  $dN/dS$  that is not, as Stoletzki et al. (2005) had suggested, merely the spurious effect of a correlation between  $dN/dS$  and preference for codons that match the most abundant tRNA.

Being statistically significant, however, does not necessarily mean being biologically relevant. I would be cautious to generalize the possibility of using genetic robustness to detect selection (Archetti,

2006). It would be worth doing an extensive comparison of  $dN/dS$  and genetic robustness in as many species and genes as possible. Yet, the fact that a correlation exists in the cases examined here (including the cases for which volatility had been questioned), suggests that methods based on genetic robustness, given that they do not require comparative analyses, are worth investigating as an aid to detect selection.

#### Acknowledgments

Thanks to Nina Stoletzki and to Robert Friedman and Austin Hughes for providing the data used in their analyses, to Joshua Plotkin for help with the calculation of volatility and to Alan Grafen for discussion and a clarification on the statistics. I am supported by a long-term postdoctoral fellowship of the Human Frontier Science Program Organization.

#### References

- Archetti, M., 2004a. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *J. Mol. Evol.* 59, 258–266.
- Archetti, M., 2004b. Selection on codon usage for error minimization at the protein level. *J. Mol. Evol.* 59, 400–415.
- Archetti, M., 2006. Genetic robustness and selection at the protein level for synonymous codons. *J. Evol. Biol.* 19, 353–365.
- Archetti, M., 2009. Survival of the steepest: hypersensitivity to mutations as an adaptation to soft selection. *J. Evol. Biol.* 22, 740–750.
- Chen, Y., Emerson, J.J., Martin, T.M., 2005. Codon volatility does not detect selection. *Nature* 433, E6–E7.
- Dagan, T., Graur, D., 2005. The comparative method rules! Codon volatility cannot detect positive Darwinian selection using a single genome sequence. *Mol. Biol. Evol.* 22, 496–500.
- Friedman, R., Hughes, A., 2005. Codon volatility as an indicator of positive selection, data from eukaryotic genome comparisons. *Mol. Biol. Evol.* 22, 542–546.
- Goodarzi, H., Najafabadi, H.S., Torabi, N., 2005. On the coevolution of genes and genetic code. *Gene* 362, 133–140.
- Hahn, M.W., Mezey, J.G., Begun, D.J., et al., 2005. Codon bias and selection on single genomes. *Nature* 433, E5.
- Kimura, M., 1983. *The Neutral Theory of Natural Selection*. Cambridge University Press, Cambridge, United Kingdom.
- Li, W.H., 1997. Rates and Patterns of Nucleotide Substitution. *Molecular Evolution*. In: Sinauer Associates, Sunderland, Massachusetts, USA, pp. 177–213.
- Marquez, R., Smit, S., Knight, R., 2005. Do universal codon-usage patterns minimize the effects of mutation and translation error? *Genome Biol.* 6 (11), R91 Art. No.
- Modiano, G., Battistuzzi, G., Motulsky, A.G., 1981. Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes, an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc. Natl. Acad. Sci. U. S. A.* 78, 1110–1114.
- Nei, M., Kumar, S., 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, United Kingdom.
- Nielsen, R., Hubisz, M.J., 2005. Detecting selection needs comparative data. *Nature* 433, E6.
- Pillai, S.K., Kosakovsky, P., Woelk, C.H., Richman, D.D., Smith, D.M., 2005. Codon volatility does not reflect selective pressure on the HIV-1 genome. *Virology* 336, 137–143.
- Plotkin, J.B., Dushoff, J., 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc. Natl. Acad. Sci. U. S. A.* 2003 (100), 7152–7157.
- Plotkin, J.B., Dushoff, J., Fraser, H.B., 2004. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *Nature* 428, 943–945.
- Plotkin, J.B., Dushoff, J., Fraser, H.B., 2005. Reply. *Nature* 433, E7.
- Plotkin, J.B., Dushoff, J., Desai, M.M., Fraser, H.B., 2006. Codon usage and selection on proteins. *J. Mol. Evol.* 63, 553–635.
- Sharp, P., 2005. Gene “volatility” is most unlikely to reveal adaptation. *Mol. Biol. Evol.* 22 (4), 807–809.
- 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.
- Sharp, P.M., Cowe, E., 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* 7, 657–678.
- Stoletzki, N., Welch, J., Hermisson, J., Eyre-Walker, A., 2005. A dissection of volatility in yeast. *Mol. Biol. Evol.* 22 (10), 2022–2026.
- Wright, F., 1990. The ‘effective number of codons’ used in a gene. *Gene* 87, 23–29.
- Zhang, J., 2005. On the evolution of codon volatility. *Genetics* 169, 495–501.