Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation.

Michael A. Gilchrist*, Premal Shah*, and Russell Zaretzki†

Initial submission on 7 Aug 2009
Accepted pending minor revisions on 8 Sept 2009
Revised submission on 15 Sept 2009
Accepted for publication 26 Sept 2009

*Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610
†Department of Statistics, University of Tennessee, Knoxville, TN 37996-0532
Abstract

Codon usage bias (CUB) has been documented across a wide range of taxa and the subject of numerous studies. While most explanations of CUB invoke some type of natural selection, most measures of CUB adaptation are heuristically defined. In contrast, we present a novel and mechanistic method for defining and contextualizing CUB adaptation to reduce the cost of nonsense errors during protein translation. Using a model of protein translation, we develop a general approach for measuring the protein production cost in the face of nonsense errors of a given allele as well as the mean and variance of these costs across its coding synonyms. We then use these results to define the nonsense error adaptation index (NAI) of the allele or a contiguous subset thereof. Conceptually, the NAI value of an allele is a relative measure of its elevation on a specific and well defined adaptive landscape. To illustrate its utility, we calculate NAI values for the entire coding sequence and across a set of non-overlapping windows for each gene in the *Saccharomyces cerevisiae* S288c genome. Our results provide clear evidence of adaptation to reduce the cost of nonsense errors and increasing adaptation with codon position and expression. The magnitude and nature of this adaptation is also largely consistent with simulation results in which nonsense errors are the only selective force driving CUB evolution. Because NAI is derived from mechanistic models, it is both easier to interpret and more amenable to future refinement than other commonly used measures of codon bias. Further, our approach can also be used as a starting point for developing other mechanistically derived measures of adaptation such as for translational accuracy.

Codon usage bias (CUB) is defined as the non-uniform use of synonymous codons within a gene (Ikemura, 1981; Bennetzen and Hall, 1982; Sharp and Li, 1987). CUB has been extensively documented across a wide range of organisms and varies greatly both within and between genomes (Grantham et al., 1980; Ikemura, 1981, 1982, 1985; Bennetzen and Hall, 1982; Sharp and Li, 1987; Ghosh et al., 2000; Carbone et al., 2003; Mougel et al., 2004; Subramanian, 2008). Most explanations of CUB involve a mixture of factors including mutational bias, intron splicing, recombination, gene conversion, DNA packaging, and selection for increased translational efficiency or accuracy (Bernardi and Bernardi, 1986; Bulmer, 1988; Shields et al., 1988; Bulmer, 1991; Kliman and Hey, 1993, 1994; Akashi, 1994; Xia, 1996; Akashi and Eyre-Walker, 1998; Xia, 1998; Musto et al., 1999; McVean and Charlesworth, 1999; Ghosh et al., 2000; Wagner, 2000; Birdsell, 2002; Comeron and Kreitman, 2002; Akashi, 2003; Musto et al., 2003; Elf et al., 2003; Chen et al., 2004; Chamary and Hurst, 2005b,a; Comeron, 2006; Lin et al., 2006; Warnecke and Hurst, 2007; Warnecke et al., 2008; Drummond and Wilke, 2008). As a result, CUB has played an important role in the neutralist-selectionist debate (e.g. Wolfe and Sharp, 1993; Duret and Mouchiroud, 1999; Musto et al., 2001; Urrutia and Hurst, 2003; Plotkin et al., 2004; Smon et al., 2008).
2005; Chamary et al., 2006; Lynch, 2007), interpretations of molecular clocks (e.g. Long and Gillespie, 1991; Tamura et al., 2004; Xia, 2009), and phylogenetics (e.g. Goldman and Yang, 1994; Nielsen et al., 2007a; Mooers and Holmes, 2000; Nielsen et al., 2007b; Anisimova and Kosiol, 2009).

Currently, multiple indices are available for measuring the average CUB of a gene (e.g. RSCU, CAI, $F_{op}$, E(g), $N_c$, CBI, CodonO, and RCB (Sharp et al., 1986; Sharp and Li, 1987; Ikemura, 1981; Karlin and Mrazek, 2000; Wright, 1990; Bennetzen and Hall, 1982; Wan et al., 2006; Roymondal et al., 2009)). However, directly relating any of these measures to a specific biological process is difficult. As a result, while certain measures of CUB are more popular in some circles than others, there is no clear ‘correct’ measure of codon bias. These shortcomings are due, in part, to the fact that these indices are either heuristic or statistical in origin.

An alternative approach which avoids these shortcomings is to develop mechanistically based indices which are based on specific biological processes that can drive the evolution of CUB. As an example of such an approach, we present a new CUB index specifically formulated to measure the degree of adaptation an allele exhibits to reduce the cost of nonsense errors during protein translation. We refer to this index as the Nonsense error Adaptation Index or NAI value of an allele. While NAI is specifically formulated to measure adaptation to reduce the cost of nonsense errors, the approach we present could be altered to measure adaptation in CUB for other aspects like translational efficiency, translational accuracy, or, ideally, a composite measure that evaluates and partitions the importance of these and other selective forces.

Nonsense errors, also referred to as processivity errors, occur when a ribosome terminates the translation of an mRNA transcript prematurely. Nonsense errors have a number of different causes such as ribosome drop-off, improper translation of release factors, frame shifts, and even missense errors (Kurland, 1992; Hooper and Berg, 2000; Zaher and Green, 2009). Direct estimates of the per codon nonsense error rates are rare, but those that do exist for E. coli suggest that they are on the order of $10^{-4}$ per codon (Manley, 1978; Tsung et al., 1989; Jorgensen and Kurland, 1990). For S. cerevisiae, no direct estimates of nonsense error rates exist. However, Arava et al. (2003) and Arava et al. (2005) provide indirect measures of nonsense error rates for S. cerevisiae which are on the order of $10^{-3}$ to $10^{-4}$ per codon. Although there is still great uncertainty, together these data imply that the probability of a nonsense error occurring during the translation of an average length protein of $\sim 400$ amino acids is greater than 20%.

Because most incomplete peptides are expected to be non-functional, such peptides impose various costs to the cell. For example, the production of incomplete peptides can tie up essential cell resources such as tRNAs and ribosomes (Dincbas et al., 1999; Cruz-Vera et al., 2004). Further, the recognition and
breakdown of incomplete peptides may require additional resources as well and the peptides themselves may be toxic to the cell (Menninger, 1978). Another important cost of a nonsense error is the amount of energy invested during protein production into the assembly of the polypeptide chain (Bulmer, 1991; Kurland, 1992; Eyre-Walker, 1996). Because the cell must expend energy during each elongation step of translation, the cost of a nonsense error will increase with codon position at which it occurs. As a result, selection against nonsense errors leads to the unique prediction that CUB should increase intragenically with codon position. Numerous researchers have shown either directly or indirectly that CUB does indeed increase with codon position in E. coli and other microorganisms (Hooper and Berg, 2000; Qin et al., 2004; Gilchrist and Wagner, 2006; Stoletzki and Eyre-Walker, 2007).

In this study we specifically define adaptation as the degree to which an allele reduces the cost of nonsense errors during protein translation. Variation in adaptation between alleles is the result of different synonymous codons having different nonsense error rates. We evaluate an allele’s nonsense error cost relative to the coding synonyms of its synonymous genotype space, i.e. the set of alleles which differ in the synonymous codons they use but not the amino acid sequence they encode. It is worth noting that other measures of CUB also restrict their focus to that of coding synonyms. Because we restrict our focus to only synonymous changes in a coding sequence, we avoid the more complex question of how non-synonymous substitutions affect protein function. Instead, we are able to focus solely on the effect synonymous substitutions have on the expected cost of nonsense errors during protein translation.

To define our Nonsense error Adaptation Index (NAI) we use a simple model of protein elongation to generate a genotype to phenotype mapping function. In our mapping function, a genotype is the specific codon usage of an allele and a phenotype is the expected amount of high energy phosphate bonds $\sim P$ that must be broken to generate the benefit equivalent of one functional protein (i.e. a protein production cost-benefit ratio or production cost, for brevity). We then contextualize the production cost of a given allele as a Z-score using the analytically derived measures of the mean and variance of these costs for its entire set of coding synonyms. The resulting NAI score of an allele is a relative measure of its elevation on a nonsense error cost adaptive landscape for the protein it encodes. More precisely, the NAI score of an allele measures the relative deviation of an allele’s nonsense error cost from its expected value scaled by the standard deviation of these costs. Because NAI is well defined from both a statistical and biological standpoint, it has a number of important advantages over other measures of CUB.

For example, because it is based on a Z-score, an allele’s NAI score is easy to interpret statistically since it can be directly related to the cumulative distribution function (CDF) of a standard normal distribution.
As a result, an allele with an NAI score of 0 indicates that it is more adapted to reduce the cost of nonsense errors than half of all of the other alleles in its synonymous genotype space. The fact that the CDF of a standard normal distribution is 95% at 1.645 means that an allele with an NAI score of 1.645 is more adapted than 95% of its coding synonyms and, therefore, could be classified as showing statistically significant signs of adaptation. None of the other commonly used CUB indices have such a clear statistical interpretation. Further, because the null expectation of NAI is a standard normal distribution, NAI measurements meet the assumptions of most standard statistical approaches such as general linear regression. Taking advantage of this property through the use of a hierarchical regression model, we are able to detect significant signals of adaptation to reduce the cost of nonsense errors across most of the *Saccharomyces cerevisiae* genome. Specifically, we find that NAI increases with both codon position and an allele’s protein production rate. Using permutation techniques and a simulation model of codon evolution, we show that these observations are consistent with the hypothesis that selection to reduce the cost of nonsense errors plays an important role in driving the evolution of CUB.

MATERIALS AND METHODS

The calculation of an allele’s NAI score can be broken down into four steps. The first step is the calculation of the codon specific elongation and nonsense error probabilities. The second step is to use these probabilities to calculate the nonsense error cost-benefit ratio, $\eta$, for the codon sequence of any given allele of a gene. The third step is to calculate the mean and variance in $\eta$ values across the synonymous genotype space of that gene. The fourth and final step is to combine a given allele’s $\eta$ value and the moments of $\eta$ across its synonymous genotype space to calculate the NAI score for that allele. Because these calculations are based on explicit biological processes, any one of them can be expanded upon or refined in future studies.

**Step 1) Calculating the per codon elongation and nonsense error probabilities:** Here we use a simple model we developed previously in Gilchrist and Wagner (2006) to calculate codon specific elongation probabilities. In this model, each elongation step is viewed as an exponential waiting process with two possible outcomes: successful elongation or the occurrence of a nonsense error. Conceptually, we assume that abundances vary between tRNA species and, following the law of mass action, this variation in tRNA abundances leads to variation in elongation rates between codons. We also assume that other factors such as codon wobble can affect the elongation rate of a codon as well. We represent the elongation rate of a particular codon as $c(NNN)$, where $N \in \{A,T,G,C\}$. While elongation rates can vary between
codons, conversely we assume that all codons experience the same universal nonsense error rate $b$. Given these assumptions, the probability a ribosome will successfully complete an elongation step at some codon $NNN$ is

$$p(\text{NNN}) = \frac{c(\text{NNN})}{c(\text{NNN}) + b}. \quad (1)$$

Consequently, the probability a ribosome will experience a nonsense error at the same codon is $1 - p(\text{NNN})$.

For simplicity, we will also assume that stop codons always lead to termination of translation, i.e. $p(TAA) = p(TAG) = p(TGA) = 0$. We wish to emphasize that there is ample room for the development of more complex and biologically accurate relationships between a codon $NNN$ and its elongation probability $p(\text{NNN})$ as defined in Equation (1). However, such elaborations are beyond the scope of this study. As long as the calculation of $p(\text{NNN})$ at each codon is independent of the other codons, any refinement of the model underlying the calculation of $p(\text{NNN})$ will not alter how these values are used in the calculations that follow.

**Step II) Calculating the cost of nonsense errors $\eta$:** In general, we expect natural selection to favor alleles that produce protein functionality more efficiently than others. Therefore, we define adaptation as reduction in the expected cost of producing the equivalent of one functional protein, $\eta$. More specifically, $\eta$ describes a cell’s expected cost in high energy phosphate bonds $\sim P$ for translating an allele divided by the expected benefit the cell gains from the translation product. The use of a cost-benefit ratio as opposed to the difference between the cost and benefit of an allele is well justified since the units of cost are different than the units of benefit. More importantly, if we assume that an organism requires a certain amount of protein to be produced at some target rate, the cost-benefit ratio can be used to calculate the expected cost, in $\sim Ps$, for meeting that target. Both Gilchrist (2007) and the simulations we use here provide a clear illustration of this concept.

We explicitly measure peptide utility in relative terms such that one unit of relative utility is equal to the functionality provided by a complete and error free peptide encoded by a given gene. Measuring the utility of a peptide in this way allows us to focus on how translational errors affect the expected performance of a protein relative to an error free version as opposed to having to understand the specific function of the encoded protein. Thus, even though the importance of a protein to the organism varies between different genes, because we consider only relative, not absolute, utility, the NAI measure we produce is independent of that importance.

When considering an entire coding sequence, we use subscripts to indicate the position of a codon
relative to its start codon. By definition, the start codon is at position 0. Given the fact the first amino acid of a sequence is part of the ribosomes initiation complex, a ribosome cannot experience a nonsense error at position 0. Thus we represent a codon sequence \( \overline{NNN} = \{NNN_1, NNN_2, \ldots, NNN_n\} \) where \( n \) is the number of elongation steps required to make a peptide. Because of the start codon is at position 0, \( n \) is one less than the number of amino acids in a complete peptide. We use the notation \( p(\overline{NNN}) = \{p(\overline{NNN}_1), p(\overline{NNN}_2), \ldots, p(\overline{NNN}_n)\} \) to represent the allele specific vector of elongation probabilities for a given codon sequence. To make the notation more compact we will drop the codon itself from our notation and, instead, simply index elongation probabilities by their position within a sequence, i.e. \( p_i = p(\overline{NNN}_i) \) and, in a similar vein, leave the codon sequence itself implicit, i.e. \( \vec{p} = p(\overline{NNN}) \).

Using this notation, we will now calculate the expected energetic cost per translational initiation event for a given codon sequence based on its corresponding elongation probability vector, \( \vec{p} \). We begin by noting that in order to reach the \( i \)th codon, a ribosome must first successfully translate the preceding \( i - 1 \) codons. Using \( \sigma_i \) to represent the probability that a ribosome will successfully translate the first \( i \) codons of an allele, it follows that,

\[
\sigma_i(\vec{p}) = \prod_{j=1}^{i} p_j. \tag{2}
\]

Successful translation occurs when a ribosome translates all \( n \) codons of an allele. The probability of successful translation can therefore be denoted \( \sigma_n(\vec{p}) \). The probability a nonsense error will occur some where between codon 0 and \( n \) is simply \( 1 - \sigma_n(\vec{p}) \).

If a nonsense error occurs at the \( i \)th codon, then translation by the ribosome terminates and the amount of energy that has been expended up to this point equals \( \beta_i \). The simplest scenario and the one we employ here is to define \( \beta_i = a_1 + a_2(i - 1) \), where \( a_1 \) represents the cost of charging the fMet-tRNA (\( 2 \sim P \)) and the assembly of the ribosome on the mRNA (\( 2 \sim P \)) and \( a_2 \) represents the cost of tRNA charging (\( 2 \sim P \)) and translocation of the ribosome during each elongation step (\( 2 \sim P \)) (Bulmer, 1991; Wagner, 2005). As with calculating \( p(\overline{NNN}) \), more complex cost functions that include additional costs, such as the overhead cost of ribosome usage, could also be used to define \( \beta_i \) (see DISCUSSION).

To calculate the expected protein production cost per initiation event, \( E(Cost | \vec{p}) \), one simply sums up the cost of each possible outcome weighted by its probability of occurring. Doing so gives,

\[
E(Cost | \vec{p}) = \sum_{i=1}^{n+1} \beta_i \sigma_{i-1}(\vec{p})(1 - p_i). \tag{3}
\]
Note that the summation is taken up to \( n+1 \) to account for the stop codon where, by definition, \( 1 - p_{n+1} = 1 \).

To calculate the expected protein utility per initiation event, we first define the function \( u_i \) as the utility of a peptide for which translation has terminated at codon \( i \). We then calculate the expected utility of a gene, \( E(Benefit|\vec{p}) \), by simply summing up the utility of each peptide given its length weighted by the probability of producing it. Doing so gives,

\[
E(Benefit|\vec{p}) = \sum_{i=1}^{n+1} u_{i-1}(\vec{p}) (1 - p_i). \tag{4}
\]

Here and in our previous work we assume that \( u_i \) follows a step function where \( u_i = 0 \) for all \( i < j \), i.e. all nonsense errors lead to a non-functional protein. In the case of such a step function it follows that \( E(Benefit|\vec{p}) = \sigma_j(\vec{p}) \) and in the current study we assume that \( j = n \). As with other aspects of this work, our assumptions about \( u_i \) can easily be relaxed in future studies. For example, \( u_i \) could be assumed to be a logistic function of \( i \). Alternatively, one could expand the formulation of \( E(Benefit|\vec{p}) \) further to include the possibility of missense errors and their effects.

Combining our results from Equations (3) and (4), the expected cost over expected benefit of a coding sequence is,

\[
\eta(\vec{p}) = \frac{E(Cost|\vec{p})}{E(Benefit|\vec{p})} = \frac{\sum_{i=1}^{n+1} \beta_i \sigma_{i-1}(\vec{p}) (1 - p_i)}{\sigma_n(\vec{p})} = \sum_{i=1}^{n} \beta_i \sigma_{i-1}(\vec{p}) (1 - p_i) + \beta_{n+1}, \tag{5}
\]

where the first term on the r.h.s. of Equation 5 represents the cost-benefit of the incomplete proteins and the \( \beta_{n+1} \) term represents the cost-benefit of translating one complete protein (GILCHRIST, 2007). In summary, the term \( \eta(\vec{p}) \) represents the expected amount of \( \sim P \) that must be spent to get the benefit of one unit of utility from an allele with a given codon sequence.

**Step III) Calculating the central moments of \( \eta \):** We now shift our focus from calculating the protein production cost \( \eta \) for a specific allele to calculating the central moments of \( \eta \) across the entire set of coding synonyms. For simplicity we will focus on calculating these moments for the entire length of an allele. In the Supporting Materials A we present the details on carrying out similar calculations for the set of coding synonyms that differ from the observed allele only over a restricted window of codons, e.g. from codons 1-20, 21-40, 41-60, and so on. These calculations of NAI over a window explicitly take into account the codon usage outside of the window, thus providing a context specific measure of adaptation to reduce the cost of
nonsense errors within the window. These moments can also be estimated through simulation and we have
exploited this fact to verify that our analytic estimates of η’s central moments are correct.

Calculating the Expected Cost-Benefit Ratio: \( \bar{\eta} \)

We begin by computing the expected cost-benefit ratio \( \bar{\eta} \) for the coding synonyms of an allele. Beginning with the definition of \( \eta \) in Equation (5) and assuming that the choice of codon at each position is independent of the other, we can calculate the expected value of \( \eta \) for the entire set of coding synonyms \( S \) as

\[
\bar{\eta} = \sum_{i \in S} \eta(\vec{p}_i) \Pr(\vec{p}) = \sum_{i=1}^{n} \beta_i \mathbb{E} \left[ \frac{1 - p_i}{p_i} \right] \prod_{j=i+1}^{n} \mathbb{E} \left[ \frac{1}{p_j} \right] + \beta_{n+1}.
\]

(6)

The expectations over functions of \( p_i \), such as \( \mathbb{E}[1/p_i] \), are taken over the set of synonymous codons \( J \) for a given amino acid. Using the amino acid tyrosine \( (Y) \) as an example, \( J_Y = \{TAT, TAC\} \). Similarly, for proline \( (P) \), \( J_P = \{CCT, CCC, CCA, CCG\} \). The set of synonymous codons for the amino acid serine are unique because they occur in two distinct subsets which cannot be connected via a single nucleotide substitution. Thus, we treat each subset as a distinct amino acid, i.e., \( J_{S_1} = \{AGT, AGC\} \) and \( J_{S_2} = \{TCT, TCC, TCA, TCG\} \).

Generally speaking, the expected elongation probability for an amino acid is \( \mathbb{E}(p) = \sum_{j \in J} \Pr(NNN_j)p(NNN_j) \) where \( \Pr(NNN_j) \) is the probability of codon \( NNN_j \) occurring in a given genome. In the absence of any mutational bias all codons have equal probability of occurring, \( \Pr(NNN_j) = 1/|J| \) for all \( j \) in \( J \). If mutational bias does occur then the \( \Pr(NNN_j) \) for each codon is simply the equilibrium frequency of each codon given the biased mutation rate. For example, in order to include the effect of a genome wide AT bias in our calculation of \( \mathbb{E}(p) \), we simply set

\[
\Pr(NNN_i) = \frac{(x/(1-x))^{g(NNN_i)}}{\sum_{j \in J}(x/(1-x))^{g(NNN_j)}},
\]

where \( x \) is the degree of observed AT bias and \( g(NNN_i) \) is the number of \( A \) or \( T \) nucleotides in codon \( NNN_i \).

Calculating the Variance in the Cost-Benefit Ratio: \( \text{Var}(\eta) \)

Given our assumption of independence between \( p_i \) values at different positions, the variance in \( \eta \) values across the synonymous genotype space \( S \)
is,

$$\text{Var} \ (\eta) = \left( \sum_{i=1}^{n} \beta_i^2 \text{Var} \ (Y_i) + 2 \sum_{i=1}^{n} \beta_i \sum_{j=i+1}^{n} \beta_j \text{Cov} \ (Y_i, Y_j) \right), \quad (7)$$

where,

$$Y_i = \left( \frac{1-p_i}{p_i} \right) \prod_{k=i+1}^{n} \left( \frac{1}{p_k} \right),$$

$$\text{Var} \ (Y_i) = \mathbb{E} \left[ \left( \frac{1-p_i}{p_i} \right)^2 \right] \prod_{j=i+1}^{n} \mathbb{E} \left[ \left( \frac{1}{p_j} \right)^2 \right] - \left( \mathbb{E} \left[ \frac{1-p_i}{p_i} \right] \prod_{j=i+1}^{n} \mathbb{E} \left[ \frac{1}{p_j} \right] \right)^2, \quad (8)$$

$$\text{Cov} \ (Y_i, Y_j) = \mathbb{E} \left[ (Y_i - \mathbb{E}[Y_i]) (Y_j - \mathbb{E}[Y_j]) \right] = \left( \mathbb{E}[Y_i] - \mathbb{E}[Y_i] \mathbb{E}[Y_j] \right), \quad (9)$$

and

$$\mathbb{E}[Y_i Y_j] = \mathbb{E} \left[ \frac{1-p_i}{p_i} \right] \left( \prod_{k=i+1}^{j-1} \mathbb{E} \left[ \frac{1}{p_k} \right] \right) \left( \mathbb{E} \left[ \frac{1-p_j}{p_j} \right] \right) \left( \prod_{k=j+1}^{n} \mathbb{E} \left[ \frac{1}{p_k} \right]^2 \right). \quad (10)$$

Substituting the results from Equations (8)-(10) into (7) allows for the direct calculation of the variance in the $\eta$ values around $\bar{\eta}$ for all $S$ genotypes.

**Step IV) Calculating the NAI score of a gene:** We define $\eta_{\text{min}}$ as the cost-benefit ratio of producing a protein that uses only the most optimal codons and hence represents the minimum cost of producing that protein. Random samples of synonyms for different genes indicate that, after subtracting off $\eta_{\text{min}}$ from each value, the distribution of $\eta$ values for the synonyms of an allele are approximately Gamma distributed (see Supporting Materials B for details). Under the assumption of a Gamma distribution, we can solve for the shape and inverse scale parameters, $\alpha$ and $\beta$, respectively, based on the $\bar{\eta}$ and $\text{Var} \ (\eta)$ values calculated in Step III. Doing so gives, $\alpha = (\bar{\eta} - \eta_{\text{min}})^2 / \text{Var} \ (\eta)$ and $\beta = (\bar{\eta} - \eta_{\text{min}}) / \text{Var} \ (\eta)$ (Rice, 1995).

As mentioned earlier, the NAI of a gene is essentially a Z score which, in turn, is based on a standard normal distribution. For a given mean and variance, the Gamma distribution has greater skewness than the normal distribution. Because of this skewness, if we calculate a gene’s NAI score using the raw, untransformed moments, $\bar{\eta}$ and $\text{Var} \ (\eta)$, we will underestimate the true proportion of genotype space with alleles that are less adaptive in their $\eta$ values. To reduce this effect, we apply a Box-Cox transformation with $h = 1/3$ which is the best skewness reducing transformation for the Gamma distribution among the Box-Cox family (PACE...
and Salvan, 1997) (see Supporting Materials C for details). Using ′ to denote the Box-Cox transformed values, we define the NAI score of a gene as,

\[ \text{NAI} = -\frac{\bar{\eta}'_{\text{obs}} - \bar{\eta}'}{\sqrt{\text{Var}(\eta')}} \]  

(11)

The negative sign is included in the definition of the NAI score because while natural selection favors a reduction in η values, adaptation is generally defined as being increased by natural selection. Thus the inclusion of the negative sign means that NAI is an increasing function of adaptation to reduce the cost of nonsense errors.

**NAI Calculations & Simulations** We have developed a computer program called NAI which can be used to calculate the NAI score of an allele as well as the NAI score for a series of moving windows within an allele. This program allows NAI to be calculated in two different ways, either using Equations (6) - (10) to calculate the exact moments analytically or through random sampling of synonymous genotype space weighted by a given AT bias. The sample population is then used to estimate the moments of η for the entire set of coding synonyms. The results we present here were generated by calculating the moments analytically, but we have used the random sampling approach to verify our results.

We have also developed an additional stochastic, simulation program called CES (Codon Evolution Simulation). CES simulates the evolution of a locus where the resident allele is allowed to evolve across its synonymous genotype space following the allele substitution model described in Sella and Hirsh (2005). The same substitution probabilities were also independently derived by Iwasa (1988) and Berg and Lässig (2003). For further details refer to Supporting Materials D. Both NAI and CES are written in ANSI standard C and released under the GNU Public License 2.0. Both pre-compiled *nix binaries and source code are available at www.tiem.utk.edu/~mikeg/Software.

**Application to the S. cerevisiae genome** In order to illustrate the utility and behavior of NAI, we applied it to the S. cerevisiae strain S288c genome based on SGD’s June 6, 2008 release (Dolinski et al., 2008). Default parameter values for both the simulation of the S. cerevisiae genome using CES and NAI calculations are given in Table 2. We restricted our analysis of gene level NAI scores to the 4674 verified nuclear genes that lack internal stops and our analysis of how NAI changes with window position to the 4377 genes within that set with at least 100 codons. The per codon elongation rates were generated as in Gilchrist and Wagner (2006). These rates are proportional to the tRNA abundance and take into
account a set of wobble penalties based on Curran and Yarus (1989). Exact values used are given in the Supporting Materials Table S1. Since no reliable empirical measurements of the nonsense error rate in S. cerevisiae exists, based on experiments with E. coli by Jorgensen and Kurland (1990) we used a nonsense error rate of \( b = 0.0012/\text{sec} \). Given an average per codon elongation rate of 10 per sec, this value is consistent with indirect empirical estimates of \( b \) in S. cerevisiae by Arava et al. (2005). Because of the uncertainty in the model parameters and the fact that such parameters are largely unknown for most organisms, we evaluated the sensitivity of NAI to changes in the key parameters: the background nonsense error rate \( b \), the cost of translation initiation \( a_1 \), the cost of each elongation step \( a_2 \), and the elongation probability for each codon \( p_i \) after Hamby (1994).

Because selection pressure against nonsense errors increases with codon position along a sequence, we also evaluated how NAI changes with codon position in both the S288c and simulated genomes. Since we are more interested in general trends across the entire S. cerevisiae genome than the behavior of specific genes, we used a hierarchical regression model approach when analyzing the data. In this approach, we began by calculating the NAI values for successive, non-overlapping windows of 20 codons. We then fitted a linear regression model to these intragenic NAI values of an allele. The regression model allows us to estimate the initial degree of adaptation to reduce the cost of nonsense errors through the regression intercept \( A_i \) as well as to how such adaptation changes with position through the regression slope \( B_i \) of a given allele \( i \). That is,

\[
\text{NAI}_i(x) = A_i + B_i x + \varepsilon
\]

(12)

where \( \varepsilon \) here and below represents a noise term, \( x \) is the codon position at the center of the window, i.e. \( x = \{10, 30, 50, \ldots, x_{\text{max}} - 10\} \), and \( x_{\text{max}} \) is the largest multiple of 20 less than or equal to the length \( n \) of the coding sequence.

In order to quantify the general behavior of the output generated by the regression model of Equation (12) across the S. cerevisiae genome, we then fitted second order regression models to the intercept \( A_i \) and slope \( B_i \) values for each gene as a function of its log protein production rate, \( \ln(\phi_i) \). In other words, we fitted the following models

\[
A_i = A_0 + A_1 \ln(\phi_i) + A_2 \ln(\phi_i)^2 + \varepsilon
\]

(13)

\[
B_i = B_0 + B_1 \ln(\phi_i) + B_2 \ln(\phi_i)^2 + \varepsilon
\]

(14)

to the set of MLE estimates for the model parameters \( A_i \) and \( B_i \) in Equation (12) weighted by their standard
error. Thus, the hierarchical analysis measures how NAI changes with the two dependent variables: codon position, $x$ and log protein production rate, $\ln(\phi)$. Although higher order functions could be fitted to the MLE data, our goal is to capture the general behavior of the system as simply as possible. The gene specific protein production rates $\phi_i$ were based on a combination of mRNA measurements from AffyMetrix datasets as presented in Beyer et al. (2004) and translation rates per mRNA based on ribosome occupancy data given in Arava et al. (2003) and MacKay et al. (2004). See Gilchrist (2007) for further details.

To better contextualize our results and control for the fact that other selective forces could be driving some or all of the adaptation we observe, we repeated the above analysis on multiple artificial datasets that were generated either by randomizing the codon order of each gene in the S. cerevisiae genome, or randomly assembled genes. In terms of randomly reordered datasets, two different types of datasets were generated: partial and complete reorderings. Partial reordering involved randomly reordering codons on a per amino acid basis such that the codon order of a gene was randomized but the amino acid order of the sequence encoded was not altered. In contrast, complete reordering involved the random reordering of codons independent of the amino acids encoded. While both partial and complete reordering change the codon order of an allele, neither approach alters the set of codons used. Since nonsense errors are the only selective force that is expected to result in increasing CUB with position, these reordered datasets serve as a control. We also generate a random population of alleles for each gene by randomly sampling the population of coding synonyms independent of their specific cost-benefit values $\eta$ but weighted by an AT bias of 0.62. These random samples serve as controls for when there is no selection on CUB evolution.

In order to understand how NAI behaves when CUB evolves solely under selection to reduce the cost of nonsense errors, we simulated the evolution of each gene using CES. The simulation was run assuming that the protein production rate of a gene was equal to its empirically estimated value $\phi_i$ multiplied by log-normal random variable centered around 1 and with a standard deviation equal to the standard error of the log transformed mRNA abundance values given in Beyer et al. (2004). The use of this additional noise factor mimics the uncertainty in the estimates of $\phi_i$ inherent in the analysis of the S288c and reordered genomes. The simulation was run for $20/\mu$ generations (i.e. sufficiently long such that we expect 20 substitutions per nucleotide under a pure mutation-drift process) so that the simulation results should represent samples from the stationary distribution of allele fixation. We applied the same hierarchical regression analysis to our CES simulated genome as we used with the S. cerevisiae 288c genome.
RESULTS

General Behavior of NAI  We begin by analyzing the behavior of NAI for entire coding sequences. First and as expected, we find that the NAI scores for random samples of coding synonyms are normally distributed around a mean of 0 and a standard deviation of 1 (cf. Figure 1) (Shapiro-Wilk’s Test, p-value=0.916).

We also see that the distribution of S288c and simulated NAI scores across the entire genome behave in a similar and predictable manner, where genes with greater protein production rates have higher NAI scores (Figure 2). The S288c NAI scores for genes were also well correlated for with the NAI scores in our CES generated genome ($\rho \sim 0.74$), which provides further evidence that selection against nonsense errors can explain much of the CUB observed in *S. cerevisiae*.

Indeed, if there was no adaptation to reduce the cost of nonsense errors, the distribution of NAI scores across the *S. cerevisiae* genome would follow a standard normal distribution. Looking at the *S. cerevisiae* genome, we find that the distribution of these NAI values does not mimic the standard normal distribution, but instead is highly skewed towards higher values (Figure 3). This distribution of NAI scores as summarized in Table 3 indicates that most genes show some degree of adaptation to reduce the cost of nonsense errors.

For example we find that 92.1% of the *S. cerevisiae* genes have NAI values greater than 0. Under a pure mutation-drift process only 50% of the genes would be expected to have NAI scores above 0. In fact, about 68% of *S. cerevisiae*’s genes have NAI scores greater than 1.645. Again, under a pure mutation-drift process we would expect to only see 5% of the genes with NAI scores in this range. These results clearly demonstrate that most genes are significantly more adapted to reducing their cost-benefit ratio $\eta$ than their coding synonyms. In fact, over half of all *S. cerevisiae* genes have an NAI score greater than 2.326, indicating that they are more adapted than 99% of their coding synonyms. More striking, a full 33% are more adapted than 99.99% of their coding synonyms (NAI > 3.719). We observed similar levels of adaptation in our simulated sequences and less adaptation in our reordered datasets (Table 3). For example, in a typical simulated dataset we found that 86% had NAI values greater than 0 and 55.5% had NAI values greater than 1.645.

The sensitivity of NAI to changes in parameter values is presented in the Supporting Materials E. To summarize, NAI scores are relatively insensitive to changes in almost all of the parameters underlying its formulation. This insensitivity is especially strong for the initiation cost, the elongation cost, and the nonsense error rate, $a_1$, $a_2$, and $b$, respectively.
Changes in NAI with Codon Position  One specific prediction about CUB driven by the cost of nonsense errors is that the strength of selection and, therefore, the degree of adaptation should increase with codon position. To test this prediction, we calculated the NAI value for each successive, non-overlapping windows of 20 codons for each gene and then fit the first order regression model in Equation (12). We then fit the weighted second order regression model in Equations (13) and (14) to the maximum likelihood estimates of the $A_i$ and $B_i$ parameters in Equation (12). The overall behavior of how NAI changes with codon position and log protein production rate $\ln(\phi)$ is summarized by the curves of the hierarchical model. Specifically, the parameters for given in Table 4 and plotted in Figure 4 describe how the intercept $A_i$ and slope $B_i$ of the NAI vs. position regressions model in Equation (12) change with $\ln(\phi)$. In general, all hierarchical model parameters differed significantly from zero. We also find that the initial NAI value $A_i$ and the rate at which NAI increases with codon position $B_i$ both increase in an accelerating manner with $\ln(\phi)$.

Quantitatively, looking at the estimates of $A_0$ in Table 4 we see that the $S.\ cervisiae$ S288c genome shows greater adaptation or higher NAI values at the start of low expression genes than the genome simulated using CES. However, inspection of $A_1$ and $A_2$ values indicate that for both the observed and simulated sequences, the NAI values at the beginning of a gene increase with $\ln(\phi)$ in a similar manner. Given the fact that the 288c alleles tend to have greater than expected NAI values, it is then perhaps not surprising that the rate at which NAI increases with position is lower in the S288c sequences when compared to the simulated sequences. This is illustrated by the fact that $B_0$, $B_1$, and $B_2$ are lower in the observed sequences than the genome simulated using CES. Although upon first glance the magnitude of the regression slopes for of NAI vs. position $B_i$ may appear slight, these values are substantial as the slope increases in units of standard deviation and the alleles generally encode hundreds of codons. Figure 5 allows us to see these differences in behavior between the the S288c and simulated and datasets for six genes from across a wide range of protein production rates.

Stepping back, we note that qualitatively the distribution of $A_i$ and $B_i$ values around the hierarchical regression curves are quite noisy in both observed and simulated datasets (Figure 4). This results from the fact that while the selective forces against nonsense errors are consistent and increase with the protein production rate of a gene, mutation and drift clearly play important roles in the evolution of CUB.

To test whether our rather strong assumption that only completely translated proteins have any functionality, we repeated the hierarchical analyses where the the final 10 or 20 amino acids were excluded from our NAI. This is roughly equivalent to assuming a (0,1) functionality threshold at $n - 10$ or $n - 20$. The results of these analyses showed no statistically significant change in the parameter estimates of our hierarchical
model (Supporting Table S2).

Given the complexity of our NAI calculations, it is possible that the observed changes in NAI values with codon position and log protein production rate, $\ln(\phi)$, in the *S. cerevisiae* genome are actually an artifact of selection for translational accuracy or translational efficiency irrespective of nonsense errors. If this were the actual case then we would expect to see no difference between how NAI changes with position in the observed and reordered genomes. Instead we find that 66% of the genes in the *S. cerevisiae* genome have slope parameters $B_i$ greater than zero, a frequency that is significantly greater than what was observed in each set of partially reordered, completely reordered, and randomly assembled genomes which are all distributed around 0.5 (see Table 5). The difference in behavior between the observed and reordered genomes can also be observed in the hierarchical regression curves, a typical example of which is illustrated in Figure S2. Note that both partially and completely reordered genomes do show a very slight upward bias in their individual regression slopes $B_i$ and this bias is unexpected, but increases with protein production rate. The cause of this bias and its relationship to $\ln(\phi)$ is not understood at this point. For example, it may be caused by subtle compositional effects or the fact that the true distribution of $\eta$ values calculated over a window are not actually gamma distributed as we assume. Whatever the cause, the effect is small when compared the fraction of genes with positive regression slopes $B_i$ and the magnitude of the parameters of our hierarchical model.

**DISCUSSION**

In this study we developed a method for quantifying the adaptation an allele displays to reduce the cost of nonsense errors during protein translation $\eta$ relative to its set of alternative coding synonyms. The approach presented here is a generalization of the definition of $\eta$ derived in Gilchrist (2007) and provides analytic expressions for the mean and variance of $\eta$ values in a given synonymous genotype space. These values are used to contextualize the $\eta$ value for any given point in synonymous genotype space into a single NAI value. To our knowledge, our work is the first to provide a general means of surveying a biologically meaningful adaptive landscape and then use these results to contextualize adaptation at the molecular level.

NAI scores are based on a mechanistic model of protein translation and functionality. As a result, the nature of adaption NAI measures is clearly defined. In addition, because of its mechanistic derivation, our approach is flexible enough to allow for future refinement or expansion of NAI as our understanding of the processes involved in protein translation and other processes progresses. This flexibility is best illustrated in the seamless manner with which we are able to incorporate mutation bias into the calculation of NAI. The
fact that we can easily incorporate codon specific nonsense error rates into our model by simply using codon specific \( b \) values in Equation (1) also illustrates this point. Another example of this flexibility is in the cost of function \( \beta(i) \) used in Equation (3).

In our current formulation \( \beta(i) \) only includes the direct assembly cost of a protein. However, \( \beta(i) \) could be expanded to include other costs as well. For example, given our assumption about exponential waiting times during each elongation step, the combined assembly and ribosome overhead costs at a given codon \( i \) could be calculated as \( \beta(i) = a_1 + a_2(i - 1) + k(a_3 + \sum_{j}^{i-1} 1/c(NNN_j)) \) where \( k \) is a scaling constant representing the per second overhead cost of the ribosome cost in \( \sim Ps \) and \( a_3 \) is the expected time it takes for a ribosome to intercept an mRNA and initiate translation. Expanding our cost function \( \beta(i) \) in this way highlights the potentially large effect nonsense errors can have on the overall translational efficiency of a ribosome, a point overlooked in most discussions of CUB.

In a similar manner, the utility function \( u_i \) could be defined by a continuous, sigmoidal function rather than the step function we use here. Doing this would allow calculations of \( \eta \) to include the contribution of incomplete, but partially functional peptides. More generally, \( u_i \) could take on negative values, allowing it to describe any toxic or interference effects some short, incomplete peptide may have. A final example of how our approach can be extended can be found by examining one of the strongest assumptions we make, the lack of interactions between ribosomes on the same mRNA. Because the presence of a ribosome at one position can interfere with the behavior of another ribosome upstream from it, this assumption is clearly violated. Relaxing this assumption would make the computation of NAI substantially more challenging (e.g., see CHO, 2003; BASU AND CHOWDHURY, 2007). While we believe that our current approach is a reasonable first approximation, other research suggests relaxing this assumption might be useful. For example, BULMER (1991) showed this type of inter-ribosomal interference can lead to additional selection for increased translational efficiency at the start of a coding sequence. This idea of inter-ribosomal interference is consistent with our own observation of greater observed NAI values at the start of a sequence than expected based on our simulation.

It is important to recognize that one of the most common uses of CUB is to predict relative gene expression levels from the coding sequence of a gene. While the primary purpose of NAI is not to predict gene expression, the concepts underlying it can be used for that purpose (GILCHRIST, 2007). Nevertheless, NAI values are well correlated with the most commonly used CUB indices (\( \rho = 0.793 - 0.822 \), Supporting Figure S1).

Most CUB indices are based on some sort of distance measure in synonymous genotype space. For
example, consider Sharp and Li (1987)’s Codon Adaptation Index (CAI) which is probably the most commonly used measure of CUB. The CAI value of an allele is based on the geometric mean frequency of codons it uses relative to the usage within a subset of highly expressed genes. Conceptually, CAI is a multiplicative distance measurement of an observed allele from an allele whose CUB mirrors that of a subset of highly expressed genes. Another, more recent example can be found in Roymondal et al. (2009)’s Relative Codon Bias (RCB) index. RCB measures the CUB of an allele relative to its expected position based on mutation and drift alone. Similarly, the CUB indices $E(g)$ (Karlin and Mrazek, 2000), $F_{op}$ (Ikemura, 1981), and $N_c$ (Wright, 1990) can be thought of as providing similar types of distance measures.

While NAI is similar to these other indices in that it is a measure of relative distance, the distance measurement is not in terms of positioning within an allele’s synonymous genotype space, but based on the relative altitude of an allele on an adaptive landscape. The fact that NAI is based on the phenotypic adaptation of an allele rather than its position in genotype space makes it fundamentally different from these other measures. Unlike most other commonly used indices, NAI explicitly takes into account the effect of genome-wide AT bias in shaping the codon usage of a gene. In addition, because NAI is based on a Z score, it is easier to interpret and has desirable statistical properties that are absent in most CUB indices.

Despite the fact that selection against codon usage bias is generally thought to be a rather weak selective force (Stoletzki and Eyre-Walker, 2007; Hershberg and Petrov, 2008), the NAI scores for the *S. cerevisiae* genome indicate that most (>92%) of its coding sequences are more adapted to reduce the cost of nonsense errors than expected. Indeed, more than 67% of the *S. cerevisiae* alleles are at a higher point on the adaptive landscape than 95% of their coding synonyms. Perhaps more striking is the finding that more than 33% of the *S. cerevisiae* alleles found above the 99.99 percentile of the adaptive landscape. Thus, almost a third of all alleles can be found in the far upper reaches of the nonsense error adaptive landscape. In addition, we also observe that NAI values generally increase with codon position of a gene, a unique pattern only expected to result from selection against nonsense errors. Indeed, the increases we see are consistent with the increases observed in our simulations where nonsense errors are the sole selective force. Taken together, our results provide additional evidence that nonsense errors play an important role in CUB evolution of *S. cerevisiae*.

One shortcoming of our study is that we only consider one source of selection: nonsense errors. In reality, many other selective forces contribute to the evolution of CUB. Indeed, the emerging picture from the field clearly indicates that any synonymous change in the coding region of a gene is likely to be pleiotropic, possibly affecting mRNA folding, translational accuracy, translational efficiency, protein folding, etc. This
suggests that there are many places for these sources of selection to either conflict with or reinforce one another. Even though some of the adaptation we observe is likely due to these other forces, the fact that we do not consider these other forces when calculating an NAI value does not invalidate its meaning. NAI is simply a measure of adaptation to the cost of nonsense errors and does not depend on the forces ultimately responsible for that adaptation. While the tendency of NAI to increase with codon position can currently only be explained by selection against nonsense errors, the exact degree to which the overall adaptation of an allele can be directly attributed to this selective force is still open to debate. One way of resolving this debate would be to expand the approach developed here to include other potential selective forces. Indeed, we hope that the approach we develop here will serve as a starting point for generating other measures of adaptation. Only when we have a combination of such measures will researchers be able to evaluate the importance of the different selective forces driving the evolution of CUB.

ACKNOWLEDGMENTS

We would like to thank Drs. Sergey Gavrilets, Michael Saum, and Hong Qin for providing helpful suggestions and comments on this manuscript. We would also like to thank two anonymous reviewers for their constructive criticisms and suggestions which have greatly improved this manuscript.

REFERENCES


Ikemura, T., 1981 Correlation between the abundance of *Escherichia-coli* transfer-rnas and the occurrence of the respective codons in its protein genes - a proposal for a synonymous codon choice that is optimal for the *Escherichia-coli* translational system. Journal of Molecular Biology **151**: 389–409.

Ikemura, T., 1982 Correlation between the abundance of yeast transfer rnas and the occurance of the repective codons in protein genes. Journal of Molecular Biology **158**: 573–597.

**Key:** Ikemura82

**Annotation:** subtitle: Differences in sysnonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abunance of isoaccepting transfer RNAs


Warnecke, T., N. N. Batada, and L. D. Hurst, 2008 The impact of the nucleosome code on protein-

Warnecke, T. and L. D. Hurst, 2007 Evidence for a trade-off between translational efficiency and splicing
regulation in determining synonymous codon usage in drosophila melanogaster. Molecular Biology and

Wolfe, K. H. and P. M. Sharp, 1993 Mammalian gene evolution: nucleotide sequence divergence between


Xia, X., 2009 Information-theoretic indices and an approximate significance test for testing the molecular


Xia, X. H., 1998 How optimized is the translational machinery in Escherichia coli, salmonella typhimurium
and Saccharomyces cerevisiae? Genetics 149: 37–44.

136: 746–762.
Table 1: List of symbols used in this study

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c(NNN)$</td>
<td>Elongation rate of codon $NNN$.</td>
</tr>
<tr>
<td>$b$</td>
<td>Background nonsense error rate.</td>
</tr>
<tr>
<td>$p(NNN)$</td>
<td>Elongation probability of codon $NNN$.</td>
</tr>
<tr>
<td>$\overrightarrow{NNN}$</td>
<td>Codon sequence of an allele.</td>
</tr>
<tr>
<td>$\vec{p}$</td>
<td>Vector of elongation probabilities for codon sequence $\overrightarrow{NNN}$.</td>
</tr>
<tr>
<td>$\sigma_i(\vec{p})$</td>
<td>Probability a ribosome will translate up to and including the $i$th codon of the sequence $\overrightarrow{NNN}$.</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>Amount of energy expended by ribosome in translating up to and including the $i$th codon.</td>
</tr>
<tr>
<td>$a_1$, $a_2$</td>
<td>Energetic cost of translation initiation and elongation, respectively.</td>
</tr>
<tr>
<td>$\mathbb{J}$</td>
<td>Set of synonymous codons of a given amino acid.</td>
</tr>
<tr>
<td>$E(p)$</td>
<td>Expected elongation probability of a given amino acid.</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Target production rate of a given protein.</td>
</tr>
<tr>
<td>$\bar{\eta}$</td>
<td>Untransformed expected cost-benefit ratio for the coding synonyms of an allele.</td>
</tr>
<tr>
<td>$\text{Var}(\eta)$</td>
<td>Untransformed variance in cost-benefit ratio for the coding synonyms of an allele.</td>
</tr>
<tr>
<td>$\alpha$, $\beta$</td>
<td>Shape &amp; inverse scale parameters of the Gamma distribution, respectively.</td>
</tr>
<tr>
<td>$\eta_{\text{obs}}$</td>
<td>Transformed cost-benefit ratio of the observed allele.</td>
</tr>
<tr>
<td>$\hat{\eta}$</td>
<td>Transformed expected cost-benefit ratio for the coding synonyms of an allele.</td>
</tr>
<tr>
<td>$\text{Var}(\hat{\eta})$</td>
<td>Transformed variance in cost-benefit ratio for the coding synonyms of an allele.</td>
</tr>
<tr>
<td>$A_i$</td>
<td>Intercept of regression of NAI with position.</td>
</tr>
<tr>
<td>$B_i$</td>
<td>Slope of regression of NAI with position.</td>
</tr>
<tr>
<td>$A_i$</td>
<td>Coefficient of hierarchical regression describing how the intercept $A_i$ changes with $\ln(\phi)$.</td>
</tr>
<tr>
<td>$B_i$</td>
<td>Coefficient of hierarchical regression describing how the slope $B_i$ changes with $\ln(\phi)$.</td>
</tr>
<tr>
<td>$\mathbb{S}$</td>
<td>The set of alleles that make up the synonymous genotype space of a gene.</td>
</tr>
</tbody>
</table>
Table 2: List of default parameter values used for all NAI calculations and the CES simulation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b$</td>
<td>0.0012/sec</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$4 \sim P$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$4 \sim P$</td>
</tr>
<tr>
<td>$q$</td>
<td>$4.19 \times 10^{-7}$</td>
</tr>
<tr>
<td>$N_e$</td>
<td>$1.36 \times 10^7$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$10^{-9}$/generation</td>
</tr>
<tr>
<td>AT Bias</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3: Distribution of NAI scores and their corresponding percentile in a standard normal distribution

<table>
<thead>
<tr>
<th>NAI Score</th>
<th>Percent genome above NAI Score</th>
<th>S. cerevisiae</th>
<th>CES Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50 %</td>
<td>92.14 %</td>
<td>85.95 %</td>
</tr>
<tr>
<td>1.282</td>
<td>90 %</td>
<td>74.59 %</td>
<td>61.57 %</td>
</tr>
<tr>
<td>1.645</td>
<td>95 %</td>
<td>67.92 %</td>
<td>55.54 %</td>
</tr>
<tr>
<td>2.326</td>
<td>99 %</td>
<td>54.63 %</td>
<td>45.67 %</td>
</tr>
<tr>
<td>2.576</td>
<td>99.5 %</td>
<td>50.42 %</td>
<td>42.45 %</td>
</tr>
<tr>
<td>3.090</td>
<td>99.9 %</td>
<td>43.27 %</td>
<td>37.38 %</td>
</tr>
<tr>
<td>3.291</td>
<td>99.95 %</td>
<td>39.84 %</td>
<td>35.43 %</td>
</tr>
<tr>
<td>3.719</td>
<td>99.99 %</td>
<td>33.79 %</td>
<td>32.30 %</td>
</tr>
</tbody>
</table>
Table 4: Hierarchical Regression Analysis Results

**S288c Genome**

Parameters for regression intercept $A$

<table>
<thead>
<tr>
<th>MLE</th>
<th>SE</th>
<th>$t$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>2.102</td>
<td>0.029</td>
<td>71.97</td>
</tr>
<tr>
<td>$A_1$</td>
<td>0.793</td>
<td>0.016</td>
<td>50.12</td>
</tr>
<tr>
<td>$A_2$</td>
<td>0.084</td>
<td>$2.3 \times 10^{-3}$</td>
<td>36.68</td>
</tr>
</tbody>
</table>

Parameters for regression slope $B$

<table>
<thead>
<tr>
<th>MLE</th>
<th>SE</th>
<th>$t$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_0$</td>
<td>$2.72 \times 10^{-2}$</td>
<td>$1.635 \times 10^{-3}$</td>
<td>16.63</td>
</tr>
<tr>
<td>$B_1$</td>
<td>$8.74 \times 10^{-3}$</td>
<td>$7.92 \times 10^{-4}$</td>
<td>11.03</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$7.94 \times 10^{-4}$</td>
<td>$9.6 \times 10^{-5}$</td>
<td>8.27</td>
</tr>
</tbody>
</table>

**CES Simulated Genome**

Parameters for regression intercept $A$

<table>
<thead>
<tr>
<th>MLE</th>
<th>SE</th>
<th>$t$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>1.021</td>
<td>0.025</td>
<td>40.48</td>
</tr>
<tr>
<td>$A_1$</td>
<td>0.447</td>
<td>0.0145</td>
<td>30.89</td>
</tr>
<tr>
<td>$A_2$</td>
<td>0.047</td>
<td>$2.05 \times 10^{-3}$</td>
<td>22.94</td>
</tr>
</tbody>
</table>

Parameters for regression slope $B$

<table>
<thead>
<tr>
<th>MLE</th>
<th>SE</th>
<th>$t$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_0$</td>
<td>0.135</td>
<td>$1.82 \times 10^{-3}$</td>
<td>74.5</td>
</tr>
<tr>
<td>$B_1$</td>
<td>$4.48 \times 10^{-2}$</td>
<td>$8.97 \times 10^{-4}$</td>
<td>49.96</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$3.74 \times 10^{-3}$</td>
<td>$1.08 \times 10^{-4}$</td>
<td>34.66</td>
</tr>
</tbody>
</table>

Table 5: Frequency of individual regression slopes $B_i > 0$

<table>
<thead>
<tr>
<th>Mean Frequency</th>
<th>SD</th>
<th>$t$-statistic</th>
<th>$p$ value</th>
<th>vs. S288c genome</th>
<th>vs. CES simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>49.9%</td>
<td>0.0086</td>
<td>$-604.630$</td>
<td>$&lt; 2 \times 10^{-16}$</td>
<td>$-1021.83$</td>
</tr>
<tr>
<td>Partially Reordered</td>
<td>51.6%</td>
<td>0.0083</td>
<td>$-560.143$</td>
<td>$&lt; 2 \times 10^{-16}$</td>
<td>$-990.58$</td>
</tr>
<tr>
<td>Completely Reordered</td>
<td>50.2%</td>
<td>0.0087</td>
<td>$-586.051$</td>
<td>$&lt; 2 \times 10^{-16}$</td>
<td>$-996.92$</td>
</tr>
</tbody>
</table>

Frequency of genes with positive regression slopes based on a sample of 1000 genomes for each category as well as a comparison of these populations to the frequency for the S288c genome (77%) and the CES simulation genome (66%).
Figure 1: Distribution of NAI values for 1000 randomly assembled alleles (open bars) and 1000 CES simulated alleles (closed bars) for the gene YBL068W. CES simulations were done using the empirically estimated protein production rate of $\phi = 0.21s^{-1}$. The dashed line at 4.015 indicates the NAI value for the YBL068W allele in *S. cerevisiae* S288c which corresponds to the 99.997th percentile of a standard normal distribution which is represented by the solid curve. Note that the mean of the NAI scores of the random population does not significantly differ from zero ($t = 0.1917$, $p$-value > 0.8) while the mean of the NAI scores of the simulated population does ($t = 172.2$, $p$-value < $10^{-15}$).
Figure 2: Correlation of NAI values with log protein production rates \( \ln(\phi) \) for the S288c and CES simulated genomes. \( \rho \) represents the correlation coefficient between the two variables.
Figure 3: Distribution of NAI values of S288c and CES simulated genomes. Solid gray bars represent the distribution of NAI scores across all genes in S288c genome and striped bars represent NAI scores of CES simulated genes. The solid black line represents the standard normal distribution which is the expected distribution of NAI scores in the absence of any selection.
Figure 4: Hierarchical analysis results. Each data point represents the maximum likelihood estimates of the regression parameters of NAI vs. position (Equation (12)) for each individual gene. Curves represent the weighted second order hierarchical regression curves of $A$ and $B$ given in Equations (13) and (14). Weighting is based on the standard error of data point.
Figure 5: Change in NAI with position and protein production rate, $\phi$. The figure shows how NAI changes with position for six genes chosen from a wide range of empirically estimated protein production rates, $\phi$. Solid circles (●) represent NAI values for the S288c coding sequence of the gene while open circles (○) represent NAI values for a simulated coding sequence. The regression lines through the NAI values are given by (− − −) for S288c sequence and (···) for simulated sequence.