Differential bicodon usage in lowly and highly abundant proteins

Luis A Diambra Corresp. 1

1 Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata, CONICET, La Plata, Argentina

Corresponding Author: Luis A Diambra
Email address: ldiambra@gmail.com

Degeneracy in the genetic code implies that different codons can encode the same amino acid. Usage preference of synonymous codons has been observed in all domains of life. There is much evidence suggesting that this bias has a major role on protein elongation rate, contributing to differential expression and to co-translational folding. In addition to codon usage bias, other preference variations have been observed such as codon pairs. In this paper I report that codon pairs have significant different frequency usage for coding either lowly or highly abundant proteins. These usage preferences cannot be explained by the frequency usage of the single codons. The statistical analysis of coding sequences of nine organisms, reveals that in many cases bicodon preferences are shared between related organisms. Furthermore, it is observed that misfolding in the drug-transport protein, encoded by *MDR1* gene, is better explained by a big change in the pause propensity due to the synonymous bicodon variant, rather than by a relatively small change in codon usage. These findings suggest that codon pair usage can be a more powerful framework to understand translation elongation rate, protein folding efficiency, and to improve protocols to optimize heterologous gene expression.

Differential bicodon usage in lowly and highly abundant proteins.

Luis Diambra[1]

[1]**Centro Regional de Estudios Genómicos, UNLP, CONICET La Plata, CP:1900, Argentina**

Corresponding author:
Luis Diambra[1]

Email address: ldiambra@gmail.com

ABSTRACT

Degeneracy in the genetic code implies that different codons can encode the same amino acid. Usage preference of synonymous codons has been observed in all domains of life. There is much evidence suggesting that this bias has a major role on protein elongation rate, contributing to differential expression and to co-translational folding. In addition to codon usage bias, other preference variations have been observed such as codon pairs. In this paper I report that codon pairs have significant different frequency usage for coding either lowly or highly abundant proteins. These usage preferences cannot be explained by the frequency usage of the single codons. The statistical analysis of coding sequences of nine organisms, reveals that in many cases bicodon preferences are shared between related organisms. Furthermore, it is observed that misfolding in the drug-transport protein, encoded by *MDR1* gene, is better explained by a big change in the pause propensity due to the synonymous bicodon variant, rather than by a relatively small change in codon usage. These findings suggest that codon pair usage can be a more powerful framework to understand translation elongation rate, protein folding efficiency, and to improve protocols to optimize heterologous gene expression.

INTRODUCTION

The central dogma of molecular biology establishes that the information that specifies which amino acid monomers will be added next during protein synthesis is coded in one or more nucleotide triplets known as codons (Watson et al., 2003). The genetic code establishes a set of rules that associate the 20 amino acids and a stop signal with 64 codons. This code is almost universal with few exceptions (Jukes and Osawa, 1993). As there are more codons than encodable signals (amino acids and stop signal) the genetic code is considered degenerated. However, it is well known that synonymous codons are not used with the same frequency. Biased codon usage is a pervasive feature of the information encoded in genomes, but it is not universal because different species have different associated preferences (Watson et al., 2003). The existence of selective pressures to promote codon usage bias highlights the complex nature of synonymous codon choices (Hershberg and Petrov, 2008; Quax et al., 2015). Early reports have pointed out that the bias in prokaryotes is towards codons with high translation rates (Gouy and Gautier, 1982; Pan et al., 1998). In this sense, Guimares *et al.* established that, in *E. coli*, the elongation rate is affected by specific amino acid composition, as well as by codon bias (Guimaraes et al., 2014). On the other hand, the impact of codon usage on translational rates in eukaryotes, where mRNA processing can also affect the overall translational rate, is an active topic of research (Tuller et al., 2007; Vogel et al., 2010; Tuller et al., 2010; Waldman et al., 2010; Camiolo et al., 2012; Pop et al., 2014). However, the role of codon usage has gone beyond translational rates, because new experimental findings suggest that codons with slow translation rates temporally separate the synthesis of defined protein portions and can synchronize the synthesis with the concurrently folding process of the proteins domains (Lemm and Ross, 2002; Zhang et al., 2009; Kimchi-Sarfaty et al., 2007; Buhr et al., 2016). It has been shown that translational pauses can schedule the sequential folding schemes, leading to different protein conformations (Buhr et al., 2016), and that the functionality of translated proteins can be affected by replacing rare codons with more frequently used codons (Komar et al., 1999; Spencer et al., 2012; Bali and Bebok, 2015).

In addition to the use of rare codons associated with scarce tRNA usage, other mechanisms exist that modulate the speed of translation or cause pauses. Among these, one can mention the blocking of ribosomal transit due to secondary structure elements in mRNAs (Nackley et al., 2006), and interactions of basic residues in the nascent polypeptides with the wall of the ribosomal exit tunnel (Gloge et al., 2014). However, in the last years emerging evidence has shown that translational rate could be encoded by a sequence longer than a triplet of nucleotides, in particular by bicodons (Guo et al., 2012). In this sense, a study encompassing 16 genomes has revealed that bicodons formed by two rare codons are frequently found in prokaryotes but rarely used in eukaryotes (Buchan et al., 2006). More recently, Gamble *et al.* showed that some adjacent codon pairs modulate translational elongation in *S. cerevisiae* (Gamble et al., 2016). By studying the expression of 35811 synthetic variants of a superfolder GFP, they identified 17 adjacent codon pairs (hereafter bicodons) associated with a substantial reduction of the translation elongation rate. The observed reduction in the tranlation efficiency could be consequence of the wobble interactions involved with these bicodons (Gamble et al., 2016). In addition, bicodons such as NNUANN are universally underrepresented, whereas others such as NNGCNN are mostly preferred (Tats et al., 2008). It was also reported that rare arginine codons, followed by proline codons, were among the slowest translated bicodons (Chevance et al., 2014). This evidence could be consequence of the codon co-occurrence bias mechanism (Cannarozzi et al., 2010; Zhang et al., 2013), or the kinetics of the mRNA translocation from the A-site to the P-site (Khade and Joseph, 2011). Codon pair bias was also observed in several viral genomes, which matched the codon usage bias of the host (Wong et al., 2010). This fact has been used to produce synthetic viruses with attenuated virulence as a new strategy for vaccine development (Coleman et al., 2008).

Thus, coding sequences seem to carry more information than that strictly needed for specifying the linear sequence of amino acids in a protein. This additional information is linked with the overall synthesis rate of the associated protein, and the pauses required for it to acquire its correct native structure. Despite the enormous impact that this subliminal code has on biotechnology, there are few systems biology methods to associate nucleotide sequences with the rate of protein synthesis (Pop et al., 2014). Among these, one can mention the sequencing of ribosome-protected mRNA fragment or ribosome profiling. This methodology has been used to correlate mRNA levels with codon decoding times (Dana and Tuller, 2014; Gamble et al., 2016; Mohammad et al., 2016). In this paper, I present an alternative manner to identify coding sequences that can modulate ribosomal transit on the mRNA. In this comprehensive survey I performed a statistical analysis of bicodon usage frequencies over two sets of proteins: the low protein abundance (PA) set and the high PA set, for nine organisms. The main finding is that there is an important bias of bicodon usage depending on protein abundance. In this sense, I determine which bicodons are statistically associated with low or high translational rates, and in which cases this bias can be explained or not by codon usage bias. I found that the bicodons reported in (Gamble et al., 2016) are almost never used to encode highly abundant proteins in yeast.

Furthermore, I present suggesting evidence for the role of bicodons in encoding translational rate. In this sense, I found that the alteration in the structure and function of the MDR1 protein (Kimchi-Sarfaty et al., 2007) associated with a synonymous single polymorphism can be better explained by a relatively big change in a pause propensity score, than by a moderate change in codon usage.

MATERIAL AND METHODS

Data sources

In this work I have used two kinds of data: (i) genome-wide protein abundance (PA) across nine model organisms, and (ii) nucleotide sequences associated with the proteins indicated above. The absolute protein abundance data from three prokaryotes (*M. aeruginosa B. subtilis* and *E. coli*), one unicellular fungus (*S. cerevisiae*), one plant (*A. thaliana*), two multicellular eukaryotes (*D. melanogaster and C. elegans*) and two mammals (*M. musculus* and *H. sapiens*) were downloaded from the PaxDb web site (http://pax-db.org/) on May 2015 (Wang et al., 2012). From these comprehensive data sets I selected one sample of the most abundant proteins, and another sample of the less abundant ones. When more than one isoform was present in the comprehensive data set, only one was included in the samples. It is important to point out that protein abundance correlates negatively with coding-sequence length in yeast (Coghlan and Wolfe, 2000). As PA distributions are generally biased, i.e., short proteins should be more abundant than larger ones (see Fig. 1), I selected two sets of 500 sequences, in which the sequence length distribution of both sets was similar. The procedure for sampling sequences with similar length
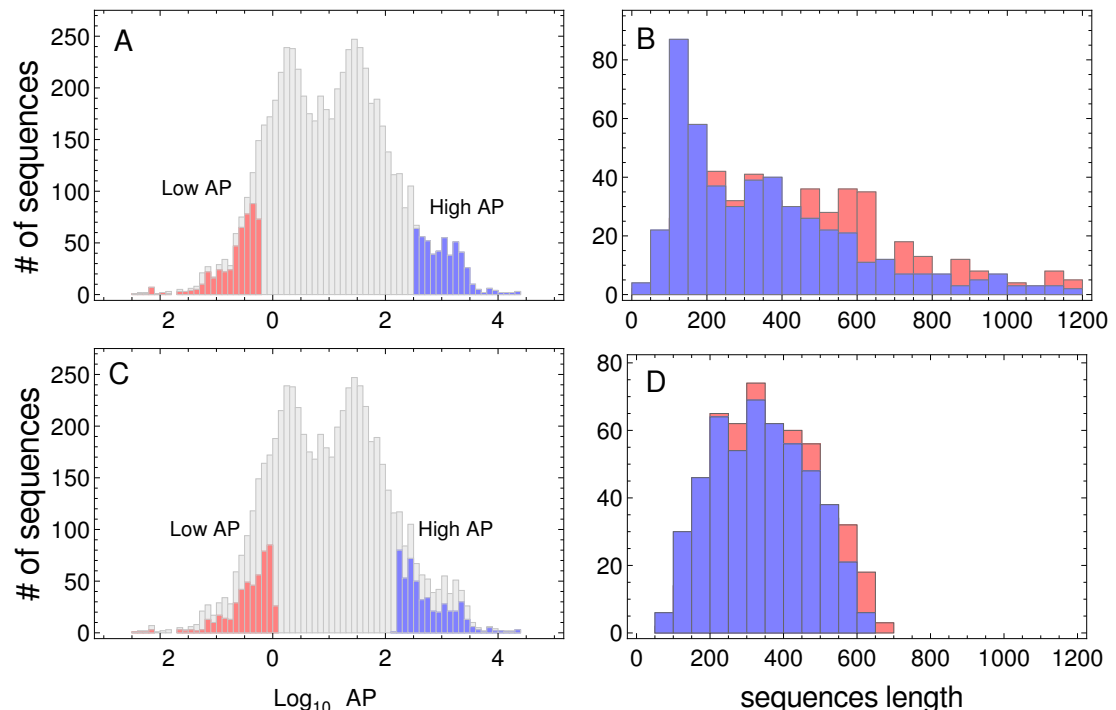
Figure 1. Protein abundance and sequence length distributions. Protein abundance distributions of the whole dataset of *S. cerevisiae*, low PA and high PA subsets are indicated in red and blue colors, respectively (A). Sequence length distributions of the subsets of sequences are shown in the left panel (B). Protein abundance distributions of the whole dataset, the selected low and high PA subsets of sequences used in the study (listed in Supplementary Tables S1 and S2), are indicated in red and blue colors, respectively (C). Sequence length distributions corresponding to the subsets of sequences are shown in the left panel (D).

distributions, consisted in ordering all the sequences which corresponded to a given organism in a PA increasing order. To select the sequences of the low PA samples, I began from the lowest PA extreme of the sequence list, and I compared $r_l = \exp[-(l - l_o)^2/2\sigma]$ (where l is the length of the sequence, l_o and σ are the mean length and standard deviation, respectively, of the target distribution) with a random number uniformly distributed r. If $r_l > r$, the sequence was added to the set of low PA sequences. Then, I tested the second sequence in a similar manner and so on, until 500 sequences were selected. To select the high PA sequence set I performed the same procedure, but beginning from the highest PA extreme of the sequence list. The resulting lists of coding sequences corresponding to these unbiased samples are given in Supplementary Tables S1 and S2. In Fig. 1 (and also in Supplementary Figs. S1-S8) I have plotted the distributions of the whole PA for all organisms used in this study, and the PA and the sequence length distribution of the selected data sets. The nucleotide coding sequences corresponding to the selected proteins were downloaded from Ensembl web sites (four eukaryote organisms from ftp://ftp.ensembl.org/pub/ and four prokaryote organisms from http://bacteria.ensembl.org), while *Arabidopsis thaliana* coding sequences were downloaded from www.arabidopsis.org.

I also used the cumulative ribosome occupancy computed by (Gamble et al., 2016) from the yeast ribosome profiling data set obtained by (Jan et al., 2014). The cumulative ribosome occupancy is defined as the ratio between the sum of joint counts at positions with the bicodon in the ribosomal P-, A-sites and E-, P-sites and the joint counts sum across all surrounding window positions, which were extracted from Table S5 of (Gamble et al., 2016).

Statistical analyses

Bicodon bias was studied in the context of the low and high PA samples. I counted all consecutive pairs of codons on the same reading frame of the coding sequences belonging to a given sample, which allowed us to compute the occurrence of each bicodon ij for all the sequences of each sample. The index

i indicates the codon corresponding to P-site, while j indicates the one corresponding to the A-site. The occurrence of the codon pair ij will be denoted by o_{ij}. I also computed the number of single codons f_i for all the sequences of each sample.

Summarizing, I analyzed the bias of bicodon usage in the two samples using three complementary measures: (i) the pause propensity score, which is based on the differential bicodon usage in both samples; (ii) the Fisher's exact test, which establishes whether the bicodon usage bias is significant; and (iii) the residual score proposed in (Gutman and Hatfield, 1989), which establishes whether the bias in bicodons can be explained by the codon usage bias, or not.

The pause propensity score, denoted by π, is defined as the difference between the relative synonymous bicodon usage computed over the low PA sequences (RSBU^L), and the one computed over the sequence sample associated with high PA, (RSBU^H). Mathematically,

$$\pi_{ij} = \text{RSBU}^L_{ij} - \text{RSBU}^H_{ij} = q_{ap}(f^L_{ij} - f^H_{ij})/N_{ap}. \tag{1}$$

Here, f^X_{ij} is the frequency of the bicodon ij computed over the sequence sample X, q_{ap} is the number of bicodons encoding for the same amino acid pair, and N_{ap} is the frequency of that amino acid pair for both samples. Thus, a large, or small, value of π_{ij} indicates the preference of bicodon ij for encoding low, or high, PA sequences, respectively. These values were clustered using a hierarchical average linkage over P-site and A-site codons according to patterns of similar bicodon preference.

Further, I use Fisher's exact test to examine whether the number of occurrences of bicodon o^L_{ij}, observed in the sample of sequences associated with lowly abundant proteins, was significantly different to the number of occurrences observed in the sample of sequences associated with highly abundant proteins o^H_{ij}. Thus, I constructed a 2×2 contingency table for each bicodon, as shown for an illustrative purposes in Supplementary Fig. S9 for the particular case of bicodon AAGAAG (Agresti, 1992). In order to compute the p-value, I approximated the factorial operator with Stirling's formula, $n! \approx \sqrt{2\pi n}(n/e)^n$ for $n \geq 25$. To improve visualization the colors in the heat maps are related to the quantity $-S\log_{10}(p\text{-value})$ where S takes value $+1$, or -1, when the bicodon has preference for sequences with low or with high PA, respectively.

Furthermore, I computed the expected number of occurrences of each codon pair, as $e_{ij} = f_i f_j N_p/N^2_{tot}$, where N_{tot} is the total number of codons in the set of sequences and N_p is the number of bicodons. Following (Gutman and Hatfield, 1989), I removed the contribution due to the nonrandomness of amino acid pairs by normalizing the former expected values as:

$$\hat{e}_{ij} = e_{ij} \times \frac{\sum^*_{kl} o_{kl}}{\sum^*_{kl} e_{kl}}, \tag{2}$$

where the * indicates that the sum is only over codon pairs kl encoding the same amino acid pair encoded by the bicodon ij. From the observed and normalized expected bicodon counts recorded in a given sample X, I computed the residual scores χ_{Xij} for each codon pair as:

$$\chi^2_{Xij} = \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}, \tag{3}$$

where X indicates the sequence samples, i.e., $X = L$ for low PA sample, or $X = H$ for high PA sequence sample. These residual scores can be used to assess whether the bias in a given codon pair can be explained, or not, by the bias in codons and amino acids.

In order to statistically assess the residual scores, I performed a random shuffling control. From each sample of sequences I generated a second random set of sequences by shuffling the order of codons (but preserving the stop codon at the end of sequence). This procedure removed the codon correlation but not the codon usage. Then, I computed the residual score of bicodons associated with this random sample. I repeated the above procedure 200 times and, finally for each bicodon ij I computed the mean value $\langle \chi^2_{ij} \rangle_{ran}$, and associated standard deviation $\text{SD}_{ran}(\chi^2_{ij})$. Thus, the residual scores of bicodon ij can be expressed as the number of standard deviations from the mean, i.e., $(\chi^2_{ij} - \langle \chi^2_{ij} \rangle_{ran})/\text{SD}_{ran}(\chi^2_{ij})$. This procedure was performed for the low and high PA samples of sequences independently.

The above analysis was performed for all possible bicodons ($61 \times 64 = 3904$), excluding stop:sense bicodons and the stop:stop bicodons. The values of frequencies, π, p-value, and χ^2 for all bicodons and the nine organism are listed in Supplementary Table S3. The scripts for each analysis in the manuscript are available in Supplementary Code S1.

RESULTS

Bicodon preferences

In the light of recent findings (Gamble et al., 2016), I am interested here in studying bicodons as determinants of translational elongation rate modulation in nine organisms. The working hypothesis was that sequences encoding abundant proteins should be optimized, in the sense of translation efficiency. This optimization could be reflected in the usage frequency of both codons and bicodons which encode sequences associated with high and low PA. Thus, I expected to capture this bias by statistical analysis of these sequence samples. To check this hypothesis, I selected a set of 500 coding sequences associated with highest abundance proteins, and another 500 coding sequences associated with lowest abundance proteins, in nine model organisms from different kingdoms. Before showing the whole analysis across several organisms, I begin with an illustrative example. Fig. 2A shows the histogram of the bicodons (red bars) which encode for the amino acid pair VR, and which were obtained from 500 low PA sequences from *S. cerevisiae*.

One can observe that bicodon usage is not uniform, i.e., it is biased, and this could be due to the known bias observed at the codon level. However, the expected frequency associated with such bicodons (black bars, obtained by the product of each codon frequency) shows that, although some bicodon frequencies can be explained by the bias in codon usage (for example, bicodon GTGAGA), other bicodons have an associated usage frequency that is underrepresented (such as bicodon GTTAGA), or overrepresented (as bicodon GTACGA). This means that two consecutive codons used to encode a given amino acid pair can be correlated. A similar analysis can be performed with sequences associated with high PA, as shown in Fig. 2B, and in all other amino acid pairs. Evidence for nonrandom associations between codon pairs, even once codon bias and bias against specific amino acid pairings were subtracted, was previously reported in *E. coli* (Gutman and Hatfield, 1989), and in many other genomes (Buchan et al., 2006). However, what is a new remarkable fact in Fig. 2, is the strong difference between the histograms computed for the low PA and high PA samples. Fisher's exact test allows one to reject, with a high significance level, the null hypothesis that bicodons are equally used in sequences from the low PA and high PA samples. In the particular case of bicodon GTTAGA, the *p*-value is 1.73×10^{-9}. By comparison of the histograms corresponding to both low PA (red bars) and high PA (orange bars) it can be seen that some bicodons, such as GTTAGA or GTCCGT, are much more frequently used in sequences with high PA than in sequences with low PA, while the frequency usage of bicodon GTAAGA has an inverse relationship (see Supplementary Fig. 10 for a direct comparison between both samples). More interestingly, the author found that some bicodons, such as: GTACGG, GTACGA or GTGCGA, are poorly or never used to encode VR amino acid pair in high PA sequences. In particular, the last two bicodons have been identified as inhibitors of translation elongation in yeast (Gamble et al., 2016). Figure 2 only illustrates the particular case of VR pair which is not a special case. There are many other bicodons with low frequency usage in high PA, and the 20 bicodons which mediate strong inhibition of translation elongation reported in (Gamble et al., 2016) are a subset. Figure 3 depicts a raster plot of relative synonymous bicodon usage computed over the low PA sequences (RSBU^L) *vs.* relative synonymous bicodon usage computed over the high PA sequences (RSBU^H). The inset illustrates the region where inhibitor bicodons are. All of them, with the exception of CTTCTG, have an associated RSBU^H value which is very small or zero. Besides these known inhibitor bicodons, many other bicodons exhibit usage bias between both samples, and these bicodons are far from diagonal line in the RSBU^L-RSBU^H plane (dashed line in Fig. 3). Bicodons with similar features could be expected in other organisms and, in order to assess the existence of this bicodon bias observed in *S. cerevisiae* in other organisms, I devised two alternative heat maps: one based on the pause propensity score, π, and another based on the *p*-value provided by the Fisher's exact test.

Pause propensity score across nine organisms

Fig. 4 depicts a heat map for *S. cerevisiae* associated with the pause propensity score of 3904 bicodons. This score quantifies bicodon preference for coding low or high PA sequences. The position of each bicodon in the grid map was determined by clustering codons with similar pairing preference on both axes to facilitate the data visualization. Thus, bicodons with a clear preference for sequences associated with high PA (blue cells) have been grouped on the right-bottom corner of the grid, while red cells (high π score) indicate bicodons with preference for coding sequences associated with low PA. The author noticed that there are more bicodons with preference for encoding low PA sequences, than bicodons with

preference for encoding high PA sequences, but the latter group has a stronger bias. Inhibitor bicodons identified in (Gamble et al., 2016) are indicated by asterisks. All of them are associated with positive π scores, but not the highest π scores, which reflects the fact that these bicodons also have a smaller frequency associated with low PA sequences than other bicodons with preference for low PA. Another interesting feature of the heat map depicted in Fig. 4, is that some bicodons have an opposite preference depending on the order of the codons. For example, this is the case, to mention a few, for bicodons AGA-TGT and AGA-GCA (indicated by black circles), which have a preference for encoding low PA sequences (yellow cell), while the reverse ordered codon pairs TGT-AGA and GCA-AGA, respectively, have a preference for encoding high PA sequences (light blue cell). This fact cannot be explained solely in terms of codon usage bias.

I also elaborated similar heat maps to the one depicted in Fig. 4, for the other eight organisms which are shown in Fig. 5. For the sake of comparison, the position of each cell has been preserved over all organisms. These maps reveal that discrepancy in bicodon usage between low and high PA samples is not the same across the studied organisms, but is a particular feature of each organism, the same as codon bias. However, *C. elegans* and *S. cerevisiae* have many bicodons with the same preference, a feature that is also shared to a lesser extent by *D. melanogaster* and *A. thaliana*. These similarities and discrepancies are more apparent when clustering bicodons for all nine organisms (see Supplementary Fig. 11). Regarding the inhibitor bicodons identified in yeast (indicated by an asterisk in Fig. 5), one can see that many of them share a preference for low PA sequences with other organisms, in particular in the cases of *C. elegans*, *A. thaliana* and *D. melanogaster*. On the other hand, both studied mammals, (*H. sapiens* and *M. musculus*), and *E. coli*, have less apparent preferences. Interestingly, it was in *E. coli* where bicodon bias was first reported (Gutman and Hatfield, 1989). However, it is clear that bicodon usage bias across the ORFeome does not imply a different usage preference between highly and lowly expressed proteins. The low preference of bicodon usage observed in *E. coli* can be a consequence of the fact that, in this particular case, there is not a clear distinction between protein abundance distributions for both samples (see Supplementary Fig. 6).

The above heat maps show that bicodon usage bias exists in lowly and highly abundant proteins, but they do not indicate whether this bias is significant. For this, I applied Fisher's exact test to compute the p-value against the null hypothesis of equal distribution in both low and high PA samples. Fig. 6 depicts the corresponding heat map associated with *S. cerevisiae*, where the color of each cell in the grid is determined by the quantity $-S\log_{10}(p\text{-value})$, where S takes the values $+1$ or -1 when the bicodon has preference for sequences with low or with high PA, respectively. This means that red cells indicate bicodons with significant bias toward low PA sequences, while blue cells indicate bicodons with significant bias to high PA sequences. Again, to facilitate data visualization, the position of each bicodon in the map was determined by clustering codons with similar pairing preference on the P-site and A-site axes. Inhibitor bicodons are indicated by asterisks. Some of them (like bicodons CGAGCG, CTCCCG and CGACGG) are associated with a not significant bias (p-value < 0.01). This is due to the fact that these bicodons have very small frequency in low PA sequences and almost zero in high PA sequences. Something similar occurs for the sense:stop bicodons, which are grouped in three almost white rows at the center of the plot. For the sake of comparison, I also elaborated similar heat maps for other organisms where the positions of each cell has been preserved over all organisms (Supplementary Fig. 12). A couple of white rows can be observed in all organisms. These cells correspond to bicodons which do not exhibit any preference or they are usually poorly used in both sequence samples and have poor associated statistics.

The ribosome profiling method has been used for obtaining valuable information about protein synthesis, such as the location and strength of translational attenuation. Thus, sites associated with high levels of ribosome occupancy can be linked to translational pauses (Ingolia et al., 2009, 2011). To determine if ribosome occupancy correlates with the proposed pause propensity score, I used the cumulative ribosome occupancy for each bicodon, which was computed using the yeast ribosome profiling data (Table S5 of (Gamble et al., 2016)). Fig. 7 depicts a scatter plot of the cumulative ribosome occupancy *versus* the pause propensity score. Despite the fact that a clear correlation between these descriptors is not apparent in these plots, one can see that in addition to the inhibitor bicodons identified in (Gamble et al., 2016), indicated by red circles, there are some other bicodons which have high levels of cumulative ribosome occupancy (> 0.03) and a significant preference for encoding low PA sequences ($\pi > 0.75$ and p-value < 0.01). These bicodons could be also candidates for modulating transcript elongation rate in

270 yeast (they have been indicated by ∗∗ in Supplementary Table S3).

Codon bias cannot explain bicodon bias.

271 The heat maps shown above are very useful to see some common features among organisms, but they
272 do not show whether bicodon preference is explained or not by the preference of the codon in the pair
273 for sequences associated with low or high PA. In order to study this, I have computed residual scores for
275 each bicodon over sequences with low PA, χ_L^2, and over sequences with high PA, χ_H^2. When the residual
276 score is high, bicodon usage cannot be explained by codon usage in the same sample of sequences, as
277 has been previously established in (Gutman and Hatfield, 1989; Buchan et al., 2006). The value of these
278 residuals scores for all codon pairs and organisms are listed in Supplementary Table S3. In Fig. 8 a
279 raster plot of these residual scores can be seen for all bicodons in *B. subtilis*, yeast, humans and *E. coli*
280 (residual plots associated with the other five organisms are displayed in Supplementary Fig. S13). As
281 there are two sequence samples for each bicodon, it is convenient to take the quantity $\chi^2 = \chi_L^2 + \chi_H^2$ as
282 a whole residual score. For all organisms I have considered that those bicodons with $\chi^2 \geq 3$ standard
283 deviation (SD) above the mean, are bicodons whose preference for encoding low or high PA sequences
284 is not explained by the preference of the individual codons in the pair (at a significance level of 0.003).

285 Thus, four types of bicodons can be distinguished in Fig. 8 : (i) codon pairs that are significantly
286 more used in sequences associated with low PA than in sequences associated with high PA (*p*-value
287 < 0.01), and whose preferences cannot be explained by the codon usage bias (red dots); (ii) codon pairs
288 which are significantly more used in sequences associated with high PA than in sequences associated
289 with low PA (*p*-value < 0.01), and whose preferences cannot be explained by codon usage bias (blue
290 dots); (iii) codon pairs with a significantly different usage frequency in low and high PA samples, but
291 whose preferences can be explained by codon usage bias, i.e., $\chi^2 < 3\times$ SD (green dots) and, finally,
292 (iv) codon pairs whose usage frequencies in low and high PA samples are not significantly different, i.e.,
293 *p*-value > 0.01 (black dots).

294 These plots indicate that while there are many bicodons with evident preference for low and high
295 PA sequences in *C. elegans* and *S. cerevisiae*, there are only few in *H. sapiens* at this significance level.
296 Humans, mouse and, surprisingly, also *E. coli*, have few bicodons with evident preference. However,
297 even in these organisms there are numerous bicodons with similar usage frequencies in both samples,
298 but whose associated frequency usage cannot be explained by codon usage bias (black dots out of the
299 grey quadrant). This fact reveals that, even in these organisms, there is a complex correlation between
300 two consecutive codons.

301 On the other hand, *S. cerevisiae* has more codon pairs than *B. subtilis*, with a significant different
302 usage frequency in low and high PA samples. Nevertheless, the bias of many bicodons is explained by
303 codon usage bias (green dots in *S. cerevisiae* panel).

304 Regarding the inhibitor bicodons, I found that only five of them have a significantly different usage
305 frequency in low PA sample and only 1 in high PA sample (see Table S3). I searched for bicodons with
306 the same preferences shared by *S. cerevisiae*, *C. elegans* and *D. melanogaster*, by selecting those with
307 *p*-value < 0.01 and $\chi^2 \geq 3\times$ SD above the mean. Table 1 lists 16 shared bicodons with preference for
308 low PA, while Table 2 lists 40 bicodons with preference for high PA sequences. It is noteworthy that,
309 even within each species, the number of bicodons with preference for encode high PA sequences (blue
310 cells in Fig. 5) are smaller than the number of bicodons with preference for encoding low PA sequences
311 (yellow cells in Fig. 5), and the number of shared bicodons in Table 2 is substantially greater than those
312 listed in Table 1. This fact seems to indicate that bicodons with preference for high PA sequences are
313 more conserved across these organisms than bicodons with preference for low PA sequences.

Pause propensity score and cotranslational protein folding

315 The above statistical analysis is able to determine which bicodons are associated with lowly or highly
316 abundant proteins. I hypothesized that bicodons associated with lowly abundant proteins could have
317 a key role in programming translation pauses of the ribosomal machinery. As a proof of principle,
318 I have identified one example that illustrates how even small biases as those observed in *H. sapiens*,
319 could explain documented protein misfolding and a pathological condition in humans. It was previously
320 reported that the single polymorphism (SNP) rs1045642 in the gene *MDR1*, which encodes the ABCB1
321 drug-transport protein, alters the structure and function of the protein with the consequent change in its
322 substrate specificity (Kimchi-Sarfaty et al., 2007).

dipeptide	bicodon	$\chi_L^2 + \chi_H^2$			$-\log_{10}(p\text{-value})$		
		elegans	fly	yeast	elegans	fly	yeast
NA	AATGCA	3.57	21.46	4.88	7.06	4.57	4.07
EL	GAACTA	3.20	3.23	4.58	2.45	6.25	2.30
QK	CAGAAA	27.33	10.55	9.09	3.02	2.66	9.20
SK	AGTAAG	19.71	51.08	10.35	2.31	4.40	2.46
YK	TATAAA	12.21	6.43	4.01	4.00	5.23	6.05
IK	ATTAAA	41.66	3.60	25.18	7.26	3.01	2.22
AF	GCATTT	32.70	11.19	3.30	12.62	2.51	7.86
FF	TTTTTT	3.52	8.28	4.598	13.53	3.72	12.19
YP	TATCCG	8.87	69.04	15.73	2.54	3.38	7.40
FQ	TTTCAG	7.52	4.82	3.79	7.90	5.29	6.46
QS	CAAAGT	10.62	10.78	9.88	2.79	2.57	3.42
ES	GAAAGT	9.73	7.50	33.32	5.48	2.58	5.31
IG	ATAGGT	9.29	6.09	4.42	2.39	2.20	5.02
NW	AATTGG	3.87	25.44	6.00	5.25	3.66	2.93
KI	AAAATA	5.13	24.35	7.88	21.12	4.69	26.20
SV	AGTGTG	25.90	11.45	3.66	2.16	2.98	3.78

Table 1. Shared bicodons in *C. elegans, D. melanogaster* and *S. cerevisiae* that have high preference for low PA (p-value < 0.01), but for which this preference cannot be explained by codon usage bias ($\chi^2 > 3\text{SD}$ above the mean).

The SNP in exon 26 at position 3435 changes codon ATC to the synonymous ATT, which reduces codon usage from 47% to 35%. It was argued that the presence of a rare codon affects the timing of co-translational folding and insertion of P-glycoprotein into the membrane. Although it is difficult to consider the ATT codon as rare, it is clear that the SNP alters the timing of ribosomal transit. Here I offer an alternative cause for translational attenuation. In this sense, I have observed that this SNP is also associated with a large change in the bicodon pause propensity score π. Specifically, bicodon ATCGTG has preference for low PA sequence with $\pi = 0.1$, while bicodon ATTGTG has preference for high PA sequence $\pi = -0.8$. This means a large change in comparison with the change in codon usage. Further, the other synonymous bicodon ATAGTG has a low pause propensity score, $\pi = -0.55$, in agreement with Kimchi-Sarfaty *et al.* observation that associates this haplotype to a larger decrease in the inhibitory effect (Kimchi-Sarfaty et al., 2007). In addition to the SNP mentioned above, there are other synonymous SNPs related to human diseases that could be explained by a large change in the pause propensity. Among these one can mention SNP rs34533956 in the gene CFHR5, which is associated with age-related macular degeneration (Narendra et al., 2009). In this case, the mutation changes bicodon GACGTG to GATGTG, and the associated change in pause propensity is from 0.03 to -0.78, while the change in the relative synonymous codon usage is only 13%. Another example corresponds to SNP rs11615 in the gene ERCC1, which was associated with colorectal cancer (Liang et al., 2010), where the pause propensity change is also large, against a small change in the relative synonymous codon usage. These relationships suggest that some pathological synonymous mutations could be understood in terms of the change in timing needed for co-translational folding programmed by bicodons.

DISCUSSION

Organisms use a small fraction of the number of options offered by the genetic code redundancy. This is due to several constraints operating to optimize many important biological features such as expression level (Guimaraes et al., 2014), translational accuracy (Zaher and Green, 2009), protein solubility (Vasquez et al., 2016), folding accuracy (Sauna and Kimchi-Sarfaty, 2011; Spencer et al., 2012; Bali and Bebok, 2015), and protein stability, among others. In this sense, it has been shown that codon usage in *E. coli* is biased to reduce the cost of translational errors (Stoletzki and Eyre-Walker, 2007). In addition, codons that bind to their cognate tRNA more rapidly are used preferentially in highly expressed genes (Curran and Yarus, 1989). In this sense, Lian *et al.* have identified codons that regulate translation speed in human cell lines (Lian et al., 2016). A bias has also been reported in bicodon usage frequency in

several organisms (Buchan et al., 2006). More recently, Gamble *et al.* identified 17 bicodons associated with reduced translation efficiency from over 35,000 GFP variants in which three consecutive codons were randomized (Gamble et al., 2016). Analysis of ribosome profiling data suggested that this translation modulation also operates in endogenous genes (Gamble et al., 2016). Many studies agree on the key role of ribosomal pause, in orchestrating the hierarchical co-translational folding of single domains (Zhang et al., 2009; Kimchi-Sarfaty et al., 2007; Buhr et al., 2016; O'Brien et al., 2014; Fluman et al., 2014; Sander et al., 2014; Tanaka et al., 2015). In summary, there is increasing evidence that many relevant features, other than the linear sequence of amino acids, are also coded at nucleotide sequence level. These features should considerably reduce the amount of alternative ways to correctly convey the message from genes to functional proteins, despite the redundancy of the genetic code.

Among the above biological constraints determining codon usage, I have focused the attention on translational speed, i.e., the sequential process of protein elongation(Quax et al., 2015). Briefly, each iterative step of this process involves recruitment of the tRNA charged anticodons, tRNA association/dissociation to mRNA, assembly of the residue to the nascent peptide, and translocation of tRNA-mRNA from the A-site to the P-site. Each step has a particular rate, and it has been shown that disruption of the interaction between the mRNA codon in the A-site from the decoding center is a rate-limiting process (Khade and Joseph, 2011). In fact, there are evidences that such rates are codon dependent in *E. coli* (Gardin et al., 2014). Further contributions to the translational rate, not linked to the tRNAs' abundance, are the non-Watson-Crick (wobble) interactions. These interactions are usually associated with higher dissociation rates between the mRNA and the decoding center (Spencer and Barral, 2012).

Even though the results provided here suggest that some bicodons could regulate translational attenuation, it is important to remark on the limitations of the present approach in assigning each bicodon with one value of the pause propensity index. The more evident limitation is that protein abundances are not uniquely dictated by a quick translation, since transcription levels also have important roles in prokaryotes (Guimaraes et al., 2014). Thus, it is possible that the observed bicodon bias could not only be due to an intrinsic translation efficiency, but also to the bias introduced by sequences associated with low or high transcript levels present in the sequence samples. To be able to dissociate the effect of transcript abundance in this present analysis, it would necessary to known the transcript abundance of the sequence in the sample is, in order to be able to normalize protein abundance. On the other hand, the RNAi pathway is a common way to regulate expression in mammalians (Friedman et al., 2009). This factor could introduce an undesired bias and overshadow some bicodon bias. It is likely that the reduced number of bicodons with an evident preference which was observed in mammals, is due to the fact that protein abundance is not majorly determined by the bicodon usage, with the consequent poor performance of the π score in these organisms. However, the residual score seems to be more robust analysis. Alternative methods based on ribosome density profile (Gamble et al., 2016; Mohammad et al., 2016; Hussmann et al., 2015), together with statistical analysis as the one proposed here, could overcome these drawbacks.

CONCLUSIONS

In this paper, I considered that ribosomal pauses are encoded by bicodons, and examined the bicodon frequency usage in nine organisms. I found that some codons have an evident preferential usage in sequences that code for highly abundant proteins, while many others have preference for encoding proteins that are scarce. The latter bicodons can be understood as short sequences linked to translational pauses. The observed bias cannot be explained by codon usage in many bicodons.

It is worth noting that few bicodons with differential bicodon usage in low and high PA sequences were found in *E. coli*, which clearly contrasts with the other prokaryotes studied here. For example, I reported almost 450 bicodons in *B. subtilis* which are preferentially used to encode either low or high abundant proteins without a codon usage correlation. Bicodon preference is also found in a plant, a fungus and two invertebrates. These results indicate that many bicodon preferences are shared by *S. cerevisiae, C. elegans* and *D. melanogaster* and, to a lesser extent by *A. thaliana* and *B. subtilis*. Regarding the inhibitor bicodons identified in yeast, I found that they have a preference for encoding low PA sequence in agreement with (Gamble et al., 2016). However, I found that there are additional bicodons associated with a high pause propensity score and high ribosome occupancy which deserve to be considered for further wet lab essays.

In the case of the mammalian species (*H. sapiens* and *M. musculus*), I found a small number of

bicodons with a clear preference. However, the number of bicodons whose frequency usage is not explained by codon usage is comparable to other organisms. To illustrate, I show an example of synonymous mutations of clinical relevance, in which the exchange of two codons with opposite preferences, even when such preferences are moderate, can alter translation ribosomal traffic. This example suggests that single mutations that change bicodon preference, can trigger pathological phenotypes by altering the translational attenuation program of the protein.

Summarizing, I report here that bicodon usage frequency depends on protein abundance. This preference cannot be explained by the traditional codon usage for many bicodons. This empirical evidence supports the hypothesis that bicodons encode translation pauses. Such a scenario allowed us to contrast this hypothesis in various circumstances where translation rates could be altered. Like the naive codon usage, the bicodon usage can empower novel strategies for rational transcript design that minimize misfolding, while simultaneously maximizing co-translational folding for foreign proteins in heterologous hosts.

ACKNOWLEDGMENTS

I thank Alejandra Carrea and Christina McCarthy for critical reading of the manuscript. LD is member of CONICET (Argentina).

REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables. *Stat Sci*, 7(1):131–153.

Bali, V. and Bebok, Z. (2015). Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *International Journal of Biochemistry and Cell Biology*, 64:58–74.

Buchan, J. R., Aucott, L. S., and Stansfield, I. (2006). tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Research*, 34(3):1015–1027.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M., and Komar, A. A. (2016). Synonymous codons direct cotranslational folding toward different protein conformations. *Molecular Cell*, 61(3):341–351.

Camiolo, S., Farina, L., and Porceddu, A. (2012). The relation of codon bias to tissue-specific gene expression in Arabidopsis thaliana. *Genetics*, 192(2):641–649.

Cannarozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. *Cell*, 141(2):355–367.

Chevance, F. F. V., Le Guyon, S., and Hughes, K. T. (2014). The effects of codon context on in vivo translation speed. *PLoS Genetics*, 10(6):e1004392.

Coghlan, A. and Wolfe, K. H. (2000). Relationship of codon bias to mRNA and concentration protein length in Saccharomyces cerevisiae. *Yeast*, 16(12):1131–1145.

Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–7.

Curran, J. F. and Yarus, M. (1989). Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *Journal of Molecular Biology*, 209(1):65–77.

Dana, A. and Tuller, T. (2014). The effect of tRNA levels on decoding times of mRNA codons. *Nucleic acids research*, 42(14):9171–9181.

Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. (2014). mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. *eLife*, 3:e03440.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105.

Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S., and Grayhack, E. J. (2016). Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell*, 166:679–90.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, 3:e03735.

Gloge, F., Becker, A. H., Kramer, G., and Bukau, B. (2014). Co-translational mechanisms of protein maturation. *Current Opinion in Structural Biology*, 24(1):24–33.

Gouy, M. and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10(22):7055–7074.

Guimaraes, J. C., Rocha, M., and Arkin, A. P. (2014). Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Research*, 42(8):4791–4799.

Guo, F. B., Ye, Y. N., Zhao, H. L., Lin, D., and Wei, W. (2012). Universal pattern and diverse strengths of successive synonymous codon bias in three domains of life, particularly among prokaryotic genomes. *DNA Research*, 19(6):477–485.

Gutman, G. and Hatfield, G. (1989). Nonrandom utilization of codon pairs in Escherichia coli. *Proceedings of the National Academy of Sciences of the USA*, 86(10):3699–3703.

Hershberg, R. and Petrov, D. A. (2008). Selection on Codon Bias. *Annual Review of Genetics*, 42(1):287–299.

Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., and Press, W. H. (2015). Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS genetics*, 11(12):e1005732.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23.

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802.

Jan, C. H., Williams, C. C., and Weissman, J. S. (2014). Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science*, 346:748–751.

Jukes, T. and Osawa, S. (1993). Evolutionary changes in the genetic code. *Comparative biochemistry and physiology B*, 106(3):489–494.

Khade, P. K. and Joseph, S. (2011). Messenger RNA interactions in the decoding center control the rate of translocation. *Nature Structural & Molecular Biology*, 18(11):1300–1302.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007). A silent polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811):525–528.

Komar, A., Lesnik, T., and Reiss, C. (1999). Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS letters*, 462(3):387–91.

Lemm, I. and Ross, J. (2002). Regulation of c-myc mRNA decay by translational pausing in a coding region instability determinant. *Molecular and cellular biology*, 22(12):3959–3969.

Lian, X., Guo, J., Gu, W., Cui, Y., Zhong, J., Jin, J., He, Q.-Y., Wang, T., and Zhang, G. (2016). Genome-wide and experimental resolution of relative translation elongation speed at individual gene level in human cells. *PLOS Genetics*, 12(2):e1005901.

Liang, J., Jiang, T., Yao, R., Liu, Z., M., Lv, H., and Qi, W. (2010). The combination of ERCC1 and XRCC1 gene polymorphisms better predicts clinical outcome to oxaliplatin-based chemotherapy in metastatic colorectal cancer. *Cancer Chemoth. Pharm.*, 66:493–500.

Mohammad, F., Woolstenhulme, C. J., Green, R., and Buskirk, A. R. (2016). Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Reports*, 14:686–694.

Nackley, A., S.A., S., Tchivileva, I., Satterfield, K., Korchynskyi, O., Makarov, S., Maixner, W., and Diatchenko, L. (2006). Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, 314:1930–1933.

Narendra, U., Pauer, G., and Hagstrom, S. (2009). Genetic analysis of complement factor H related 5, CFHR5, in patients with age-related macular degeneration. *Mol. Vis.*, 15:731–736.

O'Brien, E. P., Vendruscolo, M., and Dobson, C. M. (2014). Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nature communications*, 5:2988.

Pan, A., Dutta, C., and Das, J. (1998). Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: translational selection versus mutational bias. *Gene*, 215(2):405–413.

Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular Systems Biology*, 10(12):770–770.

Quax, T. E., Claassens, N. J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Molecular Cell*, 59(2):149–161.

Sander, I. M., Chaney, J. L., and Clark, P. L. (2014). Expanding Anfinsen's Principle: contributions of synonymous codon selection to rational protein design. *Journal of the American Chemical Society*,

514   136(3):858–861.

515   Sauna, Z. E. and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations
516   to human disease. *Nature Reviews Genetics*, 12(10):683–691.

517   Spencer, P. S. and Barral, J. M. (2012). Genetic code redundancy and its influence on the encoded
518   polypeptides. *Computational and Structural Biotechnology Journal*, 1(1):1–8.

519   Spencer, P. S., Siller, E., Anderson, J. F., and Barral, J. M. (2012). Silent substitutions predictably
520   alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology*,
521   422(3):328–335.

522   Stoletzki, N. and Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: Selection for
523   translational accuracy. *Molecular Biology and Evolution*, 24(2):374–381.

524   Tanaka, T., Hori, N., and Takada, S. (2015). How co-translational folding of multi-domain pro-
525   tein is affected by elongation schedule: molecular simulations. *PLOS Computational Biology*,
526   11(7):e1004356.

527   Tats, A., Tenson, T., and Remm, M. (2008). Preferred and avoided codon pairs in three domains of life.
528   *BMC Genomics*, 9(1):1–15.

529   Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation effi-
530   ciency in S. cerevisiae. *PLoS Computational Biology*, 3(12):2510–2519.

531   Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. (2010). Translation efficiency is determined by
532   both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the USA*,
533   107(8):3645–3650.

534   Vasquez, K. A., Hatridge, T. A., Curtis, N. C., and Contreras, L. M. (2016). Slowing Translation be-
535   tween protein domains by increasing affinity between mRNAs and the ribosomal anti-shine–dalgarno
536   sequence improves solubility. *ACS Synthetic Biology*, 5(2):133–145.

537   Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.-Y., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R.,
538   Marcotte, E. M., and Penalva, L. O. (2010). Sequence signatures and mRNA concentration can explain
539   two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6(1).

540   Waldman, Y. Y., Tuller, T., Shlomi, T., Sharan, R., and Ruppin, E. (2010). Translation efficiency in hu-
541   mans: Tissue specificity, global optimization and differences between developmental stages. *Nucleic
542   Acids Research*, 38(9):2964–2974.

543   Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S., Hengartner, M., and von Mering, C.
544   (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Molecular
545   & Cellular Proteomics*, 11(8):492–500.

546   Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., and Losick, R. (2003). *Molecular Biology
547   of The Gene*. Springer-Verlag, Berlin/Heidelberg.

548   Wong, E. H., Smith, D. K., Rabadan, R., Peiris, M., and Poon, L. L. (2010). Codon usage bias and the
549   evolution of influenza A viruses. codon usage biases of influenza virus. *BMC Evolutionary Biology*,
550   10(1):1–14.

551   Zaher, H. S. and Green, R. (2009). Quality control by the ribosome following peptide bond formation.
552   *Nature*, 457(7226):161–166.

553   Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates pro-
554   tein synthesis and co-translational folding. *Nature structural & molecular biology*, 16(3):274–280.

555   Zhang, Y.-M., Shao, Z.-Q., Yang, L.-T., Sun, X.-Q., Mao, Y.-F., Chen, J.-Q., and Wang, B. (2013). Non-
556   random arrangement of synonymous codons in archaea coding sequences. *Genomics*, 101(6):362–367.
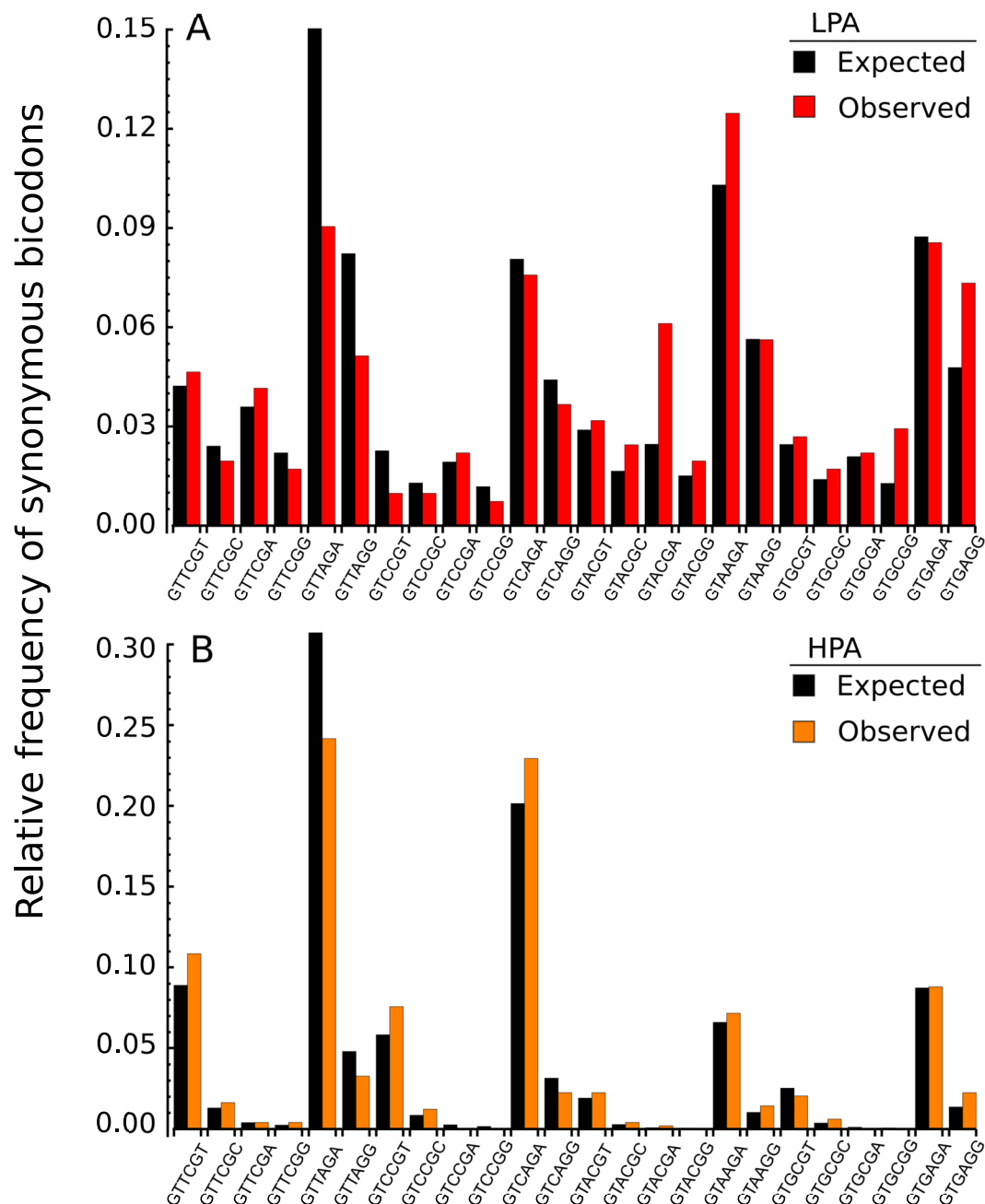
Figure 2. Bicodon usage for the VR amino acid pair. Red and orange bars denote the frequency of bicodons observed in a set of 500 coding sequences with low PA (A), and in a set of 500 coding sequences with high PA (B), respectively. Black bars represent the expected frequency obtained by the product of each codon frequency.

Figure 3. Bicodon preference in *S. cerevisiae*. The scatter plot of RSBU^L *vs.* RSBU^H illustrates the preference of bicodons in yeast. The inhibitor bicodons reported by (Gamble et al., 2016) are indicated in red.
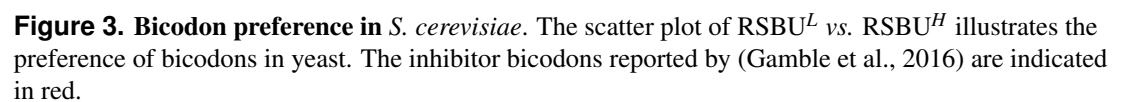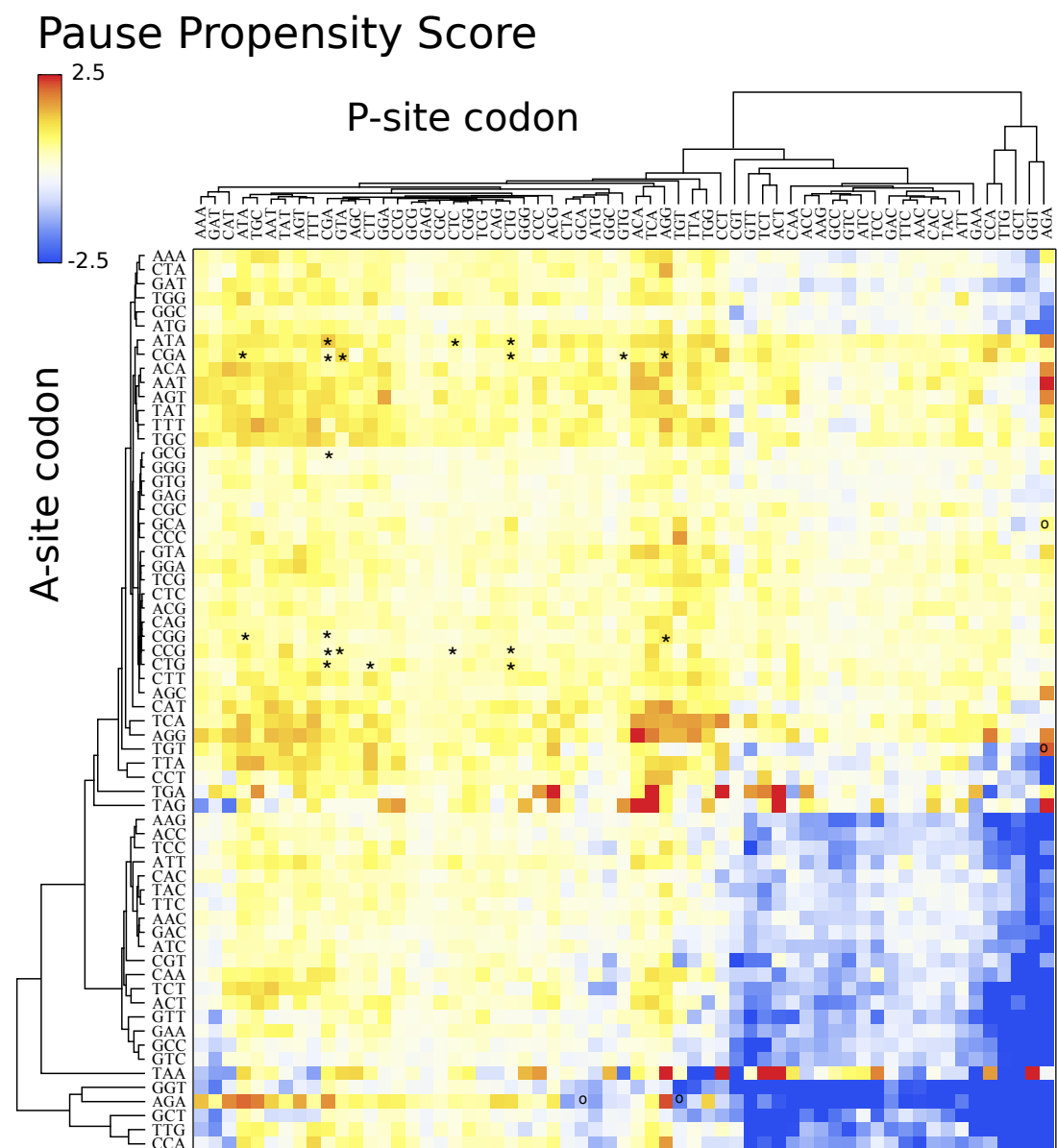
Figure 4. Pause propensity heat maps for *S. cerevisiae*. Each cell corresponds to the bicodon composed by the P-site codon (horizontal axis) and the A-site codon (vertical axis). Cell color is determined by the pause propensity score π associated with the bicodon. Red cells indicate bicodons with a clear preference for sequence associated with low PA, while blue cells indicate bicodons with high PA preference. Bicodons were grouped according to π similarity using average linkage clustering on each axis. Asterisks indicates the inhibitor bicodons identified in yeast (Gamble et al., 2016).
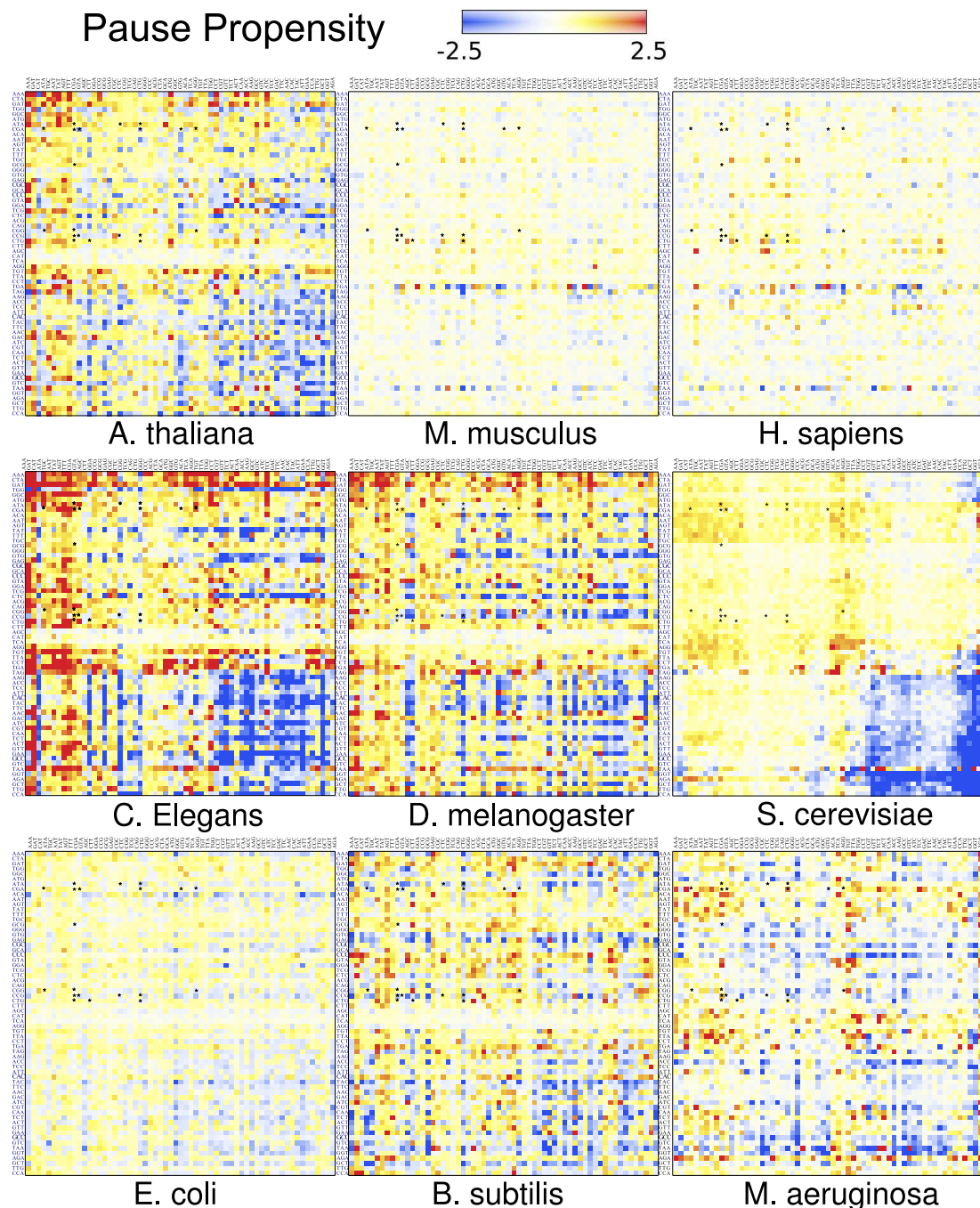
Figure 5. Pause propensity heat maps for all nine organisms. Each cell corresponds to the bicodon composed by the P-site codon (horizontal axis) and the A-site codon (vertical axis). The position of bicodons in all the maps is the same as in Fig. 4. The color of the cell is determined by the pause propensity score π associated with the bicodon. Red cells indicates bicodons with a clear preference for sequences associated with low PA, while blue cells indicate bicodons with high PA preference. Bicodons were grouped according to π similarity using average linkage clustering on each axis. Asterisks indicate the inhibitor bicodons identified in yeast (Gamble et al., 2016).
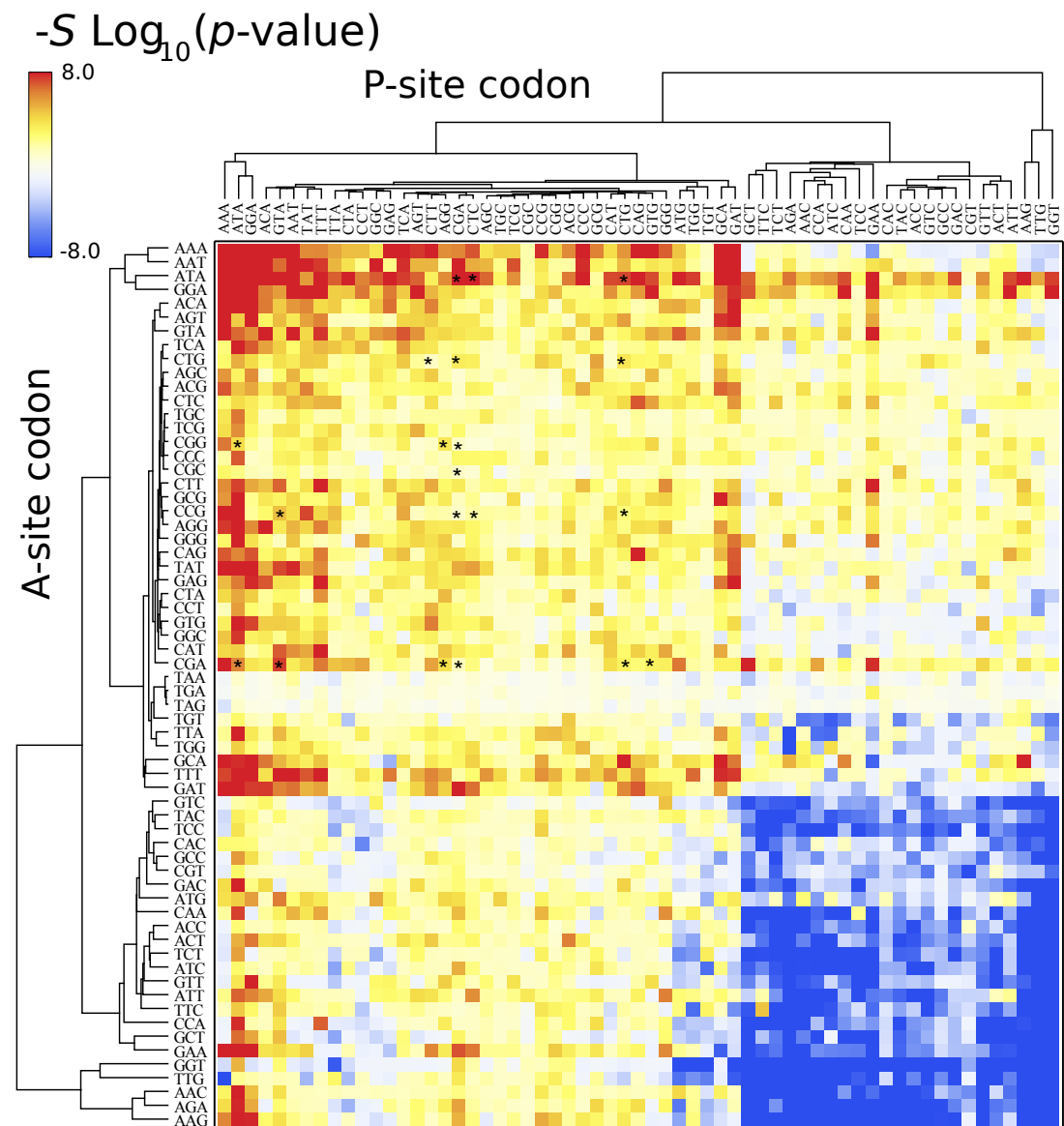
Figure 6. Fisher's exact test for *S. cerevisiae*. The color of each cell is determined by the quantity $-S\log(p\text{-value})$, where p-value is that provided by the Fisher's exact test, S takes the values $+1$ or -1 when the bicodon has preference for sequences with low or with high PA, respectively. Thus, red cells indicate bicodons with a clear preference for sequences associated with low PA, while blue cells indicate bicodons with high PA preference.

Figure 7. Cumulative ribosome occupancy *vs.* **π score and** *p***-value for** *S. cerevisiae***.** (A) Scatter plot of the cumulative ribosome occupancy and the pause propensity scores. (B) Scatter plot of the cumulative ribosome occupancy and quantity $-S \log_{10}(p\text{-value})$. Red circles correspond to the known inhibitor bicodons. Other candidates for modulating transcript elongation rate are indicated in Supplementary Table S3 by ∗∗ ($\pi > 0.75$, p-value < 0.01 and cumulative ribosome occupancy > 0.03
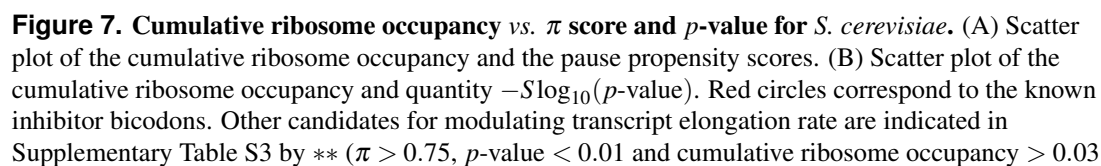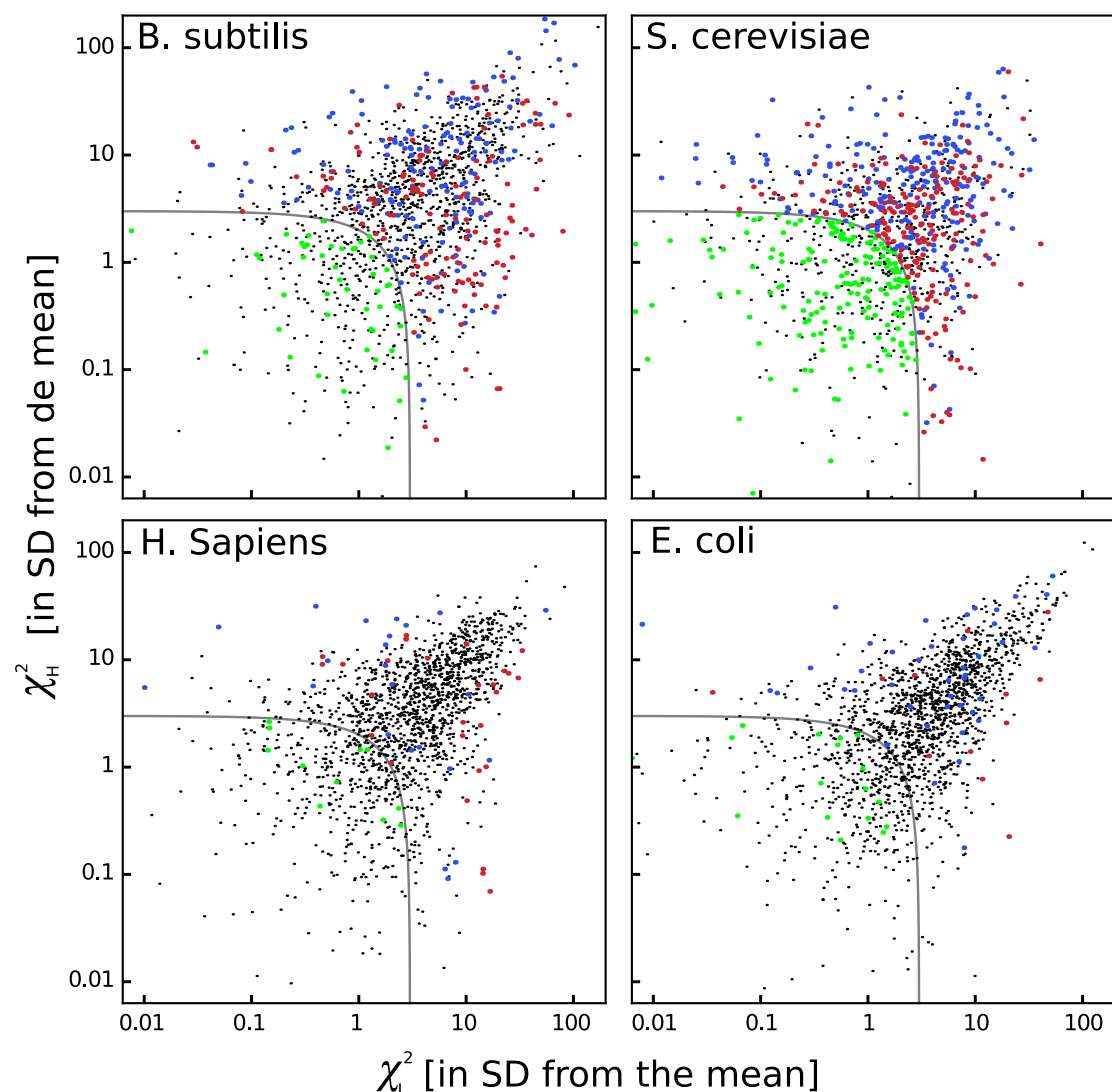
Figure 8. Bicodon bias cannot be explained by codon bias. Scatter plots indicating the residual scores χ^2_L and χ^2_H computed over low PA and high PA samples, respectively, for *B. subtilis, S. cerevisiae, H. sapiens*, and *E. coli*. The axes units are the number of standard deviations from the means obtained as indicated in methods. Raster plots of other organisms are displayed in Supplementary Fig. S13. The codon pairs whose preference for sequences with low or high PA cannot be explained by codon usage bias are outside the grey quadrant (i.e., $\chi^2 \geq 3$ SD above the mean). Among these, it is possible to distinguish bicodons more frequently used in low PA sequences (red dots), or in high PA sequences (blue dots). Inside the quadrant, there are codon pairs with a significantly different usage frequency in low and high PA samples, but whose bias can be explained by codon usage bias (green dots). Codon pairs whose usage frequencies in low and high PA samples are not significantly different aer indicated with black dots.

dipeptide	bicodon	$\chi_L^2 + \chi_H^2$			$-\log_{10}(p\text{-value})$		
		elegans	fly	yeast	elegans	fly	yeast
AA	GCCGCC	11.85	40.65	3.17	7.60	7.90	2.37
DA	GATGCT	15.28	9.96	16.66	2.13	2.03	2.25
TA	ACCGCC	12.25	7.80	4.02	7.41	3.13	5.70
IA	ATTGCC	3.05	143.23	35.23	4.45	9.06	4.18
VA	GTTGCC	10.63	5.47	10.16	5.26	3.15	6.78
RR	CGTCGT	3.62	4.25	34.28	11.91	7.12	8.11
AK	GCCAAG	56.94	232.59	81.31	30.33	12.85	7.94
NK	AACAAG	76.90	53.88	5.93	17.40	17.28	28.86
DK	GACAAG	33.90	144.15	25.88	11.51	12.82	11.80
FK	TTCAAG	34.07	14.43	39.10	18.98	4.90	10.80
TK	ACCAAG	8.56	51.21	30.44	17.016	13.45	14.38
YK	TACAAG	26.80	38.75	17.08	10.78	9.69	10.39
IK	ATCAAG	92.63	73.15	31.05	27.48	14.95	22.05
VK	GTCAAG	22.94	71.67	27.28	18.17	4.83	16.28
AN	GCCAAC	48.52	34.31	5.87	13.30	11.83	3.34
NN	AACAAC	19.27	52.78	9.09	5.03	6.21	15.14
DN	GACAAC	35.91	77.21	15.15	5.41	10.01	4.75
FN	TTCAAC	25.22	6.25	11.16	9.31	4.63	13.52
TN	ACCAAC	29.56	26.29	13.69	9.64	8.03	15.67
YN	TACAAC	25.17	26.30	8.26	2.74	3.23	8.62
IN	ATCAAC	58.97	51.28	21.56	11.88	5.47	15.06
TF	ACCTTC	3.71	10.08	3.77	7.68	6.84	3.09
PP	CCACCA	6.17	60.99	3.47	16.27	3.65	3.65
KS	AAGTCC	27.85	6.26	5.37	5.84	4.31	8.17
AT	GCCACC	24.28	102.57	4.98	9.58	10.90	4.08
NT	AACACC	14.60	6.294	5.81	7.15	5.43	12.79
DT	GACACC	11.14	48.32	5.38	6.27	6.27	7.15
FT	TTCACC	13.13	36.00	14.68	18.63	6.05	6.92
ST	TCCACC	5.21	22.41	3.85	6.89	5.70	5.38
TT	ACCACC	5.94	56.13	8.62	11.92	5.32	10.75
IT	ATCACC	19.35	25.86	8.31	18.03	8.72	3.96
VT	GTCACC	30.70	49.61	4.60	8.83	2.92	4.72
DY	GACTAC	11.84	62.43	4.73	6.56	8.66	4.94
AI	GCCATC	5.91	33.39	29.66	17.33	9.26	3.84
FI	TTCATC	5.64	12.57	9.23	12.91	4.21	5.65
TI	ACCATT	3.62	60.64	32.32	3.02	2.57	10.12
TI	ACCATC	4.43	22.69	8.51	15.17	6.35	10.03
II	ATCATC	12.55	27.02	14.52	27.31	8.41	12.78
VI	GTCATC	15.21	71.68	5.70	11.29	2.76	8.11
GV	GGTGTC	10.47	11.41	13.41	2.18	2.36	13.04

Table 2. Shared bicodons in *C. elegans, D. melanogaster* and *S. cerevisiae* that have high preference for high PA sequences (*p*-value < 0.01), but for which this preference cannot be explained by codon usage bias ($\chi^2 > 3\text{SD}$ above the mean).