Supplementary material for: Synonymous codon usage influences the local protein structure observed

Rhodri Saunders

Department of Statistics
Oxford University, 1 South Parks Road, Oxford, OX1 3TG, UK

saunders@stats.ox.ac.uk

Charlotte M. Deane

Department of Statistics
Oxford University, 1 South Parks Road, Oxford, OX1 3TG, UK

deane@stats.ox.ac.uk

Abstract

Here we present all our materials, methods and results to support findings in the main paper. The supplementary material is ordered to corresponds to the order of referrals in the main paper.

1 A1 - Other measures used

1.1 Secondary structure preference and translation speed.

It was observed in [1] that codon translation speed and secondary structure propensity were correlated. In general, slow translating codons were found have a higher propensity to form beta strand and coil, whereas fast translating codons were found predominantly in helix. We investigated this using CSandS and all measures of codon usage available to us. For each organism the full range of translation speeds was split in to B equal sized bins, where B was 8, 10, 12 and 20. For example, using MinMax the minimal speed is -100 and the maximal +100; with 10 bins the first will be from -100 to -80. For each bin the propensity of forming Helix, Strand or Coil is calculated. In all cases we find no general consensus.

1.2 RNA structure

Using the Vienna RNA tools (version 1.8.4 `http://www.tbi.univie.ac.at/RNA/`), the structural significance of codons on local mRNA structure was measured through $RNALfold$. This program identifies local secondary structure elements within a particular base-pairing window and calculates the overall energy of the RNA sequence. Here, we allow base-pairing within 15 nucleotides to analyse local secondary structures. The energy, E, of lowest energy conformation of this mRNA sequence is returned by $RNALfold$. Next, we make all synonymous mutations to the codon of interest and for

each one calculate the lowest energy conformation. Taking $log(E_{new}/E_{old})$ a positive result indicates an increase in local structure stability on synonymous mutation. No results are obtained for Met or Trp as they have no synonymous codons. Likewise, no results are obtained for codons within 20 nucleotides of either terminus. These tests were only carried out on the `Ecoli` data set due to time constraints. Other measures of RNA structural change were also assessed and are described briefly below.

In addition to $RNALfold$ this study made use of $RNAdist$ and $RNAheat$. In both cases we first utilise $RNAfold$ to calculate a global structure, G, for the native mRNA sequence, giving G_{old}. This process is repeated for each synonymous codon mutation to give G_{new}. As a basic test of structural similarity we use $RNAdist$ to calculate the number of base pair changes between G_{old} and G_{new}. Similarly, $RNAheat$ is used to calculate the specific heat capacity of G_{old} at 37 . Subsequently the specific heat capacity of each G_{new} at 37 is attained and compared to that of G_{old}. An increase in specific heat capacity indicates in increase in structural stability. For both $RNAdist$ and $RNAheat$ we identify no significant differences between `Ecoli` codons identified as significant and all other `Ecoli` codons.

1.3 Slow and fast translating segments.

We investigate whether the observed slower translation of N-terminal domains compared to C-terminal domains is a whole protein effect or caused by particular regions of the protein. For `Ecoli` two-domain proteins we examine every window of 10 codons. If it contains at least X of the slowest N codon translation speeds the window is designated slow. Fast windows are assigned in an analogous fashion. Here we take $X = 4$ and $N = 12$. The N slowest/fastest codon translation speeds are taken from the protein as a whole; however, the two domains are analysed separately.

Here data is presented for tRNA concentration. In general we find that the slowest and fastest translating regions are confined to a single domain (Table 1); not necessarily the same domain. In general, the larger the window size used to calculate translation speed the more slow/fast regions are observed per domain. Additionally, the percentage of slow/fast regions confined to a single domain drops but it is still high. In all cases, the first domain has more slow regions per domain than fast regions per domain. Conversely, there are more fast regions per domain than slow regions per domain in the second domain.

1.4 Shannon entropy information.

Secondary structure propensity, as described above, provides information about the structural preference of a particular codon. In each instance, it relates a particular codon to a particular secondary structure assignment. Some amino acids and amino acid motifs are known to effect the protein structures around them [2, 3, 4]; for example a run of non-polar residues is more likely to be found in a beta-strand or trans-membrane helix. Such effects are not assessed using the method described above. Local sequence can be taken into account using the Shannon entropy. Using Shannon entropy, Brunak [2] observed significant information content in codons around the start and end of beta strands. For each of our six terminal secondary structure classifications (H1/E1/C1 and H3/C3/E3) we calculate the information content (for details see [2]) contained in the surrounding nucleotide sequence.

For the Shannon entropy, only a forward (C-terminal) reading frame, is used to assign the secondary structure classification. That is, secondary structure is assigned to a codon taking into account only that codon and its following codon. In this way, on a secondary structure transition, one codon

Measure	Window size		
	3	9	19
% slow	92.5	86.8	84.9
% fast	90.0	81.4	73.6
Domain 1			
Mean slow	1.43	5.60	6.87
Mean fast	1.31	5.12	6.26
Domain 2			
Mean slow	1.28	4.65	6.13
Mean fast	1.31	5.30	7.70

Table 1: Slow and fast segments are confined to one domain in two-domain proteins. Column 2 shows the results obtained when calculating codon translated speed from a window of size 3. Column 3 uses a window of size 9 and column 4 a window of size 19. Row 2 (% slow) gives the percentage of proteins in which the slow translating regions are confined to a single domain. Row 3 (% fast) provides the same data for fast translating codons. Mean slow (rows 5 and 8) gives the mean number of slow segments per domain. Mean fast (rows 6 and 9) provides the same data for fast translating codons.

is responsible for both ending the current secondary structure and starting the following secondary structure. Using the example above (TGCATGTTGCAGAAA $->$ HHHCC), `TTG` would be the transition codon for both H3 and C1.

As in [2] the information content of nuncleotides (`A`, T, C and G) around secondary structure boundaries is calculated. We also calculate the information content in translation speed around secondary structure transitions. In this case, each speed is compared to the maximal translation speed and binned in 5 speed categories accordingly.

1.4.1 Burial of optimal codons.

We test whether buried amino acids, those with low solvent accessibility, are encoded for by optimal codons as suggested by Zhou [5]. JOY assesses solvent accessibility on a per amino acid basis via the OOI number [6]. In the JOY output OOI numbers are presented as values between 0 and 9. The most buried residues have a value of nine. We take buried amino acids to be those with an OOI number of 7 to 9. Taking all residues in a group the mean translation speed is calculated.

1.5 Extrusion effects.

The ribosome exit tunnel is known to shield portions of the synthesised protein from the folding environment. Current research suggests that the 30 to 72 most recently synthesised residues can be contained in the exit tunnel [7]. It is thus possible that the timings and speed of translation may result in protein structural effects not centred on the currently translating codon. For this reason we assess all our measures incorporating a frame-shift of 30, 40 or 50 residues. For example, with a frame shift of 30 we assign the secondary structure at residue 1 to the codon at position 31. However, no coherent results are obtained.

1.6 The Mantel-Haenszel test

To examine whether secondary structure classification affected the synonymous codons observed for each amino acid we used the Mantel-Haenszel procedure. Results are shown in Table 3.

		Ecoli									Human								
		H1	H2	H3	C1	C2	C3	E1	E2	E3	H1	H2	H3	C1	C2	C3	E1	E2	E3
GCA	A					**O**			**U**						u			o	
GCC	A	**O**	**U**						**O**										
GCG	A	**U**	**O**			**U**				u									
GCT	A		**U**	o															
TGC	C																		
TGT	C																		
GAC	D	o														u			
GAT	D	u														o			
GAA	E	**O**	**U**		**O**														
GAG	E	**U**	**O**		**U**														
TTC	F				**O**				**U**						o			u	
TTT	F				**U**				**O**						u			o	
GGA	G				**U**	o			**U**	u		o							
GGC	G																	u	
GGG	G		**O**																**U**
GGT	G								**O**										
CAC	H																		
CAT	H																		
ATA	I					**O**			u						u		**O**	o	
ATC	I														o				
ATT	I								o								**U**		
AAA	K					u												o	
AAG	K					o												u	
CTA	L																		
CTC	L	o	**U**					u											
CTG	L		**O**			**U**			**O**	**U**			u				u		
CTT	L		**U**							**O**									
TTA	L					**O**			**U**										
TTG	L																		
AAC	N	**O**								**O**									
AAT	N	**U**								**U**									
CCA	P																		
CCC	P	u																	
CCG	P						o												
CCT	P																		
CAA	Q								**U**									**U**	
CAG	Q								**O**									**O**	
AGA	R																	o	
AGG	R		o																o
CGA	R												o						
CGC	R										**U**	o							
CGG	R					**U**		**O**							o				
CGT	R		u			**O**													
AGC	S																		
AGT	S																	**U**	
TCA	S										**O**								
TCC	S						**U**						**O**					**O**	
TCG	S					u	**O**												
TCT	S																		
ACA	T								u									**U**	
ACC	T																	**O**	
ACG	T		**O**		**U**					**U**	o								
ACT	T							u		**O**			o						
GTA	V		**O**	o		**O**			**U**										
GTC	V		**U**		**O**					**O**	**U**								
GTG	V	**U**	**O**			**U**		o		**U**						**U**			
GTT	V	**O**	**U**					u			**O**					**O**	o		
TAC	Y		u																
TAT	Y		o																

Table 2: Results of the Mantel-Haenszel test with significance at the 5% level (p-value ¡= 0.05). A **O** indicates that the result is over-expressed (both MH and CHi tests) and a **U** indicates that the result is under-expressed (both MH and CHi tests). Significant results in the MH test not corroborated by the Chi test are indicated with an "o" or "u". Column 1 shows the codon and column 2 the encoded amino acid (by which codons are grouped). Columns 3 to 11 give data for `Ecoli` and 12 to 30 `Human`.

		Yeast								
		H1	H2	H3	C1	C2	C3	E1	E2	E3
GCA A	-									
GCC	A									
GCG	A									
GCT	A									
TGC	C								u	
TGT	C								o	
GAC	D						u			
GAT	D						o			
GAA	E									
GAG	E									
TTC	F									
TTT	F									
GGA	G					**O**				
GGC	G									
GGG	G									
GGT	G					**U**			o	
CAC	H									
CAT	H									
ATA	I									
ATC	I									
ATT	I									
AAA	K									
AAG	K									
CTA	L									
CTC	L									
CTG	L									
CTT	L									
TTA	L									
TTG	L				o					
AAC	N									
AAT	N									
CCA	P									
CCC	P									
CCG	P									
CCT	P				o					
CAA	Q									
CAG	Q									
AGA	R									
AGG	R									
CGA	R									
CGC	R									
CGG	R					u				
CGT	R									o
AGC	S									
AGT	S									
TCA	S								o	
TCC	S									
TCG	S									
TCT	S									
ACA	T									
ACC	T									
ACG	T									
ACT	T									
GTA	V			u						
GTC	V									
GTG	V						u			
GTT	V									
TAC	Y								**O**	
TAT	Y								**U**	

Table 3: Results of the Mantel-Haenszel test with significance at the 5% level (p-value ¡= 0.05). A **O** indicates that the result is over-expressed (both MH and CHi tests) and a **U** indicates that the result is under-expressed (both MH and CHi tests). Significant results in the MH test not corroborated by the Chi test are indicated with an "o" or "u". Column 1 shows the codon and column 2 the encoded amino acid (by which codons are grouped). Columns 3 to 11 give data for `Yeast`.

2 A2 - Codon frequency and tRNA abundance are not correlated

Codon usage measures based on large data sets do not correlate well with the original codon adaptation index as published in [8]. Neither our measures of codon usage nor the original CAI correlate with the abundance of cognate tRNA in `Ecoli` (Table 4). Nor do the RCSU values of Sharp and Li [8] correlate to tRNA concentration or our MinMax values (Figure 1).

Measure	MM HE	CAI All	CAI HE	tRNA	Original CAI
MM All	0.939	0.607	0.624	0.037	0.109
MM HE	-	0.578	0.635	0.038	0.134
CAI All	-	-	0.956	0.064	0.387
CAI HE	-	-	-	0.068	0.389
tRNA	-	-	-	-	0.1179

Table 4: Correlation (R^2) between measures of codon translation speed. Data presented is for `Ecoli`. For MinMax and CAI 'All' represents speeds calculated from all EMBL coding sequences while 'HE' includes only coding sequences of highly expressed genes. tRNA data is taken from Dong [9] and the Original CAI data is from [8]

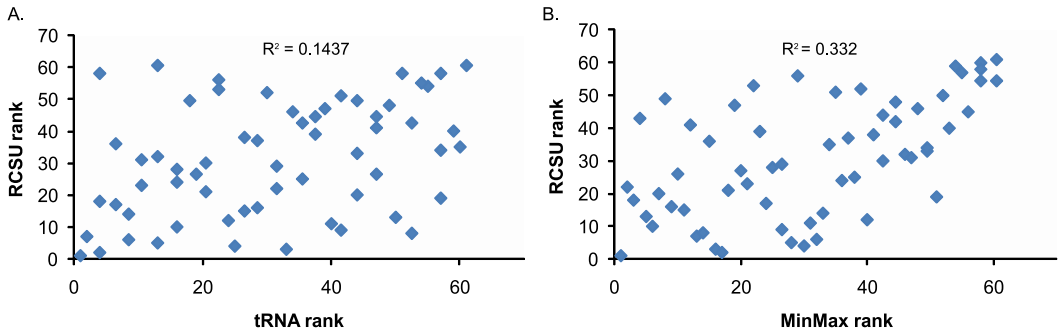

Figure 1: As with CAI adaptivity, the RCSU values published by Sharp and Li [8] do not correlate with tRNA concentration ($R^2 = 0.1437$). The correlation between RCSU and our MinMax scores is $R^2 = 0.332$.

3 A3 - Gene expression

Gene expression data is difficult to compare over different experiments. For this reason data was normalised relative to the mean expression level detected within each experiment. A value of 1 means that the gene was expressed at the mean level; greater than 1 the gene had above average expression and below 1 below average expression. In each of the organisms the number of available gene expression experiments varies. In the case of `Yeast` there are 131 different experiments. For each experiment we identify the top 5% of expressed genes. Then in order to create our final set of

highly expressed genes we select the top 5% of genes which are most commonly identified as highly expressed across the 131 experiments.

Data sets were obtained from the gene expression omnibus on 4^th June 2009.

3.1 `Ecoli`

data sets GDS2584 (11 experiments) and GDS2587 (7 experiments).

3.2 `Human`

data sets GDS594 (158 experiments) and GDS596 (158 experiments).

3.3 `Yeast`

data set GDS1116 (131 experiments).

Further tests were carried out taking the top 10% of genes and, for `Human` the top 5% of genes from three experiments at random. Neither significantly changed the results.

4 A4 - Buriedness and codon speed

We find that more buried residues as measured by Ooi number [6] are translated more quickly as measured by tRNA concentration. JOY [10] scales Ooi number to be between zero (exposed) and 9 (buried). On average, residues with an Ooi score of zero are encoded for by codons with a speed of 4.044. For an Ooi score of nine this is 4.65. If we consider residues in the extreme two bins, zero/one and eight/nine, the mean speeds are 4.013 and 4.24 respectively.

5 A5 - Translation speed is linked to large differences in propensity

We note that translation speed as measured by tRNA concentration is related to differences in propensity (Figure 2). The propensity of a codon for secondary structure SS is compared to the propensity of its encoded amino acid for the same secondary structure. Taking the square of the difference for each structural class the median and mean difference in propensity between a codon and its amino acid is calculated. The relationship to translation speed is maintained if we examine tRNA concentration or relative tRNA concentration, where the tRNA concentration (TC) of a codon is related to the mean tRNA concentration for all synonymous codons (M) by $\log \frac{TC}{M}$.

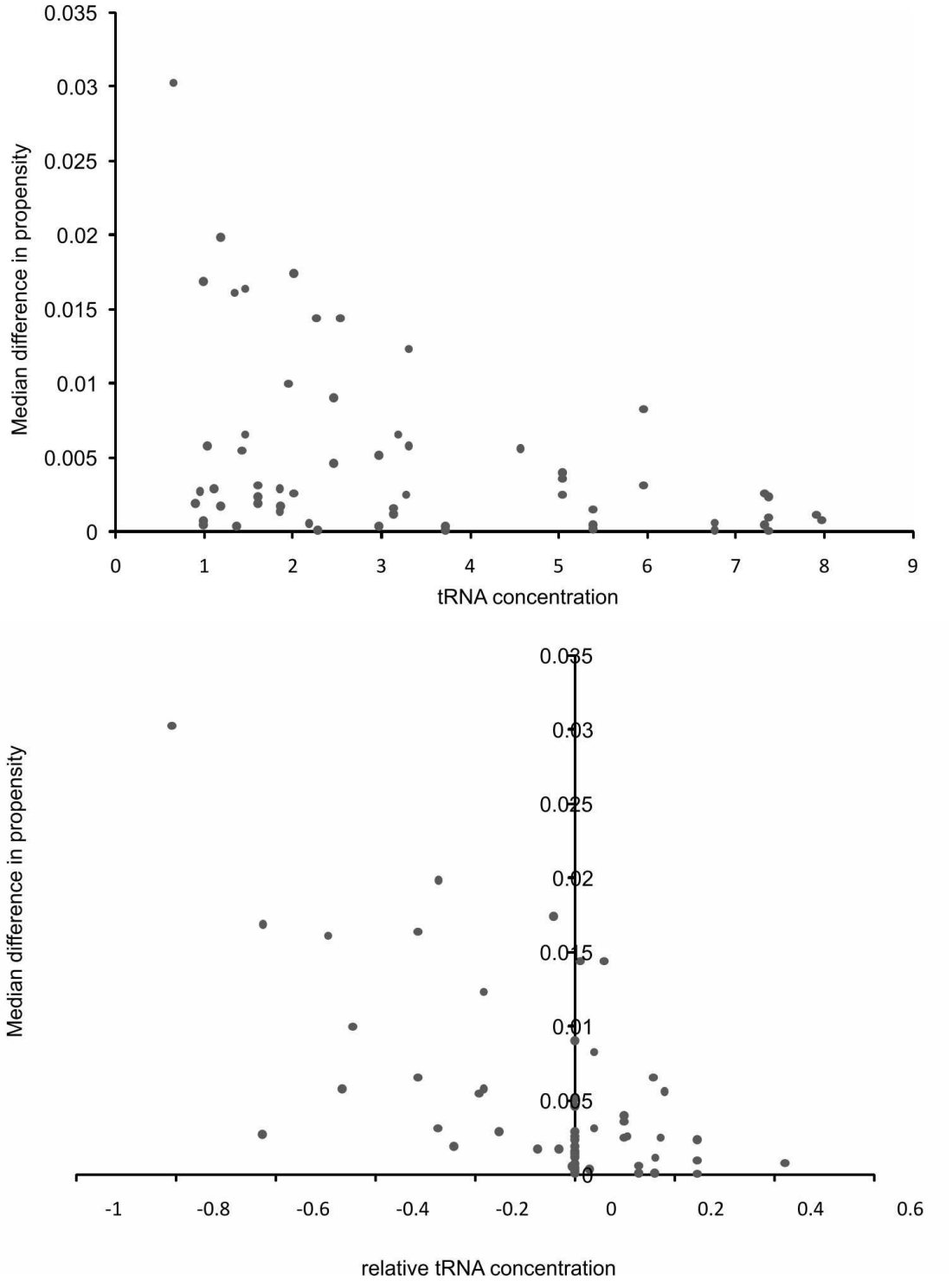

Figure 2: Codons that exhibit large differences in secondary structure propensity relative to the amino acid they encode are, in general, slow translating. The median difference in propensity over the 9 structural classes (Y-axis) is plotted against the translation speed (X-axis). Translation speed is measured by the tRNA concentration (top) and the relative tRNA concentration with synonymous codon familes (bottom). Data is presented for `Ecoli`; each point on the graph represents a codon. Using the mean difference in propensity over the 9 structural classes does not alter the results. Similar trends are observed for MinMax and CAI.

Codon	Amino acid	Helix	Coil	Strand	Codon	Amino acid	Helix	Coil	Strand
GCC	A	0	+-	0	AAT	N	0	0	-+
GCA	A	-+	+-	0	CCG	P	+-	-+	0
GCG	A	-+	0	0	CCA	P	+-	-+	-+
GCT	A	-+	+-	0	CCC	P	+-	-+	0
TGT	C	-+	+-	-+	CCT	P	+-	-+	-+
TGC	C	0	0	0	CAG	Q	-+	0	0
GAC	D	+-	0	-+	CAA	Q	-+	0	0
GAT	D	+-	0	-+	CGA	R	0	-+	+-
GAG	E	+-	0	0	AGA	R	-+	+-	+-
GAA	E	0	0	0	CGC	R	-+	+-	+-
TTT	F	-+	+-	+-	AGG	R	0	0	0
TTC	F	-+	+-	0	CGT	R	-+	+-	0
GGA	G	0	0	0	CGG	R	-+	-+	+-
GGT	G	0	+-	-+	AGT	S	+-	-+	0
GGG	G	-+	+-	-+	AGC	S	0	0	-+
GGC	G	0	+-	-+	TCG	S	0	0	0
CAC	H	-+	+-	-+	TCA	S	+-	-+	-+
CAT	H	-+	0	0	TCC	S	+-	0	-+
ATC	I	0	0	0	TCT	S	0	0	-+
ATT	I	0	0	0	ACA	T	+-	-+	0
ATA	I	0	0	0	ACC	T	+-	0	0
AAG	K	-+	0	+-	ACG	T	+-	-+	0
AAA	K	-+	0	+-	ACT	T	+-	0	0
CTG	L	-+	0	0	GTG	V	0	-+	+-
CTC	L	-+	+-	0	GTT	V	0	0	0
TTG	L	-+	+-	0	GTA	V	0	0	0
CTT	L	-+	+-	-+	GTC	V	0	0	0
TTA	L	-+	+-	0	TGG	W	0	-+	+-
CTA	L	-+	+-	0	TAC	Y	-+	+-	+-
ATG	M	-+	0	+-	TAT	Y	-+	0	+-
AAC	N	0	0	-+					

Table 5: **Codons that exhibit opposed propensities for secondary structure termini**. Data taken from `Ecoli` proteins. Column 1 gives the codon, column 2 the amino acid it encodes and columns 3,4,5 the relative terminal propensity for helix, coil and strand respectively. +- indicates the codon is over-represented at the amino- and under-represented at the carboxy-terminus; while -+ indicates the codon is under-represented at the amino- and over-represented at the carboxy-terminus. Zero means the codon has equivalent propensity for the two termini.

6 A6 - Shannon entropy information results

Using the Shannon entropy we investigate whether there are informative signatures in the nucleotide sequence surrounding the termini of secondary structure elements. As in [2] we observe an increase in the information content of Thymine, `T`, at the start of beta strands. For the three codons immediately upstream of the beta strand start point, the central nucleotide of codons is more likely to be `T`. As suggested by Brunak [2], this may be due to the hydrophobic amino acids often found in beta strands. Taking the `Ecoli` data set we further investigate whether there are nucleotide signals around those codons shown (above) to be significant in their secondary structure preferences; e.g. `CGA` (Arg). No signatures are observed which suggests that structural significance originates from the specific codon rather than other local sequence effects.

Codon translation speed has been linked to secondary structure: with fast codons having a preference for helix and slow codons a preference for strand [1]. Undertaking equivalent tests as those used in [1] we found no such correlation (supplementary material). However, in grouping codons by speed and calculating the Shannon entropy of these speed bins around secondary structure transitions a slight bias was observed. Around the start of helices (H1), the information content of slow codons increases at the point of transition into Helix.

6.1 Relative codon speed

On calculating the Shannon entropy to analyse information content we calculate the relative codon speed within a single mRNA sequence. These relative speeds are scaled to lie between zero (minimal translation speed for that sequence) and 10 (maximal translation speed for that sequence). These scaled speeds are then binned in to 5 sets; to produce a normal distribution (Figure 3).

In all figures the codon assigned to the signal transition is placed in position 6 and indicated with a $*$. Prior, N-terminal, to the transition are residues -1 to -5. Likewise, After, C-terminal, to the transition are the residue 1 to 5. Slow codons are A and B, fast codons are D and E, median speed codons are assigned as C. In general, fast codons are more prevelant in coil than in either helix or strand. Additionally, at, or just before, all transition points the information conent of slow codons increases. This is most prominent in on the transition into helix (H1: Figure 4). On average, we found that the first domain of two domain proteins is the slower transcribed (see Translation speed of domains) and consider that the observed bias toward slow codons may come from N-terminal helices. Comparing H1 transitions that occur in the first quarter (N-terminal, centre logo in figure 4) of the protein to H1 transitions that occur in the final quarter (C-terminal, bottom logo in figure 4) of the protein we observe a stronger signal in N-terminal H1 transitions. C-terminal transitions to helix avoid fast translating codons.

Below details for all secondary structure transition boundaries are provided. In all cases data is presented for relative codon translation speed using tRNA concentration in `Ecoli`; where A is the slowest codons and E the fastest codons. Opposed to the work of Thanaraj and Argos [1] (described above) our results indicate that coil is the fastest transcribed secondary structure. As described in the main paper, the start of a helix (H1 transition) is signed by a relative decrease in translation speed. An increase in the information content of slow codons (A) is also observed (Figure 4). Our results additionally suggest general differences between the properties of secondary structure transitions occurring in the amino-quartile compared to those in the carboxy-quartile. The amino-quartile transitions have an increase in slow codons. This feature could be linked to co-translational protein folding, in which amino-terminal secondary structure elements are given increased time to fold with high fidelity.

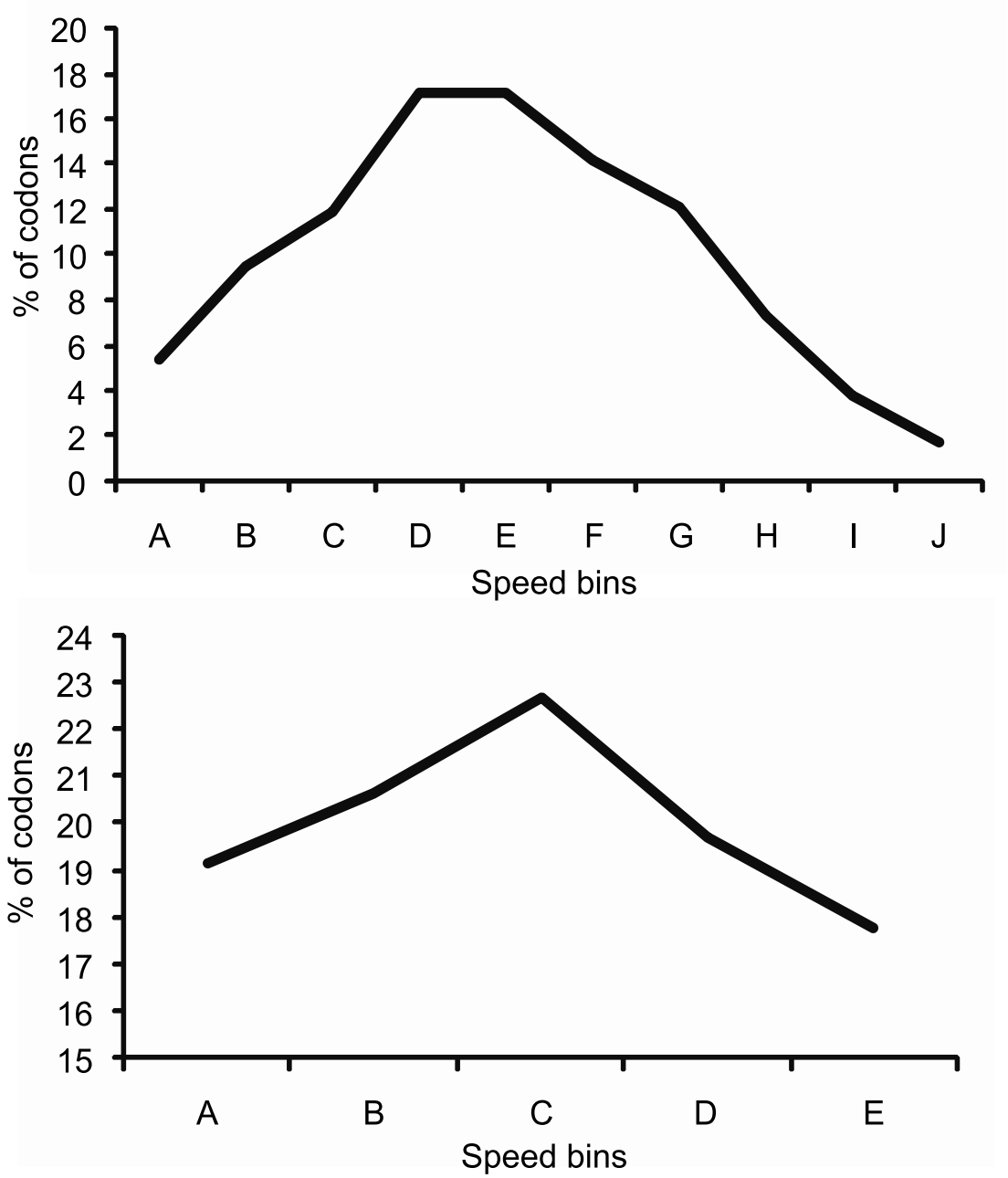

Figure 3: Relative codon translation speed as calculated from tRNA concentration is only approximately normally distributed. Using standard sized bins (Top) 60% of data lies in the first half of distribution. Scaled bins (Bottom) produce a normal distribution with 40% of data in bins A and B and 38% of data in bins D and E.

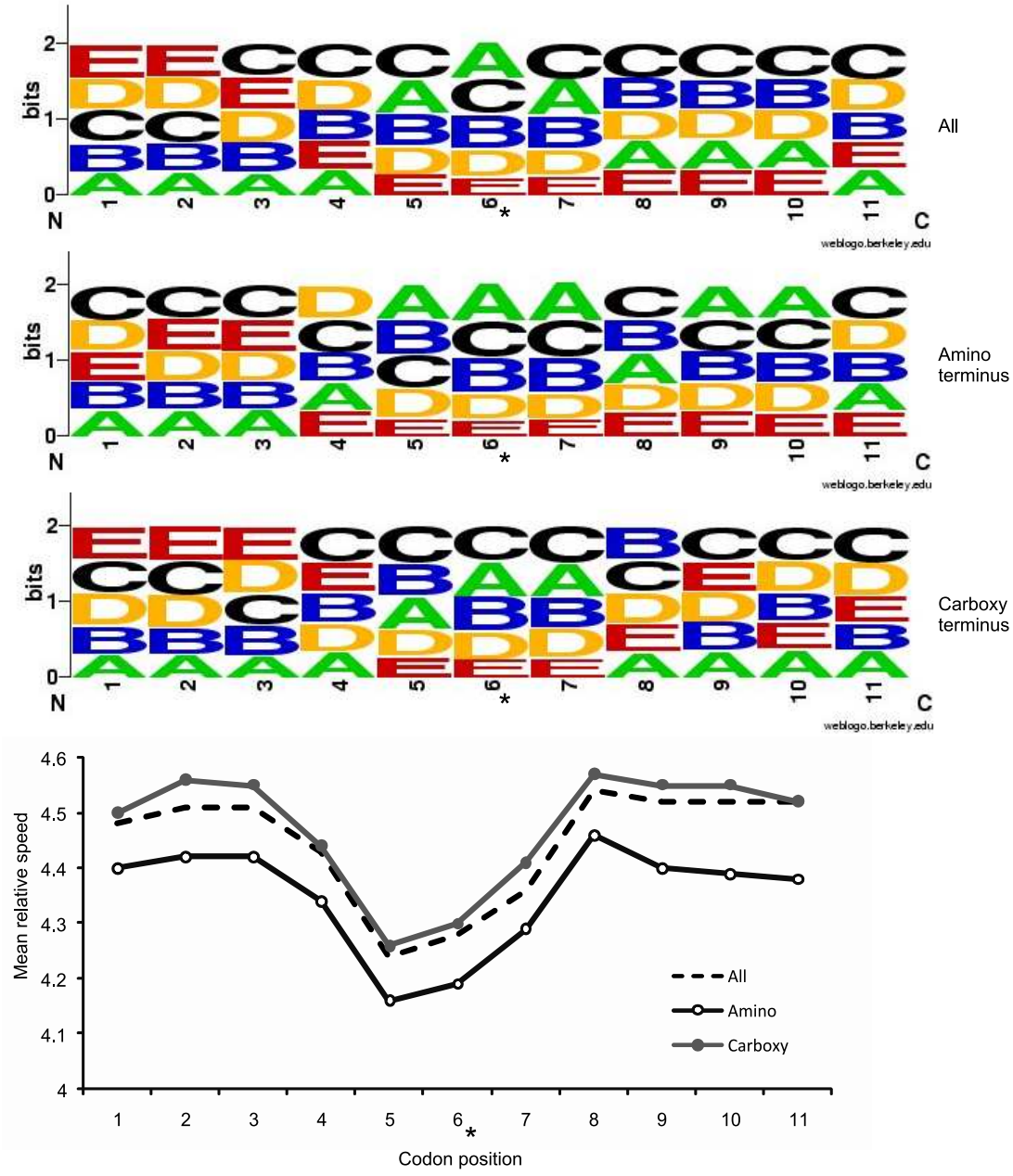

Figure 4: The H1 transition. The slowest codons, A, are enriched at the transition in to Helix (H1). This is particularly true in amino terminal quartile of proteins. Fast codons, D and E, are avoided in this area but are enriched in codon positions 1 to 3. The enrichment of slow codons is clearly observed in the graph where mean translation speed (Y-axis) drops around the transition boundary (codon 6, X-axis). Figure created using Weblogo [11].

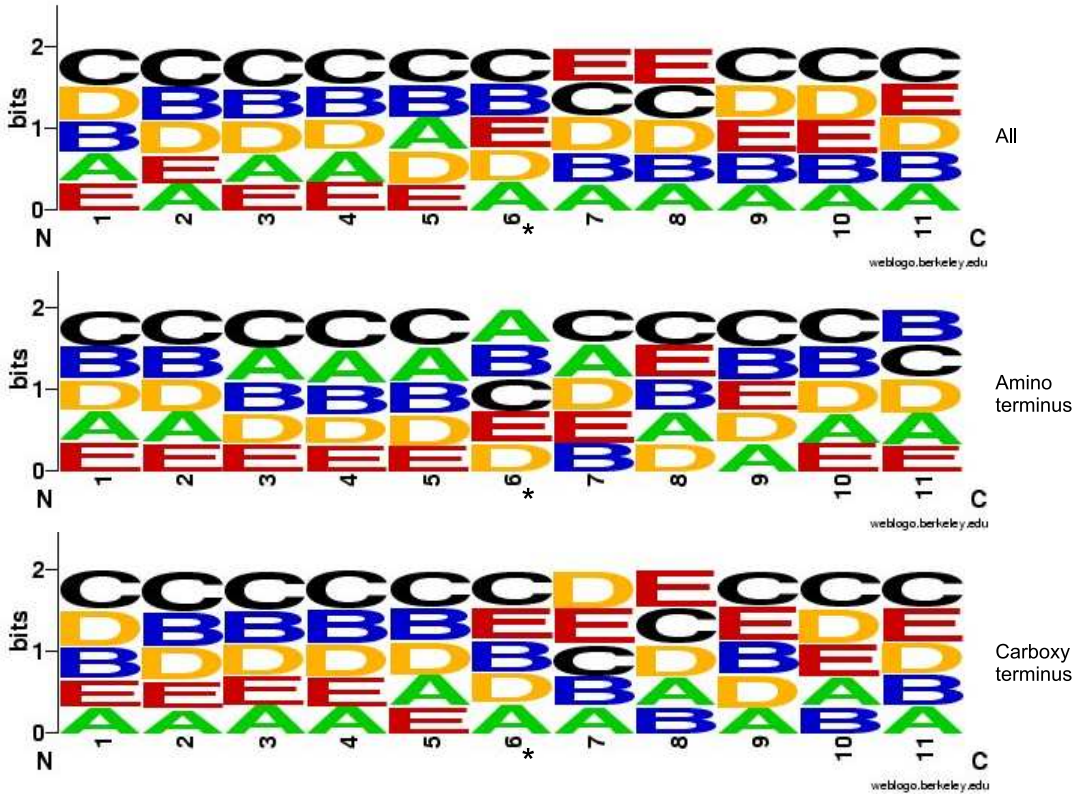

Figure 5: The H3 transition. Median speed codons, C, dominate just before the helix end (H3). In general, an increase in fast codons, D and E, is seen just after the transition (codons 6 to 8). Figure created using Weblogo [11].

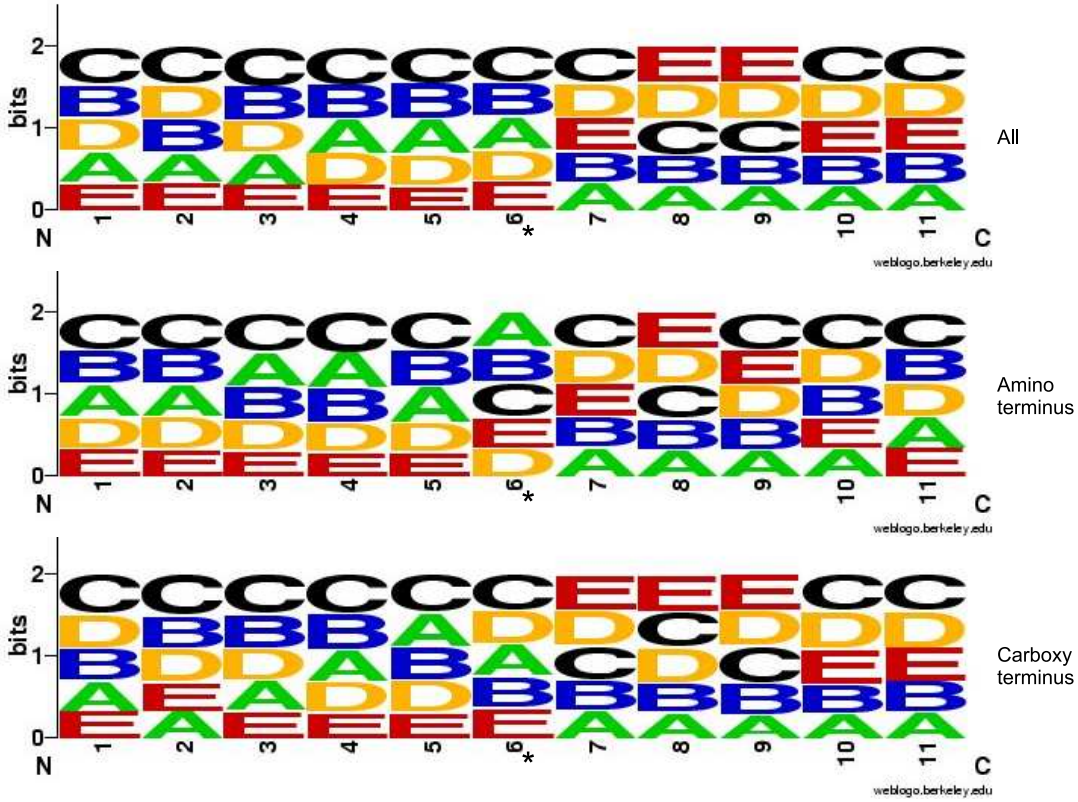

Figure 6: The C1 transition. In all cases we see a prominence of fast codons, E and D, at the start of the coil. Whereas slowest codons, A, are avoided. Figure created using Weblogo [11].

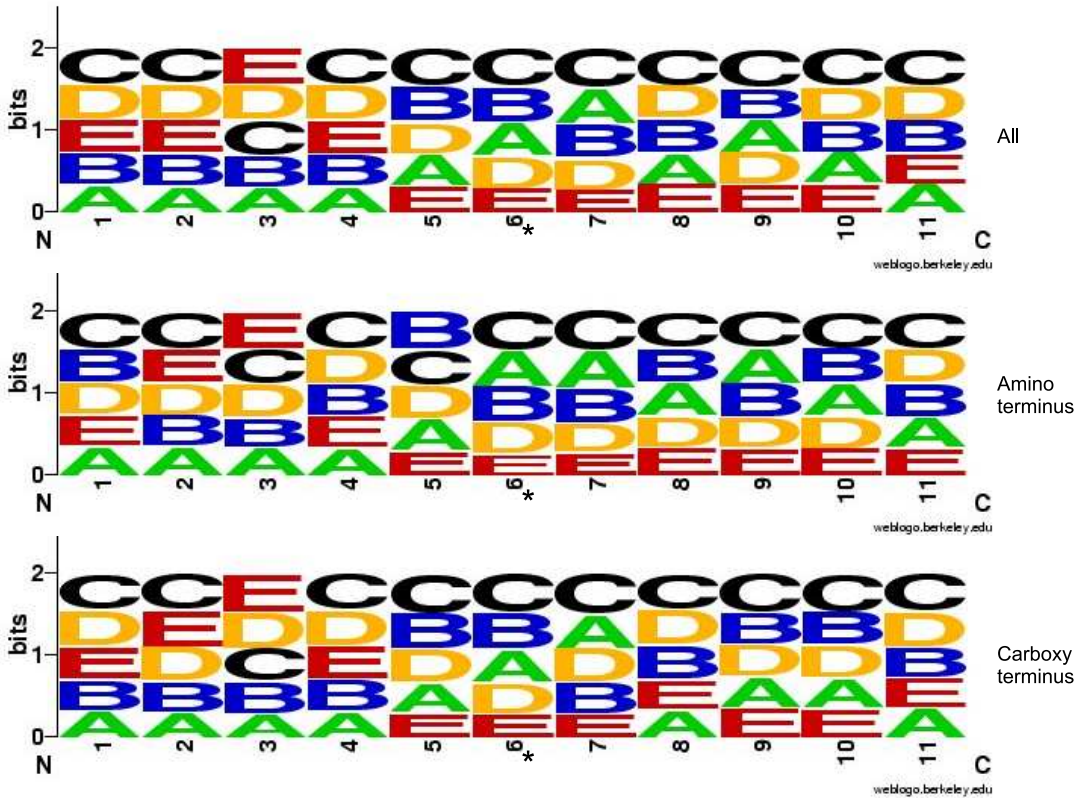

Figure 7: The C3 transition. The slowest codons, A, are avoided in the coil; however, as the coil terminates (codon position 5) fast codons, E, are avoided. Figure created using Weblogo [11].

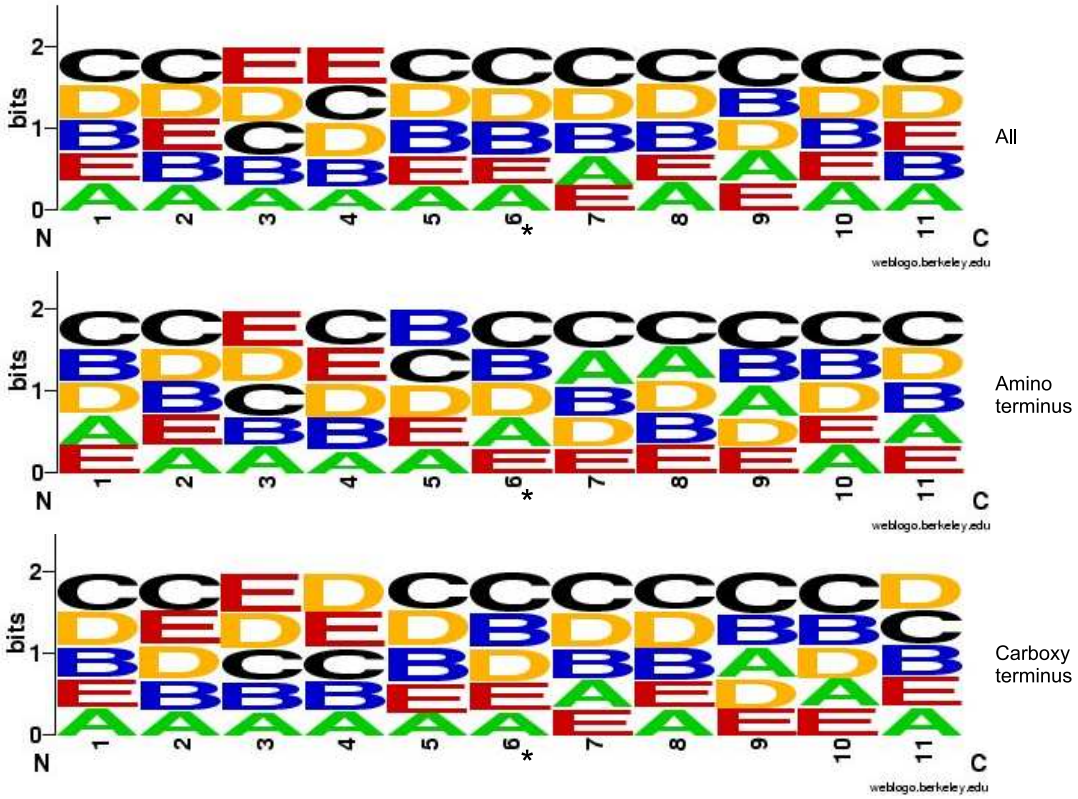

Figure 8: The E1 transition. A peak in fast codons, D and E, just before the beta strand starts is observed in position 3. Slow codons are generally avoided in and around strands with median speed codons (C) dominating. Figure created using Weblogo [11].

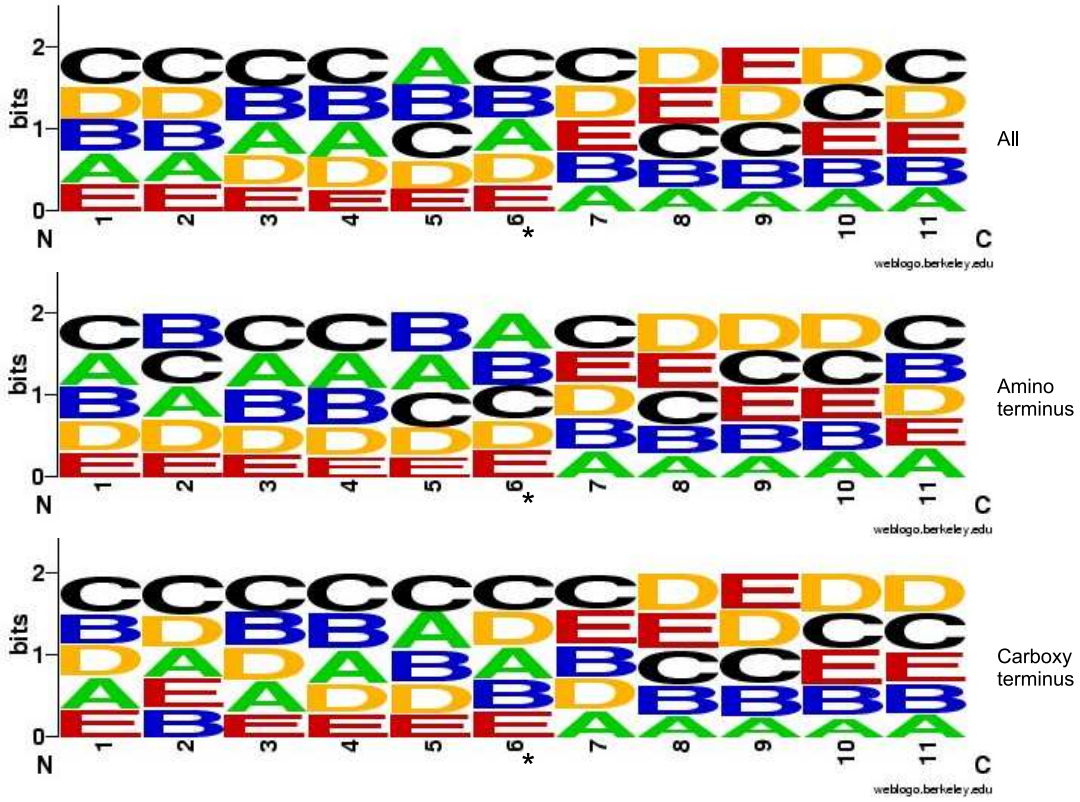

Figure 9: The E3 transition. Fast codons, E, are avoided before the strand ends, codon positions 1 to 6. Slow codons, A, are avoided immediately after the E3 transition. Slow codons, though, are often preferred immediately at the transition boundary in amino terminal helices. Figure created using Weblogo [11].

7 A7 - Codon speed at domain boundaries

Domains are commonly thought of as structuraly stable, individual folding units within the protein. The placement of domain boundaries is based on knowledge of protein structure and folding. However, the particular amino acid assigned to define the domain boundary is in essence an arbitary selection and varies even between well regarded databases such as SCOP, CATH [12] and Pfam [13]. Domain boundary definitions are also updated and change over time. Here, structurally defined domain boundaries as defined by SCOP (release 1.75) are used.

Like Brunak [2], we found no evidence that slow codons are clustered around domain boundaries in any of the three organisms in the study. This is true for all tested window sizes and codon speed definitions. However, we observe some evidence that domain boundaries avoid slow codons and that they are enriched in fast codons (Figure 10). Three positions about the domain boundary are examined. These codon windows are:

1. downstream of (N-terminal to) the domain boundary and end at the domain boundary
2. centred around the domain boundary
3. upstream of (C-terminal to) the domain boundary and start at the domain boundary

Using tRNA concentration (left side of Figure 10) we find that slow codons are avoided and fast codons slightly favoured around (centre) and upstream (bottom) of domain boundaries. Downstream (top) of domain boundaries the results are less clear. In general the domain boundary is thought to be less structurally conserved than intra domain loops and as such less codon selection is perhaps expected. This is what we observe downstream of domain boundaries and it may be that domain boundary placement is currently biased towards the C-terminus. Using the original CAI (right side of Figure 10) we do observe, for `Yeast`, a significant increase in slow codons downstream of domain boundaries (top). Similarly, using tRNA concentration there are a number of specific examples of proteins where slow codons are clustered around domain boundaries and we are able to reproduce the example figures found in [14]. We suggest that translational pausing at domain boundaries is not a general trait, but that pauses are incorporated in the nucleotide sequence where required for high-fidelity folding. At times this will be the domain boundary.

Our measures have so far utilised domain boundaries as presented in SCOP (release 1.75). Our tests have been rerun using domain definitions from CATH and PFAM. In general results are consistent with those of SCOP with more variance observed for PFAM (Figure 11). This may be due to domains in PFAM being defined by sequence rather than structure as in CATH and SCOP.

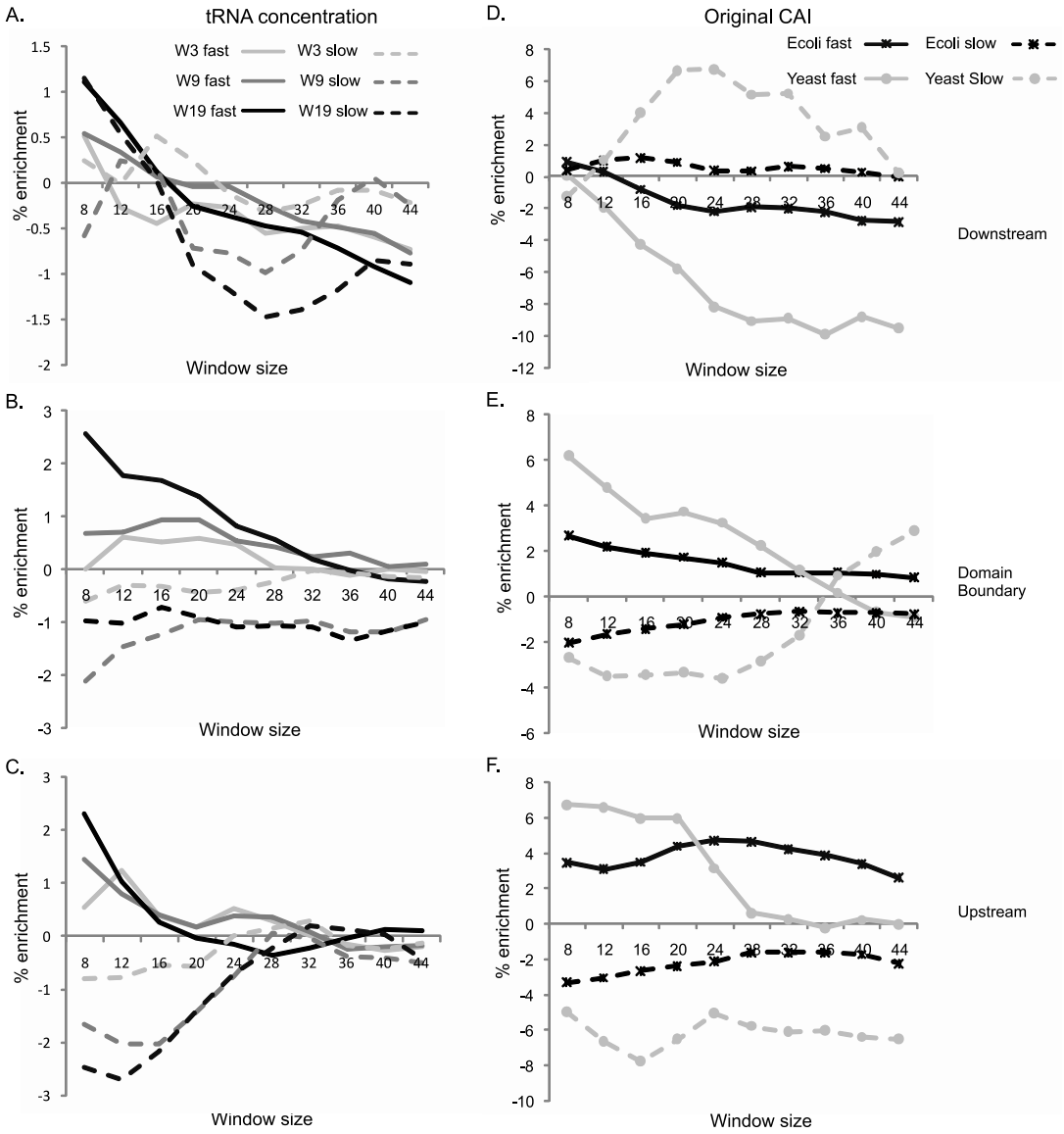

Figure 10: Domain boundaries are deficient in slow codons and enriched in fast codons. Data is shown using translation speed calculated from tRNA concentration (A, B, C) and the original CAI (D, E, F) for the window immediately downstream (A and D), around (B and E) and upstream (C and F) of the domain boundary. Solid lines represent fast codons and dotted lines slow codons. Enrichment (Y-axis), the percentage increase over that found in the protein as a whole, is shown for each window size (X-axis). For tRNA concentration (A, B, C) only `Ecoli` data is available, with data displayed for all three codon speed windows considered in this study (3, 9 and 19). In E to F data is displayed for `Ecoli` (Black) and `Yeast` (Grey) using a codon speed window of 19.

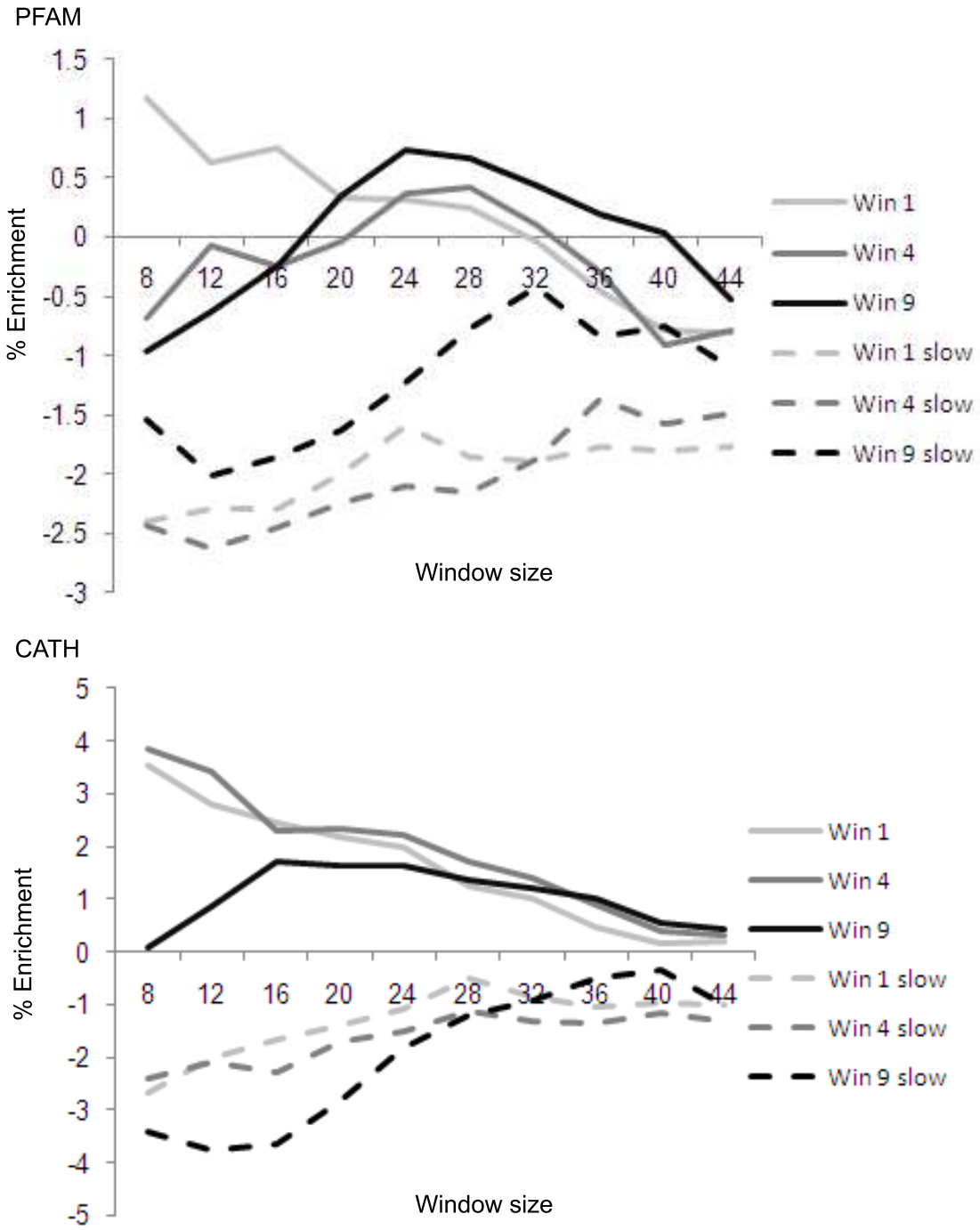

Figure 11: Secondary structure transitions are signed by a relative decrease in translation speed. Here we demonstrate that this is also observed when translation speed is ranked and that it is most clearly observed when analysing the modal value for each codon position (A). It is also evident when the mean (blue) or median (red) rank are used (B). Both A and B present data from the H1 transition (CCCHHH) using a fragment of 16 codons centred on the transition. In C. modal data is presented for both the H1 (black) and E1 (grey) transitions and in D. the modal data for H3 (black dashed) and E3 (grey dashed) transitions are shown. Again a fragment of 16 codons is used. In all cases a window of size 3 is used to estimate translation speed.

8 A8 - Buried residues are translated more quickly

Zhou *et al.* suggest that optimal codons, those with near maximal translation speeds, associate with buried residues [5]. They also conclude that these codons encode amino acids that are structurally important to the stability of the protein as a whole. This work is based on the GTOP database that provides a theoretical mapping between many (sometimes putative) gene sequences and a single solved protein structure. Buried residues should not alter significantly between sequences of high homology and thus the GTOP mapping is suitable for this study. CSandS provides a very detailed cross-reference between coding sequence and structure. By comparing the codon translation speed to the encoded residue's OOI number we support the link between faster translation and buried residues (supplementary material).

Some codons have a high propensity to encode buried amino acids as measured by OOI number. As an example of an 'over-buried' codon: Ala is encoded for by four codons, in `Human` it is far more likely that Ala is buried if encoded for by `GCG` than its three other synonymous codons. `GCG` encodes only 8.9% of Ala residues but 22% of buried Alanines. Using `Ecoli` data, we found that synonymous codon changes to 'over-buried' codons resulted in stabilisation of local mRNA structures. In 63.1% of cases the mRNA was stabilised by synoymous mutation of 'over-buried' codons. For all other codons the value was 44%. A student t-test indicated significance at the 2% level. We suggest that buried residues, and the fast translating codons that encode them, are situated in areas of the mRNA with low structural stability. The structure of mRNA can be critical to translation [15, 16, 17, 18] and tt has been shown that mRNA structures must be unfolded to enter the ribosome [19].

9 A9 - Translation speed at secondary structure transitions

For each secondary strcuture transition of the form YYYZZZ, where X and Z are different secondary structures (e.g. CCCHHH), we calculate the translation speed of each codon relative to the mean translation speed of the fragment. A relative decrease in translation speed is observed at the transition point in both CCCHHH and CCCEEE transitions. Calculating significance in terms of standard deviation or standard error is not applicable in this instance due to the distributions. We can however use non-parametrtic test to investigate whether the distribution of speeds for position A varies significantly from the distribution of speeds at position B. Here we use both the Wilcoxon Signed Ranks test and the KolmogorovSmirnov test as implemented in the **R** statistical package as *wilcox.test* and *ks.test* respectively. Table 6 shows that the transition codons (CH) behave significantly differently to nearly all other codon positions. When ranking the translation speeds (from 1 (slowest) to 16 (fastest)) for each fragment and taking the mean rank for each codon position the same general trend as shown in the main paper (Figure 4) is observed. Further, the same trend is found for the median rank. When the modal rank for each codon is analysed the transition codons are clearly different from the other codon position (Figure 12).

10 ACKNOWLEDGEMENTS

Charlotte M. Deane is partially funded by the Oxford University Doctoral Training Centres (Industrial and Systems Biology).

10.0.1 Conflict of interest statement.

None declared.

Codon	-5	-4	-3	-2	-1	C	C	C	H	H	H	1	2	3	4	5
-5		0.60	0.35	0.32	0.84	0.0010	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	6.9e-13	1.4e-06	0.00090	0.045
-4	0.91		0.15	0.13	0.47	0.0059	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	1.9e-15	4.7e-11	1.9e-05	0.0056	0.14
-3	0.22	0.27		0.96	0.48	3.2e-05	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	1.5e-15	1.1e-08	2.6e-05	0.0037
-2	0.31	0.43	0.97		0.43	2.1e-05	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	5.5e-16	6.3e-09	1.8e-05	0.0028
-1	0.83	0.52	0.47	0.78		4.9e-4	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	1.7e-13	4.7e-07	0.00042	0.027
C	0.0098	0.011	0.0019	0.0012	0.0066		1.4e-15	2.2e-16	2.2e-16	3.947e-15	4.8e-10	1.8e-07	0.00012	0.12	0.96	0.20
C	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.1e-12		1.9e-05	2.1e-08	0.99	0.093	0.0070	4.3e-05	1.5e-10	2.2e-15	2.2e-16
C	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	0.00021		0.18	2.5e-05	4.4e-09	5.6e-12	2.2e-16	2.2e-16	2.2e-16	2.2e-16
H	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	8.0e-07	0.38		3.4e-08	6.8e-13	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
H	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	3.12e-14	0.54	0.00046	4.6e-06		0.099	0.0075	5.2e-05	3.4e-10	8.3e-15	2.2e-16
H	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	1.8e-09	0.02	1.0e-06	2.9e-09	0.44		0.30	0.016	3.5e-06	9.0e-10	8.1e-14
1	2.1e-14	5.4e-14	2.2e-16	4.4e-16	2.0e-13	1.7e-06	0.022	8.3e-09	3.0e-12	0.027	0.38		0.17	0.00028	2.7e-07	9.7e-11
2	1.0e-10	1.0e-10	2.1e-12	3.5e-13	1.7e-10	0.00025	0.00067	4.2e-12	1.6e-15	0.00061	0.053	0.51		0.022	0.00016	3.2e-07
3	2.4e-05	3.0e-05	2.8e-06	9.1e-07	4.6e-05	0.14	6.1e-09	2.2e-16	2.2e-16	2.6e-08	7.5e-06	0.0011	0.02		0.13	0.0053
4	0.0014	0.015	0.0024	0.00041	0.00098	0.93	8.6e-13	2.2e-16	2.2e-16	4.2e-13	9.6e-09	1.1e-06	0.00022	0.23		0.19
5	0.12	0.078	0.025	0.025	0.093	0.38	4.5e-15	2.2e-16	2.2e-16	2.2e-16	2.9e-13	1.3e-08	3.2e-07	0.0091	0.22	

Table 6: Non-parametric tests of significance. Taking the distribution of speeds for the H1 transition (XXXXXCCCHHHXXXXX) in `Ecoli` we demonstrate that the transition codons (CH) vary significantly in their speed distribution than all other codon positions in the fragment. Data (p-values) for the Wilcoxon Signed Ranks test is displayed in the top-right half of the table. The bottom-left half provides data (p-values) from the Kolmogorov-Smirnov test. Tests carried out using the R statistical package (*ks.test* and *wilcox.test*) where the results do not go lower than $2.2e^{-16}$. Where the tabled value is 2.2e-16 this refers to "<2.2e-16".

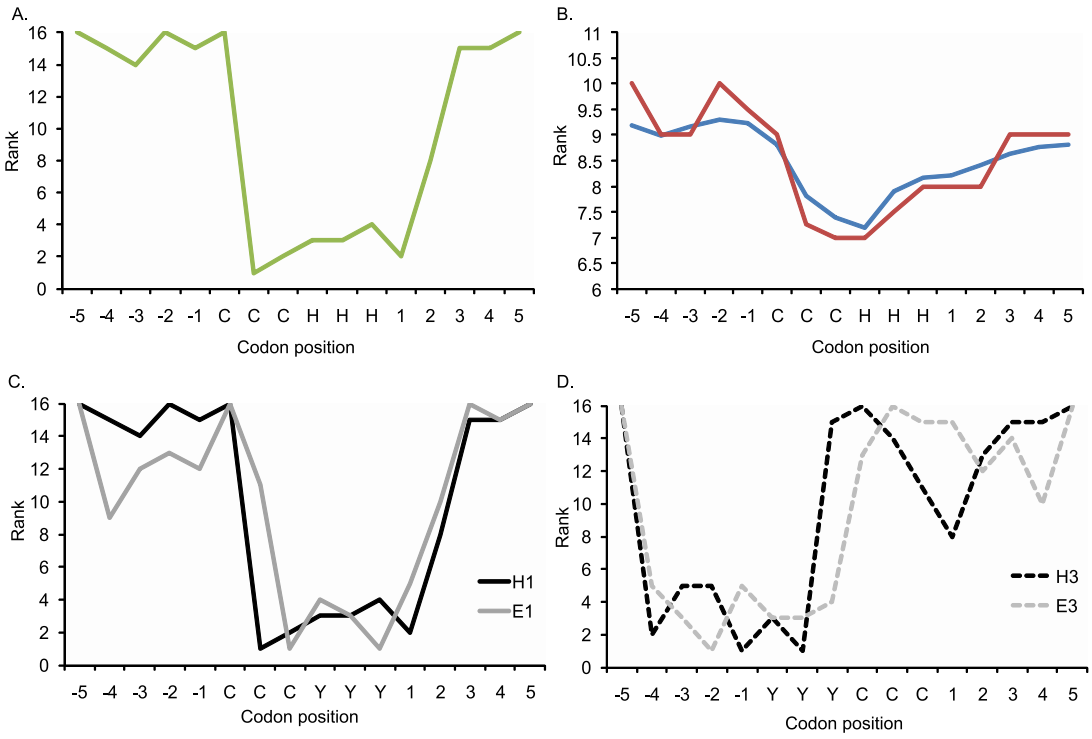

Figure 12: Secondary structure transitions are signed by a relative decrease in translation speed. Here we demonstrate that this is also observed when translation speed is ranked and that it is most clearly observed when analysing the modal value for each codon position (A). It is also evident when the mean (blue) or median (red) rank are used (B). Both A and B present data from the H1 transition (CCCHHH) using a fragment of 16 codons centred on the transition. In C. modal data is presented for both the H1 (black) and E1 (grey) transitions and in D. the modal data for H3 (black dashed) and E3 (grey dashed) transitions are shown. Again a fragment of 16 codons is used. In all cases a window of size 3 is used to estimate translation speed.

References

[1] Thanaraj, T. A. and Argos, P. (1996) Protein secondary structural types are differentially coded on messenger rna *Protein Sci* **5(10)**, 1973–83.

[2] Brunak, S. and Engelbrecht, J. June 1996 Protein structure and the sequential structure of mrna: alpha-helix and beta-sheet signals at the nucleotide level. *Proteins* **25(2)**, 237–252.

[3] Wilson, C. L., Hubbard, S. J., and Doig, A. J. July 2002 A critical assessment of the secondary structure alpha-helices and their termini in proteins *Protein Eng.* **15(7)**, 545–554.

[4] Saunders, R. and Deane, C. M. (2009) Protein structure prediction begins well but ends badly *Proteins: Structure, Function, and Bioinformatics* **78(5)**, 1282 – 1290.

[5] Zhou, T., Weems, M., and Wilke, C. O. July 2009 Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular biology and evolution* **26(7)**, 1571–1580.

[6] Nishikawa, K. and Ooi, T. July 1980 Prediction of the surface-interior diagram of globular proteins by an empirical method. *International journal of peptide and protein research* **16(1)**, 19–32.

[7] Kramer, G., Ramachandiran, V., and Hardesty, B. (2001) Cotranslational folding–omnia mea mecum porto? *Int J Biochem Cell Biol* **33(6)**, 541–53.

[8] Sharp, P. M. and Li, W. H. February 1987 The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research* **15(3)**, 1281–1295.

[9] Dong, H. August 1996 Co-variation of trna abundance and codon usage inescherichia coliat different growth rates *Journal of Molecular Biology* **260(5)**, 649–663.

[10] Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998) Joy: protein sequence-structure representation and analysis *Bioinformatics* **14(7)**, 617–23.

[11] Crooks, G. E., Hon, G., Chandonia, J.-M. M., and Brenner, S. E. June 2004 Weblogo: a sequence logo generator. *Genome research* **14(6)**, 1188–1190.

[12] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) Cath–a hierarchic classification of protein domain structures *Structure* **5(8)**, 1093–108.

[13] Sonnhammer, E. L., Eddy, S. R., and Durbin, R. July 1997 Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28(3)**, 405–420.

[14] Zhang, G., Hubalewska, M., and Ignatova, Z. February 2009 Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology* **16**, 274–280 doi:10.1038/nsmb.1554.

[15] Fukami, H. and Imahori, K. March 1971 Control of translation by the conformation of messenger rna. *Proceedings of the National Academy of Sciences of the United States of America* **68(3)**, 570–573.

[16] Shpaer, E. G. January 1985 The secondary structure of mrnas from escherichia coli: its possible role in increasing the accuracy of translation. *Nucleic acids research* **13(1)**, 275–288.

[17] Carlini, D. B., Chen, Y., and Stephan, W. October 2001 The relationship between third-codon position nucleotide content, codon bias, mrna secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes adh and adhr. *Genetics* **159(2)**, 623–633.

[18] Frugier, M., Bour, T., Ayach, M., Santos, M. A., Rudinger-Thirion, J., and Théobald-Dietrich, A. November 2009 Low complexity regions behave as trna sponges to help co-translational folding of plasmodial proteins. *FEBS letters* **Epub** doi:10.1016/j.physletb.2003.10.071.

[19] Marzi, S., Myasnikov, A. G., Serganov, A., Ehresmann, C., Romby, P., Yusupov, M., and Klaholz, B. P. September 2007 Structured mrnas regulate translation initiation by binding to the platform of the ribosome. *Cell* **130(6)**, 1019–1031.