The relationship between synonymous codon usage and protein structure

Xie Tao, Ding Dafu*

Shanghai Institute of Biochemistry, Academia Sinica, 320 Yue Yang Road, Shanghai 200031, People's Republic of China

Received 26 June 1998

Abstract The hypothesis that synonymous codon usage is related to protein three-dimensional structure is examined by investigating the correlation between synonymous codon usage and protein secondary structure. All except two codons in *E. coli* show the same secondary structural preference for alpha-helix, beta-strand or coil as that of amino acids to be encoded by the respective codons, while 17 codons show secondary structural bias in mammalian proteins. The results indicate that there is no significant correlation between synonymous codon usage and protein secondary structure in *E. coli*, but there is a correlation in mammals. It could be deduced that synonymous codons carry much less structural information in prokaryotes than in eukaryotes due to their divergent evolutionary mechanism.

© 1998 Federation of European Biochemical Societies.

Key words: Synonymous codon usage; Protein secondary structure; Translation rate; Protein folding

1. Introduction

Coding sequences of DNA do not use synonymous codons with equal frequency, and codon usage bias is species-specific [1]. It has been known for some time that there is a high correlation between codon, tRNA abundance, translation rate and gene expression level [2–4]. When directly compared to protein structures, their coding mRNA sequences have in principle an additional potential to carry structural information concerning the encoded protein due to the degeneracy of the genetic code, which can be at the level of a single codon or a nucleotide fragment [5]. The first, second and third base of the codon have been associated with, respectively, the biosynthetic pathway, the hydrophobicity pattern and the alpha-helix or beta-strand forming potentiality of the coded amino acid [6–8]. Thanaraj et al. [9,10], using codon fractional frequencies as measures of translation speed, showed for *E. coli* that protein domain boundaries are largely coded by translationally slow mRNA regions, and protein helices are preferentially coded by fast codons while beta-strands and coils are preferentially coded by slow codons. However, Brunak et al. [11] came to a dissimilar conclusion that the correlation between the positioning of rare codons and the location of structural units of protein cannot be confirmed. Recently, Adzhubei et al. [12] clearly reported for a mammalian database [13] that synonymous codon usage in an amino acid residue is related to protein secondary structure. The analysis performed in this paper indicated that all but two synonymous codons in *E. coli* display the same propensity for alpha-helix, beta-strand or coil as that of amino acids to be encoded by the respective codons, while the synonymous codon bias against the secondary structure classes is indeed significant in the mammalian database.

2. Materials and methods

The sequence-structure dataset of 54 proteins from *E. coli* used in the study comprises mRNA sequence [9], amino acid sequence and the three-dimensional structure of each given protein. Protein sequences and structures were extracted from PDB database [14] while secondary structural assignments of individual residues are available from DSSP [15]. Alpha-helices are annotated by G and H in the relative DSSP file, beta-strands by E and B, and coils by the rest. The mRNA sequences were taken from EMBL database [16]. After alignment between each protein sequence and its mRNA sequence (after the translation), 37 mismatches were detected and deleted. Pairwise alignment was done in the 54-protein dataset, and all sequence identities were below 25%. Finally, the total sum of residues (codons) for the non-redundant dataset is 14 107.

For a given amino acid in the dataset, N(ss,sc) denotes the observed occurrence of a synonymous codon sc relative to a secondary structure type ss. The total observed occurrence of a secondary structure ss is $N(\mathrm{ss}) = \Sigma_{\mathrm{sc}} N(\mathrm{ss, sc})$ and the total observed occurrence of a synonymous codon ss is $N(\mathrm{sc}) = \Sigma_{\mathrm{ss}} N(\mathrm{ss, sc})$. Then it follows that the total occurrence of the given amino acid in the dataset is $N = \Sigma_{\mathrm{sc}} \Sigma_{\mathrm{ss}} N(\mathrm{ss, sc})$. So the conditional probability of a synonymous codon sc for a secondary structure type ss is $P(\mathrm{ss|sc}) = N(\mathrm{ss,sc})/N(\mathrm{sc})$. Also, the probability for a secondary structure ss is $P(\mathrm{ss}) = N(\mathrm{ss})/N$; the probability for a synonymous codon sc is $P(\mathrm{sc}) = N(\mathrm{sc})/N$. Here, ss = {alpha-helix, beta-strand, coil}, sc = {each codon in a synonymous codon family}, for example, sc = GCU, GCC, GCA and GCG for the amino acid Ala.

The null hypothesis [17] is tested below:

H0: sc's secondary structure bias is the same as that of the amino acid it encoded, i.e. $P(\mathrm{ss|sc}) = P(\mathrm{ss})$.
H1: $P(\mathrm{ss|sc}) \neq P(\mathrm{ss})$.

If H0 is true, there is no correlation between sc and ss. By shuffling the relation between sc's and ss's in the original codons of a given amino acid, 200 random samples of N codons were created so that the original P(sc) and P(ss) were maintained. So, in each of these samples, codon usage and secondary structural bias of the given amino acid is kept, but the correlation between them is random. And this is what we want to investigate.

Then, each $P_{\mathrm{i}}(\mathrm{ss|sc})$, i = 1,2,...200, was calculated in these 200 samples. It was demonstrated through the statistical test that $P_{\mathrm{i}}(\mathrm{ss|sc})$ as a random variable is normally distributed with the mean $m(\mathrm{ss|sc})$ and the standard derivation $\sigma(\mathrm{ss|sc})$. Then, the statistic for testing the two hypotheses is the value $Z(\mathrm{ss|sc}) = (P(\mathrm{ss|sc}) - m(\mathrm{ss|sc}))/\sigma(\mathrm{ss|sc})$. Taking a statistical significance $\alpha = 0.01$, the two-trail-threshold of Z-statistic is $Z_0 = 2.33$. Hence, if $|Z| \leq Z_0$, then with the significance α, the null hypothesis is accepted, i.e. we can claim that the synonymous codon bias against the secondary structure classes is not significant statistically; otherwise if $|Z| > Z_0$ then the null hypothesis is rejected and hypothesis H1 is accepted.

3. Results and discussion

The obtained results are shown in Table 1. It can be seen from Table 1 that only Z(beta-strand|GGU) = 2.63 and Z(beta-strand|GGC) = −2.45 are statistically significant, but

*Corresponding author. Fax: (86) (21) 64338357.
E-mail: dingdafu@server.shcnc.ac.cn

PII: S0014-5793(98)00955-7

the other 59 sense codons are not. So, as a whole, the synonymous codon usage in *E. coli* does not show a significant correlation with protein secondary structure. The average *Z*(alpha-helix|sc) and average *Z*(beta-strand|sc) for 18 common codons (items with '#' in Table 1) are 0.397 and −0.142, respectively, and the average *Z*(alpha-helix|sc) and average *Z*(beta-strand|sc) for 8 rare codons (items with '*' in Table 1) of which RSCU (relative synonymous codon usage) < 0.05 [18], are −0.036 and 0.103, respectively. The analysis shows that neither of the common codons nor the rare codons has a preference for a protein secondary structural type, although an opposite result has been reported by Thanaraj et al. [9] that protein secondary structural types are differentially coded on messenger RNA in *E. coli*. Adzhubei [12] found by chi square

Table 1
Statistical results of the relationship between synonymous codons and protein secondary structure in *E. coli*

Amino acid	Codon	$N(\mathrm{H}\mid\mathrm{sc})$	$P(\mathrm{H}\mid\mathrm{sc})$	$Z(\mathrm{H}\mid\mathrm{sc})$	$N(\mathrm{B}\mid\mathrm{sc})$	$P(\mathrm{B}\mid\mathrm{sc})$	$Z(\mathrm{B}\mid\mathrm{sc})$	$N(\mathrm{C}\mid\mathrm{sc})$	$P(\mathrm{C}\mid\mathrm{sc})$	$Z(\mathrm{C}\mid\mathrm{sc})$	$N(\mathrm{sc})$	$P(\mathrm{sc})$
Ala	GCU	161	0.55	0.51	42	0.14	−1.04	92	0.31	0.32	295	0.21
	GCC	172	0.53	0.09	56	0.17	0.87	94	0.29	−0.78	322	0.22
	GCA	148	0.48	−2.25	51	0.17	0.37	110	0.36	2.22	309	0.22
	GCG#	283	0.56	1.34	80	0.16	−0.10	145	0.29	−1.34	508	0.35
Cys	UGU	14	0.26	−0.91	17	0.31	−0.38	23	0.43	1.28	54	0.39
	UGC#	28	0.34	0.91	29	0.35	0.38	26	0.31	−1.28	83	0.61
Asp	GAU#	148	0.32	−1.20	48	0.11	−1.14	261	0.57	1.78	457	0.54
	GAC	142	0.36	1.20	51	0.13	1.14	199	0.51	−1.78	392	0.46
Glu	GAA	348	0.50	−1.45	121	0.18	0.83	221	0.32	0.90	690	0.72
	GAG#	149	0.56	1.44	41	0.15	−0.83	77	0.29	−0.90	267	0.28
Phe	UUU	76	0.36	−0.71	69	0.33	0.24	66	0.31	0.56	211	0.46
	UUC#	98	0.40	0.71	76	0.31	−0.24	71	0.29	−0.56	245	0.54
Gly	GGU	82	0.18	−1.13	102	0.22	2.63	281	0.60	−1.04	465	0.39
	GGC#	112	0.20	0.40	88	0.16	−2.45	365	0.65	1.44	565	0.48
	GGA*	12	0.21	0.31	12	0.21	0.49	34	0.59	−0.63	58	0.05
	GGG	23	0.23	1.09	16	0.16	−0.65	59	0.60	−0.37	98	0.08
His	CAU	39	0.35	0.18	25	0.22	0.36	48	0.43	−0.51	112	0.41
	CAC#	55	0.34	−0.18	33	0.20	−0.36	73	0.45	0.51	161	0.59
Ile	AUU	141	0.35	−0.78	161	0.40	0.12	96	0.24	0.72	398	0.44
	AUC#	182	0.38	0.75	192	0.40	−0.21	107	0.22	−0.58	481	0.53
	AUA*	8	0.38	0.13	9	0.43	0.30	4	0.19	−0.48	21	0.02
Lys	AAA#	304	0.43	1.62	122	0.17	0.35	284	0.40	−1.92	710	0.82
	AAG	58	0.36	−1.62	26	0.16	−0.35	77	0.48	1.92	161	0.18
Leu	UUG	62	0.53	0.73	22	0.19	−1.56	32	0.28	0.85	116	0.09
	UUA	44	0.47	−0.75	22	0.23	−0.37	28	0.30	1.31	94	0.07
	CUU	39	0.41	−1.82	31	0.33	1.96	24	0.26	0.11	94	0.07
	CUC	69	0.58	1.87	20	0.17	−2.17	29	0.25	−0.04	118	0.09
	CUA*	15	0.54	0.35	3	0.11	−1.79	10	0.36	1.36	28	0.02
	CUG#	415	0.50	−0.30	222	0.27	1.84	192	0.23	−1.56	829	0.65
Met	AUG	141	0.44	0.00	70	0.22	0.00	109	0.34	0.00	320	1.00
Asn	AAU	60	0.31	0.61	19	0.10	−1.94	112	0.59	0.65	191	0.31
	AAC#	120	0.29	−0.61	67	0.16	1.94	231	0.55	−0.65	418	0.69
Pro	CCU	15	0.22	0.01	7	0.10	−0.04	45	0.67	0.02	67	0.12
	CCC*	9	0.25	0.56	4	0.11	0.05	23	0.64	−0.52	36	0.06
	CCA	18	0.18	−1.35	11	0.11	0.10	72	0.71	1.05	101	0.18
	CCG#	82	0.23	0.77	37	0.11	−0.08	231	0.66	−0.60	350	0.63
Gln	CAA	75	0.43	−2.00	29	0.17	1.18	70	0.40	1.27	174	0.30
	CAG#	211	0.52	2.00	52	0.13	−1.18	140	0.35	−1.27	403	0.70
Arg	CGU#	151	0.47	−0.47	58	0.18	−0.68	115	0.35	1.09	324	0.52
	CGC	129	0.49	0.74	49	0.19	−0.01	84	0.32	−0.82	262	0.42
	CGA*	5	0.42	−0.48	4	0.33	1.22	3	0.25	−0.57	12	0.02
	CGG*	11	0.48	0.08	6	0.26	0.90	6	0.26	−0.83	23	0.04
	AGA*	2	0.50	0.19	1	0.25	0.28	1	0.25	−0.44	4	0.01
	AGG*	0	0.00	−1.43	0	0.00	−0.63	2	1.00	2.08	2	0.00
Ser	UCU#	52	0.35	1.10	31	0.21	−0.53	66	0.44	−0.57	149	0.23
	UCC	44	0.34	0.89	30	0.23	0.16	55	0.43	−0.99	129	0.20
	UCA	22	0.39	1.34	12	0.21	−0.21	23	0.40	−1.13	57	0.09
	UCG	21	0.27	−0.84	24	0.31	1.76	33	0.42	−0.72	78	0.12
	AGU	13	0.20	−1.88	15	0.23	0.12	36	0.56	1.79	64	0.10
	AGC	50	0.29	−0.68	35	0.20	−0.91	87	0.51	1.42	172	0.27
Thr	ACU	56	0.34	0.10	40	0.24	−0.22	70	0.42	0.08	166	0.21
	ACC#	130	0.31	−1.06	101	0.24	−0.14	182	0.44	1.23	413	0.52
	ACA	18	0.36	0.46	13	0.26	0.15	19	0.38	−0.61	50	0.06
	ACG	58	0.36	1.01	40	0.25	0.32	61	0.38	−1.17	159	0.20
Val	GUU	103	0.30	−1.26	146	0.42	−0.09	99	0.28	1.41	348	0.33
	GUC	64	0.34	0.50	82	0.43	0.31	44	0.23	−0.88	190	0.18
	GUA	53	0.34	0.19	60	0.38	−0.94	45	0.28	0.94	158	0.15
	GUG#	125	0.34	0.74	159	0.43	0.54	86	0.23	−1.36	370	0.35
Trp	UGG	66	0.46	0.00	36	0.25	0.00	40	0.28	0.00	142	1.00
Tyr	UAU	83	0.38	0.81	73	0.33	−0.33	62	0.28	−0.45	218	0.50
	UAC#	75	0.35	−0.81	75	0.35	0.33	65	0.30	0.45	215	0.50

H, alpha-helix; B, beta-strand; C, coil; *, rare codons; #, preferred codons.

test that 9 synonymous codon families in 109 mammalian proteins carry additional structural information regarding the encoded proteins. These families were also investigated here. The results are shown in Table 2. There are 17 codons (items with '*' in Table 2) in 8 families biased against secondary structure types. As to the Glu family, individual codons pass neither in the Z-statistic test nor in the chi square test. So, the results of the two methods are consistent. But this method gives more detailed information of individual codons in synonymous families.

Our data were extracted from *E. coli*, a typical organism of prokaryotes, while Adzhubei et al. [12] got data from mammalian proteins. Obviously there are great genetic differences between prokaryotes and eukaryotes. So, it is rational to deduce that codons in eukaryotes carry more structural information than codons in prokaryotes. It is interesting that GGU prefers beta-strand in Table 1 ($Z(\text{beta-strand}\mid\text{GGU}) = 2.63$), but prefers alpha-helix in Table 2 ($Z(\text{alpha-helix}\mid\text{GGU}) = 2.48$). So, even if there are a few codons in *E. coli* that carry structural information, their preferences may be different from those of mammal's. Adzhubei et al. [12] suggested that in order to achieve a high level of expression of active protein, secondary structural preferences of codons must be considered in addition to species-specific synonymous codon usage optimization. It is right for those translation systems whose codons show these preferences. But it seems that the *E. coli* translation system lacks this mechanism. Brunak et al. [11] built a mixed database including both prokaryotes and eukaryotes, but the data were not shown in their paper. So a comparison cannot be made with their results. The difference between the conclusion of this study and that of Thanaraj et al. can be attributed to two reasons. First, Thanaraj et al. [9] used fractional frequency values as the measure of the translation rate, and hence the structural biases [19] of amino acids were not considered. Second, the scales of measuring codon frequencies in the two studies are also different in that they use tricodon fractional frequency values but our scale is single codon frequency. Zhang et al. [20] pointed out that the rare codons that are tightly clustered should have a much greater impact on ribosome movement than codons that are widely separated.

Research on prion-like (Ure2p) protein [21] suggested that the nature of the translation process can affect the intracellular folding pathway. It is a good model for further study on the relationship between synonymous codon usage and protein structure in vivo. Brunak et al. [11] found that a codon-based network does not give a better performance than the amino acid networks. From these results, it seems that at least there is not a universal codon-structure dictionary. Since so many factors affect synonymous codon usage and translation rate [22–24], much more work needs to be done in theory and in experiment to answer the question: 'Is the 3-dimensional

Table 2
Statistical results of the relationship between synonymous codons and protein secondary structure in mammalian (original data is taken from Adzhubei et al. [12])

Amino acid	Codon	$N(\mathrm{H}\mid\mathrm{sc})$	$P(\mathrm{H}\mid\mathrm{sc})$	$Z(\mathrm{H}\mid\mathrm{sc})$	$N(\mathrm{B}\mid\mathrm{sc})$	$P(\mathrm{B}\mid\mathrm{sc})$	$Z(\mathrm{B}\mid\mathrm{sc})$	$N(\mathrm{C}\mid\mathrm{sc})$	$P(\mathrm{C}\mid\mathrm{sc})$	$Z(\mathrm{C}\mid\mathrm{sc})$	$N(\mathrm{sc})$	$P(\mathrm{sc})$
Leu	UUG	75	0.31	−0.85	80	0.33	−0.91	90	0.37	1.70	245	0.13
	UUA	34	0.35	0.36	31	0.32	−0.65	33	0.34	0.36	98	0.05
	CUU	71	0.35	0.72	56	0.27	−2.43*	77	0.38	1.70	204	0.11
	CUC	112	0.28	−2.44*	164	0.41	2.82*	124	0.31	−0.54	400	0.21
	CUA	43	0.37	1.17	48	0.42	1.65	24	0.21	−2.61*	115	0.06
	CUG	298	0.35	1.25	295	0.34	−0.70	270	0.31	−0.57	863	0.45
Val	GUU	75	0.28	1.95	105	0.40	−3.12*	84	0.32	1.49	264	0.17
	GUC	83	0.21	−1.45	204	0.51	1.11	113	0.28	0.10	400	0.26
	GUA	21	0.16	−1.87	61	0.47	−0.23	47	0.36	2.18	129	0.08
	GUG	184	0.25	0.94	375	0.50	1.40	192	0.26	−2.41*	751	0.49
Gly	GGU	42	0.16	2.48*	37	0.14	−2.21	186	0.70	0.27	265	0.16
	GGC	75	0.12	0.54	133	0.21	1.89	413	0.67	−1.91	621	0.38
	GGA	38	0.10	−1.46	66	0.17	−1.23	287	0.73	1.95	391	0.24
	GGG	38	0.10	−1.08	76	0.20	0.89	257	0.69	−0.05	371	0.23
Pro	CCU	35	0.12	1.43	41	0.15	0.42	204	0.73	−1.39	280	0.28
	CCC	28	0.08	−2.24	66	0.18	2.72*	276	0.75	−0.65	370	0.37
	CCA	30	0.12	0.80	22	0.09	−2.74*	201	0.79	1.67	253	0.25
	CCG	11	0.11	0.15	11	0.11	−1.01	79	0.78	0.76	101	0.10
Glu	GAA	196	0.33	−0.31	126	0.21	−2.18	274	0.46	2.14	596	0.41
	GAG	290	0.34	0.31	225	0.26	2.18	349	0.40	−2.14	864	0.59
Phe	UUU	100	0.25	0.45	152	0.38	−2.44*	143	0.36	2.20	395	0.41
	UUC	137	0.24	−0.45	261	0.46	2.45*	167	0.30	−2.20	565	0.59
Ile	AUU	103	0.27	−0.32	170	0.45	−0.54	105	0.28	0.93	378	0.33
	AUC	168	0.27	−0.51	305	0.49	2.61*	148	0.24	−2.23	621	0.55
	AUA	44	0.32	1.31	48	0.35	−3.04*	45	0.33	1.91	137	0.12
Ser	UCU	51	0.20	0.86	52	0.21	−1.70	150	0.59	0.81	253	0.17
	UCC	72	0.19	0.58	110	0.29	2.66*	192	0.51	−2.89*	374	0.26
	UCA	21	0.13	−1.70	39	0.25	0.26	96	0.62	1.07	156	0.11
	UCG	12	0.17	−0.45	25	0.35	2.16	35	0.49	−1.58	72	0.05
	AGU	42	0.20	0.59	41	0.19	−1.80	130	0.61	1.12	213	0.15
	AGC	69	0.18	−0.41	89	0.23	−1.07	236	0.60	1.23	394	0.27
Thr	ACU	72	0.24	2.01	80	0.27	−2.92*	149	0.50	1.33	301	0.23
	ACC	106	0.18	−1.53	223	0.38	2.57*	257	0.44	−1.22	586	0.44
	ACA	64	0.20	0.02	106	0.33	−0.65	150	0.47	0.54	320	0.24
	ACG	23	0.19	−0.18	45	0.38	0.91	51	0.43	−0.65	119	0.09

H, alpha-helix; B, beta-sheet; C, coil; *, rare codons, $|Z(\mathrm{ss}\mid\mathrm{sc})| > 2.33$.

structure of a protein related to the specific coding region in its mRNA?'

Acknowledgements: We gratefully acknowledge Drs. Tang Haixu and A.A. Adzhubei for helpful comments on this work.

References

[1] Nakamura, Y., Gojobori, T. and Ikemura, T. (1998) Nucleic Acids Res. 26, 334.

[2] Ikemura, T. (1985) Mol. Biol. Evol. 2, 13–34.

[3] Sorensen, M., Kurland, C. and Pederson, S. (1989) J. Mol. Biol. 207, 365–377.

[4] Li, H. and Luo, L.F. (1996) J. Theor. Biol. 181, 111–124.

[5] Guisez, Y., Robbens, J., Remaut, E. and Fiers, W. (1993) J. Theor. Biol. 162, 243–252.

[6] Taylor, F. and Coates, D. (1989) Biosystems 22, 177–187.

[7] Volkenstein, M. (1966) Biochim. Biophys. Acta 119, 421–424.

[8] Siemion, I. and Siemion, P. (1994) Biosystems 33, 139–148.

[9] Thanaraj, T.A. and Argos, P. (1996) Protein Sci. 5, 1973–1983.

[10] Thanaraj, T.A. and Argos, P. (1996) Protein Sci. 5, 1594–1612.

[11] Brunak, S. and Engelbrecht, J. (1996) Proteins 25, 237–252.

[12] Adzhubei, A.A., Adzhubei, I.A., Krasheninnikov, I.A. and Neidle, S. (1996) FEBS Lett. 399, 78–82.

[13] Adzhubei, I.A., Adzhubei, A.A. and Neidle, S. (1998) Nucleic Acids Res. 26, 327–331.

[14] Bernstein, F.C., Lake, J.A., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.

[15] Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. and Cameron, G.N. (1993) Nucleic Acids Res. 21, 2967–2971.

[16] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.

[17] Freund, J.E. and Walpole, R.E. (1981) Mathematical Statistics, 3rd Edn., Prentice Hall.

[18] Sharp, P.M. and Li, W.H. (1986) Nucleic Acids Res. 14, 7737–7749.

[19] Chou, P.Y. and Fasman, G.D. (1978) Annu. Rev. Biochem. 47, 251–276.

[20] Zhang, S., Goldman, E. and Zubay, G. (1994) J. Theor. Biol. 170, 339–354.

[21] Komara, A.A., Lesnika, T., Cullina, C., Guillemet, E., Ehrlich, R. and Reiss, C. (1997) FEBS Lett. 415, 6–10.

[22] Karlin, S. and Mrazek, J. (1996) J. Mol. Biol. 262, 459–472.

[23] Solomovici, J., Lesnik, T. and Reiss, C. (1997) J. Theor. Biol. 185, 511–521.

[24] Osawa, S. (1995) Evolution of the Genetic Code, Oxford University Press, Oxford.