Protein Science (1996), *5*:1973–1983. Cambridge University Press. Printed in the USA.

Protein secondary structural types are differentially coded on messenger RNA

T.A. THANARAJ[1,2] AND PATRICK ARGOS[1]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-69012, Heidelberg, Germany
[2]Centre for Cellular & Molecular Biology, Hyderabad, India

(RECEIVED June 18, 1996; ACCEPTED July 18, 1996)

Abstract

Tricodon regions on messenger RNAs corresponding to a set of proteins from *Escherichia coli* were scrutinized for their translation speed. The fractional frequency values of the individual codons as they occur in mRNAs of highly expressed genes from *Escherichia coli* were taken as an indicative measure of the translation speed. The tricodons were classified by the sum of the frequency values of the constituent codons. Examination of the conformation of the encoded amino acid residues in the corresponding protein tertiary structures revealed a correlation between codon usage in mRNA and topological features of the encoded proteins. Alpha helices on proteins tend to be preferentially coded by translationally fast mRNA regions while the slow segments often code for beta strands and coil regions. Fast regions correspondingly avoid coding for beta strands and coil regions while the slow regions similarly move away from encoding alpha helices. Structural and mechanistic aspects of the ribosome peptide channel support the relevance of sequence fragment translation and subsequent conformation. A discussion is presented relating the observation to the reported kinetic data on the formation and stabilization of protein secondary structural types during protein folding. The observed absence of such strong positive selection for codons in non-highly expressed genes is compatible with existing theories that mutation pressure may well dominate codon selection in non-highly expressed genes.

Keywords: alpha helices; beta strands; coils; codon usage; protein folding; translation speed

Coding nucleotide sequences carry an integral message containing several different types of information for the various molecular mechanisms involved in gene expression (Kypr, 1986; Trifonov, 1989). Degeneracy in the genetic code allows an additional potential for messenger RNA to carry structural information regarding the encoded protein that can be at the level of a single codon or at a contiguous nucleotide region (Brunak et al., 1994). Existence of a correspondence between the physico-chemical characteristics of the coded amino acids and the nucleotide composition of the corresponding codons has been discussed in the literature (Volkenstein, 1966; Woese et al., 1966; Taylor and Coates, 1989; Siemion and Siemion, 1994). Translational constraints appear sufficient to affect the global amino acid composition of proteins (Lobry and Gautier, 1994). The first, second, and third base of the codon have been respectively connected to the biosynthetic pathway (Taylor and Coates, 1989), residue hydrophobicity (Volkenstein, 1966; Woese et al., 1966), and the helix or beta strand-forming potential (Siemion and Siemion, 1994) of the coded amino acid. Neural network algorithms for protein secondary structure prediction based on mRNA sequence rivaled those relying on amino acid representation (Brunak et al., 1994). It has been suggested that protein folding and mRNA sequence patterns may be correlated by the differential translation speed caused by the choice of codons (Sharp et al., 1986; Kypr and Mrazek, 1987; Liljenstrom and von Heijne, 1987; Purvis et al., 1987; Candelas et al., 1989; McNally et al., 1989; Brunak et al., 1994; Lobry and Gautier, 1994). Rare codon clusters have been found at the boundaries of polypeptide chain fragments of the same secondary structure type during cotranslational protein folding (Krasheninnikov et al., 1989, 1991). We showed in a previous study (Thanaraj and Argos, 1996) relying on codon usage frequencies and cellular availability of cognate transfer RNAs that protein domain boundaries are largely encoded by translationally slow mRNA regions; such a hypothesis also finds support in reported mutagenic data. In the present study, using the same codon frequency measures (referred to as *codfreqsum* in [Thanaraj and Argos, 1996]) as indices of translation speed, the translation rates of mRNA regions corresponding to helix, beta strands, and loops on proteins were analyzed. The results show that

Reprint requests to: T.A. Thanaraj, EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England; e-mail: thanaraj@ebi.ac.uk.

the protein secondary structural types are differentially coded on mRNA.

Results and discussion

Rationale

The rate at which an mRNA region is translated by the ribosome depends on several parameters such as codon usage, codon–anticodon interactions (Grosjean and Fiers, 1982), codon context (Lipman and Wilbur, 1983; Yarus and Folley, 1985; Shpaer, 1986; Varenne et al., 1989), and the secondary structure of mRNA regions (Chaney and Morris, 1979; Tu et al., 1992). The supposition connecting codon usage and translation rate is supported by various lines of published experimental evidence, as will be subsequently described. Theoretical models have also been developed to explain the so-called ribosome 'queue' caused by clusters of low-usage codons and the overall effect on translation efficiency (Zhang et al., 1994) when such codons are located at different regions on the mRNA. Modulation of ribosome 'traffic' on mRNA by substituting a string of rare or abundant synonymous codons in *E. coli bla* and *ompA* genes has been used to explain changes in RNase activity on mRNA and subsequent effects on the levels of mRNA present (Deana et al., 1996).

Cellular levels of the isoaccepting tRNA molecules for amino acid types are not identical (Ikemura, 1985), and it has been suggested that these differences control the relative translational rate of the cognate synonymous codons (Ikemura and Ozeki, 1983; Ikemura, 1985; Sorensen et al., 1989; Sorensen and Pedersen, 1991). The mRNAs of highly expressed genes, where the encoded proteins are synthesized abundantly, preferentially use a subset of codons that promotes the rate of translation (Ikemura and Ozeki, 1983; Ikemura, 1985; Sharp and Li, 1987; Andersson and Kurland, 1990; Lobry and Gautier, 1994). Some codons are translated more slowly than others, and abundant codons are mostly translated at a higher rate (Bonekamp et al., 1985; Carter et al., 1986; Harms and Umbarger, 1987; Curran and Yarus, 1989; Sorensen et al., 1989; Sorensen and Pedersen, 1991).

Optimization has been suggested also in the selection of amino acid types appearing in abundant proteins (Lobry and Gautier, 1994); highly expressed genes show a bias towards amino acids having abundant major tRNAs. In order to assess the correspondence between the sum of frequencies of the synonymous codons for individual amino acids and the sum of the cellular contents of isoaccepting tRNAs (Ikemura, 1985), the frequencies of the 61 sense codons for the 20 amino acids as they occur in the mRNAs of highly expressed genes from *Escherichia coli* were examined using a data base consisting of 9512 codons from 60 mRNA sequences of highly expressed genes (Sharp and Li, 1986, 1987). Regression to a linear function for the amino acids val, gly, ile, lys, glu, asp, gln, asn, tyr, his, and phe (for others tRNA content has not been measured) resulted in a correlation coefficient of 0.94, with a standard error of 0.008. The total amount of tRNA for a particular amino acid has been experimentally shown to parallel the total usage of that amino acid in proteins for *Escherichia coli* and *Mycoplasma capricolum* (Yamao et al., 1991). Given this strong correspondence as well as observations on correlation between the differential translation speed of synonymous codons and the cellular content of cognate tRNAs, the codon frequency values given in Table 1 were used to quantify the slowness in translated mRNA regions. Such values reflect both the synonymous codon and amino acid frequencies as found in highly expressed genes.

Distribution of observed tricodons with translation speed

The *codfreqsum* value, a measure of the speed of translation (see Materials and methods), was calculated for tricodons starting at each codon of the mRNAs from a data set containing 40 proteins with known tertiary structure (Table 2). We had shown earlier from the same *codfreqsum* values (Thanaraj and Argos, 1996) that codon usage in slow regions was consistent with the experimentally determined translation rates. The observed minimum *codfreqsum* was found to be 0.0004, while 0.1833 represented the maximum. Tricodons whose *codfreqsum* values occurred within an interval of 0.01 were grouped. The distribution of the observed number of tricodons versus *codfreqsum* values in the range of 0.000 to 0.1900 with plotted intervals 0.01 showed a gaussian profile (Fig. 1); the mean (m) and standard deviation (σ) were, respectively, 0.088 and 0.035.

Differential coding of protein structural types

The tricodon regions on the mRNA were grouped into four classes, depending on their *codfreqsum* values: class 1 = tricodons with *codfreqsum* in the range of 0 to ($m - \sigma$); class 2 = tricodons with *codfreqsum* in the range of ($m - \sigma$) to m; class 3 = tricodons with *codfreqsum* in the range of m to ($m + \sigma$); and class 4 = tricodons with *codfreqsum* in the range of ($m + \sigma$) to the maximum value.

Each range of *codfreqsum* values in incremental steps of 0.005 were scrutinized for the conformation of the encoded residues in the tertiary structure of the corresponding protein. Overlapping tricodons with *codfreqsum* values in a given range were joined. The preference/avoidance (observed minus expected as a number of standard deviations (*sd*) as described in Materials and methods) of helix and beta strand and coil in the corresponding protein spans over the various increments were calculated (Table 3A and Fig. 2).

It can be seen from Figure 2 that in moving from translationally slow to fast regions, helices (shown by red lines) are preferred on mRNA by fast regions and avoided by slow regions; beta strands (shown by black lines) show an opposite trend with preference for slow regions and avoidance at fast regions; coil (shown by green lines) displays an unbiased behavior, although there is a tendency

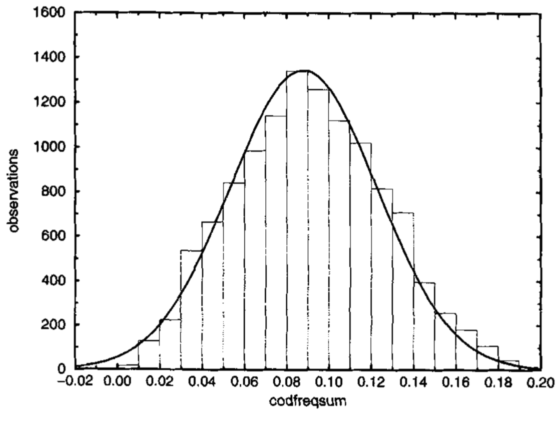


Fig. 1. Distribution of the observed number of tricodons against *codfreqsum* values follows a Gaussian profile.

Table 1. *Codon fractional frequency values as calculated from mRNAs of highly expressed genes from Escherichia coli*

Amino acid	Codon	Frequency value[a]	Amino acid	Codon	Frequency value	Amino acid	Codon	Frequency value
met	aug	.0197	pro	ccu	.0038	ser	agu	0012
trp	ugg	.0061		ccc	.0003		agc	.0091
				cca	.0038		ucu	.0169
phe	uuu	.0069		ccg	.0264		ucc	.0126
	uuc	.0260					uca	.0009
tyr	uau	.0055	thr	acu	.0253		ucg	.0005
	uac	.0192		acc	.0269			
cys	ugu	.0016		aca	.0011	arg	aga	.0001
	ugc	.0030		acg	.0034		agg	.0000
his	cau	.0044					cgu	.0488
	cac	.0138	ala	gcu	.0470		cgc	.0187
gln	caa	.0046		gcc	.0080		cga	.0003
	cag	.0304		gca	.0272		cgg	.0001
asn	aau	.0032		gcg	.0201			
	aac	.0371				leu	uua	.0005
lys	aaa	.0600	val	guu	.0499		uug	.0026
	aag	.0191		guc	.0063		cuu	.0029
asp	gau	.0185		gua	.0227		cuc	.0023
	gac	.0385		gug	.0120		cua	.0003
glu	gaa	.0542					cug	.0611
	gag	.0155	gly	ggu	.0536			
				ggc	.0354			
ile	auu	.0108		gga	.0005			
	auc	.0475		ggg	.0015			
	aua	.0000						

[a] Calculation of these values is described in detail in Thanaraj and Argos (1996).

towards avoidance in fast regions. A deviation from an expected value by ±3 *sd* was considered significant. Further classification of the extreme regions in the Gaussian profile (in terms of 2σ) yielded similar results (Table 3B). The slowest regions, with *codfreqsum* in the range of 0 to $(m - 2\sigma)$, showed avoidance of helices by a statistical significance of -5.4 *sd* and preference for beta sheet by 4.2 *sd* , both very significant. The fastest regions, with *codfreqsum* in the range of $(m + 2\sigma)$ to the maximum value, showed a clear preference for helix (4.4 *sd*) and avoidance of coils (-4.5 *sd*) . The avoidance of beta sheet is shown by fast regions with the *codfreqsum* in the range of $(m + \sigma)$ to $(m + 2\sigma)$ with a statistical significance of -3.7. Further restriction of slow regions to a range of *codfreqsum* from 0.00 to 0.01 revealed the preference for coil in the slowest regions (3.0 *sd*) and avoidance of helix by -3.3 *sd.*

To examine the effect of protein size, different data sets were constructed from the original 54 entries based on the 1σ profile: (i) all 54 entries, (ii) protein entries whose individual secondary structural element varied from the average of such proteins by no more than 25% for each of the three types (Fig. 3); and (iii) protein entries whose individual secondary structural content varied by no more than 15% from the mean of such proteins. The overall pattern of preference/avoidance for protein structural types remained nearly the same irrespective of the data set used (see Fig. 3 for exemplary case ii). However, the preference of coil by slow regions is not shown emphatically by any of the data sets. To probe

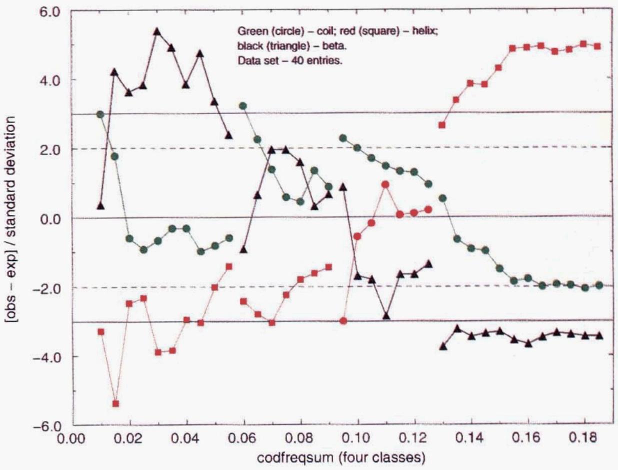


Fig. 2. Preference/avoidance pattern for the protein secondary structural types in the data set of 40 entries (see Table 2). The values of [observed-expected] expressed in the units of standard deviations (see Materials and methods) for each of the three structural types is shown against the *codfreqsum* values of the corresponding tricodon spans on the mRNA. The *codfreqsum* values are divided into four classes (see text). As the *codfreqsum* values increase, so does the effective mRNA translation speed. Green lines with circles correspond to coil; red lines with squares to helix; and black lines with triangles to beta. Base lines are given for the 2 and 3 *sd* levels.

Table 2. *List of the protein structures used in the study*

PDB/EMBL identifiers[a]	Percentage of secondary structural content Coil	Helix	Beta	CAI value[b]
1. 1emd/ECMDH1	34.3	47.8	17.9	0.592
2. 1dra-A/ECFOLX	43.4	24.5	32.1	0.412
3. 1rnr-A/ECNRDA[c]	26.2	69.7	4.1	0.542
4. 1trb/ECTRXB	42.9	28.9	28.3	0.473
5. 1tre-A/ECTPI	35.8	48.8	15.4	0.781
6. 3icd/ECICD	41.8	40.1	18.1	0.592
7. 1dsb-A/ECDSF	36.6	52.2	11.3	0.614
8. 1ltp-L/ECLACI	46.9	39.7	13.4	0.320
9. 1aam/ECASPC[c]	58.1	33.1	8.8	0.503
10. 1acm-A/ECPYRBIA	41.3	43.2	15.5	0.428
11. 1acm-B/ECPYRBIA[c]	48.6	15.8	35.6	0.400
12. 1ake-A/ECADK	34.1	48.6	17.3	0.697
13. 1gla-F/ECPTSHI[c]	50.3	11.2	38.5	0.663
14. 1poh/ECPTSHI	38.8	34.1	27.1	0.696
15. 1gla-G/ECGLYK	41.7	34.8	23.5	0.500
16. 1pfk-A/ECPFKA	34.4	46.9	18.8	0.691
17. 1cdd-A/ECPURMN	45.5	33.3	21.2	0.393
18. 3tms/ECTHYA	36.7	36.7	26.5	0.469
19. 1kfd/ECPOLA	32.9	53.0	14.1	0.423
20. 2pol-A/ECDNAAN[c]	32.0	21.6	46.4	0.476
21. 1abe/ECARAFGH	30.5	47.5	22.0	0.489
22. 2dri/ECRBSP	32.5	45.0	22.5	0.505
23. 2liv/ECLIVHMG	35.4	44.7	19.9	0.498
24. 1cma-A/ECMETJA[c]	44.2	45.2	10.6	0.427
25. 1wrp-R/ECTRPR1[c]	23.5	76.5	0.0	0.294
26. 1fia-A/ECFIS[c]	31.6	65.8	2.5	0.507
27. 1cgp-A/ECCRP	32.5	40.1	27.4	0.529
28. 1bia/ECBIRA	38.4	32.5	29.1	0.268
29. 2reb/ECRECA	31.0	43.9	25.1	0.636
30. 2abk/ECNTH[c]	39.3	60.7	0.0	0.376
31. 1goa/ECQRNH	37.8	34.0	28.2	0.404
32. 1mat/ECMAP	45.6	26.6	27.8	0.499
33. 1pda/ECHEMC	39.2	36.8	24.0	0.350
34. 3eca-A/ECANSBA	43.8	32.7	23.5	0.558
35. 1alk-A/ECPHOAA	47.2	31.7	21.1	0.371
36. 2glt/ECGSHII	34.8	35.1	30.1	0.465
37. 3chy/ECCHEY	37.5	45.3	17.2	0.350
38. 2trx-A/ECRHOA	38.7	33.0	28.3	0.628
39. 256b-A/ECCYBC[c]	20.8	79.2	0.0	0.491
40. 1ctf/ECRPOBC[c]	17.6	55.9	26.5	0.842
41. 1etu/ECTGTUFB	35.6	44.1	20.3	0.806
42. 2omf/ECOMPF[c]	36.2	4.4	59.4	0.689
43. 1pho/ECPHOE[c]	41.5	2.1	56.4	0.386
44. 1pii/ECTRPC	41.6	38.3	20.1	0.337
45. 1dmb/ECUW89	38.6	42.7	18.6	0.586
46. 2rsl-A/ECTN1000	28.9	51.8	19.3	0.194
47. 1eip-A/ECPPA1	50.3	21.3	28.4	0.733
48. 1bmt-A/ECMETH[c]	34.6	56.5	8.9	0.435
49. 1hjr-A/ECRUVC	34.2	41.8	24.1	0.403
50. 1ger-B/ECGOR	37.6	34.9	27.5	0.487
51. 1qor-A/ECDNABA	40.2	34.7	25.2	0.335
52. 1scu-A/ECGLTA01	40.3	37.2	22.6	0.657
53. 1scu-B/ECGLTA02	34.5	41.8	23.7	0.633
54. 1isa-A/ECSODB	37.5	50.5	12.0	0.603
Average	38.4	38.9	22.7	

[a] PDB 4-character file names are followed by chain identifiers.
[b] Codon adaptation index as defined in Materials and methods.
[c] The coil, helix, or beta content exceeds that of the average by 50% or more.

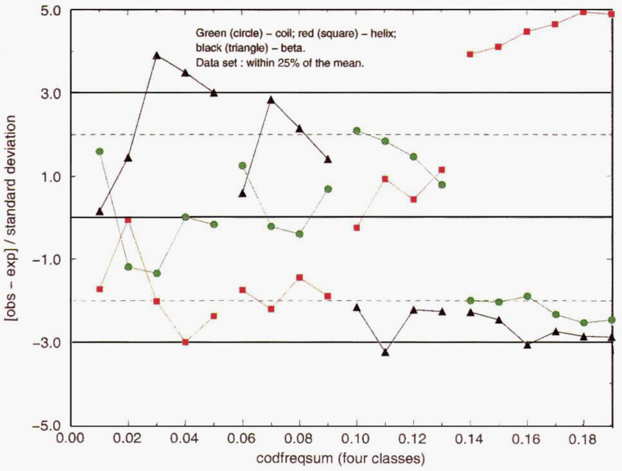


Fig. 3. Preference/avoidance pattern for the protein secondary structural types in the data set containing only those proteins whose coil, helical, or beta content do not deviate by more than 25% from the respective averages for coil, helical, or beta content of the data set. The conditions of Figure 2 also apply here.

sensitivity for structural biases in translation speed, a hypothetical data set was constructed containing only the mRNA regions that code for coil in those proteins whose coil content is not less than the average over the entire 54-protein data set; data sets for mRNA regions encoding helix and beta structures were also similarly constructed. Tricodons from such mRNA regions were classified according to their *codfreqsum* values and statistics were performed with the expected frequencies determined from the sum of helix, coil, and beta content as coded by all the above mRNA regions. The results of such an analysis are shown in Figure 4. The preference/avoidance patterns for all the three protein secondary structural types are significantly maintained. The preference/avoidance pattern of coil regions is now resolved such that coils prefer slowly translated mRNA segments. The smooth transition

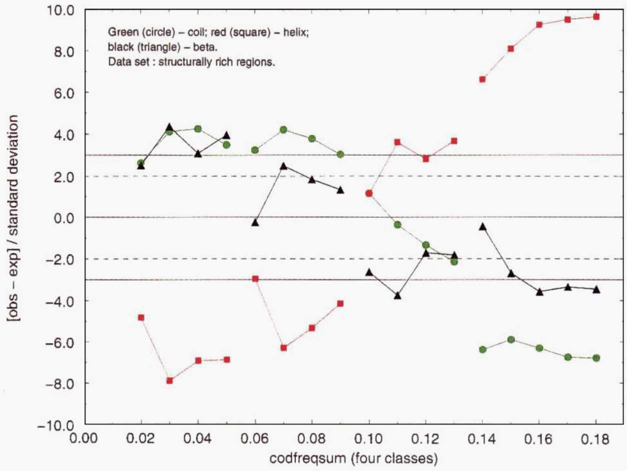


Fig. 4. Preference/avoidance pattern for the protein secondary structural types in the data set of proteins whose respective coil, helical and beta content is equal to or more than that of the average found in the 54-protein data set. The conditions of illustration are the same as those of Figure 2.

Table 3. *Correlation of regions from gaussian distribution with protein structure*

codfreqsum range[a]	No. of participating residues[b]	Preference/avoidance of structural elements expressed in standard deviations (*sd*)[c]		
		Coil	Helix	Beta
A. Classification based on 1σ profile				
1. Range: 0 to ($m - \sigma$)				
0.000 to 0.00	535	—	—	—
0.000 to 0.010	135	2.99	−3.28	0.37
0.000 to 0.015	306	1.79	−5.38	4.22
0.000 to 0.020	512	−0.60	−2.47	3.62
0.000 to 0.025	829	−0.92	−2.32	3.82
0.000 to 0.030	1220	−0.67	−3.90	5.39
0.000 to 0.035	1758	−0.32	−3.85	4.91
0.000 to 0.040	2277	−0.31	−2.95	3.84
0.000 to 0.045	2827	−0.98	−3.04	4.74
0.000 to 0.050	3442	−0.82	−2.02	3.35
0.000 to 0.055	4038	−0.59	−1.42	2.38
2. Range: ($m - \sigma$) to m				
0.055 to 0.060	1202	3.22	−2.42	−0.92
0.055 to 0.065	2372	2.25	−2.79	0.65
0.055 to 0.070	3529	1.39	−3.04	1.96
0.055 to 0.075	4495	0.59	−2.24	1.95
0.055 to 0.080	5493	0.45	−1.80	1.60
0.055 to 0.085	6483	1.35	−1.62	0.32
0.055 to 0.090	7221	0.89	−1.45	0.67
3. Range: m to ($m + \sigma$)				
0.090 to 0.095	1659	2.28	−3.00	0.88
0.090 to 0.100	2885	2.00	−0.57	−1.68
0.090 to 0.105	3966	1.71	−0.18	−1.79
0.090 to 0.110	4808	1.49	0.93	−2.84
0.090 to 0.115	5542	1.33	0.06	−1.64
0.090 to 0.120	6030	1.30	0.10	−1.64
0.090 to 0.125	6454	0.95	0.20	−1.35
4. Range: ($m + \sigma$) to maximum value				
0.125 to 0.130	930	0.54	2.63	−3.74
0.125 to 0.135	1674	−0.64	3.36	−3.22
0.125 to 0.140	2058	−0.91	3.83	−3.45
0.125 to 0.145	2386	−0.97	3.80	−3.35
0.125 to 0.150	2634	−1.49	4.28	−3.30
0.125 to 0.155	2727	−1.85	4.84	−3.55
0.125 to 0.160	2823	−1.77	4.87	−3.67
0.125 to 0.165	2878	−1.99	4.91	−3.46
0.125 to 0.170	2922	−1.93	4.74	−3.33
0.125 to 0.175	2932	−1.96	4.81	−3.38
0.125 to 0.180	2939	−2.06	4.96	−3.44
0.125 to 0.185	2944	−1.98	4.88	−3.44
B. Classification based on 2σ profile				
1. Range: 0 to ($m - 2\sigma$)				
0.000 to 0.015	306	1.79	−5.38	4.21
2. Range: ($m - 2\sigma$) to ($m - \sigma$)				
0.015 to 0.050	3357	−0.95	−1.74	3.17
3. Range: ($m + \sigma$) to ($m + 2\sigma$)				
0.125 to 0.160	2823	−1.77	4.87	−3.67
4. Range: ($m + 2\sigma$) to maximum value				
0.160 to 0.185	456	−4.53	4.39	0.13

[a] The conformation of residues encoded by mRNA tricodon regions with *codfreqsum* values in the given range. (The symbols m and σ refer to the mean and standard deviation of the gaussian profile of Figure 1 (see text for details)).

[b] Total number of residues encoded by mRNA tricodon regions with *codfreqsum* values in the given range.

[c] Statistical significance of the preference/avoidance of secondary structural types assumed by residues coded by tricodon mRNA regions expressed in number of standard deviations, *sd* (see Materials and methods).

from preference to avoidance on either side of the mean value of the gaussian profile is clearly seen.

Folding of the nascent chain in the peptide channel

The recently published structure of the *E. coli* ribosome using cryo-electron microscopy (Frank et al., 1995; Stark et al., 1995) shows that the peptide channel contains solvent and bifurcates toward a site known to bind to cytoplasmic membrane forming part of the secretory pathway and toward the cytoplasm. The main pathway has a linear length of 85 Å and a diameter of about 20 Å. Because the average helix diameter and beta sheet thickness is almost 10 Å (Robson and Garnier, 1986), and because the radius of gyration of a single protein chain averages 16 ± 4 Å (calculated from a data set of soluble proteins), smaller folding units have sufficient space to adopt conformation in the solvent-filled ribosome channel.

Protease protection studies (Blobel and Sebatini, 1970), antibody binding studies (Bernabeu and Lake, 1982), and crosslinking studies (Brimacombe, 1995; Stade et al., 1995) indicate that 30–40 residues of the nascent chain are protected inside the peptide channel. Given that the rise per helical residue is 1.5 Å and the maximal separation of successive main chain C_α atoms in extended conformation is 3.3 Å (Robson and Garnier, 1986), residues in a 40-peptide nascent chain could adopt various helical, extended, and coil conformations in the 85 Å long ribosome channel. A 40-residue segment in an unfolded or fully extended conformation would require a length of about 144 Å. Other supporting observations for nascent chain structure arise from crosslinking studies in the large ribosomal subunit (Brimacombe, 1995; Stade et al., 1995) where progressively longer peptides were found to crosslink to similar ribosomal RNA sites as shorter peptides, implying a peptide channel sufficiently wide to allow the nascent chain to fold back upon itself. Further, stereochemical analysis of the transpeptidation event indicate that the ribosome can generate a regular alpha helical conformation at the carboxyl end of the nascent peptide (Lim and Spirin, 1986). Further, Hardesty and co-workers (Kudlicki et al., 1995) observe that reactions directly related to the folding of the nascent peptide can influence ribosomal peptidation and that a significant part of protein folding process takes place within the 50S subunit (Hardesty et al., 1993). An analogous channel that occurs in *in vivo* protein folding is that of the central channel of the 'folding cage' of the chaperonin family that accommodates a partially folded polypeptide in its central cavity and promotes folding of the peptide (Hartl et al., 1994; Sigler and Horwich, 1995; Hunt et al., 1996).

Kinetic data on protein folding

Hydrogen bonding interactions involved in helix formation are local and extend over only four residues for each turn of a helix. They can form within a low millisecond time scale when unfolded proteins are transferred into refolding conditions (Roder et al., 1988). Alpha helices often form before other levels of protein structure (Anfinsen, 1973; Ptitsyn, 1995). Williams et al. (1996) infer a folding rate constant of $6 \times 10^7\ s^{-1}$ ($t_{1/2} = 16$ ns) for helix formation in a small 21-residue alanine peptide and further indicated that such a nanosecond time scale of helix formation is at least three orders of magnitude faster than the characteristic time scale of intramolecular tertiary contact formation. The folding pathway may well involve states in which part of a native fold has been formed in a compact core, while other parts of the protein remain disordered on quite long time scales (Doniach et al., 1995) and indications are that such a compact domain can be largely helical as in the cases of lysozyme (Buck et al., 1993, 1994) and lactalbumin (Chyan et al., 1993). While a major fraction of partially folded states of cytochrome *c*, a helix protein, fold rapidly to the native state on a 15 ms time scale (Sosnick et al., 1994), only a minor fraction of the states of lysozyme, a helix + beta protein, fold to the native state in ~20 ms and certainly have not formed within ~2 ms (Dobson et al., 1994) in contrast to the cytochrome *c* results. Thus, there is at least a 10-fold difference between the folding rates of a helix + beta and a helix protein. Studies on the partially folded state of hen egg white lysozyme, which contains two beta sheets and five helices, in trifluoroethanol indicated significant protection from proton/deuterium exchange for 25 backbone amides, the majority of which are located in helical regions of the protein (Buck et al., 1993). Experimental observations that alpha helical intermediates can accumulate during refolding of some beta strand dominated proteins (Thomas and Dill, 1993; Liu et al., 1994; Logan et al., 1994; Shiraki et al., 1995) suggest that helices can be formed rapidly (Carlsson and Jonsson, 1995), albeit not all. Such a suggestion is substantiated by the observation that the alpha helical content of 23 different proteins in their trifluoroethanol-induced state (which corresponds to the stable molten globule state [Buck et al., 1993]) showed better correlation with the predicted than the native state (Shiraki et al., 1995). No significantly increased density of intrahelical sidechain–sidechain contacts were observed for residues separated by more than four residues in helices (Walther and Argos, 1996). Narayana and Argos (1984) observed that a residue in helical conformation can find 53.3% of the total number of atoms interacting with it in the ±4 residue flanks, while a residue in beta conformation completes only 32.1% of its interactions with the same flanking residues.

While helices often utilize local interactions, formation of beta sheet involves sequentially distant parts of the protein. The formation time for the growth of helices has been estimated to be as small as 10^{-8} seconds (Williams et al., 1996), while beta structures may require as long as 10^{-2} seconds (Zana, 1975). However, recent refolding experiments from denatured polypeptides of ribonuclease A (Udgaonkar and Baldwin, 1990), ribonuclease T1 (Mullins et al., 1993), ubiquitin (Briggs and Roder, 1992), lysozyme (Dobson et al., 1994), and staphylococcal nuclease (Jacobs and Fox, 1994) suggested that beta sheets can be formed at a rate comparable to that of alpha helices, and sometimes the residues in beta sheet are protected from hydrogen/deuterium exchange even before protection is observed in helices (Jacobs and Fox, 1994). However, the nuclease structures depend on several disulfide bridges and the helical domain of lysozyme as well as lactalbumin (Chyan et al., 1993) folds first. Theoretical arguments have been put forward to the effect that beta structure may be formed rapidly only if it is stabilized (Finkelstein, 1991). The data of Udgaonkar and Baldwin (1990) supports this by showing that the beta sheet is only moderately stable when it is formed rapidly and only subsequent folding events stabilize it, possibly through interactions involving hydrophobic side chains. In the cases of all-beta proteins such as interleukin-1β (Varley et al., 1993) and SH3 (Src homology region 3 [Farrow et al., 1995]), stable beta-sheet formation takes about one second, indicating that the structural context is important for the kinetics of beta-sheet folding, whereas helix can form locally close-packed structures (Yang and Honig, 1995). Fersht and co-workers (Prat Gay et al., 1995) find that the conformational pathway of the *in vitro* growing polypeptide chain of chymotrypsin inhibitor-2 parallels the protein folding pathway. They observe that molten globule-like states occur as intermediates in the elongation of the polypeptide chain. The structure observed in the fragments of different lengths pinpointed the requirements for forming a stable folding nucleus, which in the case of chymotrypsin inhibitor-2 contained helix while the rest of the structure (the beta-sheet) formed only weakly.

Suggestions have also been made that the stabilization of loop regions is a later (subsequent) event in protein folding. In the partially folded state of hen egg white lysozyme in trifluoroethanol (Buck et al., 1993), only little protection from deuterium/hydrogen exchanges could be observed for the amides located in a long loop while a majority of the observed protection occurred in helical regions. Experiments on the refolding reaction of ubiquitin (Briggs and Roder, 1992) indicated that while the amide protons from the beta sheet and the alpha helix as well as the protons involved in hydrogen bonding at the helix/sheet interface become 80% protected in an initial 8-ms folding phase, somewhat slower protection rates were observed for the residues from a surface loop. In the partially folded species of lactalbumin, native secondary structures are largely formed while the loop regions remain disordered (Feng et al., 1994; Redfield et al., 1994).

Correlation of the secondary structural preference/avoidance profile with protein folding

The role of ribosomal context in the folding of a protein has been discussed both for eukaryotes and prokaryotes (Phillips, 1966; Baldwin, 1975; Fedorov et al., 1992; Yonath, 1992; Tsalkova et al., 1993; Friguet et al., 1994; Kudlicki et al., 1994, 1995; Wiedmann et al., 1994; Tokatlidis et al., 1995). *In vivo* protein folding on the ribosome might be aided by kinetics of translation and also by interactions with the ribosome; such an involvement of the ribosome might limit the conformational space to be searched and act as a guide to the final native structure. Since in protein structural folding the formation and stabilization of helices probably occurs earlier when compared to beta sheets, the observation here that helices are avoided by slow regions and preferred by fast regions on mRNA and that the opposite is observed for beta strands is consistent. The observation that loops exhibit a similar avoidance/preference pattern as that of beta strands also points to general loop stabilization as a later event in protein folding.

Amino acid propensity in fast and slow regions

Different amino acids possess different intrinsic preferences for specific secondary structures, and such preferences are often understood on structural grounds. The results presented here indicate that peptide regions corresponding to translationally fast codons tend to adopt alpha helical conformation while those in slow regions are often in beta strand conformation. Such correlations arose from two independent parameters; namely, the codon frequency on the mRNA, which is coupled to translational efficiency, and the secondary structural conformational preferences of the corresponding residues in the encoded proteins. The independence of these two statistics is supported by correlation calculations. The propensities of the amino acids to be part of fast and slow translated regions were calculated respectively as $p_{f,i} = f_i/t_i$ and $p_{s,i} = s_i/t_i$, where t_i are the fractional frequencies of the 20 residue types in all regions of the 40-protein set, while f_i and s_i are, respectively, fractional frequencies in fast and slow regions. The propensities of amino acids to form alpha helix ($p_{h,i}$) or beta sheet ($p_{b,i}$) were taken from Palau et al. (1981), who studied many different protein structures from various species. Linear regression analysis on $p_{f,i}$ and $p_{h,i}$ and on $p_{s,i}$ and $p_{b,i}$ revealed low correlation coefficients of 0.24 (for fast regions and helix propensities) and 0.29 (for slow regions and beta sheet propensities). Thus, there is not a strong coupling for a residue to be preferred in a fast (slow) region and also to adopt a helical (strand) conformation in proteins generally.

Regulated speed of translation and optimization of co-translational folding efficiency

In the present work we observe that alpha helical regions are translated faster than beta sheet regions. Studies by others on refolding kinetics of denatured proteins indicated that alpha helices fold more quickly than beta sheets. Thus, it is concluded here that speed of translation is regulated to optimize co-translational folding efficiency. Such a conclusion gets support from the following observation. Komar and Jaenicke (1995) observed that N-terminal domain of YB-crystallin is translated faster than C-terminal domain, and such an observation agreed with the finding that N-terminal domain in natural crystallin is known to fold faster than C-terminal domain (Mayr et al., 1994). It has also been reported (Komar and Jaenicke, 1995) that the relative frequencies of codons in N-terminal domain are about 2.5-fold higher than those in C-terminal domain. Thus, the regulated speed of translation through codon usage seems to have an effect on the structural events in the encoded protein sequence. A discussion on the relative rates of translation of synonymous codons is in order here. Curran and Yarus (1989) have reported translation speed of synonymous codons. The reported speeds indicate the relative rates of association between the codon and the cognate aminoacyl-tRNA. We had discussed this data earlier in relation to codon usage in translationally slow mRNA regions (Thanaraj and Argos, 1996). Examination of the ratios of these reported values for the slowest translated codon with the fastest translated synonymous codon indicated a maximum value of 24-fold, while the average value (considering only pro, ser, arg, and leu while for other amino acids with more than two synonymous codons such values were not available) was 12-fold. Such a 12-fold difference in the translation rates of the fastest and slowest synonymous codons agrees with a 10-fold difference in the folding rates of a typical helix + beta protein and a helix protein (see the section on kinetic data on protein folding). Thus, the reported correlation between the speed of translation and optimization of co-translational folding efficiency is quite plausible.

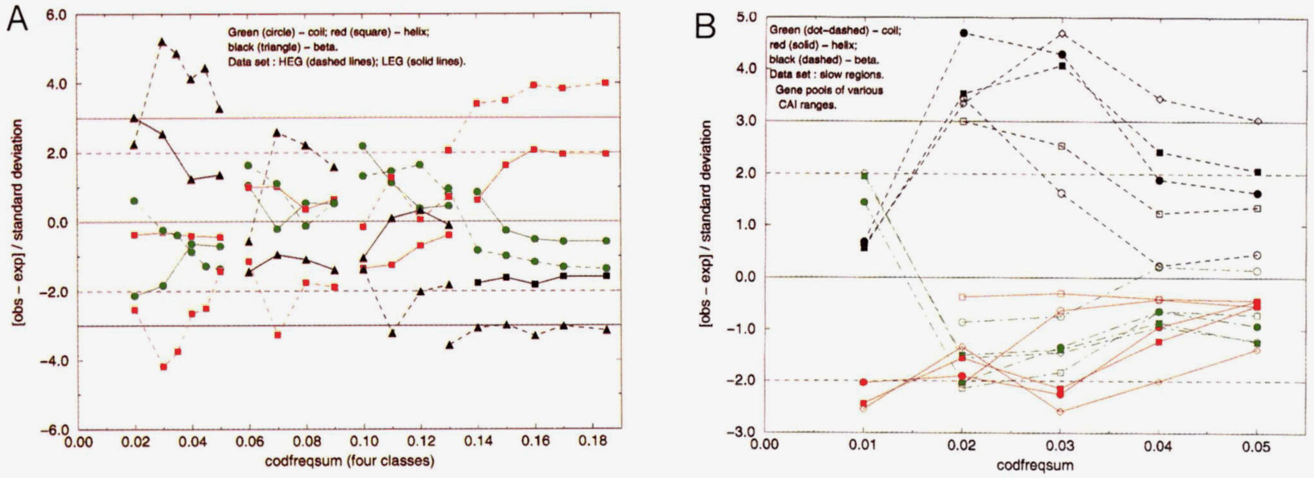
Lowly expressed genes (LEG) versus highly expressed genes (HEG)

Codon usage studies have indicated that the choice of synonymous codons in highly expressed genes is to optimize the translation rate of the mRNA, whereas in lowly/moderately expressed genes the choice is more uniform with clear relaxation of positive selection for translation optmization (Sharp and Li, 1986). In very lowly expressed genes there is also no preferential selection for rare codons such that the pattern of codon usage reflects the mutation pattern of the genome (Sharp and Li, 1986; Sharp et al., 1993). It, thus, might be expected that the observed differential coding of the protein secondary structural types on mRNA may be more prominent in highly rather than lowly expressed genes. To examine this supposition, the following analysis was performed.

The Codon Adaptation Index (CAI, see Materials and methods) as formulated by Sharp and Li (1987), which indicates the level of expression of genes, was used to classify the genes in the data set used here. The calculated CAI values for the 40 entries in Table 2 ranged from 0.194 to 0.805. Accordingly, these genes were divided into two sets: CAI values ≤ 0.40 (LEG proteins), and CAI values > 0.40 (HEG proteins). The preference/avoidance patterns for the protein secondary structural types were determined in the usual manner for the two sets and the results are shown in Figure 5A, which clearly illustrates that the preference/avoidance pattern is much more prominent in HEG proteins. As a next step, five different gene pools of progressively higher CAI values were created. Set 1 included genes of CAI values ≤ 0.35; set 2 with CAI values ≤ 0.40; set 3 with CAI values ≤ 0.45; set 4 with CAI values ≤ 0.50; and set 5 with CAI values ≤ 0.60. The structural preference/avoidance pattern for slow regions in each of the 5 sets is shown in Figure 5B and that for fast regions in each of the 5 sets is shown in Figure 5C. Figure 5B shows that the preference for beta sheets by slow regions attains ≥3 *sd* in the sets 3–5 as compared to the sets 1–2. Similarly, Figure 5C shows that the preference for alpha helices and avoidance for beta sheets by fast regions attain ≥3 *sd* in the sets 3–5 as compared to the sets 1–2. The results presented so far in this section indicate that the preference/avoidance pattern improves as genes of higher CAI values are used. Thus, the relaxation of positive selection towards optimal codons in lowly expressed genes is also reflected in the preference/avoidance pattern of the encoded protein structural types and further supports the proposed correlation between differential coding of mRNA regions and encoded protein structure resulting in selection constraints acting on codon usage in expressed genes.

Conclusion

The study has made salient an intriguing correlation between structural features on proteins and codon frequency in the corresponding mRNA molecules; namely, that protein alpha helices are preferentially coded by translationally fast mRNA regions while beta strands and coils are preferentially coded by slow mRNA regions. Such correlations become significant in the context of the differential kinetics in the formation and stabilization of helices and beta sheets during the co-translational folding of proteins. The results imply that translational kinetics are important factors in the *in vivo* structural organization of encoded proteins. The study should also be an aid in the design and mutation of genes such that fast (abundant) codons are selected for encoding critical helices while slow (low-usage) codons are utilized for critical beta strands.



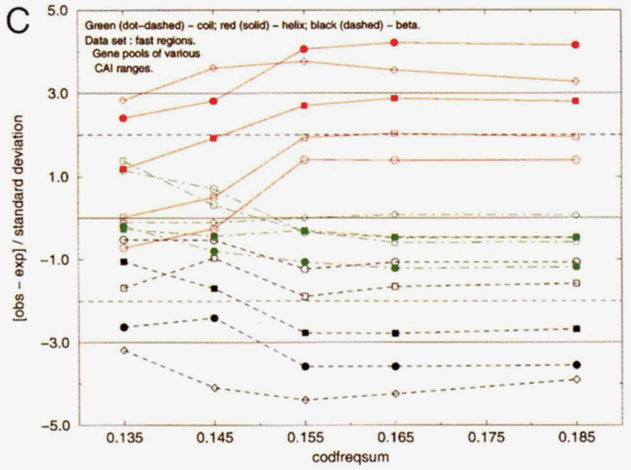


Fig. 5. Preference/avoidance pattern for the protein secondary structural types as a function of gene expressivity given by the codon adaptation index, CAI. The analysis was performed on the data set containing 40 proteins (Table 2). **(A)** Two protein sets were delineated. Solid lines correspond to lowly expressed genes (LEG) with CAI ≤ 0.40 and dashed lines correspond to highly expressed genes (HEG) with CAI > 0.40. Green lines with circles correspond to coil; red lines with squares to helix; and black lines with triangles to beta. **(B)** Consideration of slow mRNA regions for data sets that progressively include genes with increasing CAI values from the standard 40-protein bank (Table 2). Green dot-dashed lines correspond to coil; red solid lines to helix; and black dashed lines to beta. Lines with open circles correspond to genes of CAI ≤ 0.35; lines with open squares to genes of CAI ≤ 0.40; lines with closed circles to genes of CAI ≤ 0.45; dot-dashed lines with closed squares to genes of CAI ≤ 0.50; and long dashed lines with diamonds to genes of CAI ≤ 0.60. **(C)** Consideration of fast mRNA regions. The conditions of illustration are the same as those of (B).

Materials and methods

Data set used

A collection of 54 proteins from *Escherichia coli* was utilized. Their atomic coordinates (determined through crystallography) were taken from the Protein Data Base [PDB (Bernstein et al., 1977)], while secondary structural assignments of individual residues are available from DSSP (Kabsch and Sander, 1983) and nucleotide sequence data for the corresponding messenger RNA from the EMBL data bank (Rice et al., 1993). The sequences were extracted using the SRS information retrieval software (Etzold and Argos, 1993a, 1993b). The PDB/DSSP and EMBL identifiers of the selected entries in this study are given in Table 2. The residues in helix conformation are annotated by G or H in the corresponding DSSP file, beta structures by E or B, and coil by S, T, or blank. A total of 14 (marked by * in Table 2) of the 54 entries available from PDB were ignored because they displayed no helix or no beta or no coil residues or the content of the three secondary structural types deviated from the average calculated from the 54 protein entries by more than 50%. Such skewed structures may not utilize normal and consistent folding mechanisms (Carlsson and Jonsson, 1995).

Codon adaptation index

The codon adaptation index (CAI), as formulated by Sharp and Li (1987), is an indicative measure of the expression level of a gene. The index is based on the codon usage in a reference set of highly expressed genes. A score for a given gene is calculated from codon usage in the gene as compared to that in the reference set. A high (low) value for the index indicates that the gene is highly (lowly) expressed. The indices were determined by using the formalism of Sharp and Li (1987) and the codon frequency values calculated here and in Thanaraj and Argos (1996).

Recognizing differential translation speeds

The codon frequency values were used to quantify the speed of translation of a codon on the mRNA. In each mRNA sequence, the speed of translation of all individual tricodons starting at each codon position was scrutinized by taking the sum (denoted *codfreqsum*) of the codon frequency values (Table 1), for the constituent codons. A tricodon was considered as a unit of nucleotide region because three successive residues generally constitute basic protein structural motifs; namely, a small beta strand at 3 residues, one helical turn at 3.6, and a reverse beta turn at 2–4 amino acids (Colloc'h et al., 1993).

Calculation of secondary structural preference/avoidance within specific mRNA regions

The fractional content of helix, beta strand, and coil (denoted as F_h, F_b, and F_c) in all the 40 protein structures used here (Table 2) represented their expected frequencies. The number of residues (corresponding to the codons constituting the mRNA regions of interest) with the conformation of helix or beta strand or coil were respectively designated as O_h, O_b, and O_c, and their sum assigned as N such that the expected respective counts (E) in the specific mRNA regions are $E_h = N \times F_h$, $E_b = N \times F_b$, and $E_c = N \times F_c$. The values for F_h, F_b, and F_c were derived from only those proteins from which the corresponding mRNA spans related to the *codfreqsum* under consideration. The preference/avoidance (P) of an individual structural feature in the protein spans corresponding to the regions of *codfreqsum* under consideration are given by $P_h = (O_h - E_h)/s_h$; $P_b = (O_b - E_b)/s_b$; and $P_c = (O_c - E_c)/s_c$, where $s_h = [E_h \times (1 - F_h)]^{1/2}$, $s_b = [E_b \times (1 - F_b)]^{1/2}$ and $s_c = [E_c \times (1 - F_c)]^{1/2}$. The latter represent standard deviations (*sd*). A positive value for preference/avoidance indicates the preferential coding of that particular structure, while a negative value shows avoidance and a zero value points to expected behavior.

A specific sample calculation is shown below to ensure clarity. We consider here the preference/avoidance pattern for the regions of *codfreqsum* range 0.000 to 0.045 (see Table 3). The values of F_h, F_b, and F_c are 0.3965, 0.2195, and 0.3839 respectively. The number of residues, N, is 2827. The observed numbers of such residues with the conformation of helix, beta, and coil are respectively $O_h = 1042$, $O_b = 725$, and $O_c = 1060$, while the expected counts are $E_h = N \times F_h = 2827 \times 0.3965 = 1121.02$, $E_b = N \times F_b = 2827 \times 0.2195 = 620.64$, and $E_c = N \times F_c = 2827 \times 0.3839 = 1085.34$. The standard deviations are $s_h = [E_h \times (1 - F_h)]^{1/2} = 26.01$, $s_b = [E_b \times (1 - F_b)]^{1/2} = 22.01$, and $s_c = [E_c \times (1 - F_c)]^{1/2} = 25.86$. The preference/avoidance of individual structural features is $P_h = (O_h - E_h)/s_h = -3\ .038$, $P_b = (O_b - E_b)/s_b = 4.741$, and $P_c = (O_c - E_c)/s_c = -0.979$.

Acknowledgments

T.A.T. gratefully acknowledges support from the Human Frontier Science Program for a long-term fellowship. The authors thank Dirk Walther for the calculation of radius of gyration of soluble proteins.

References

Andersson SGE, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev 54*:198–210.

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science 181*:223–230.

Baldwin RL. 1975. Intermediates in protein folding reactions and the mechanism of protein folding. *Annu Rev Biochem 44*:453–475.

Bernabeu C, Lake JA. 1982. Nascent polypeptide chains emerge from the exit domain of the large ribosomal subunit: Immune mapping of the nascent chain. *Proc Natl Acad Sci USA 79*:3111–311.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Blobel G, Sebatini DD. 1970. Controlled proteolysis of nascent polypeptides in rat liver cell fractions. I. Location of the polypeptides within ribosomes. *J Cell Biol 45*:130–145.

Bonekamp F, Andersen HD, Christensen T, Jensen KF. 1985. Codon-defined ribosomal pausing in *Escherichia coli* detected by using the *pyrE* attenuator to probe the coupling between transcription and translation. *Nucleic Acids Res 13*:4113–4123.

Briggs MS, Roder H. 1992. Early hydrogen-bonding events in the folding reaction of ubiquitin. *Proc Natl Acad Sci USA 89*:2017–2021.

Brimacombe R. 1995. The structure of ribosomal RNA: A three-dimensional jigsaw puzzle. *Eur J Biochem 230*:365–383.

Brunak S, Engelbrecht J, Kesmir C. 1994. Correlation between protein secondary structure and the mRNA nucleotide sequence. In: Bohr H, Brunak S, eds. *Protein Structure by Distance Analysis*. Amsterdam: IOS Press. pp 327–334.

Buck M, Radford SE, Dobson CM. 1993. A partially folded state of hen egg white lysozyme in trifluoroethanol: Structural characterisation and implications for protein folding. *Biochemistry 32*:669–678.

Buck M, Radford SE, Dobson CM. 1994. Amide hydrogen exchange in a highly denatured state. Hen egg-white lysozyme in urea. *J Mol Biol 237*:247–254.

Candelas GC, Ortiz A, Ortiz N. 1989. Features of the cell-free translation of a spider fibroin mRNA. *Biochem Cell Biol 67*:173–176.

Carlsson U, Jonsson B-H. 1995. Folding of β-sheet proteins. *Curr Opin Struct Biol 5*:482–487.

Carter PW, Bartkus JM, Calvo JM. 1986. Transcription attenuation in *Salmonella typhimurium*: The significance of rare leucine codons in the *leu* leader. *Proc Natl Acad Sci USA 83*:8127–8131.

Chaney WG, Morris AG. 1979. Non-uniform size distribution of nascent peptides—The effect of messenger RNA structure upon the rate of translation. *Arch Biochem Biophys 194*:283–291.

Chyan CL, Wormald C, Dobson CM, Evans PA, Baum J. 1993. Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: A hydrogen exchange study. *Biochemistry 32*:5681–5691.

Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon J-P. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. *Protein Eng 6*:377–382.

Curran JF, Yarus M. 1989. Rates of aa-tRNA selection at 29 sense codons *in vivo*. *J Mol Biol 209*:65–67.

Deana A, Ehrlich R, Reiss C. 1996. Synonymous codon selection controls *in vivo* turnover and amount of mRNA in *Escherichia coli bla* and *ompA* genes. *J Bacteriol 178*:2718–2720.

Dobson CM, Evans PA, Radford SE. 1994. Understanding how proteins fold: The lysozyme story so far. *Trends Biochem Sci 19*:31–37.

Doniach S, Bascle J, Garel T, Orland H. 1995. Partially folded states of proteins: Characterisation by X-ray scattering. *J Mol Biol 254*:960–967.

Etzold T, Argos P. 1993a. Transforming a set of biological flat file libraries. *Comput Appl Biosci 9*:59–64.

Etzold T, Argos P. 1993b. SRS—An indexing and retrieval tool for flat file data libraries. *Comput Appl Biosci 9*:49–57.

Farrow NA, Zhang O, Forman-Kay JD, Kay LE. 1995. Comparison of the backbone dynamics of a folded and an unfolded SH3 domain existing in equilibrium in aqueous buffer. *Biochemistry 34*:868–878.

Fedorov AN, Friguet B, Djavadi-Ohaniance L, Alakhov YB, Goldberg ME. 1992. Folding on the ribosome of *Escherichia coli* tryptophan synthase beta subunit nascent chains probed with a conformation-dependent monoclonal antibody. *J Mol Biol 228*:351–358.

Feng Y, Sligar SG, Wand AJ. 1994. Solution structure of apocytochrome b562. *Nature Struct Biol 1*:30–35.

Finkelstein AV. 1991. Rate of β-structure formation in polypeptides. *Proteins* 9:23–27.

Frank J, Zhu J, Penczek P, Li Y, Srivastava S, Verschoor A, Radermacher M, Grassucci R, Lata RK, Agrawal RK. 1995. A model of protein synthesis based on cryo-electron microscopy of the *E. coli* ribosome. *Nature 376*:441–444.

Friguet B, Djavadi-Ohaniance L, King J, Goldberg ME. 1994. *In vitro* and ribosome-bound folding intermediates of P22 tailspike protein detected with monoclonal antibodies. *J Biol Chem 269*:15945–15949.

Grosjean H, FiersW. 1982. Preferential codon usage in prokaryotic genes: The optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene 18*:199–209.

Hardesty B, Odom OW, Kudlicki W, Kramer G. 1993. Extension and folding of nascent peptides on ribosomes. In: Nierhaus KH, Franceschi F, Subramanian AR, Erdmann VA, Wittmann-Liebold B, eds. *The translational apparatus*. New York: Plenum Press. pp 347–358.

Harms E, Umbarger HE. 1987. Role of codon choice in the leader region of the *ilvGMEDA* operon of *Serratia marcescens*. *J Bacteriol 169*:5668–5677.

Hartl F-U, Hlodan R, Langer T. 1994. Molecular chaperones in protein folding: The art of avoiding sticky situations. *Trends Biochem Sci 19*:20–25.

Hunt JF, Weaver AJ, Landry SJ, Gierasch L, Deisenhofer J. 1996. The crystal structure of the GroES co-chaperonin at 2.8 Å resolution. *Nature 379*: 37–45.

Ikemura T, Ozeki H. 1983. Codon usage and transfer RNA contents: Organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harbor Symp Quant Biol* 47:1087–1097.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13–34.

Jacobs MD, Fox RO. 1994. Staphylococcal nuclease folding intermediate characterised by hydrogen exchange and NMR spectroscopy. *Proc Natl Acad Sci USA 91*:449–453.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Komar AA, Jaenicke R. 1995. Kinetics of translation of YB-crystallin and its circularly permuted variant in an *in vitro* cell-free system: Possible relations to codon distribution and protein folding. *FEBS Lett 376*:195–198.

Krasheninnikov IA, Komar AA, Adzhubei IA. 1989. Role of the code redundancy in determining cotranslational protein folding. *Biokhimiia 54*:187–200.

Krasheninnikov IA, Komar AA, Adzhubei IA. 1991. Nonuniform size distribution of nascent globin peptides, evidence for pause localisation sites, and a cotranslational protein-folding model. *J Prot Chem 10*:445–453.

Kudlicki W, Odom OW, Kramer G, Hardesty B. 1994. Chaperone-dependent folding and activation of ribosome-bound nascent rhodanese. *J Mol Biol 244*:319–331.

Kudlicki W, Kitaoka Y, Odom OW, Kramer G, Hardesty B. 1995. Elongation and folding of nascent ricin chains as peptidyl-tRNA on ribosomes: The effect of amino acid deletions on these processes. *J Mol Biol 252*:203–212.

Kypr J. 1986. A part of codon bias in genes protects protein spatial structures from destabilisation by random single point mutations. *Biochem Biophys Res Commun 139*:1094–1097.

Kypr J, Mrazek J. 1987. Occurrence of nucleotide triplets in genes and secondary structure of the coded proteins. *Int J Biol Macromol* 9:49–53.

Liljenstrom H, von Heijne G. 1987. Translation rate modification by preferential codon usage: Intragenic position effects. *J Theor Biol 124*:43–55.

Lim VI, Spirin AS. 1986. Stereochemical analysis of ribosomal transpeptidation. Conformation of nascent peptide. *J Mol Biol 188*:565–574.

Lipman DJ, Wilbur WJ. 1983. Contextual constraints on synonymous codon choice. *J Mol Biol 163*:363–376.

Liu Z-P, Rizo J, Gierasch LM. 1994. Equilibrium folding studies of cellular retinoic acid binding protein, a predominantly β-sheet protein. *Biochemistry 33*:134–142.

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res 22*:3174–3180.

Logan TM, Theriault Y, Fesik SW. 1994. Structural characterisation of the FK506 binding protein unfolded in urea and guanidine hydrochloride. *J Mol Biol 236*:637–648.

Mayr E-M, Jaenicke R, Glockshuber R. 1994. Domain interactions and connecting peptides in lens crystallins. *J Mol Biol 235*:84–88.

McNally T, Purvis IJ, Fothergill-Gilmore LA, Brown AJP. 1989. The yeast pyruvate kinase gene does not contain a string of non-preferred codons: Revised nucleotide sequence. *FEBS Lett 247*:312–316.

Mullins LS, Pace CN, Raushel FM. 1993. Investigation of ribonuclease T1 folding intermediates by Hydrogen-Deuterium amide exchange—Two dimensional NMR spectroscopy. *Biochemistry 32*:6152–6156.

Narayana SV, Argos P. 1984. Residue contacts in protein structures and implications for protein folding. *Int J Pept Protein Res 24*:25–39.

Palau J, Argos P, Puigdomenech P. 1981. Protein secondary structure. *Int J Pept Protein Res 19*:394–401.

Phillips DC. 1966. The three-dimensional structure of an enzyme molecule. *Sci Am 215*:78–90.

Prat Gay G-de, Ruiz-Sanz J, Neira JL, Corrales FJ, Otzen DE, Ladurner AG, Fersht AR. 1995. Conformational pathway of the polypeptide chain of chymotrypsin inhibitor-2 growing from its N-terminus *in vitro* parallels with the protein folding pathway. *J Mol Biol 254*:968–979.

Ptitsyn OB. 1995. How the molten globule became. *Trends Biochem Sci 20*:376–379.

Purvis IJ, Bettany AJE, Santiago TC, Coggins JR, Duncan K, Eason R, Brown AJP. 1987. The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. *J Mol Biol 193*:413–417.

Redfield C, Smith RAG, Dobson CM. 1994. Structural characterisation of a highly ordered 'molten globule' at low pH. *Nature Struct Biol 1*:23–29.

Rice CM, Fuchs R, Higgins DG, Stoehr PJ, Cameron GN. 1993. The EMBL data library. *Nucleic Acids Res 21*:2967–2971.

Robson B, Garnier J. 1986. The concept of conformation. In *Introduction to proteins and protein engineering*. New York: Elsevier. pp 47–118.

Roder H, Elove GA, Englander SW. 1988. Structural characterisation of folding intermediates in cytochrome *c* by H-exchange labelling and proton NMR. *Nature 335*:700–704.

Sharp PM, Li W-H. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection. *Nucleic Acids Res 14*:7737–7749.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res 14*:5125–5143.

Sharp PM, Li W. 1987. The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res 15*:1281–1295.

Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage: Mutational bias, translation selection, or both? *Biochem Soc Trans 21*:835–841.

Shiraki K, Nishikawa K, Goto Y. 1995. Trifluoroethanol-induced stabilisation of the α-helical structure of β-lactoglobulin: Implication for non-hierarchical protein folding. *J Mol Biol 245*:180–194.

Shpaer EG. 1986. Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. *J Mol Biol 188*:555–564.

Siemion IZ, Siemion PJ. 1994. The informational context of the third base in amino acid codons. *Biosystems 33*:139–148.

Sigler PB, Horwich AL. 1995. Unliganded GroEL at 2.8 Å: Structure and functional implications. *Philos Trans R Soc Lond [Biol] 348*:113–119.

Sorensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol 207*:365–377.

Sorensen MA, Pedersen S. 1991. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*—The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol 222*:265–280.

Sosnick TR, Mayne K, Hiller R, Englander SW. 1994. The barriers in protein folding. *Nature Struct Biol 1*:149–156.

Stade K, Junke N, Brimacombe R. 1995. Mapping the path of the nascent peptide chain through the 23S RNA in the 50S ribosomal subunit. *Nucleic Acids Res 23*:2371–2380.

Stark H, Mueller F, Orlova EV, Schatz M, Dube P, Erdemir T, Zemlin F, Brimacombe R, van Heel M. 1995. The 70S *Escherichia coli* ribosome at 23 Å resolution: Fitting the ribosomal RNA. *Structure 3*:815–821.

Taylor FJR, Coates D. 1989. The code within codons. *Biosystems 22*:177–187.

Thanaraj TA, Argos P. 1996. Ribosome mediated translational pause and protein domain organisation. *Protein Sci 5*:1594–1612.

Thomas PD, Dill KA. 1993. Local and non-local interactions in globular proteins and mechanism of alcohol denaturation. *Protein Sci 2*:2050–2065.

Tokatlidis K, Friguet B, Deville-Bonne D, Baleux F, Fedorov AN, Navon A, Djavadi-Ohaniance L, King J, Goldberg ME. 1995. Nascent chains: Folding and chaperone interaction during elongation on ribosome. *Philos Trans R Soc Lond [Biol] 348*:89–95.

Trifonov, EN. 1989. The multiple codes of nucleotide sequences. *Bull Math Biol 51*:417–432.

Tsalkova T, Zardeneta G, Kudlicki W, Kramer G, Horowitz PM, Hardesty B. 1993. GroEL and GroES increase the specific enzymatic activity of newly-synthesised rhodanese if present during *in vitro* transcription/translation. *Biochemistry 32*:3377–3380.

Tu C, Tzeng TH, Bruenn JA. 1992. Ribosomal movement impeded at a pseudoknot required for frameshifting. *Proc Natl Acad Sci USA 89*:8636–8640.

Udgaonkar JB, Baldwin RL. 1990. Early folding intermediates of ribonuclease A. *Proc Natl Acad Sci USA 87*:8197–8201.

Varenne S, Baty D, Verheij H, Shire D, Lazdunski C. 1989. The maximum rate of gene expression is dependent on the downstream context of unfavourable codons. *Biochimie 71*:1221–1229.

Varley P, Gronenborn AM, Christensen H, Wingfield PT, Pain RH, Clore GM. 1993. Kinetics of folding of the all-β sheet protein interleukin-1β. *Science 260*:1110–1113.

Volkenstein MV. 1966. The genetic coding of the protein structure. *Biochim Biophys Acta 119*:421–424.

Walther D, Argos P. 1996. Intrahelical side chain–side chain contacts, the consequences of restricted rotameric states and the implications for helix engineering and design. *Protein Eng 9*:471–478.

Wiedmann B, Sakai H, Davis TA, Wiedmann M. 1994. A protein complex required for signal-sequence-specific sorting and translocation. *Nature 370*:434–440.

Williams S, Causgrove TP, Gilmanshin R, Fang KS, Callender RH, Woodruff WH, Dyer RB. 1996. Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry 35*:691–697.

Woese CR, Dugre DH, Saxinger WC, Dugre SA. 1966. The molecular basis of the genetic code. *Proc Natl Acad Sci USA 55*:966–974.

Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S. 1991. Levels of tRNAs in bacterial cells as affected by aminoacid usage in proteins. *Nucleic Acids Res 19*:6119–6122.

Yang A-S, Honig B. 1995. Free energy determinants of secondary structure formation: I. α-helices. *J Mol Biol 252*:366–376.

Yarus M, Folley LS. 1985. Sense codons are found in specific contexts. *J Mol Biol 182*:529–540.

Yonath A. 1992. Approaching atomic resolution in crystallography of ribosomes. *Annu Rev Biophys Biomol Struct 21*:77–93.

Zana R. 1975. On the rate determining step for helix propagation in the helix-coil transition of polypeptide in solution. *Biopolymers 14*:2425–2428.

Zhang S, Goldman E, Zubay G. 1994. Clustering of low usage codons and ribosome movement. *J Theor Biol 170*:339–354.