

Submit to CSB2003

**Protein secondary structural type is correlated with codon translation efficiency and mRNA structure**

Liaofu Luo\* Mengwen Jia Xiaoqin Li  
(*Laboratory of Theoretical Biophysics, Faculty of Science and Technology,  
Inner Mongolia University, Hohhot, 010021 China*)

**Abstract**

The possible influence of codon usage and mRNA structure on protein secondary structure is discussed through statistical analysis of sequence-structure data. It is demonstrated that the influences exist through two approaches – the translation efficiency expressed by tRNA copy numbers of codons and the mRNA stem-loop structure. From statistical analyses of 401 human and *E.coli* polypeptide chains we have found that the messenger RNA segment of *m*-codons (for *m*=3 to 8) with averagely high tRNA copy number preferably code for alpha helix but less code for coil and this structural preference / avoidance turns out contrary to the codons with low tRNA copy number. For beta strand the preference / avoidance tendency is not obvious and it shows oscillation with tRNA copy number of codons. On the other hand, by calculating mRNA folding and studying the relation between RNA stem-loop frequencies and protein secondary structure in a database of 648 protein sequences we have found that helices and strands on proteins tend to be preferably “coded” by mRNA stem region, while coils tend to be preferably “coded” by mRNA loop region. The deduced correlations between protein secondary structure and mRNA-related information can hardly be explained by the stochastic fluctuation effect. The possible mechanisms related to the influence of tRNA copy number and mRNA folding on the protein secondary structure are discussed briefly.

**Key words**

protein secondary structure; sequence-structure database; oligo-codon segment ; mRNA sequence; tRNA copy number; translation efficiency; stem / loop structure.

Postal mailing address : Laboratory of Theoretical Biophysics, Faculty of Science and Technology, Inner Mongolia University, Hohhot, 010021 China.

Electronic mailing address : lfluo@mail.imu.edu.cn

\* To whom correspondence should be addressed.

Following Anfinsen's principle on the folding of protein chains, the protein spatial structure is fully determined by information contained in its amino acid sequence. Has the codon usage and mRNA structure nothing to do with the protein secondary structure? Several authors suggested the possible connections between protein secondary structure and mRNA information [Guisez *et al*, 1993; Brunak & Engelbrecht, 1996; Thanaraj & Argos, 1996; Adzhubei *et al*, 1998; Xie & Ding, 1998; Oresic & Shalloway, 1998]. However, to our knowledge, there have not been direct experiments which show protein structural change by synonymous codon replacement. Due to the importance of the problem, a more convincing statistical analysis of up-to-date sequence-structure data is necessary.

ISSD database [Adzhubei *et al*, 1999] holds 119 *Homo sapiens* and 92 *E. coli* polypeptide chains (sequence identity < 50%). Apart from ISSD 2.0, to give a better statistics a new integrated sequence-structure database has been constructed, called IADE (Integrated ASTRAL- DSSP- EMBL) [Jia, Luo & Liu, 2003]. In IADE1 it includes 2269 protein sequence-structures with sequence identity less than 40%. After matching with nucleotide sequence the database is called IADE2. The latter includes 648 protein domains. We shall use it to analyze the influence of mRNA structure on protein secondary structure. After deleting those overlapped with ISSD, IADE2 includes 102 polypeptide chains for human and 88 polypeptide chains for *E.coli*. We shall use it to check the relation between codon usage and protein structure obtained from ISSD.

### **The relation between codon translation efficiency and protein secondary structural type**

Since the regular secondary structure (helix and extended strand) occurs in the very early epoch of protein folding and the tRNA molecule is the adaptor of mRNA sequence to amino acid sequence we first study if the tRNA molecule can exert some influence on the formation of protein secondary structure. In many organisms the tRNA abundance appears to be roughly correlated with tRNA gene copy number, so tRNA gene copy number has been used as a proxy for tRNA abundance. We shall study the possible influence of tRNA copy number of codons on protein secondary structure formation through statistical analysis of sequence- structure data of human and *E. coli*. [Li, Luo & Liu, 2003]. The tRNA copy numbers (denoted as  $v$ ) for human and *E.coli* are given in Table 1.

Consider  $m$ -codon segment,  $m=3, 4, 5, 6, 7, \text{ or } 8$  (hereafter called  $m$ -mer). The average of  $v$  values for codons in an  $m$ -mer is the indicative measure of translation efficiency parameter (TEP) for the corresponding segment in mRNA sequence. Consider a sliding window of width of  $m$  codons shifted along the mRNA sequence (with step 1 codon). The  $m$ -mer number in the  $k$ -th TEP interval  $v_k$  which codes for protein secondary structure  $\alpha \beta$  or  $c$  is denoted as  $n_k^{(j)}$  (obs) ( $j = \alpha \beta c$ ). The total number of  $m$ -mers in the  $k$ -th TEP interval is denoted by  $n_k$ . Set the normalized residue numbers (probability) in structure  $j = \alpha \beta$  or  $c$  in the database denoted by  $q_j$ . Theoretically,  $n_k^{(j)}$  ( $j = \alpha \beta c$ ) in three structures obey polynomial distribution

$$p(\{n_k^{(j)}\}) = \frac{n_k!}{\prod_j n_k^{(j)}!} q_\alpha^{n_k^{(\alpha)}} q_\beta^{n_k^{(\beta)}} q_c^{n_k^{(c)}} \quad (1)$$

if no structural preference exists for  $m$ -mers in the  $k$ -th TEP interval. The expectation value of  $m$ -mer number coding for structure  $j$  in TEP interval  $\nu_k$  is  $n_k^{(j)}(\text{exp}) = n_k \cdot q_j$  and the deviation is  $\sigma_k^{(j)} = \sqrt{n_k \cdot q_j \cdot (1 - q_j)}$ . Calculating

$$F_k^{(j)} = \frac{\{n_k^{(j)}(\text{obs}) - n_k^{(j)}(\text{exp})\}}{\sigma_k^{(j)}} \quad (2)$$

in each TEP interval we obtain  $F_k^{(j)} - \nu_k$  relations.

The statistical investigations are carried out based on two databases – ISSD 2.0 (211 polypeptide chains) and IADE 2 (190 polypeptide chains), respectively. The results on  $F^{(j)}(m) - \nu$  relation are plotted in Figure 1 (for human) and 2 (for *E.coli*). For human the interval of TEP (tRNA copy number) is taken to be 1, and for *E.coli* the corresponding interval is 0.1. To make a clear appearance in these figures we have omitted those points which are related to  $m$ -mer number occurred within a TEP interval smaller than 30 where the fluctuation may be too large.

The contribution from stochastic fluctuations to the frequency  $n_k^{(j)}$  of  $m$ -codon fragments (denoted as  $O_k^j$ ) can be estimated by polynomial distribution (Eq. 1). By use of the frequency data one deduces  $O_k^j$  smaller than  $10^{-4}$  for  $F_k^{(j)} \geq 4$ , smaller than  $2 \times 10^{-3}$  for  $F_k^{(j)} \geq 3$ , and taking  $10^{-1}$  or  $10^{-2}$  for  $F_k^{(j)} \approx 2$ . So, if  $F_k^{(j)} \geq 2$  then the codon fragment frequency  $n_k^{(j)}$  within the  $k$ -th interval of TEP that codes for secondary structure  $j$  can hardly be explained by stochastic fluctuations and a statistically meaningful conclusion can be drawn. As  $F_k^{(j)} \geq 3$  the conclusion is meaningful at confidence level 99.7%.

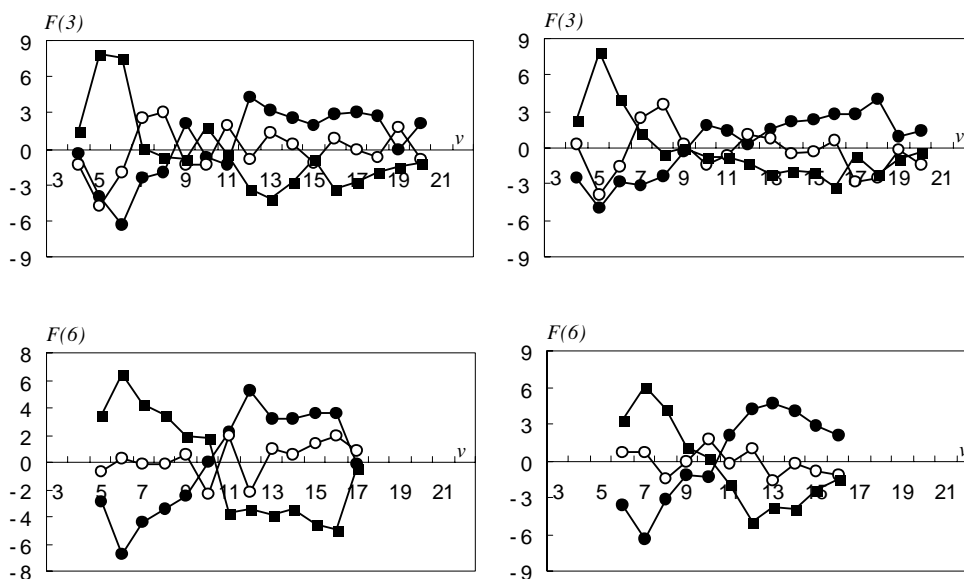
If only cases with  $F_k^{(j)} \geq 3$  or  $\leq -3$  are retained in  $F^{(j)}(m) - \nu$  relations the statistical results can be tabulated in a more clear way. Results calculated based on two databases are separately listed and aligned in Table 2 and 3. From the third column to the last column these  $m$ -mers with strong structural preference/avoidance are arranged in a line following the order of TEP  $\nu$ . For example, 5c,N $\alpha$  $\beta$  in the first line (labeled by  $F(3)$ ) of Table 2 means 3-mers in the 5-th TEP interval ( $4 < \nu \leq 5$ ) preferably code for random coil but less code for helix and strand; 6c,N $\alpha$  in this line means 3-mers in the 6-th TEP interval ( $5 < \nu \leq 6$ ) preferably code for coil but less code for helix; 8 $\beta$  in this line means 3-mers in the 8-th TEP interval ( $7 < \nu \leq 8$ ) preferably code for strand, etc.

**Table 1 The tRNA copy number of codons for Human and *E. coli***

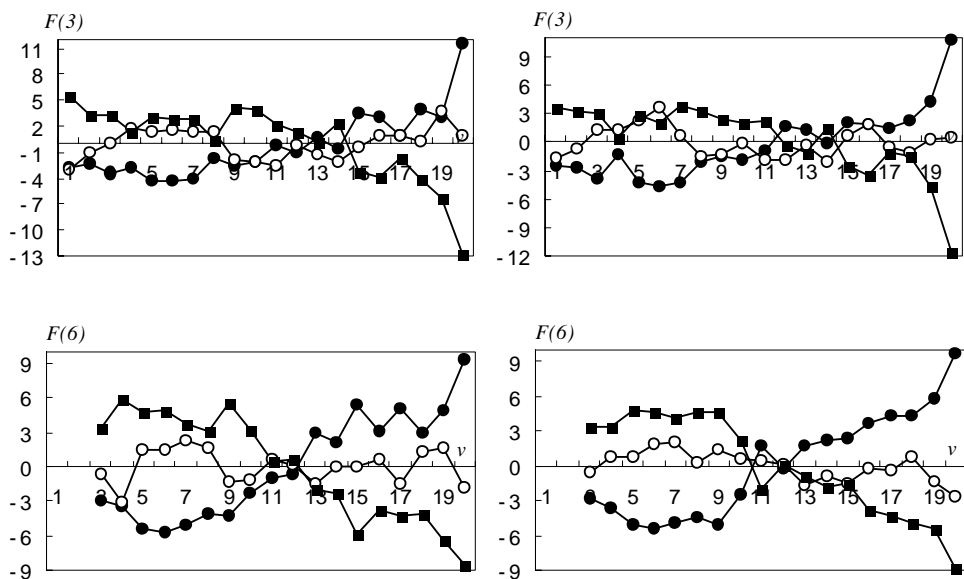
AMINO ACID	CODON	Human tRNA-copy no	ECO tRNA-copy no	AMINO ACID	CODON	Human tRNA-copy no	ECO tRNA-copy no
Phe	UUU	6.4	0.94	Ala	GCU	9.9	0.92
	UUC	7.6	1.06		GCC	15.1	1.08
Leu	CUU	5.3	0.49		GCA	10	1.15
	CUC	7.7	0.51		GCG	5	1.85
	CUA	2	1	Tyr	UAU	5.84	1.51
	CUG	6	4		UAC	6.16	1.49
	UUA	8	1	His	CAU	4.97	0.49
	UUG	6	1		CAC	7.03	0.51
Ile	AUU	5.6	1.39	Gln	CAA	11	2
	AUC	8.4	1.61		CAG	21	2
	AUA	5	1	Asn	AAU	16.39	1.11
	AUG	17	5		AAC	17.61	1.89
Val	GUU	8.64	1.25	Lys	AAA	16	2.28
	GUC	11.36	0.75		AAG	22	0.72
	GUA	5	1.40	Asp	GAU	4.67	1.70
	GUG	19	2.60		GAC	5.33	1.30
Ser	UCU	4.61	1.06	Glu	GAA	14	2.82
	UCC	5.39	0.94		GAG	8	1.18
	UCA	5	1	Cys	UGU	13.62	0.42
	UCG	4	1		UGC	16.38	0.58
	AGU	2.71	0.30	Trp	UGG	7	1
	AGC	4.29	0.70		AGA	5	1
Pro	CCU	5.17	0.63		AGG	4	1
	CCC	5.83	0.37		CGU	2.75	2.28
	CCA	10	1		CGC	6.25	1.72
	CCG	4	1		CGA	7	0.42
Thr	ACU	3.24	0.64		CGG	5	0.58
	ACC	4.76	1.36		Gly	GGU	3.6
	ACA	10	1			GGC	7.4
	ACG	7	1		GGA	5	1
					GGG	8	1

The tRNA copy numbers are taken from International Human Genome Sequencing Consortium (2001) for human and from Komine, Adachi, Inokuchi & Ozeki (1990) for *E. coli* respectively. For example, for human Leu there are 5 anticodons, UAA,CAA,AAG,UAG and CAG. The corresponding number of tRNA genes found with these anticodons in human are 8,6,13,2 and 6 respectively. Anticodon AAG pairs to two codons CUU and CUC with

frequency ratio 127:187.



**Figure 1**  $F(m) - v$  relation for Human To save space only  $F(m)$  with  $m=3$  and 6 are shown. Left figures give the statistics based on database ISSD 2.0 and right figures based on IADE2. In figures  $\bullet$  indicates the preference for  $\alpha$  helix  $\circ$  the preference for  $\beta$  strand and  $\blacksquare$  the preference for coil. The abscissa gives the number of tRNA copies.



**Figure 2**  $F(m) - v$  relation for *E.coli* To save space only  $F(m)$  with  $m=3$  and 6 are shown. Left figures give the statistics based on database ISSD 2.0 and right figures based on IADE2. In figures  $\bullet$  indicates the preference for  $\alpha$  helix  $\circ$  the preference for  $\beta$  strand and  $\blacksquare$  the preference for coil. The abscissa gives the number of tRNA copies. The first interval refers to copy number  $\leq 0.7$ , next (from 2<sup>nd</sup> to 19<sup>th</sup>) refers to  $0.7 < \text{copy number} \leq 0.8$ ,  $0.8 < \text{copy number} \leq 0.9$ , ..., and  $2.4 < \text{copy number} \leq 2.5$ , respectively and the last (20<sup>th</sup>) interval refers to copy number  $> 2.5$ .

**Table 2 Secondary structural preference/avoidance of codon TEP  
(tRNA copy number) for human (only  $F(m) \geq 3$  or  $\leq -3$  listed)**

F(3)	ISSD	5c,N $\alpha\beta$	6c,N $\alpha$	8 $\beta$		12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$
	IADE	5c,N $\alpha\beta$	6c,N $\alpha$	8 $\beta$					16 $\alpha$ ,Nc	17 $\alpha$ ,N $\beta$	18 $\alpha$
F(4)	ISSD	5c,N $\alpha\beta$	6c,N $\alpha$			12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc	17 $\alpha$ ,Nc	
	IADE	5c,N $\alpha$	6c,N $\alpha$	7c,N $\alpha$	8 $\beta$ ,N $\alpha$		13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc	16 $\alpha$	
F(5)	ISSD	5c,N $\alpha\beta$	6c,N $\alpha$	7c,N $\alpha$	8c	12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	
	IADE	5c,N $\alpha$	6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$	12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$	16 $\alpha$	
F(6)	ISSD	5c,N $\alpha$	6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$	12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	
	IADE		6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$	12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc		
F(7)	ISSD	6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$		12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc		
	IADE	6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$	9c,N $\alpha$	12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc		
F(8)	ISSD	6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$	9c,N $\alpha$	11 $\alpha$ ,Nc		13 $\alpha$ ,Nc		15 $\beta$ ,Nc	
	IADE	6c,N $\alpha$	7c,N $\alpha$	8c,N $\alpha$	9c,N $\alpha$		12 $\alpha$ ,Nc	13 $\alpha$ ,Nc	14 $\alpha$ ,Nc	15 $\alpha$ ,Nc	

**Table 3 Secondary structural preference/avoidance of codon TEP  
(tRNA copy number) for *E. coli* (only  $F(m) \geq 3$  or  $\leq -3$  listed)**

<b>F(3)</b>	<b>ISSD</b>	1c,Na $\beta$	2c	3c,Na	5c,Na	9c	10c			
				15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc		
	<b>IADE</b>	1c,Na	2c,Na	3c,Na	5c,Na	6 $\beta$ ,Na	7c,Na	8c		
				19 $\alpha$ ,Nc	20 $\alpha$ ,Nc					
<b>F(4)</b>	<b>ISSD</b>	1c,N $\beta$	3c,Na	4c,Na	5c,Na	6 $\beta$ ,Na	8c,Na	9c	10c	11c
				15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
	<b>IADE</b>	3c,Na	5c,Na	6c,Na	7c,Na	8c,Na	9c			
				15 $\alpha$ ,Nc	16 $\alpha$	17 $\alpha$	18 $\alpha$	19 $\alpha$	20 $\alpha$ ,Nc	
<b>F(5)</b>	<b>ISSD</b>	1c,Na	3c	4c,Na	5c,Na	6 $\beta$ ,Na	7c,Na	8c,Na	9c,Na	10c,Na
				15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
	<b>IADE</b>	2c,Na	4 $\beta$ ,Na	5c,Na	6c,Na	7c,Na	8c,Na			
				14 $\alpha$	15 $\alpha$	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
<b>F(6)</b>	<b>ISSD</b>	3c,Na	4c,Na $\beta$	5c,Na	6c,Na	7c,Na	8c,Na	9c,Na	10c	
				15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
	<b>IADE</b>	3c	4c,Na	5c,Na	6c,Na	7c,Na	8c,Na	9c,Na		
					16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
<b>F(7)</b>	<b>ISSD</b>	4c,Na	5c,Na	6 $\beta$ ,Na	7c,Na	8c,Na	9c	10c,Na		
			13 $\alpha$ ,Nc	15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
	<b>IADE</b>	4c	5c,Na	6c,Na	7c,Na	8c,Na	9c,Na			
				15 $\alpha$	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
<b>F(8)</b>	<b>ISSD</b>	4c,Na	5c,Na	6c,Na	7c,Na	8c,Na	9c	10c,Na		
			13 $\alpha$ ,Nc	15 $\alpha$ ,Nc	16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	
	<b>IADE</b>	5c,Na	6c,Na	7c,Na	8c,Na	9c,Na				
					16 $\alpha$ ,Nc	17 $\alpha$ ,Nc	18 $\alpha$ ,Nc	19 $\alpha$ ,Nc	20 $\alpha$ ,Nc	

From Figure 1-2 and Table 2-3 we find that for human

- 1) The mRNA sequences consisting of  $m$ -codons ( $m=3$  to 8) with averagely high copy number of tRNA (averaged copy number larger than 11 for human, larger than 2 for *E. coli*, respectively) preferably code for alpha helix but less code for coil.
- 2) The structural preference / avoidance turns out contrary to the codons with low tRNA copy number. As  $4 < (\text{averaged copy number } v) \leq 9$  for human or averaged copy number smaller than 1.5 or 1.6 for *E. coli* the  $m$ -mers preferably code for coil but less code for alpha helix.
- 3) For beta strand the preference/avoidance tendency is not obvious ( $F_k^{(\beta)} < 2$  in general, apart from few cases) and shows oscillation with TEP.

- 4) The results calculated based on two databases are very similar to each other.
- 5) We have studied the correlation of protein secondary structure with codon fractional frequency for *E. coli* [Thanaraj and Argos, 1996] and found that the correlation seems also exist but weaker than that with tRNA copy number. In tRNA copy number case there are 37 (30) modes of strong alpha structural preference and 43 (35) modes of strong coil structural preference in ISSD (IADE) database (for  $m=3$  to 8, see Table 3). But in codon fractional frequency case the corresponding number is only 4 (13) for strong alpha structural preference and 7(5) for strong coil structural preference.

The comparative studies between human and *E.coli* support the view that tRNA copy number correlates with protein secondary structure. But, how the formation of protein secondary structure is influenced by the translation process, the tRNA-related properties of codons? The statistical law given above leads to the following hypothesis: to decrease the wrong assignment of protein secondary structure the alpha helix should preferentially use codons with high tRNA copy number while the coil preferentially uses codons with low copy number. In fact, the translation accuracy of a codon and its translocation time across ribosome is dependent on tRNA copy number and mRNA local structure. On the other hand, a definite probability of erroneous translation within a definite translocation time would cause different effects in forming the correct protein secondary structures. The tolerance to translational error is different for alpha helix and random coil. So these two structures are preferentially coded by codons with different tRNA copy numbers.

#### **The relation between mRNA stem-loop structure and protein secondary structural type**

The mRNA secondary structure is deduced from nucleotide sequence by use of RNAstructure3.6 [Mathews, Sabina, Zuker, *et al.* 1999]. The secondary structure of mRNA during translocation should differ significantly from the global secondary structure of free mRNA. Considering the ribosomes placed along the mRNA about 90 nucleotides from one another and the very low probability of long-range base-pairing between different parts we subdivide the mRNA in a polysome into parts and fold the mRNA in each part. More plainly, we fold the mRNA sequence through base pairing by use of RNAstructure3.6 in a window of 100 nucleotides, and shift the window along the sequence. The un-pairing part in the tail of the first 100 nucleotides is put into the shifted window and participates in the next folding. Based on the above model (called local folding) we postulate the secondary structure of mRNA as a number of hairpins or more complex structures, constructed by loops (un-pairing bases including end loop, buldge, interior loop and multi-branch loop) and stems (pairing bases). The nucleotide in loop is denoted by 0 and that in stem by 1. So, the secondary structure of a mRNA is depicted by a sequence written by two symbols, 0 and 1. To study the possible dependence of our statistical results on mRNA folding approaches, as a comparison, we also fold the whole sequence of messenger RNA by use of Zuker's program (called whole folding hereafter).

Let the normalized base numbers (probability) in the mRNA structure  $j$  ( $j= 0$  or  $1$ )

in the database be  $q_j$ . The total number of nucleotides in database corresponding to the  $k$ -th protein secondary structure ( $k= \text{H,E,T or C}$ , denoting helix, extended strand, turn or coil respectively) is  $n_k$ . The expectation value of the base number of structure  $j$  occurring in the  $k$ -th protein secondary structure is given by  $n_k^{(j)}(\text{exp}) = n_k \cdot q_j$  under the no-relation-assumption between mRNA stem / loop and protein structure. Let the observed base number of structure  $j$  occurring in the  $k$ -th protein secondary structure be  $n_k^{(j)}(\text{obs})$ . Calculating Eq.(2) by use of sequence-structure data in IADE2 database (648 proteins) we obtain  $F_k^{(1)}$  as follows (Table 4)

**Table 4 Preference of protein secondary structural types for the mRNA stems**

	$k=\text{H}$ (helix)	$k=\text{E}$ (strand)	$k=\text{T}$ (turn)	$k=\text{C}$ (coil)
$F_k^{(1)}$ (local folding)	3.40	4.21	3.43	-9.25
$\sigma_k^{(1)}$ (local folding)	186	136	104	169
$F_k^{(1)}$ (whole folding)	2.80	3.31	3.91	-8.16
$\sigma_k^{(1)}$ (whole folding)	183	134	102	166

( $q_0 = 0.447$ ,  $q_1 = 0.553$  for local folding;  $q_0 = 0.395$ ,  $q_1 = 0.605$  for whole folding. The preference for loops  $F_k^{(0)}$  is the minus of  $F_k^{(1)}$ )

The results in Table 4 show the correlation between mRNA stem-loop structure and protein secondary structural type. The stem preferably codes for helix and strand but less code for coil, while the loop preferably codes for coil but less code for helix and strand.

To obtain a better statistics we define ‘‘structural word’’ (SW) as four-residue-fragment that shows pronounced secondary structural propensity. More plainly, shifting a window of width 4 residues along each protein sequence (in IADE1 database) we numerate the frequency of a given four- residue -fragment, denoted by  $N$ . Only the fragment with  $N \geq 3$  will be considered. Suppose the fragment occurring in structure  $k$  ( $k = \text{helix, extended strand, etc.}$ )  $n_k$  times. If its occurrence in structure  $k$

is a stochastic event then the probability for occurring  $n_k$  or more times will be

$$1 - CL_k = \sum_{n \geq n_k} \frac{N!}{n!(N-n)!} P_k^n (1 - P_k)^{(N-n)} \quad (3)$$

where  $P_k$  is the relative frequency of structure  $k$  in database, namely,  $P_k = \frac{m_k}{M}$ ,  $m_k$  – the total frequency of all four-residue-fragments occurring in structure  $k$  and  $M = \sum_k m_k$ . As Eq(3) is a small quantity one may say that the fragment occurring in

structure  $k$  for  $n_k$  times should not be at random. The confidence level of this statement is  $CL_k$ . In the following we will choose  $CL_k \geq 95\%$ . When the frequency of

a four-residue-fragment occurring in structure  $k$  satisfies Eq.(3) with  $CL_k \geq 95\%$ , that is, the R.H.S of Eq (3)  $< 0.05$ , we define the four-residue-fragment as  $k$ -type SW. It means that the occurrence of the word in structure  $k$  is not at random with 95% confidence level and this word is a characteristic word of the structure  $k$ . We shall study the following types of SW by use of database IADE1(2269 proteins): H-type SW, the secondary structures encoded by four-residue-fragment are HHHH; E-type SW, the secondary structures encoded by four-residue-fragment are EEEE; T-type SW, the secondary structures encoded by four-residue-fragment are TTTT and boundary type SW, the secondary structures encoded by four-residue-fragment are T and other structures (H,E,C), for example, HHHT, TTEE, TCCC, HHTC, etc. Few four-residue-fragments are SW of two types.

Based on SWs we are able to deduce the relation between protein secondary structure and mRNA stem-loop structure in a more clear way. We calculate the occurrence frequency of stem or loop in each kind of SW. In calculation the double count is avoided. For successive SWs of same type the overlapping part should be counted only once. The results on the preference of mRNA stems/loops for the protein secondary structural words are calculated by use of IADE2 database (648 proteins ) and they are shown in Table 5.

**Table 5 Preference of protein structural words for the mRNA stems**

SW type ( $k$ )	SW number	loop frequency	stem frequency	$F_k^{(1)}$	$\sigma_k^{(1)}$
H-type word(local)	4337	40500	52026	5.73	151
E-type word(local)	2866	20901	27522	6.83	109
T-type word(local)	1273	10936	13262	-1.52	77
Boundary type word (local)	3464	30079	39911	9.21	132
H-type word(whole)	4337	35987	56539	4.00	149
E-type word(whole)	2866	18376	30047	7.15	108
T-type word(whole)	1273	9629	14569	-0.82	76
Boundary type word (whole)	3464	26648	43342	7.92	129

To gain more information on the relationship between mRNA stem/loop structure and protein secondary structure we study the occurrence frequency of a dimer, a trimer or generally, an  $n$ -mer ( $n$ -nucleotide fragment expressed by 0 and 1) in mRNA stem/loop structure and its relation to different secondary structures and SWs. For dimer one has  $j=00,01,10$  and 11, for trimer one has  $j=000,001,010,011,100,101,110$  and 111, etc. For each  $j$  we calculate  $F_k^{(j)}$ . The protein structure  $k$  is defined by four types of SW – H-type, E-type, T-type and boundary-type SW; or five types of secondary structures – helix, extended strand, turn, structural boundary (defined by  $n$ -mer including a boundary between a turn and an other structure) and other secondary structure. All  $F_k^{(j)} \geq 3$  and  $F_k^{(j)} \leq -3$  modes have been calculated and listed in Table 6.

From Table 4-6 we find the regular secondary structures of protein – helices and strands – are strongly related to the stems of the corresponding mRNA structure. These regular structures tend to be preferably “coded” by mRNA stem region. While the coils

tend to be preferably “coded” by mRNA loop region. These tendencies are more obvious if we observe the structural words. The law is also proved in  $n$ -mers (up to  $n=6$ ). All  $n$ -mers solely composed of loops very scarcely occur in H-type words, E-type words, helices and strands with a high confidence level, but they tend to code for non-regular secondary structures. However, on the other hand, the E-type SWs and strands preferably tend to be coded by  $n$ -mers solely composed of stems. The H-type words also preferably tend to be coded by  $n$ -mers mainly composed of stems but the tendency is not so obvious as in E-type words.

**Table 6 Preference of protein structural words and structural types for the stem-loop structure in mRNA oligo-nucleotide fragments ( $n$ -mers with  $n \leq 6$ ) (all modes with  $F_k^{(j)} \geq 3$  or  $\leq -3$  are indicated)**

(a)

Strc	0	1	00	11	000	101	111	0000	1011	1111	00000
H.W.	-5.7	5.7	-6.9	3.8	-7.7	3.2	----	-8.6	3.1	----	-9.5
E.W.	-6.8	6.8	-6.9	6.5	-6.6	----	7.7	-6.0	----	8.4	-5.0
B.W.	-9.2	9.2	-10.3	7.6	-12.1	----	6.8	-13.2	----	5.7	-13.6
T.W.	----	----	----	----	----	----	----	----	----	----	----
Strc	00111	10000	10111	11101	11111	000000	100000	100001	110000	111101	111111
H.W.	----	---	3.1	----	----	-10.1	----	----	----	----	----
E.W.	----	-3.1	----	3.1	8.9	-3.6	-3.4	----	-3.7	3.9	8.5
B.W.	3.3	---	---	---	4.9	-13.9	----	3.3	---	---	4.0

(b)

Strc	0	1	00	11	000	001	011	110	111
$\alpha$	-3.4	3.4	-4.2	---	-4.9	---	---	---	---
$\beta$	-4.2	4.2	-4.0	4.6	-3.6	---	---	---	6.0
turn	-3.4	3.4	---	---	-3.3	4.2	---	---	---
bd	---	---	-3.3	---	---	---	4.1	3.7	---
other	9.2	-9.2	10.4	-7.4	10.4	---	---	---	-6.1
Strc	0000	0011	1111	00000	00111	11111	000000	000110	111111
$\alpha$	-5.7	---	---	-6.0	----	----	-6.5	---	---
$\beta$	-3.2	---	6.3	----	----	6.4	---	---	6.5
turn	----	3.6	----	----	4.2	---	---	3.1	---
bd	----	----	----	----	----	---	---	----	---
other	10.4	---	-4.8	10.1	---	-3.1	9.7	---	---

The first half of the table (a) gives the preference of protein structural words for the mRNA stem/loop structure, the second half (b) gives the preference of protein secondary structural types for the mRNA stem/loop structure.

H.W. = H-type word, E.W. = E-type word, B.W. = Boundary-type word, T.W. = T-type word;

$\alpha$  = helix,  $\beta$  = strand, turn = turn, bd = structural boundary and other = other secondary structures.

Each column gives a kind of stem-loop structures of  $n$ -mer and the corresponding structural preference  $F_k^{(j)}$ .

What is the possible mechanism responsible for the influence of mRNA folding on protein secondary structure? From the observation on *E.coli* ribosome by use of cryo-electron microscopy it gives the length of peptide channel about 85 nm and its diameter about 20 nm. Considering the average diameter of an  $\alpha$  helix or a  $\beta$  strand about 10 nm there is enough space for the folding of a nascent peptide in peptide channel to form a fragment of helix or strand. This is consistent with the view of co-translational folding of nascent peptides. On the other hand, the tertiary structure of

mRNA has to be unfolded during translation and the single stranded mRNA often keeps a helical conformation dominated by base-stacking interaction. The computer modeling shows that the conformation of single stranded mRNA corresponding to the stem of a hairpin after unfolding tends to be helical while that corresponding to the loop of a hairpin tends to be less spiral or a coil. Thus, we are able to explain why the extended strand and helix preferably tend to be coded by stems. The next point is: the relatively rigid tRNA may turn and move freely in the cavity located between large and small subunits of the ribosome. One may assume that due to commutative adaption, the distances between mRNA and CCA ends of tRNAs keep constant approximately as the anticodons of tRNA pairing with the codons of mRNA. So, when the tRNAs “read” and “proofread” a mRNA chain, the nascent peptide folding will be influenced by the specific topological configuration of single stranded mRNA.

The above results show that the nascent peptide folding may not be fully determined by the amino acid sequence. Their formation is possibly influenced by the translation efficiency of codon and the stem-loop structure of mRNA.

**Acknowledgement** The work was supported by National Science Foundation of China, Project No. 90103030. The authors are indebted to Prof Liu CQ for discussion about the possible influence of mRNA structure on protein structure.

### References

- 1 Guisez Y Robbens J Remaut E, Fiers W. *J Theor Biol.* 1993; **162**: 243 - 252.
- 2 Brunak S Engelbrecht J. *Protein Sci.* 1996; **25**:237 - 252.
- 3 Thanaraj TA Argos P. *Protein Sci.* 1996; **5**: 1973 - 1983.
- 4 Adzhubei IA Adzhubei AA Neidle S. *Nucl Acids Res.* 1998; **26**: 327 - 331.
- 5 Xie T Ding DF. *FEBS Lett.*, 1998; **434**: 93-96.
- 6 Oresic M, Shalloway D. *J Mol Biol.* 1998; **281**: 31- 48.
- 7 Adzhubei IA Adzhubei AA. *Nucl Acids Res.* 1999; **27**: 268-271.
- 8 Jia MW, Luo LF, Liu CQ. *Biopolymers* (to be published).
- 9 Li XQ, Luo LF, Liu CQ. *Acta Biochim Biophys Sinica* 2003; **35**: 193-196 (in Chinese).
- 10 International Human Genome Sequencing Consortium. *Nature*, 2001; **409**: 860-921.
- 11 Komine Y, Adachi T, Inokuchi H & Ozeki H. *J. Mol. Biol.* 1990; **212**: 579-598.
- 12 Mathews DH, Sabina J, Zuker M, *et al.* *J Mol Biol.* 1999; **288**: 911-940.