doi:10.4149/gpb_2011_02_154

The influence of protein coding sequences on protein folding rates of all-β proteins

RuiFang Li[1,2] and Hong Li[1]

[1] *School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China*
[2] *College of Physics and Electronic Information, Inner Mongolia Normal University, Hohhot 010022, China*

Abstract. It is currently believed that the protein folding rate is related to the protein structures and its amino acid sequence. However, few studies have been done on the problem that whether the protein folding rate is influenced by its corresponding mRNA sequence. In this paper, we analyzed the possible relationship between the protein folding rates and the corresponding mRNA sequences. The content of guanine and cytosine (GC content) of palindromes in protein coding sequence was introduced as a new parameter and added in the Gromiha's model of predicting protein folding rates to inspect its effect in protein folding process. The multiple linear regression analysis and jackknife test show that the new parameter is significant. The linear correlation coefficient between the experimental and the predicted values of the protein folding rates increased significantly from 0.96 to 0.99, and the population variance decreased from 0.50 to 0.24 compared with Gromiha's results. The results show that the GC content of palindromes in the corresponding protein coding sequence really influences the protein folding rate. Further analysis indicates that this kind of effect mostly comes from the synonymous codon usage and from the information of palindrome structure itself, but not from the translation information from codons to amino acids.

Key words: All-β protein — Protein folding rate — Protein coding sequence — GC content of palindromes — Synonymous codon usage

Abbreviations: GC content, guanine and cytosine content of protein coding sequence; P_{GC}, GC content of palindromes; C_{GC}, GC content of protein coding sequence.

Introduction

Discovering the mechanism of protein folding is a great challenge in molecular biology. A key step is to find useful factors that are related to the protein folding rates. Baker's group made an important observation in 1998 that the folding rates of two-state folding proteins correlate with the native topologies and proposed a concept of contact order (CO) (Plaxco et al. 1998a) to predict the protein folding rates. Since then, a great deal of studies (Dill et al. 1993; Fiebig and Dill 1993; Plaxco and Baker 1998b; Alm and Baker 1999; Debe and Goddard 1999; Mounoz and Eaton 1999; Dinner and Karplus 2001; Gromiha and Selvaraj 2001; Mirny and Shakhnovich 2001; Zhou and Zhou 2002; Gong et al. 2003; Ivankov et al. 2003; Nölting et al. 2003; Zhang et al. 2003; Ivankov and Finkelstein 2004) has shown that the protein folding rates correlated significantly with protein's three-dimensional or secondary structures. However, these conclusions are all based on the knowledge of the native structure of proteins. There were also some investigations concerning predicting the protein folding rates based on amino acid sequences showed that the protein folding rates depend substantially on amino acid sequences (Shao and Zeng 2003; Kuznetsov and Rackovsky 2004; Gromiha 2005; Punta and Rost 2005; Galzitskaya and Garbuzynskiy 2006; Gromiha et al. 2006; Ouyang and Liang 2008). And some useful web-serve were proposed for protein folding rates prediction, such as the FOLD-RATE (Gromiha et al. 2006) proposed by Gromiha's group (http://psfs.cbrc.jp/fold-rate/), and the FoldRate (Chou and Shen 2009) proposed by Chou's group (www.csbio.sjtu.edu.cn/bioinf/FoldRate/).

Correspondence to: Hong Li, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
E-mail: ndlihong@imu.edu.cn

Some research also indicated that there are important relations between the mRNA sequence and the protein structures. For example, the uneven elongation rates of ribosome along with mRNA sequence impact on the synthesis rates of nascent peptide chains (Varenne et al.1984; Purvis et al. 1987; Krasheninnikov et al. 1998; Komar 2009), the synonymous codon usage correlates with the protein structure (Chiusano et al. 2000; Gupta et al. 2000; Makhoul and Trifonov 2002; Gu et al. 2004; Mukhopadhyay et al. 2007), and the protein secondary structure correlates either with the mRNA sequence or the mRNA structure (Brunak and Engelbrecht 1996; Thanaraj and Argos 1996; Adzhubei et al. 1998; Mathews et al. 1999). However, how mRNA sequence influences the protein structures and functions are still unclear. We suppose following difficulties on this problem: 1) it is hard to find a proper parameter to characterize the functional structures of protein and its variances; 2) it is also hard to define proper parameters to characterize the structures of mRNA. Additionally, the effect of protein coding sequence on its corresponding protein structures is thought to be weak, though it is very important. These two problems lead to the difficulties of studying their relations directly and obtaining reasonable results. Here, we tried to analyze this problem with an indirect method.

By investigating protein folding process, we have found that the parameter of protein folding rate carries much information of protein sequence and structures. So we selected the protein folding rate as a parameter to reflect the information of protein structures. Furthermore, we tried to seek a proper parameter in protein coding sequence which may contain the main character of mRNA structures, then to study the influence of the parameter on the protein folding rate. We think the palindrome is a quite suitable character of protein coding sequence. Palindrome is a kind of common structure in mRNA sequences and it has the potential to form mRNA stem structures (Nag and Kurst 1997). So palindrome sequences contain not only the mRNA sequence information, but also the mRNA structure information. Therefore we selected the guanine and cytosine (GC) content of palindromes (P_{GC}) as a parameter of mRNA.

To evaluate the effect of P_{GC} on the protein folding rate, we added it as a new parameter in Gromiha's model (Gromiha 2005) for predicting protein folding rates of all-β proteins and evaluated our results. If our predicted protein folding rates are improved significantly, that is to say P_{GC} is an effective parameter in protein folding process.

Materials and Methods

Materials

We used the 13 all-β proteins of Gromiha's work (Gromiha 2005) to do the indirect analysis. The 13 all-β protein sequences were taken from the database (http://psfs.cbrc.jp/fold-rate/sequence.html) constructed by Gromiha's group, and then the corresponding protein coding sequences of the proteins were taken from EMBL nucletide sequence database (Baker et al. 2000) through the cross-reference with the Protein Data Bank (PDB) (Berman et al. 2000). Some of the 13 all-β proteins in Gromiha's work are protein segments, so we intercepted these protein sequences and corresponding protein coding sequences that is the same as Gromiha's selection. The experimental values of folding rates for the 13 all-β proteins were derived from Gromiha's work. The information about the 13 all-β proteins or segments is given in Table 1.

Palindrome sequences

A palindrome sequence is a couple segments in protein coding sequence. When one segment is read from the 5' end to the 3' end, it is exactly the complement sequence of another segment read from the 3' end to the 5' end (Chew et al. 2005). More precisely, a palindrome structure in protein coding sequences can be defined as a word pattern of the form $b_1 \ldots b_L$ $n_1 \ldots n_s$ $b'_L \ldots b'_1$, where base b' is the complement of base b, the couple segments $b_1 \ldots b_L$ and $b'_L \ldots b'_1$ may form a stem structure. The region $b_1 \ldots b_L$ is its left arm and $b'_L \ldots b'_1$ is its right arm, and L is the length of the stem. The region $n_1 \ldots n_s$ is a gap sequence or a loop sequence and S is its length. We defined the segment $b_1 \ldots b_L$ and $b'_L \ldots b'_1$ as a palindrome sequence or palindrome. For example, AAGAACAnnnnUGUUCUU is a palindrome structure (L = 7 bp and S = 4 bp) and its palindrome sequence is AAGAACAUGUUCUU.

Table 1. The information of the 13 selected all-β proteins

PDB code	SWISS-PROT accession number	EMBL accession number	$\ln(k_f)_{ei}$	Structure class
1nyf	P06241	AAA36615	4.54	β
1pks	P27986	AAH94795	−1.05	β
1shg	P07751	CAA32663	1.41	β
1srl	P00523	CAA23696	4.04	β
1hng	P08921	CAA28757	2.89	β
1ten	P24821	CAA39628	1.06	β
1csp	P32081	CAA42235	6.98	β
1mjc	P0A9Y1	AAG58705	5.24	β
2ait	P01092	AAA26686	4.20	β
1fnf-9	P02751	BAD52437	−0.91	β
1fnf-10	P02751	BAD52437	5.48	β
1wit	P24821	CAA39628	0.41	β
1tit	Q8WZ42	CAA62188	3.47	β

$\ln(k_f)_{ei}$ is the experimental protein folding rate for each protein.

Table 2. The P_{GC} values and the other four P_{ave}values for the 13 all-β proteins

PDB code	K^0	P_β	R_α	ΔASA	P_{GC}
1nyf	0.392	0.473	0.352	0.422	0.396
1pks	0.386	0.442	0.325	0.406	0.413
1shg	0.383	0.481	0.342	0.442	0.523
1srl	0.403	0.489	0.325	0.415	0.531
1hng	0.418	0.479	0.358	0.420	0.406
1ten	0.398	0.455	0.331	0.395	0.58
1csp	0.382	0.45	0.392	0.394	0.319
1mjc	0.431	0.457	0.346	0.366	0.519
2ait	0.438	0.502	0.322	0.394	0.757
1fnf-9	0.466	0.479	0.345	0.410	0.582
1fnf-10	0.485	0.503	0.335	0.381	0.551
1wit	0.441	0.458	0.358	0.392	0.441
1tit	0.392	0.457	0.381	0.416	0.432

The values for the first four parameter come from the work done by Gromiha (2005), the checking computation were done for the 13 all β-proteins, the results are exactly the same as Gromiha's results. The last parameter is parameter P_{GC} calculated by Eq. (1).

For the 13 corresponding protein coding sequences, we calculated all the palindromes and their locations (L ≥ 4 bp) and calculated P_{GC} in each protein coding sequence.

GC content of palindromes

The distribution of palindromes in a protein coding sequence is complex. Some bases or base segments in a palindrome are often used many times by other palindromes. That is to say, one segment may take part in several palindromes. Though eventually these segments can form secondary structure (stem structure) in only one of these palindromes, but they make the potential variability and complexity of mRNA structures. We must consider these factors. We think these factors may be the main reason of influence on the protein folding rates. So P_{GC} is defined as follows:

$$P_{GC} = \frac{(N_{G0} + N_{C0}) + (\sum_{i=1}^{M} N_{Gi} + \sum_{i=1}^{M} N_{Ci})}{N_{P0} + \sum_{i=1}^{M} N_{Pi}} \qquad (1)$$

where N_{G0} and N_{C0} are respectively the number of base G and base C and N_{P0} is the total base number of palindromes in a protein coding sequence (not include repeatedly used bases); N_{Pi} is the total base number of the i-th time used by other palindromes, N_{Gi} and N_{Ci} are the number of base G and base C respectively of the i-th time used by other palindromes, and M is the repeat times. For example, AUGCUAnGCAUnnnU-AGC is a protein coding sequence which has 18 bases and contains two palindrome structures. The first one is AUGC-nnnGCAU and the second one is GCUAnUAGC. The base G and C in the left arm of the first palindrome structure is used one time by the second palindrome structure. So for this protein coding sequence, N_{P0} = 14, N_{G0} = 3, N_{C0} = 3, M = 1, N_{Gi} = 1, N_{Ci} = 1 and N_{Pi} = 2. P_{GC} in this sequence is P_{GC} = ((3+3) +1+1)/(14+2) = 0.50.

The values of P_{GC} are also normalized just as the other P_{ave} (average amino acid property) values. The definition of P_{GC} is different from the normal definition of GC content, because of containing the information of the repeated use of bases by palindromes.

In our analysis, the P_{GC} were taken as the fifth parameter for calculating protein folding rates in Gromiha's predicting process. The values of P_{GC} and of four P_{ave} are represented in Table 2.

Linear regression procedures

Gromiha selected four parameters (four P_{ave}) from the information of amino acid sequence to predict the protein folding rates for all-β proteins (Gromiha 2005). The four parameters are listed in Table 2 and their biological meanings are shown in Eq. (5). In our analysis, P_{GC} was taken as the fifth parameter and was added in Gromiha's model. We did the multiple linear regression analysis between the five parameters and the experimental protein folding rates and got a new linear regression equation (called our regression equation). The statistical significance of it has been verified with the jack-knife test and p-value by standard procedure.

Evaluating our predicted results

The protein folding rates were calculated by Gromiha's regression equation and our regression equation respectively. And then the two kinds of calculated results were compared and evaluated.

For the two kinds of calculated results, the population variance, the average absolute difference and the difference between the predicted and experimental values of the protein folding rates were compared.

The population variance is defined as following:

$$\sigma^2 = \sum_{i=1}^{N} (\ln(k_f)_{pi} - \ln(k_f)_{ei})^2 / N \qquad (2)$$

where σ^2 is the population variance, $\ln(k_f)_{pi}$ is predicted protein folding rate and $\ln(k_f)_{ei}$ is experimental protein folding rate for the i-th protein and N is the total number of analyzed proteins.

The average absolute difference and the difference between the predicted and experimental values are separately calculated by Eq. (3) and (4).

Table 3. The predicted protein folding rates and the difference of the 13 all-β proteins for the two models

PDB code	$\ln(k_f)$			D_i (G)	D_i
	$\ln(k_f)_{ei}$	$\ln(k_f)_{pi}$(G)	$\ln(k_f)_{pi}$		
1nyf	4.54	3.18	3.74	−1.36	−0.80
1pks	−1.05	−1.39	−0.85	−0.34	0.20
1shg	1.41	1.72	1.42	0.31	0.01
1srl	4.04	3.68	4.10	−0.36	0.06
1hng	2.89	2.79	3.30	−0.10	0.41
1ten	1.06	1.72	1.14	0.66	0.08
1csp	6.98	7.22	7.37	0.24	0.39
1mjc	5.24	4.47	4.40	−0.77	−0.83
2ait	4.20	5.54	4.67	1.34	0.47
1fnf-9	−0.91	−0.81	−1.33	0.10	−0.42
1fnf-10	5.48	4.67	5.37	−0.81	−0.10
1wit	0.41	1.27	1.36	0.86	0.95
1tit	3.47	3.70	3.10	0.23	−0.37

$\ln(k_f)_{ei}$, experimental protein folding rates; $\ln(k_f)_{pi}$(G) and $\ln(k_f)_{pi}$, predicted protein folding rates calculated by Gromiha's and our equation, respectively; D_i (G) and D_i, difference between the experimental and the predicted protein folding rate for the ith protein, respectively by Gromiha's and our equation, which were calculated by Eq. (4).

$$D_{ave} = \sum_{i=1}^{N} |D_i| / N \tag{3}$$

$$D_i = \ln(k_f)_{pi} - \ln(k_f)_{ei} \tag{4}$$

where D_{ave} is the average absolute difference, N is the total number of analyzed proteins, and D_i is the difference between predicted and experimental protein folding rate for the i-th protein.

Chi-square test was performed to evaluate the difference between predicted and experimental protein folding rates for the two models.

Results

Multiple linear regression equation

Based on the 13 all-β proteins, the multiple regression equation between the experimental protein folding rates and the five parameters was obtained as follows:

$$\ln(k_f) = -8.12P_{GC} - 80.88K^0 + 182.79P_\beta + \\ + 58.81R_a - 144.57\Delta ASA - 7.49 \tag{5}$$

where $\ln(k_f)$ is the predicted protein folding rate, P_{GC} is PGC in the protein coding sequence, K^0 is compressibility, P_β is β-strand tendency, R_α is reduction in solvent accessibility and ΔASA is solvent accessible surface area for protein unfolding.

In the linear regression equation, the p-value for the term P_{GC} is 0.029. For the other four terms, the p-values are lower than 0.003. The results of the partial correlation analysis show that the p-value is also 0.029 for the parameter P_{GC} (see Table 4). Our model also passed jack-knife test, the p-value for the term P_{GC} is 0.047, for the other four terms, the p-values are lower than 0.001. That is to say, the influence of P_{GC} on protein folding rates is significant.

Comparing our results with Gromiha's

The significance of the correlations between predicted and experimental protein folding rates was analyzed (Table 5). The correlation coefficient is 0.99 in our model and 0.96 in Gromiha's (see Fig. 1 and 2). The population variance σ^2 is 0.24 in our model and 0.50 in Gromiha's. The average absolute difference D_{ave} is 0.39 in our model and 0.58 in Gromiha's. The χ^2 value in our model is 2.86 and its p-value is 0.004; while χ^2 value is 3.53 and its p-value is 0.01 in Gromiha's model. The results of jack-knife test show that the correlation coefficient is 0.979 in our model, and 0.956 in Gromiha's model.

The coefficient of parameter P_{GC} is −8.12 (see Eq. 5), its absolute value is smaller than that of the other four parameters, it is consistent with our guesstimate, but the effect of P_{GC} or that of mRNA can not be neglected because the influence of P_{GC} is significant in protein folding process.

Comparing our predicted protein folding rates with Gromiha's, we found that the distinct improvement in our model occurred in the proteins which has greater differences between the experimental and the predicted protein folding rates calculated by Gromiha's model (see Table 3; Fig. 1 and 2), such as the protein 1nyf, 1ten, 2ait, and 1fnf-10. These distinct improved values indicate that

Table 4. The results of partial correlations between the folding rates and each parameter for the 13 all-β proteins

Variable 1	Variable 2	r_{12}
$\ln(k_f)$	K^0	-0.953^{***}
$\ln(k_f)$	P_β	0.971^{***}
$\ln(k_f)$	R_α	0.867^{**}
$\ln(k_f)$	ΔASA	-0.968^{***}
$\ln(k_f)$	P_{GC}	-0.719^{*}

r_{12}, partial correlation coefficients. Two-tailed significance: $^{*}p \leq 0.05$; $^{**}p \leq 0.01$; $^{***}p \leq 0.001$.

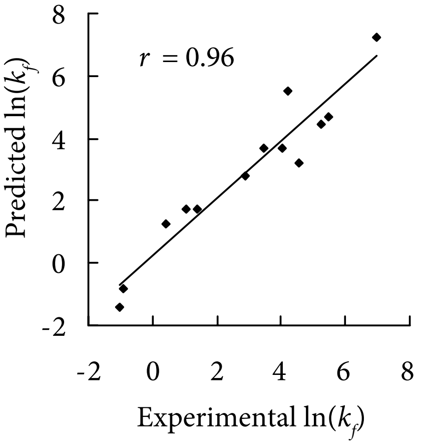

Figure 1. Correlation between the experimental and the predicted values of protein folding rate calculated by Gromiha's model.

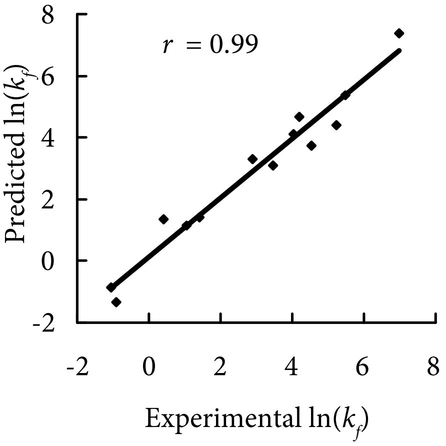

Figure 2. Correlation between the experimental and the predicted values of protein folding rate calculated by our proposed model.

P_{GC} is likely to play an important role in regulating protein folding.

The significance analysis shows that all of the indexes were improved while compared with the Gromiha's results and the parameter P_{GC} is significant. We can conclude that protein folding process is influenced by P_{GC}.

Discussion

The effect of mRNA structure complexity

Our research shows that mRNA sequence can influence protein folding rates. Its biological meaning can be explained by the process from mRNA to protein. Theoretically speaking, the protein folding rates, protein structures as well as the final functions of proteins are influenced by the uneven synthesis rates of nascent peptide chains, and the uneven synthesis rates of nascent peptide chains are affected by the uneven elongation rates of mRNA sequences along with ribosome because of the co-translational protein folding (Purvis et al. 1987; Krasheninnikov et al. 1988). Otherwise, the complex and variability of mRNA structures are the reasons of making an uneven elongation rate of mRNA along with ribosome, thus, the protein folding rates may be influenced by the complex and variability of mRNA structures.

Table 5. The results of significance analysis for the 13 all-β proteins

	r	χ^2 (p-value)	σ^2	D_{ave}
Gromiha model	0.96	3.53 (0.01)	0.50	0.58
Our model	0.99	2.86 (0.004)	0.24	0.39

r, correlation coefficient between the experimental and the predicted protein folding rates; χ^2 (p-value), result of chi-square test; σ^2, population variance; D_{ave}, average values of the absolute difference between the experimental and the predicted protein folding rates, which were calculated by Eq. (3).

Palindromes are one of the factors of forming the complex and variability of mRNA structures. The number or the length of stem structures makes the potential complexity of mRNA secondary structures or higher structures, while the bases in a palindrome which are also used by other palindromes in mRNA sequence make the potential variability of mRNA structures. So the protein folding process should be affected by palindromes.

To verify the theoretical speculation, the parameter P_{GC} was substituted with the overall GC content of the protein coding sequence ($C_{GC} = (N_{G0}+N_{C0})/N_{P0}$), we did multiple linear regression analysis again based on the same protein group (see Table 6).

The results show that the p-value for the term C_{GC} is 0.047, but, the p-value in jack-knife test is 0.087 for the term of C_{GC}. It means that, comparing with P_{GC} (p = 0.029 and jack-knife test p = 0.047), C_{GC} has a weaker correlation with protein folding rates. It is obvious that there is minor significance with regard to the parameter C_{GC} excluding the structure information of mRNA.

Otherwise, we also calculated the proportion of the palindrome sequence to the overall protein coding sequence in the 13 protein coding sequences. More than 83% of protein coding sequences belong to the palindrome region, so there is slight effect of the non-palindrome region to the C_{GC} values. That is to say, the influence of the normal defined GC content in palindrome region is nearly the same as that of C_{GC}. Thus, the difference between the influence of C_{GC} and that of P_{GC} indicated that an important influential factor containing in

P_{GC} is the palindrome structure. Considering the definition of P_{GC} is different from the normal GC content, we deduced that the complex of mRNA structures may play a key role in protein folding process.

The effect of biased synonymous codon usage

Parameters P_{GC} or C_{GC} also include the information of base distribution in protein coding sequence. Which site in codon is important in protein folding process? We know that the first two bases of codon determine the property of an amino acid and relate to the translate rule, but the third base reflects the synonymous codon usage. We guess the biased usage of third base also affects the protein folding. The problem will be discussed as follows.

Because it is difficult to analyze the base distributions of each site of codon by P_{GC} and C_{GC} has a minor correlation with protein folding rates, we used C_{GC} to discuss this problem. For each protein coding sequence, we calculated the GC contents in each site of codons, named $C^1{}_{GC}$, $C^2{}_{GC}$ and $C^3{}_{GC}$, and calculated the values of $C^{1,2}{}_{GC}$, $C^{1,3}{}_{GC}$ and $C^{2,3}{}_{GC}$. Where $C^{1,2}{}_{GC}$ is the GC content in which the third base of codons is excluded, and the definitions of $C^{1,3}{}_{GC}$ and $C^{2,3}{}_{GC}$ are similar to that of $C^{1,2}{}_{GC}$. From the above regression analysis, we added the 6 parameters respectively as the fifth parameter in Gromiha's model and did multiple linear regression analysis separately. The results are shown in Table 6. The p-value is larger than 0.05 for the term $C^1{}_{GC}$, $C^2{}_{GC}$, $C^{1,2}{}_{GC}$, and $C^{1,3}{}_{GC}$ in the corresponding multiple linear regression equation. For the term $C^3{}_{GC}$ and $C^{2,3}{}_{GC}$, the p-values are 0.05 and 0.04, respectively. It indicates that there is a weaker correlation with the protein folding rates for both $C^3{}_{GC}$ and $C^{2,3}{}_{GC}$, but the parameter $C^1{}_{GC}$, $C^2{}_{GC}$, $C^{1,2}{}_{GC}$, and $C^{1,3}{}_{GC}$ do not show significant influence on the protein folding rates. Comparing the coefficient of the 6 GC contents (see Table 6), we found that the coefficient of $C^3{}_{GC}$ is larger than of $C^1{}_{GC}$ and $C^2{}_{GC}$, and the coefficient of $C^{1,3}{}_{GC}$ and $C^{2,3}{}_{GC}$ are larger than that of $C^{1,2}{}_{GC}$, once again, the results show that, compared with the first two sites, the third site play a substantial function in the parameter of P_{GC} or C_{GC}. And the results are consistent with the conclusions in our earlier related work (Li and Li 2010). The distributions of first two bases of codons do not show significant influence on protein folding. So we conclude that the GC content of the third base and its correlation with the adjacent bases might be another main factor that influences the protein folding rates, and this influence comes from mRNA rather than from amino acid sequence.

Table 6. The results of multiple linear regression and partial correlations between the folding rates and the fifth parameter for the 13 all-β proteins

The fifth parameter	a	p	r_{12}
P_{GC}	−8.12	0.029	−0.719
C_{GC}	−12.87	0.047	-0.67^*
$C^1{}_{GC}$	−1.52	0.72	−0.14
$C^2{}_{GC}$	−2.08	0.84	−0.08
$C^3{}_{GC}$	−4.90	0.05	-0.66^*
$C^{1,2}{}_{GC}$	−4.94	0.61	−0.20
$C^{1,3}{}_{GC}$	−7.11	0.08	−0.60
$C^{2,3}{}_{GC}$	−10.55	0.04	-0.70^*

P_{GC}, GC content of palindromes in protein coding sequence; C_{GC}, GC content of protein coding sequence; $C^1{}_{GC}$, $C^2{}_{GC}$ and $C^3{}_{GC}$, GC contents in each site of codons; $C^{1,2}{}_{GC}$, GC content of first and second sites of codons (the definition of $C^{1,3}{}_{GC}$ and $C^{2,3}{}_{GC}$ are similar with $C^{1,2}{}_{GC}$); a , coefficient of the fifth terms; p, significance level in the corresponding multiple linear regression; r_{12}, partial correlation coefficients between $\ln(k_f)$ and each GC content. Two-tailed significance: * $p \le 0.05$.

Table 7. The correlation coefficient of each parameter with GC content

	P_{GC}	C_{GC}	$C^3{}_{GC}$	$C^{2,3}{}_{GC}$	$\ln(k_f)_{ei}$
K^0	0.493	0.400	0.150	0.356	−0.017
P_β	0.626^*	0.647^*	0.523	0.693^{**}	0.235
R_α	-0.696^{**}	-0.683^{**}	-0.702^{**}	-0.776^{**}	0.346
ΔASA	−0.199	−0.222	−0.049	−0.061	−0.323
$\ln(k_f)_{ei}$	−0.119	0.059	−0.036	−0.058	1.000

Two-tailed significance: * $p \le 0.05$; ** $p \le 0.01$. (See abbreviations in Table 6).

Relations between GC content and other four protein parameters

The parameters K^0, P_β, R_α and ΔASA from amino acid sequence are vital factors of influencing the protein folding rates, and from the above discussion, we found that the parameters P_{GC}, C_{GC}, $C^3{}_{GC}$ and $C^{2,3}{}_{GC}$ from mRNA sequence are also related with the protein folding rates. So we think there must be some relations between the two kinds of parameters, and these relations can uncover the detailed route of the GC content influence. In order to track the detailed route of the influence of GC content, we did the linear regression analysis directly between each of the Gromiha's parameter ($\ln(k_f)$, K^0, P_β, R_α and ΔASA) and each of the GC content (P_{GC}, C_{GC}, $C^3{}_{GC}$ and $C^{2,3}{}_{GC}$) separately, the results are shown in Table 7. P_{GC}, C_{GC}, $C^3{}_{GC}$ and $C^{2,3}{}_{GC}$ are all correlated negatively with R_α (reduction in solvent accessibility of the protein sequence), and P_{GC}, C_{GC} and $C^{2,3}{}_{GC}$ are correlated positively with P_β (β-strand tendency of the protein sequence). So we think P_{GC} is likely to influence the protein folding rates through impacting on the reduction in solvent accessibility or the β-strand tendency of the protein sequence. Otherwise, the results of the partial correlation analysis and the jack-knife test indicated that the 5 parameters (K^0, P_β, R_α, ΔASA and P_{GC}) are all individual and valid (see Table 4). It means that parameter P_{GC} is an individual and influential parameter in protein folding process. And we noticed that, for the 5 parameters (K^0, P_β, R_α, ΔASA and P_{GC}), each of them is not directly correlated with $\ln(k_f)$, which shows that the present indirect method is necessary for analyzing the influence of protein coding sequence.

In summary, we proposed an indirect method to verify that P_{GC} from mRNA sequence is a valid parameter of influencing the protein folding rates. The essential information included in P_{GC} comes mainly from the complexity and variability of the mRNA structures and from the third base usage of codons. It suggests that mRNA sequence plays a key role in regulating protein folding.

Although P_{GC} is a simple parameter to represent mRNA information, its influence is significant. If we can find better parameters to represent the information of mRNA, such as the information of base correlation in palindromes, we believe that more detailed and clear relations will be discovered between mRNA sequences and protein folding rates.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (30660044) and the Ph.D. Programs Foundation of Ministry of Education of China (20050126003).

References

Adzhubei I. A., Adzhubei A. A., Neidle S. (1998): An integrated sequence-structure database incorporating matching mRNA sequence, amino acid sequence and protein three-dimensional structure data. Nucleic Acids Res. **26,** 327–331 doi:10.1093/nar/26.1.327

Alm E., Baker D. (1999): Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc. Natl. Acad. Sci. U.S.A **96,** 11305–11310 doi:10.1073/pnas.96.20.11305

Baker W., Van Den Broek A., Camon E., Hingamp P., Sterk P., Stoesser G., Tuli M. A. (2000): The EMBL nucleotide sequence database. Nucleic Acids Res. **28,** 19–23 doi:10.1093/nar/28.1.19

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T., Weissig N. H., Shindyalov I. N., P Bourne. E. (2000): The protein databank. Nucleic Acids Res. **28,** 235–242 doi:10.1093/nar/28.1.235

Brunak S., Engelbrecht J. (1996): Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. Proteins **25,** 237–252 doi:10.1002/(SICI)1097-0134(199606)25:2<237::AID-PROT9>3.3.CO;2-Y

Chiusano M. L., Alvarez-Valin F., Di Giulio M., D'Onofrio G., Ammirato G., Colonna G., Bernardi G. (2000): Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene **261,** 63–69 doi:10.1016/S0378-1119(00)00521-7

Chew D. S., Choi K. P., Leung M. Y. (2005): Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. Nucleic Acids Res. **33,** e134 doi:10.1093/nar/gni135

Chou K. C., Shen H. B. (2009): FoldRate: A web-server for predicting protein folding rates from primary sequence. Open Biol. J. **3,** 31-50

Debe D. A., Goddard W. A., 3rd. (1999): First principles prediction of protein folding rates. J. Mol. Biol. **294,** 619–625 doi:10.1006/jmbi.1999.3278

Dill K. A., Fiebig K. M., Chan H. S. (1993): Cooperativity in protein-folding kinetics. Proc. Natl. Acad. Sci. U.S.A **90,** 1942–1946 doi:10.1073/pnas.90.5.1942

Dinner A. R., Karplus M. (2001): The roles of stability and contact order in determining protein folding rates. Nat. Struct. Biol. **8,** 21–22 doi:10.1038/83003

Fiebig K. M., Dill K. A. (1993): Protein core assembly processes, J. Chem. Phys. **98,** 3475–3487 doi:10.1063/1.464068

Galzitskaya O. V., Garbuzynskiy S. O. (2006): Entropy capacity determines protein folding. Proteins **63,** 144–154 doi:10.1002/prot.20851

Gong H., Isom D. G., Srinivasan R., Rose G. D. (2003): Local secondary structure content predicts folding rates for simple, two-state proteins. J. Mol. Biol. **327,** 1149–1154 doi:10.1016/S0022-2836(03)00211-0

Gromiha M. M., Selvaraj S. (2001): Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. **310,** 27–32 doi:10.1006/jmbi.2001.4775

Gromiha M. M. (2005): A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J. Chem. Inf. Model **45,** 494–501 doi:10.1021/ci049757q

Gromiha M. M., Thangakani A. M., Selvaraj S. (2006): FLOD-RATE: Prediction of protein folding rates from amino acid sequence. Nucleic Acids Res. **34,** W70-74 doi:10.1093/nar/gkl043

Gu W. J., Zhou T., Ma J. M., Sun X., Lu Z. H. (2004): The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. Biosystems **73,** 89–97 doi:10.1016/j.biosystems.2003.10.001

Gupta S. K., Majumdar S., Bhattacharya T. K., Ghosh T. C. (2000): Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophys. Res. Commun. **269,** 692–696 doi:10.1006/bbrc.2000.2351

Ivankov D. N., Garbuzynskiy S. O., Alm E., Plaxco K. W., Baker D., Finkelstein A. V. (2003): Contact order revisited: influence of protein size on the folding rate. Protein Sci. **12,** 2057–2062 doi:10.1110/ps.0302503

Ivankov D. N., Finkelstein A. V. (2004): Prediction of protein folding rates from the amino acid sequence-predicted

secondary structure. Proc. Natl. Acad. Sci. U.S.A. **101,** 8942–8944
doi:10.1073/pnas.0402659101

Komar A. A. (2009): A pause for thought along the co-translational folding pathway. Trends. Biochem. Sci. **34,** 16–24
doi:10.1016/j.tibs.2008.10.002

Krasheninnikov I. A., Komar A. A., Adzhubeĭ I. A. (1988): Role of the rare codon clusters in defining the boundaries of polypeptide chain regions with identical secondary structures in the process of co-translational folding of proteins. Dokl. Akad. Nauk. SSSR. **303,** 995–999 (in Russian)

Kuznetsov I. B., Rackovsky S. (2004): Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. Proteins **54,** 333–341
doi:10.1002/prot.10518

Li R. F., Li H. (2010): Study on the influences of palindromes in protein coding sequences on the folding rates of peptide chains. Protein Pept. Lett. **17,** 881-888
doi:10.2174/092986610791306652

Makhoul C. H., Trifonov E. N. (2002): Distribution of rare triplets along mRNA and their relation to protein folding. J. Biomol. Struct. Dyn. **20,** 413–420

Mathews D. H., Sabina J., Zuker M., Turne D. H. (1999): Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288,** 911–940
doi:10.1006/jmbi.1999.2700

Mirny L., Shakhnovich E. (2001): Protein folding theory: from lattice to all-atom models. Annu. Rev. Biophys. Biomol. Struct. **30,** 361–396
doi:10.1146/annurev.biophys.30.1.361

Mounoz V., Eaton W. A. (1999): A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc. Natl. Acad. Sci. USA **96,** 11311–11316
doi:10.1073/pnas.96.20.11311

Mukhopadhyay P., Basak S., Ghosh T. C. (2007): Synonymous codon usage in different protein secondary structural classes of human genes: implication for increased non-randomness of GC3 rich genes towards protein stability. J. Biosci. **32,** 947–963
doi:10.1007/s12038-007-0095-z

Nag D. K., Kurst A. (1997): A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast saccharomyces cerevisiae. Genetics **146,** 835–847

Nölting B., Schälike W., Hampel P., Grundig F., Gantert S., Sips N., Bandlow W., Qi P. X. (2003): Structural determinants of the rate of protein folding. J. Theor. Biol. **223,** 299–307
doi:10.1016/S0022-5193(03)00091-2

Ouyang Z., Liang J. (2008): Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci. **17,** 1256–1263
doi:10.1110/ps.034660.108

Plaxco K. W., Simons K. T., Baker D. (1998a): Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. **277,** 985–994
doi:10.1006/jmbi.1998.1645

Plaxco K. W., Baker D. (1998b): Limited internal friction in the rate-limiting step of a two-state protein folding reaction. Proc. Natl. Acad. Sci. U.S.A. **95,** 13591–13596
doi:10.1073/pnas.95.23.13591

Punta M., Rost M. (2005): Protein folding rates estimated from contact predictions. J. Mol. Biol. **348,** 507–512
doi:10.1016/j.jmb.2005.02.068

Purvis I. J., Bettany A. J., Santiago T. C. (1987): The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. J. Mol. Biol. **193,** 413–417
doi:10.1016/0022-2836(87)90230-0

Shao H., Zeng Z. H. (2003): A sequence function reveals new features in beta-protein folding. Protein Pept. Lett. **10,** 435–439
doi:10.2174/0929866033478690

Thanaraj T. A., Argos P. (1996): Ribosome-mediated translational pause and protein domain organization. Protein Sci. **5,** 1594–1612
doi:10.1002/pro.5560050814

Varenne S., Buc J., Lloubes R., Lazdunski C. (1984): Translation is non-uniform process -effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. **180,** 549–576
doi:10.1016/0022-2836(84)90027-5

Zhang L., Li X. J., Jiang Z. T., Xia A. (2003): Folding rate prediction based on neural network model. Polymer **44,** 1751–1756
doi:10.1016/S0032-3861(03)00021-1

Zhou H., Zhou Y. (2002): Folding rate prediction using total contact distance, Biophys. J. **82,** 458–463
doi:10.1016/S0006-3495(02)75410-6

Received: November 5, 2010
Final version accepted: January 18, 2011