Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 343 (2006) 177–182
BBRC
www.elsevier.com/locate/ybbrc

The relation between mRNA folding and protein structure

Mengwen Jia [a,b], Liaofu Luo [a,*]

[a] Laboratory of Theoretical Biophysics, Faculty of Science and Technology, Inner Mongolia University, Hohhot 010021, China
[b] MOE Key Laboratory of Bioinformatics, Department of Automation, Tsinghua University, Beijing 100084, China

Received 19 February 2006
Available online 3 March 2006

Abstract

About 200 mRNA sequences of *Escherichia coli* and human with matching protein secondary structure data were studied. The mRNA folding for each native sequence and for corresponding randomized sequences was calculated through free energy minimization. We have found that the folding energy of mRNA segments in different protein secondary structures is significantly different. The average Z score is more negative for regular secondary structure (α-helix and β-strand) than that for coil. This suggests that the codon choice in native mRNA sequence coding for protein regular structure contributes more to the mRNA folding stability.

Keywords: mRNA sequence; Folding energy; Protein secondary structure; Codon choice

Classical studies show that for many proteins, the information required for folding a protein is contained in the amino acid sequence. However, several authors recently indicated the possible connections between protein secondary structure and genetic information stored in mRNA. Statistical analyses have shown the codon usage correlated to protein secondary structures [1–7]. Localized secondary structures of single-stranded mRNA may play important functional roles in translational regulation and gene expression [8–12]. If mRNA secondary structure does effect the translation process, then it is reasonable to infer that the mRNA structure may exert some influence on the formation of protein secondary structure since regular secondary structures occur in the very early stage of the nascent peptide folding [13,14]. The possible influence of mRNA stem–loop frequency on protein secondary structure has been indicated from the study of up-to-date sequence-structure data [15]. The mRNA structure is determined in principle by the minimization of the free energy of the molecule. The aim of this study was to give a statistical analysis of mRNA folding energy contributed from codons in different protein secondary structures. We shall demonstrate that the codon choice in native mRNA sequence coding for protein regular secondary structure contributes more to the mRNA folding stability than that for protein coil region.

Materials and methods

Data. The gene sequences and structural data of two species, *Escherichia coli* and human, were analyzed. Starting from ASTRAL1.50 [16], we have set up an integrated sequence-structure database, called IADE2 [15] (Integrated ASTRAL-DSSP [17]-EMBL [18]). The database incorporates matching mRNA sequence, amino acid sequence, and protein secondary structure data of 107 *E. coli* and 125 human proteins with less than 40% sequence identity to each other. Their PDB names and corresponding EMBL accession numbers can be found in Table 1 of [15].

mRNA folding energy and secondary structure. Both secondary structure and folding free energy for a native mRNA sequence were calculated using RNAfold [19–21]. For comparison the folding free energies of 50 randomized sequences corresponding to a native mRNA sequence have also been calculated. Two ways of folding were used in this study. First, the native mRNA sequence and corresponding randomized sequences are folded in whole length (called "whole folding"). Second, the mRNA sequence is folded in a local window pattern (called "windows folding"). In local windows folding, the sequence is folded in short regions of 50 bases and shifted by 10 bases [22].

Randomization of a native mRNA sequence. For each native mRNA sequence, the random sequences were produced by use of Codonrandom (CODRAN) randomization methods, which preserves the same encoded amino acid sequence of mRNA sequence under the random codon choice.

* Corresponding author. Fax: +86 471 4951761.
E-mail address: lfluo@mail.imu.edu.cn (L. Luo).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.02.135

Table 1
Average Z score for *E. coli* and human

	Number of sequence	Average free energy of native sequences	Average free energy of random sequences	Z score
E. coli	107	−222.0	−206.2	−1.69
Human	125	−149.2	−132.4	−2.26

The folding free energy of native sequence and randomized sequence and the corresponding Z score (defined by Eq. (1)) for each gene are calculated. The average values for 107 *E. coli* genes and 125 human genes are listed.

In this procedure, the randomized codon is selected from its synonymous codon family in equal frequency. Usually, the randomization is processed in whole native mRNA sequence of a gene. However, to consider the difference between regular and irregular secondary structure of the encoded protein we should only randomize part codons in an mRNA sequence of a gene. Two different types of randomization sequences were produced based on CODRAN method. First, the native mRNA sequence is partly randomized for a given percentage of codons (termed 'Part-Random'). Second, the native mRNA sequence is partly randomized for all codons in regular protein secondary structure segments (termed 'Reg-Random') or for all codons in coil segments (termed 'Coi-Random'). That is, only those codons which code for protein regular or coil structure should be randomized, but others remain unchanged.

Z score value. The energy difference between native and randomized sequence is measured by Z score. Z score is defined by

$$Z = \{E_{\text{native}} - \langle E_{\text{random}} \rangle\}/\text{STD}, \quad (1)$$

where $\langle E_{\text{random}} \rangle$ means the energy of randomized sequence averaged over a large number of (usual 50 in this work) samples generated from the native sequence and STD means its standard deviation. When the regular secondary structure (coil) segments are randomized, $\langle E_{\text{random}} \rangle$ means the energy averaged over a large number of Reg-Random (Coi-Random) sequences, and the corresponding Z score calculated from Eq. (1) is denoted by Z^{reg} (or Z^{coil}). Likewise, when Part-Random sequences with the same randomized percentage as the regular secondary (coil) structure segments in native sequence are calculated the corresponding Z score can be served as a control of Z^{reg} (or Z^{coil}) and is denoted by $Z^{\text{reg}}_{\text{ctrl}}$ (or $Z^{\text{coil}}_{\text{ctrl}}$).

Results and discussions

Energy Z score for E. coli and human

For 125 human and 107 *E. coli* genes the average Z score values are listed in Table 1, and the histograms of Z score distribution shown in Fig. 1. The average Z score value is −1.69 for *E. coli* and −2.26 for human, respectively. The Z score values of each mRNA sequence for human and *E. coli* are given in supplementary material (Tables A1 and A2). The above energy calculation given in Table 1 was completed by using whole folding. As a comparison, we have also calculated the Z score in windows folding pattern. Both for *E. coli* and human, the results of windows folding show the basically same trend with whole folding (see below).

Consider the Z score distribution of random sequences instead of native mRNA. It is easily proved that they obey a normal distribution with mean 0 and standard deviation 1. For a set of 107 or 125 random sequences the Z score obeys the $N(0, 1/\sqrt{107})$ or $N(0, 1/\sqrt{125})$ distribution, respectively, and, at 1% significant level, the offset is $-2.33/\sqrt{107} = -0.223$ or $-2.33/\sqrt{125} = -0.208$. Comparing with Z score distribution of random sequences we find that the above Z scores for human and *E. coli* genes are negative enough and can conclude that the averagely more negative free energy of native sequences than random samples is very significant.

The dependence of energy Z score of mRNA sequence on its encoding protein secondary structure

To explore the folding property of mRNA sequence in different segments, namely, segment coding for protein regular structure (α-helix and β-strand) and that coding for coil, we have used Reg-Random, Coi-Random, and Part-Random randomization procedure. The free energy of a randomized sequence depends on the number of codons that have been randomized. To consider the background of random percentage, in calculating Z^{reg} or Z^{coil} by use of Reg-Random or Coi-Random, we always employ Part-Random to calculate $Z^{\text{reg}}_{\text{ctrl}}$ or $Z^{\text{coil}}_{\text{ctrl}}$ as a control. In all these calculations, only windows folding was adopted to ease the computational intensity.

The detailed calculation results on Z^{reg} and Z^{coil} and their control values for 125 human and 107 *E. coli* genes are given in supplementary material (Tables A1 and A2).

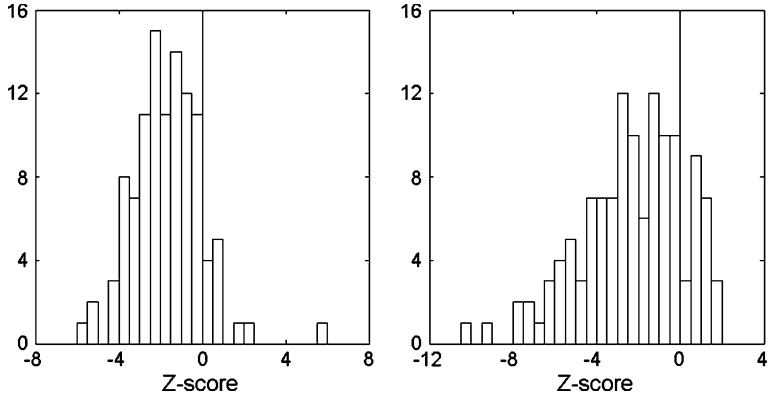

Fig. 1. The histograms of Z score distribution for *E. coli* and human. The left figure refers to Z score distribution of 107 *E. coli* genes and the right figure refers to Z score distribution of 125 human genes.

The average Z scores for mRNA sequences in different protein secondary structures are shown in Table 2. From Table 2 we find for *E. coli* the mean Z score of regular structure is −1.38 with control −1.01 (difference 0.37), and for human, the two values are −1.71 and −1.24, respectively (difference 0.47). So, both for *E. coli* and human, the Z scores of regular structure are explicitly lower than their control values. However, the case is different for coil region. For *E. coli*, the average Z score of coil region is −0.93, very near the control value −0.86 (difference 0.07), and for human, the Z score of coil is −1.03, even slightly larger than its control −1.20 (difference −0.17). So, we conclude the mRNA folding energy in protein regular structure (α-helix and β-strand) is statistically lower than that in randomized sequence, but for irregular structure (coil) no such conclusion can be deduced.

The detailed difference between Z scores in regular and irregular structures can be plotted in a diagram. We study the distributions of $Z^{\text{reg}} - Z^{\text{reg}}_{\text{ctrl}}$ and $Z^{\text{coil}} - Z^{\text{coil}}_{\text{ctrl}}$ for 107 *E. coli* and 125 human genes. The results are shown in Fig. 2. Evidently, both for *E. coli* and human the maximum distributions of Z score difference in regular structure are located at some values smaller than zero, shifted towards left as compared with those of irregular structure. These results indicate obviously that the mRNA sequence coding for protein regular structure has more negative free energy (relative to randomized samples) than the sequence segment coding for coil.

A more detailed comparison of Z scores for regular structure and irregular structure can be established through statistical test (see supplementary material). Consider 107 (125) Z^{reg}'s and $Z^{\text{reg}}_{\text{ctrl}}$'s for *E. coli* (human) as two sets of variables separately. We calculated the square deviation for each set and their ratio (F value). The F values for *E. coli* and human are shown in the lower panel of Table 3. They both exceed the percentile of F distribution, $F_{0.975}$. This means that the difference of square deviation between two sets is significant and Aspin-Welch t test should be used in comparing the averages of two sets of variables. The t value and the degree of freedom df in Aspin-Welch t test were calculated and are shown in the upper panel of Table 3. For regular structure the calculated t values are larger than $t_{0.975}$. So, the NULL hypothesis that the Z scores of regular structure, Z^{reg}, are similar to those of the control set, $Z^{\text{reg}}_{\text{ctrl}}$, should be rejected. The folding free energy of native mRNA sequence segment coding for protein regular structure (relative to the energy of randomized sequence) is significantly different from the control set both for *E. coli* and human. Likewise, introducing Z^{coil} and $Z^{\text{coil}}_{\text{ctrl}}$ for 107 (125) *E. coli* (human) genes as two sets of variables we find for irregular structure, both t-values (absolute value) for *E. coli* and human are smaller (see Table 3) and the differences of two Z scores, Z^{coil} and $Z^{\text{coil}}_{\text{ctrl}}$, are not significant.

Table 2
The dependence of energy Z score of mRNA on protein secondary structure

	Average Z^{reg}	Average $Z^{\text{reg}}_{\text{ctrl}}$	Average Z^{coil}	Average $Z^{\text{coil}}_{\text{ctrl}}$
E. coli	−1.38	−1.01	−0.93	−0.86
Human	−1.71	−1.24	−1.03	−1.20

The average Z scores for mRNA sequence segments in protein regular structure (Z^{reg}) and protein irregular structure (Z^{coil}), and their controls ($Z^{\text{reg}}_{\text{ctrl}}$ and $Z^{\text{coil}}_{\text{ctrl}}$) are shown in the table.

Table 3
Aspin-Welch t test and F test for the difference of Z scores

	Regular	Structure	Irregular	Structure
	t value	df	t value	df
Human	−2.0[a]	240	1.02	247
E. coli	−2.04[a]	200	−0.44	193
	F value	df_1, df_2	F value	df_1, df_2
Human	1.43[b]	124	1.10	124
E. coli	1.66[b]	106	1.91[b]	106

df, degree of freedom. When F value is large enough, Aspin-Welch t test for average Z scores should be used.

[a] Means the calculated t value (absolute value) $> t_{0.975}$ and the difference of average Z scores very significant between Z^{reg} and $Z^{\text{reg}}_{\text{ctrl}}$.

[b] Means the calculated F value $> F_{0.975}$ and the difference of deviations of Z scores very significant between Z^{reg} and $Z^{\text{reg}}_{\text{ctrl}}$ (or Z^{coil} and $Z^{\text{coil}}_{\text{ctrl}}$).

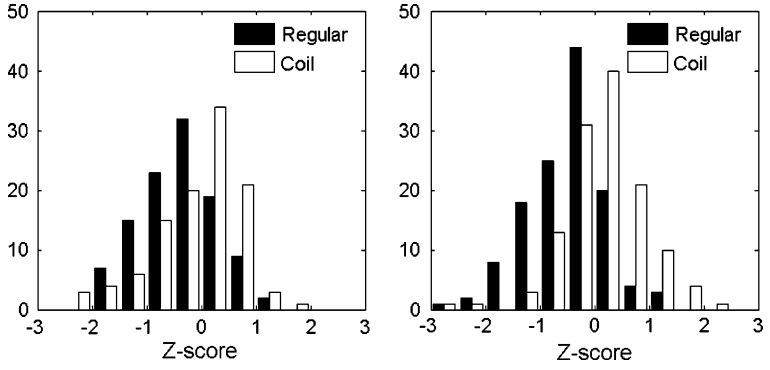

Fig. 2. The histograms for the distribution of $Z^{\text{reg}} - Z^{\text{reg}}_{\text{ctrl}}$ and $Z^{\text{coil}} - Z^{\text{coil}}_{\text{ctrl}}$ for *E. coli* and human. The left figure gives the distribution of $Z^{\text{reg}} - Z^{\text{reg}}_{\text{ctrl}}$ and $Z^{\text{coil}} - Z^{\text{coil}}_{\text{ctrl}}$ for 107 *E. coli* genes and the right figure gives the distribution of $Z^{\text{reg}} - Z^{\text{reg}}_{\text{ctrl}}$ and $Z^{\text{coil}} - Z^{\text{coil}}_{\text{ctrl}}$ for 125 human genes.

We shall give a brief discussion on the biological meaning of the correlation between protein secondary structure and mRNA folding stability. It shows the evolutionary interaction between protein structure–function and nucleic acid. In a previous work [15], we found that the regular structures on proteins tend to be preferentially coded by mRNA stem region while the coils on proteins tend to be preferentially coded by mRNA loop region. The stem/loop Z score is negatively correlated with energy Z score (see supplementary material). The present study on energy Z score of different protein regular structures is consistent with the previous work. Shpaer [8] studied the number of small unbranched hairpins in mRNA during its translation in a polysome. He found that the rare codons translated by minor tRNAs occur significantly more frequently in the position 5′ to the hairpins. This was explained by an assumption that the process of hairpin unfolding can increase the time of translocation from the A to P ribosome site thus decreasing the probability of translational error. That is, to obtain a higher accuracy, the translational process proceeds sometimes fast and sometimes slow to match the codon arrangement in mRNA sequence. More hairpins mean locally stronger folding energy. So the mRNA folding energy correlates to translational accuracy. On the other hand, erroneous translation leads to "wrong assignment" of protein secondary structures. However, the tolerance to translational error is different for three kinds of protein secondary structure. In case of weaker mRNA folding energy, the relatively frequent translational errors make more wrong assignment for coil. This gives a tentative explanation on why we have observed the mRNA folding stability correlating to protein structure [7].

The comparison between different randomization methods

We have calculated energy Z scores for *E. coli* and human based on CODRAN randomization and deduced that the averagely more negative free energy of native sequences than random samples for these two species. The results are consistent with those of Seffens and Digby [25]. But in [25] only 4 sequences of *E. coli* and 15 sequences of human were calculated. Our conclusion is deduced based on a larger database and should be more convincible. On the other hand, as indicated by [22,26], the energy Z score is dependent on randomization method. Different randomization methods were proposed. For example, the Codonshuffle (CODSHU) method shuffles synonymous codon in a given mRNA sequence and preserves the same encoded amino acid sequence and codon usage (base composition) of mRNA sequence [25]; the Dicodonshuffle (DICODSH) method preserves the di-nucleotide composition, codon usage, and encoded amino acid sequence of mRNA sequence [22,26]. CODRAN method preserves di-nucleotide composition at (1,2) position of codons only, while CODSHU preserves di-nucleotide composition at (1,2) and (2,3), positions of codons and DICODSH preserves di-nucleotide composition at (1,2), (2,3), and (3,1) positions of codons. So, DICODSH works under the most strong constraint. If the constraint subjected to randomization is adopted not only from the point of methodology but also from what really occurred in evolutionary process, then the CODRAN randomization is a good choice in coding region computation since the random codon choice under preservation of the same encoded amino acid sequence in CODRAN reflects the influence of nucleotide mutation under the functional constraint of encoded protein. To give more detailed comparison between different randomization approaches, we have also calculated energy Z scores in latter two methods (random sequences were generated by using the programs CODSHU and DICODSH provided by Katz and Burge [22]). The average Z score values for whole folding and windows folding in three randomization methods are listed, respectively, in Table 4. The detailed data are given in supplementary material (Tables A3 and A4).

From Table 4 we notice that the folding free energy in whole folding pattern is, on average, not far from the value obtained in windows folding. So, the folding pattern may have a little effect on the systematical comparison of Z scores in three kinds of randomization. We find that for *E. coli*, based on three randomization methods, the Z scores are −1.69, −0.78, and −0.53 for whole folding and −1.83, −0.60, and −0.43 for windows folding, respectively. Although these Z score values are diverse for different random methods, the overall trend is same, namely, the native sequence has more negative folding free energy than randomized sequences at a higher significance. For human, the average Z scores calculated from CODRAN and CODSHU show the same trend (−2.26 and −0.64 for whole folding and −2.17 and −0.45 for windows folding) but those calculated from DICODSH are near zero [26].

Compare the nucleotide sequence site by site between native mRNA and randomized sequence and define the number (in percentage) of nucleotide difference as bpdif. For *E. coli*, we obtain bpdif 22.4%, 17.2%, and 12% for codonrandom, codonshuffle, and dicodonshuffle, respectively. For human they are 22.1%, 16.9%, and 9.7% respectively. Bpdif indicates the degree of randomness in a randomized procedure. The larger the bpdif the higher the degree of randomness. The parameter is related to the constraint inherent in the randomization. The strong constraint would decrease the degree of randomness of

Table 4
Z score for human and *E. coli* in three randomization methods

Average Z score		Randomized method		
		CODRAN	CODSHU	DICODSH
Human	Whole folding	−2.26	−0.64	−0.04
	Windows folding	−2.17	−0.45	0.005
E. coli	Whole folding	−1.69	−0.78	−0.53
	Windows folding	−1.83	−0.60	−0.43

The Z scores in three randomization methods and in two folding patterns averaged over 125 human genes and 107 *E. coli* genes are listed.

the randomized sequence, and in turn, affect the calculated free energy difference between native and randomized sequence (Z score). For *E. coli*, the average Z scores (-1.69, -0.78, and -0.53 in whole folding or -1.83, -0.60, and -0.43 in windows folding) for three methods are in accordance with the order of bpdif (22.4%, 17.2%, and 12.6%). For human, the average Z scores in three randomization methods (-2.26, -0.64, -0.04 in whole folding or -2.17, -0.45, and 0.005 in windows folding) are also in accordance with bpdif (22.1%, 16.9%, and 9.7%). We notice that the average Z score values for DICODSH are nearly equal to 0. The latter means, although the native sequence has more negative folding energy than random samples for CODRAN and CODSHU randomization, the conclusion for DICODSH is only true for *E. coli*. Now, this apparent inconsistency between two species in DICODSH randomization can be explained by its comparison with bpdif. As indicated above, the bpdifs in codon-random and codonshuffle for *E. coli* (22.4% and 17.2%) are similar to those for human (22.1% and 16.9%), but the bpdif in di-codonshuffle is only 9.7% for human, lower than *E. coli* (12.6%) by three points. The difference in bpdif in DICODSH randomization between human and *E. coli* explains the apparent contradiction of Z score values. It also shows there may exist some difference in native mRNA sequences themselves between these two species.

D_1 and D_2 are informational parameters of DNA sequence (see Supplementary material). Systematical changes of these parameters have been found in DNA sequence evolution [23,24]. $(D_1 + D_2)$ describes the divergence of base pair numbers (neighboring base correlation) from the random value in a DNA sequence. The average $(D_1 + D_2)$ takes 0.102 for human genes but 0.056 for *E. coli*. So, the di-nucleotide composition for *E. coli* mRNA is closer to a random sequence but it is not so for human mRNA. For human only a small portion of bases can be altered in the DICODSH randomization process; so, the DICODSH randomization gives a very small Z score value.

Acknowledgments

The authors are grateful to Drs. Katz and Burge who have sent us the program codonshuffle and dicodonshuffle. The work was supported by National Science Foundation of China, No. 90403010.

Appendix Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.02.135.

References

[1] I.A. Adzhubei, A.A. Adzhuhei, S. Neidle, An integrated sequence-structure database incorporating matching mRNA sequence, amino acid sequence and protein three-dimensional structure data, Nucleic Acids Res. 26 (1998) 327–331.
[2] T. Xie, D.F. Ding, The relationship between synonymous codon usage and protein structure, FEBS Lett. 434 (1998) 93–96.
[3] X.Q. Li, L.F. Luo, C.Q. Liu, The relation between translation speed and protein secondary structure, Acta Biochim. Biophys. 35 (2003) 193–196 (in Chinese).
[4] S. Brunak, J. Engelbrecht, Protein structure and the sequential structure of mRNA: α-helix and β-sheet signals at the nucleotide level, Proteins 25 (1996) 237–252.
[5] M. Oresic, D. Shalloway, Specific correlations between synonymous codon usage and protein secondary structure, J. Mol. Biol. 281 (1998) 31–48.
[6] T.A. Thanaraj, P. Argos, Protein secondary structural types are differentially coded on messenger RNA, Protein Sci. 5 (1996) 1973–1983.
[7] L.F. Luo, M.W. Jia, X.Q. Li, Protein-structure-preference, tRNA copy number and mRNA stem/loop content, Biopolymers 74 (2004) 432–447.
[8] E.G. Shpaer, The secondary structure of mRNA from *E. coli*: its possible role in increasing the accuracy of translation, Nucleic Acids Res. 13 (1985) 275–288.
[9] M.H. de Smit, J. van Duin, Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis, Proc. Natl Acad. Sci. USA 87 (1990) 7668–7672.
[10] Y. Guisez, J. Robbins, E. Remaut, W. Fiers, Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon usage: a hypothesis, J. Theor. Biol. 162 (1993) 243–252.
[11] D.B. Carlini, Y. Chen, W. Stephan, The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr, Genetics 159 (2001) 623–633.
[12] M.W. Jia, Y.D. Li, The relationship among gene expression, folding free energy and codon usage bias in *E. coli*, FEBS Lett. 579 (2005) 5333–5337.
[13] L.F. Luo, Conformation transitional rate in protein folding, Int. J. Quant. Chem. 54 (1995) 243–247.
[14] R.L. Baldwin, G.D. Rose, Is protein folding hierarchic? I. Local structure and peptide folding, Trends Biochem. Sci. 24 (1999) 26–83.
[15] M.W. Jia, L.F. Luo, C.Q. Liu, Statistical correlation between protein secondary structure and messenger RNA stem–loop structure, Biopolymers 73 (2004) 16–26.
[16] J.M. Chandonia, N.S. Walker, L.L. Conte, P. Koehl, M. Levitt, S.E. Brenner, ASTRAL compendium enhancements, Nucleic Acids Res. 30 (2002) 260–263.
[17] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[18] W. Baker, A. van den Broek, E. Camon, et al., The EMBL nucleotide sequence database, Nucleic Acids Res. 28 (2000) 19–23.
[19] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster, Fast folding and comparison of RNA secondary structures, Monatsh. Chem. 125 (1994) 167–188.
[20] M. Zuker, P. Stiegler, Optimal computer folding of large RNA sequence using thermodynamics and auxiliary information, Nucleic Acid Res. 9 (1981) 133–148.
[21] D.H. Mathews, J. Sabina, M. Zucker, H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Mol. Biol. 288 (1999) 911–940.
[22] L. Katz, C.B. Berge, Widespread selection for local RNA secondary structure in coding regions of bacterial genes, Genome Res. 13 (2003) 2042–2051.
[23] L.F. Luo, W.J. Lee, L.J. Jia, F.M. Ji, L. Tsai, Statistical correlation of nucleotides in a DNA sequence, Phys. Rev. E 58 (1998) 861–871.
[24] L.F. Luo, Theoretic-Physical Approach to Molecular Biology, Shanghai Sci. & Tech Publisher, Shanghai, China, 2004.

[25] W. Seffens, D. Digby, mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences, Nucleic Acid Res. 27 (1999) 1578–1584.

[26] C. Workman, A. Krogh, No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution, Nucleic Acid Res. 27 (1999) 4816–4822.