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).

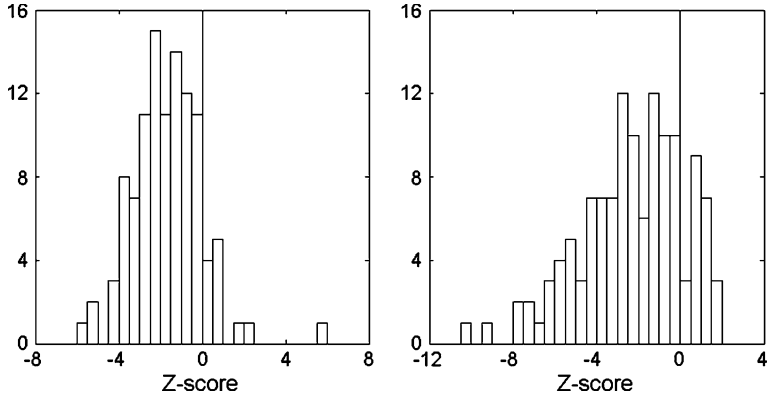

Fig. 1. The histograms of Z score distribution for *E. coli* and human. The left figure refers to Z score distribution of 107 *E. coli* genes and the right figure refers to Z score distribution of 125 human genes.