PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2018, 45(6): 672~678
www.pibb.ac.cn

mRNA 环结构对蛋白质折叠速率的影响 *

李瑞芳 [1)**] 李 宏 [2)] 郭春阳 [1)] 杨萨如拉 [1)]
([1)] 内蒙古师范大学物理与电子信息学院，呼和浩特 010022；[2)] 内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要 前期的相关研究发现 mRNA 二级结构中存在对蛋白质折叠速率的重要影响因素．而 mRNA 二级结构中普遍存在着各种复杂的环结构，这些环结构是否对蛋白质折叠速率也有重要的影响呢？不同的环结构对蛋白质折叠速率的影响是否相同呢？基于此想法，建立了一个包含 mRNA 内部环、发夹环、膨胀环和多分支环等环结构信息和相应蛋白质折叠速率的数据库．对于数据库中的每一个蛋白质，计算了 mRNA 二级结构中各种环结构碱基含量、配对碱基含量及单链碱基含量等参量，分析了各参量与相应蛋白质折叠速率的相关性．结果显示，各种环结构碱基含量与蛋白质折叠速率均呈极显著或显著正相关．说明 mRNA 环结构对蛋白质折叠速率有重要的影响．进一步，把蛋白质按照不同折叠类型或不同二级结构类型分组后，对每一组蛋白质重复上述的分析工作．结果表明，对不同类蛋白质，mRNA 的各种环结构对其相应蛋白质折叠速率的影响存在着显著差异．上述研究将为进一步开展有关 mRNA 和蛋白质折叠速率的研究奠定理论基础．

关键词 mRNA 二级结构，内部环，发夹环，膨胀环，多分支环，蛋白质折叠速率
学科分类号 Q61，Q7 **DOI**: 10.16476/j.pibb.2017.0473

众所周知，蛋白质只有折叠成特有的空间结构才能具有生物活性，才能行使其特定的生物学功能．所以，蛋白质折叠机理的研究是蛋白质分子设计及蛋白质工程的需要，也是越来越多基因工程产物复性复活的需要．而且也有研究表明多种疾病与蛋白质的错折叠有关[1]．所以蛋白质折叠机理的研究也是理解与错折叠相关疾病起源的需要．近期，蛋白质复杂的折叠过程得到了很多研究者的关注[2-5]，蛋白质的折叠大致可分解为以下两个方面的问题[6-7]：一方面是寻找蛋白质的折叠驱动力，即蛋白质折叠的热力学问题．另一方面是解释新生肽链迅速折叠成天然构象状态的机理，即蛋白质折叠的能力学或折叠速率问题．

关于蛋白质折叠速率，目前已经有一些较成熟的理论研究方法，早期的相关研究是基于蛋白质的三级结构展开的[8-11]．然后，基于蛋白质的二级结构，一些影响蛋白质折叠速率的参量相继被提出[12-14]．近年来，出现了很多基于氨基酸序列来预测蛋白质折叠速率的工作[15-20]．所有这些相关研究均有力地促进了蛋白质折叠速率研究的发展．然而，近年来，人们认识到，蛋白质折叠过程从新生肽链的生成和延长时就开始了，这种边翻译边折叠的过程被称为蛋白质的共翻译折叠[21]．由此可以断定，核糖体在 mRNA 上的移位速率必定会影响到新生肽链的增长速率，进而会影响、甚至决定蛋白质的折叠速率[22-28]．而核糖体在 mRNA 上的移位速率必定受 mRNA 的空间结构的影响，因此，蛋白质折叠速率与其相应 mRNA 二级结构之间必定存在某些相关性．相关研究也将成为蛋白质折叠速率研究中的一个重要组成部分．我们先前的研究表明，mRNA 中描述回文结构的一些参量对蛋白质折叠速率有重要的影响[26]，mRNA 的柔性和折叠自由能与蛋白质折叠速率之间有显著的相关性，当把 mRNA 的二级结构按照环结构和茎结构初略分类时，发现 mRNA 中环结构碱基含量、茎结构碱基含量与蛋白质折叠速率的相关性截然相反[28]．我们注意到

* 国家自然科学基金 (31260219)，教育部博士点基金 (20121501110006)，内蒙古自治区自然科学基金 (2016MS0362)和内蒙古高等学校科学技术研究 (NJZY17043) 资助项目．
** 通讯联系人．
Tel: 0471-4392576, E-mail: liruifang@imnu.edu.cn
收稿日期：2017-12-26，接受日期：2018-04-09

mRNA 中有较复杂的环结构，那么各种环结构对蛋白质折叠速率是否有不同的影响呢？基于此想法，本文详细统计了各蛋白质 mRNA 二级结构中内部环、发夹环、膨胀环和多分支环．在此基础上，分析和探讨 mRNA 中各种环结构对蛋白质折叠速率的影响．

1 材料与方法

1.1 数据集

本文的研究样本是我们整理得到的蛋白质折叠数据库中的 99 个蛋白质．蛋白质折叠速率数据来源于蛋白质折叠速率相关文献，相应 mRNA 的序列信息从 PDB(protein data bank)数据库和 EMBL (european molecular biology laboratory)数据库中整理和收集．本文的研究样本按照蛋白质折叠的动力学行为(即折叠类型)分类，包含 59 个二态蛋白质和 40 个多态蛋白质．按照二级结构分类，包含 21 个全 α 类蛋白质、36 个全 β 类蛋白质和 42 个混合类(α-β)蛋白质．相关信息见附件表 S1 和 S2．

1.2 mRNA 二级结构的预测

我们选用 RNAfold 软件来预测每个蛋白质的 mRNA 二级结构．预测出的结果中得到了 mRNA 二级结构中配对的碱基、单链碱基、内部环、发夹环、膨胀环和多分支环等信息．本文主要讨论 mRNA 二级结构中各种环结构对蛋白质折叠速率的影响．因此，首先统计出每个蛋白质相应 mRNA 二级结构中内部环、发夹环、膨胀环和多分支环的碱基含量，然后，为了对比 mRNA 二级结构中环结构与其他结构对蛋白质折叠速率影响的差异，统计了 mRNA 二级结构碱基对含量和单链碱基含量．以此为基础，详细分析 mRNA 二级结构中 4 种环结构碱基含量对蛋白质折叠速率的影响.

1.3 mRNA 二级结构的特征量提取

1.3.1 mRNA 的环结构碱基含量

mRNA 二级结构中环结构包括内部环、发夹环、膨胀环和多分支环．内部环指隔开 2 个螺旋的环区．发夹环是由一对规范的碱基对闭合的一组未配对碱基所组成的结构，环中一般有个数大于 2 的未配对碱基．膨胀环(凸环)是螺旋区 2 条链的其中 1 条链内部有一个或多个不配对的核苷酸构成的结构．多分支环是连接 3 个或 3 个以上的螺旋区的未配对的碱基部分构成的结构．在 mRNA 二级结构的预测结果中得到了各种环结构信息，基于此，同时为了保证不同长度蛋白质分析结果的可比性，我们按照如下方式定义了 4 种环结构碱基含量：

$$\widetilde{N}_i=\frac{\sum_{j=1}^{M_i}N_j}{N} \tag{1}$$

其中$\widetilde{N}_i$为(i 分别代表内部环、发夹环、膨胀环和多分支环)某个蛋白质相应 mRNA 二级结构中第 i 种环结构碱基含量，M_i 为 mRNA 中包含的第 i 种环结构的数量，N_j 为第 i 种第 j 个环结构中包含的碱基个数．N 为 mRNA 序列的长度(即包含核苷酸个数)．按照公式(1)计算了每个蛋白质所对应的 mRNA 二级结构中 4 种环结构碱基含量，分别表示为：$\widetilde{N}_{\text{int}}$、$\widetilde{N}_{\text{hair}}$、$\widetilde{N}_{\text{bul}}$和$\widetilde{N}_{\text{mul}}$．

1.3.2 mRNA 二级结构中碱基对含量

碱基对是形成茎结构的基本单元，为了与环结构作对比，我们统计了每个蛋白质的 mRNA 二级结构中碱基对的数量，并按照如下方式定义了碱基对含量：

$$\widetilde{N}_{\text{mat}}=\frac{N_{\text{mat}}}{N} \tag{2}$$

其中$\widetilde{N}_{\text{mat}}$为某个蛋白质相应 mRNA 二级结构中的碱基对含量，N_{mat} 为 mRNA 二级结构中包含的碱基对数，N 为 mRNA 序列的长度(即包含核苷酸个数).

1.3.3 mRNA 二级结构中单链碱基含量

为了与环结构作对比，我们也统计了每个蛋白质相应 mRNA 二级结构中单链碱基数，并按照如下方式定义了单链碱基含量：

$$\widetilde{N}_{\text{sc}}=\frac{N_{\text{sc}}}{N} \tag{3}$$

其中$\widetilde{N}_{\text{sc}}$为某个蛋白质相应 mRNA 二级结构中的单链碱基含量，N_{sc} 为 mRNA 二级结构中包含的单链碱基数，N 为 mRNA 序列的长度(即包含核苷酸个数).

对于每一个蛋白质的 mRNA 序列，使用 RNAfold 软件预测得到 mRNA 二级结构中的内部环、发夹环、膨胀环、多分支环结构，以及碱基对和单链碱基，然后，依据公式(1)～(3)分别计算了 mRNA 二级结构中 4 种环结构(内部环、发夹环、膨胀环和多分支环)碱基含量，碱基对含量和单链碱基含量．计算结果见附件表 S1 和 S2．

1.4 回归分析

本文研究 mRNA 环结构参量对蛋白质折叠速率的影响，即已确定参数之间的因果关系，所以选

择回归分析方法．在检验为显著的基础上，回归结果中相关系数的大小反应相关关系的强弱．决定系数的大小反应回归关系的强弱．而决定系数是相关系数的平方．为了能反映出各变量除作用强弱外对蛋白质折叠速率的影响方向(促进或阻碍)，使用相关系数来表征 mRNA 环结构对蛋白质折叠速率的影响．具体做法如下：首先计算出蛋白质对应 mRNA 二级结构中碱基对含量$\widetilde{N}_{\mathrm{mat}}$、单链碱基含量$\widetilde{N}_{\mathrm{sc}}$、内部环含量$\widetilde{N}_{\mathrm{int}}$、发夹环含量$\widetilde{N}_{\mathrm{hair}}$、膨胀环含量$\widetilde{N}_{\mathrm{bul}}$和多分支环含量$\widetilde{N}_{\mathrm{mul}}$，然后作蛋白质折叠速率与各参量值的回归分析，研究 mRNA 每种环结构对蛋白质折叠速率的影响．进一步，按照折叠类型，把蛋白质分为二态蛋白质和多态蛋白质，再按其二级结构分为全 α 类蛋白质、全 β 类蛋白质和混合类(α-β)蛋白质，对于每一类蛋白质，也作蛋白质折叠速率与各参量值的回归分析，研究 mRNA 环结构对不同蛋白质的折叠速率的影响．最后使用相关系数 r 值来表征参量间作用大小和方向，使用 P 值检验法来检验分析结果的可靠性．

2　结　　果

2.1　蛋白质折叠速率与其相应 mRNA 二级结构参量的相关性分析

对于数据库中所有蛋白质，分别分析了由公式(1)～公式(3)计算出的各种二级结构参量与蛋白质折叠速率的相关性，结果见表 1．

Table1　The results of linear regression between the protein folding rates and each parameter of mRNA secondary structures

	r	P
$\widetilde{N}_{\mathrm{int}}$	0.49	< 0.001
$\widetilde{N}_{\mathrm{hair}}$	0.51	< 0.001
$\widetilde{N}_{\mathrm{bul}}$	0.45	< 0.001
$\widetilde{N}_{\mathrm{mul}}$	0.29	0.003
$\widetilde{N}_{\mathrm{mat}}$	–0.23	0.02
$\widetilde{N}_{\mathrm{sc}}$	–0.069	0.496

Note: $\widetilde{N}_{\mathrm{int}}$, $\widetilde{N}_{\mathrm{hair}}$, $\widetilde{N}_{\mathrm{bul}}$ and $\widetilde{N}_{\mathrm{mul}}$ is the base content of internal loops, hairpin loops, bulge loops, and multi-branch loops respectively, $\widetilde{N}_{\mathrm{mat}}$ is the content of base pairs in mRNA secondary structures, $\widetilde{N}_{\mathrm{sc}}$ is the base content of the single strand in mRNA secondary structures, r is the correlation coefficient, P is the significance level.

结果表明，蛋白质折叠速率与 mRNA 二级结构中碱基对含量呈负相关，与单链碱基含量没有相关性．而与内部环碱基含量、发夹环碱基含量、膨胀环碱基含量呈极显著正相关，与多分支环碱基含量显著正相关．比较表 1 中 r 值 和 P 值，发现相比碱基对含量和单链碱基含量，mRNA 4 种环结构碱基含量与蛋白质折叠速率有更显著的相关性．意味着 mRNA 4 种环结构的使用会促进蛋白质的折叠，而且，相比单链碱基和碱基对，环结构可能对蛋白质的折叠速率有更大的影响．早有研究表明，mRNA 二级结构中各种环的结构特征比较配对的双链区来说具有碱基突出并有较大的柔性等特点[29]，正是由于环结构具有如较大柔性等这些特点，使得环结构的使用能够促进核糖体在 mRNA 上的移动速率，进一步促进了蛋白质的折叠．即 mRNA 二级结构中环结构碱基含量的增加可能会加快蛋白质的折叠．

2.2　蛋白质折叠速率随 mRNA 二级结构中碱基对含量及 4 种环结构碱基含量的变化关系

表 1 结果中发现，mRNA 二级结构碱基对含量和 4 种环结构碱基含量与蛋白质折叠速率均具有相关性，为了对比它们对蛋白质折叠速率影响的差别，对于所选的所有蛋白质，分别作了蛋白质折叠速率随 mRNA 二级结构中碱基对含量和 4 种环结构碱基含量的变化关系图．如图 1 和图 2．

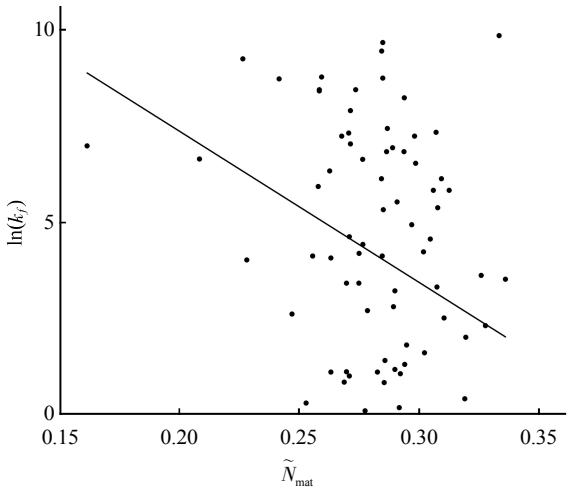

Fig. 1　Changes of the protein folding rates with the contents of base pairs in mRNA secondary structures

图 1 和图 2 结果表明，蛋白质折叠速率随相应 mRNA 二级结构中碱基对含量的增加而减小，却随着 4 种环结构碱基含量的增加而增大．碱基对是

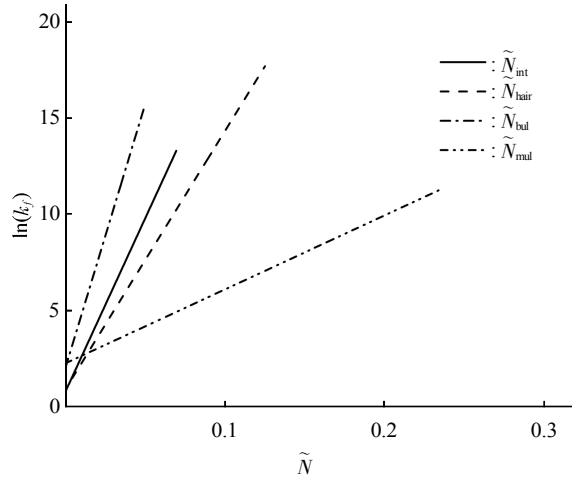

Fig. 2 Changes of the protein folding rates with the base content of the four kinds of loop structures in mRNA

mRNA 二级结构中形成茎结构的基本单元．有相关研究表明：核糖体在阅读 mRNA 茎区时，需要克服较高的自由能障碍以解链配对区，这使得核糖体对 mRNA 茎区的翻译速率慢于对单链与环区的翻译速率[30]，那么碱基对的增多就会在一定程度上减缓蛋白质的折叠进程．图 1 和图 2 的结果说明 mRNA 二级结构碱基配对区和环区碱基含量对蛋白质折叠速率的影响是截然相反的．

2.3 二态蛋白质和多态蛋白质折叠速率与其相应 mRNA 各类环结构碱基含量的相关性分析

前期的相关研究表明，当把蛋白质分为二态蛋白质和多态蛋白质时，两种蛋白质的折叠速率受同一参量的影响不尽相同[26–28]．基于此想法，把所选蛋白质按照折叠类型的不同，分为二态蛋白质和多态蛋白质，对每一类蛋白质的折叠速率与相应 mRNA 4 类环结构碱基含量进行相关性分析，结果见表 2 和表 3．

Table 2 The results of linear regression between the folding rates of two-state proteins and the base contents of each kind of loop structures in corresponding mRNA

	r	*P*
$\widetilde{N}_{int}$	0.37	0.004
$\widetilde{N}_{hair}$	0.44	0.0005
$\widetilde{N}_{bul}$	0.41	0.001
$\widetilde{N}_{mul}$	0.11	0.39

Note: $\widetilde{N}_{int}$, $\widetilde{N}_{hair}$, $\widetilde{N}_{bul}$ and $\widetilde{N}_{mul}$ is the base content of internal loops, hairpin loops, bulge loops, and multi-branch loops respectively, r is the correlation coefficient, P is the significance level.

Table 3 The results of linear regression between the folding rates of multi-state proteins and the base contents of each kind of loop structures in corresponding mRNA

	r	*P*
$\widetilde{N}_{int}$	0.51	0.0008
$\widetilde{N}_{hair}$	0.61	< 0.001
$\widetilde{N}_{bul}$	0.34	0.03
$\widetilde{N}_{mul}$	0.45	0.003

Note: $\widetilde{N}_{int}$, $\widetilde{N}_{hair}$, $\widetilde{N}_{bul}$ and $\widetilde{N}_{mul}$ is the base content of internal loops, hairpin loops, bulge loops, and multi-branch loops respectively, r is the correlation coefficient, P is the significance level.

结果表明，对于二态蛋白质，内部环碱基含量与蛋白质折叠速率呈显著相关性，发夹环碱基含量和膨胀环碱基含量均与蛋白质折叠速率呈极显著相关性，而没有发现多分支环碱基含量与蛋白质折叠速率之间的相关性．而对于多态蛋白质，内部环碱基含量和发夹环碱基含量均与蛋白质折叠速率呈极显著相关性，膨胀环碱基含量和多分支环含量与蛋白质折叠速率呈显著相关性．

2.4 不同二级结构类蛋白质的折叠速率与其相应 mRNA 各类环结构碱基含量的相关性分析

前期相关研究结果中认识到，对于不同二级结构类蛋白质，同一参量对蛋白质折叠速率影响的大小和趋势可能都有所不同[26–27]．本节按照二级结构类把蛋白质分为全 α 类蛋白质 21 个、全 β 类蛋白质 36 个和混合类(α-β)蛋白质 42 个．以每一类蛋白质为研究样本，分析了 mRNA 二级结构中内部环、发夹环、膨胀环和多分支环等各环结构碱基含量与蛋白质折叠速率之间的相关性．结果如表 4、表 5 和表 6 所示．

Table 4 The results of linear regression between the folding rates of all-α proteins and the base contents of each kind of loop structures in corresponding mRNA

	r	*P*
$\widetilde{N}_{int}$	0.32	0.15
$\widetilde{N}_{hair}$	0.52	0.015
$\widetilde{N}_{bul}$	0.51	0.02
$\widetilde{N}_{mul}$	–0.02	0.93

Note: $\widetilde{N}_{int}$, $\widetilde{N}_{hair}$, $\widetilde{N}_{bul}$ and $\widetilde{N}_{mul}$ is the base content of internal loops, hairpin loops, bulge loops, and multi-branch loops respectively, r is the correlation coefficient, P is the significance level.

Table 5 The results of linear regression between the folding rates of all-β proteins and the base contents of each kind of loop structures in corresponding mRNA

	r	*P*
$\widetilde{N}_{int}$	0.53	<0.001
$\widetilde{N}_{hair}$	0.56	<0.001
$\widetilde{N}_{bul}$	0.45	0.004
$\widetilde{N}_{mul}$	0.39	0.01

Note: $\widetilde{N}_{int}$, $\widetilde{N}_{hair}$, $\widetilde{N}_{bul}$ and $\widetilde{N}_{mul}$ is the base content of internal loops, hairpin loops, bulge loops, and multi-branch loops respectively, *r* is the correlation coefficient, *P* is the significance level.

Table 6 The results of linear regression between the folding rates of α-β proteins and the base contents of each kind of loop structures in corresponding mRNA

	r	*P*
$\widetilde{N}_{int}$	0.60	< 0.001
$\widetilde{N}_{hair}$	0.52	< 0.001
$\widetilde{N}_{bul}$	0.44	0.006
$\widetilde{N}_{mul}$	0.31	0.056

Note: $\widetilde{N}_{int}$, $\widetilde{N}_{hair}$, $\widetilde{N}_{bul}$ and $\widetilde{N}_{mul}$ is the base content of internal loops, hairpin loops, bulge loops, and multi-branch loops respectively, *r* is the correlation coefficient, *P* is the significance level.

对比分析表 4、表 5 和表 6，可以发现，对于全 β 类蛋白质，内部环碱基含量和发夹环碱基含量与相应蛋白质折叠速率均呈极显著正相关性，膨胀环碱基含量与相应蛋白质折叠速率呈显著正相关性，多分支环碱基含量与蛋白质折叠速率呈正相关．而对于全 α 类蛋白质，发夹环碱基含量和膨胀环碱基含量均与蛋白质折叠速率呈正相关，没有发现内部环碱基含量和多分支环碱基含量与蛋白质折叠速率之间的相关性．对于混合类(α-β)蛋白质，内部环碱基含量和发夹环碱基含量与相应蛋白质折叠速率均呈极显著正相关性，膨胀环碱基含量与相应蛋白质折叠速率呈显著正相关性．比较 *r* 值也可发现，对于全 β 类蛋白质，mRNA 二级结构中环结构碱基含量对相应蛋白质折叠速率的影响更大一些．

3 讨 论

本文选取了内部环碱基含量、发夹环碱基含量、膨胀环碱基含量和多分支环碱基含量等参量来研究 mRNA 各种环结构对蛋白质折叠速率的影响．总的来说，mRNA 二级结构环结构碱基含量与蛋白质折叠速率呈正相关性，意味着在 mRNA 二级结构中，环结构比例的增加可能会促进蛋白质的折叠，这是由于环结构相比茎结构有一些有利于蛋白质折叠的特性．如环区构象主要以非堆积构象为主，环区具有较大的构象柔性等．这种非堆积构象和较大柔性有利于 mRNA 沿核糖体的顺利延展，进而促进蛋白质的折叠．

4 种环都有各自不同的结构和特征．内部环是隔开 2 个螺旋区的环区，内部环的存在和比例将会影响 mRNA 的二级结构，从而会影响核糖体在 mRNA 上的移位速率，进而影响到相应蛋白质的折叠速率．一个完整发夹结构中的发夹环部分和茎区部分的特性存在很显著的差异，前期的研究中，我们已经发现茎区部分的碱基比例与蛋白质折叠速率有显著的负相关性[28]，那么发夹环部分碱基含量也必定是影响蛋白质折叠速率的一个重要因素．膨胀环是由螺旋区中的不配对的核苷酸构成的，膨胀环的多少必定会影响整个螺旋区的稳定性．由于多分支环与多个螺旋区连接，多分支环的增加就会增加 mRNA 二级结构的复杂性．这就意味着不同的环可能以不同的方式影响着 mRNA 的二级结构，进而影响到相应蛋白质的折叠速率．另外，蛋白质按照折叠的动力学行为的不同可以粗略地分成“二态”蛋白质和“多态”蛋白质，按照二级结构类大致分为全 α 类蛋白质、全 β 类蛋白质和混合类(α-β)蛋白质．由于不同类蛋白质的结构和组织形式不同，环结构对它们折叠速率的影响也很有可能不同．本文的结果证实了以上分析：当把蛋白质按照不同折叠类分组后作 4 种环结构碱基含量与蛋白质折叠速率之间的相关性分析，比较相关性分析结果中 *r* 值和 *P* 值，发现对于多态类蛋白质，二者具有更好的相关性．同时，把蛋白质按照二级结构类分组后作同样的分析，发现对于全 β 类蛋白质，mRNA 二级结构中环结构碱基含量对相应蛋白质折叠速率的影响更大一些．这些结果说明 mRNA 二级结构中内部环结构、发夹环结构、膨胀环结构和多分支环结构对不同类蛋白质折叠速率的影响是不同的．而且分析表 1～6 的结果，均可以发现不同的环结构对同一类蛋白质折叠速率的影响也是不同的．

RNA 和蛋白质是生命最基本最重要的组成物质，生命活动中，蛋白质折叠是最基本的生命过

程，本文的结果证实，在二者的相互作用中，mRNA 除通过翻译把遗传信息传递给蛋白质外，可能还在它的环结构等信息中携带了影响蛋白质折叠等方面的信息.

附件 表 S1 和表 S2 见本文网络版附录(http://www.pibb.ac.cn)

参 考 文 献

[1] 周筠梅. 蛋白质的错误折叠与疾病. 生物化学与生物物理进展, 2000, **27**(6): 579−584
Zhou J M. Prog Biochem Biophys, 2000, **27**(6): 579−584

[2] Song Y S, Zhou X, Zheng W M, *et al*. Stabilities and dynamics of protein folding nuclei by molecular dynamics simulation. Commun Theor Phys, 2017, **68**(7): 137−148

[3] Hatters D M. Protein folding: illuminating chaperone activity. Nat Chem Biol, 2017, **13**(4): 346−347

[4] Pang Y P. How fast fast-folding proteins fold in silico. Biochem Biophys Res Commun, 2017, **492**(1): 135−139

[5] Ljubetič A, Gradišar H, Jerala R. Advances in design of protein folds and assemblies. Curr Opin Chem Biol, 2017, **40**: 65−71

[6] Dill K A, Ozkan S B, Weikl T R, *et al*. The protein folding problem: when will it be solved?. Curr Opin Struct Biol, 2007, **17**(3): 342-346

[7] Dill K A, Ozkan S B, Shell M S, *et al*. The protein folding problem. Annu Rev Biophys, 2008, **37**: 289−316

[8] Plaxco K W, Simons K T, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol, 1998, **227**(4): 985−994

[9] Debe D A, Goddard W A-3rd. First principles prediction of protein folding rates. J Mol Biol, 1999, **294**(3): 619−625

[10] Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J, 2002, **82**(1): 458−463

[11] Zhang L X, Li J, Jiang Z T, *et al*. Folding rate prediction based on neural network model. Polymer, 2003, **44**(5): 1751−1756

[12] Gong H, Isom D G, Srinivasan R, *et al*. Local secondary structure content predicts folding rates for simple, two-state proteins. J Mol Biol, 2003, **327**(5): 1149−1154

[13] Mirny L, Shakhnovich E. Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct, 2001, **30**(1): 361−396

[14] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA, 2004, **101**(24): 8942−8944

[15] Gromiha M M. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model, 2005, **45**(2): 494−501

[16] Kuznetsov I B, Rackovsky S. Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. Proteins, 2004, **54**(2): 333−341

[17] Punta M, Rost B. Protein folding rates estimated from contact predictions. J Mol Biol, 2005, **348**(3): 507−512

[18] Galzitskaya O V, Garbuzynskiy S O. Entropy capacity determines protein folding. Proteins, 2006, **63**(1): 144−154

[19] Ouyang Z, liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci, 2008, **17**(7): 1256−1263

[20] Chou K C, Shen H B. FoldRate: a web-server for predicting protein folding rates from primary sequence. Open Biol J, 2009, **3**(1): 31−50

[21] Komar, A A. A pause for thought along the co-translational folding pathway. Trends Biochem Sci, 2009, **34**(1): 16−24

[22] Purvis I J, Bettany A J, Santiago T C, *et al*. The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. A hypothesis. J Mol Biol, 1987, **193**(2): 413−417

[23] Krasheninnikov I A, Komar A A, Adzhubeĭ I A. 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, 1988, **303**(4): 995−999

[24] Shpaer E G. The secondary structure of mRNAs from *Escherichia coli*: its possible role in increasing the accuracy of translation. Nucleic Acids Res, 1985, **13**(1): 275−288

[25] Li R F, Li H. The influence of protein coding sequences on protein folding rates of all-β proteins. Gen Physiol Biophys, 2011, **30**(2): 154−161

[26] Li R F, Li H. Study on the influences of palindromes in protein coding sequences on the folding rates of peptide chains. Protein Pept Lett, 2010, **17**(7): 881−888

[27] 于志芬, 李瑞芳. 同义密码子的使用偏好性对蛋白质折叠速率的影响. 生物物理学报, 2013, **29**(8): 603−613
Yu Z F, Li R F. Acta Biophys Sin, 2013, **29**(8): 603−613

[28] 李瑞芳, 于志芬, 黄俏. mRNA 的二级结构对蛋白质折叠速率的影响. 生物物理学报, 2014, **30**(7): 497−508
Li R F, Yu Z F, Huang Q. Acta Biophys Sin, 2014, **30**(7): 497−508

[29] 张静, 石秀凡, 刘次全. Yeast 基因下游二级结构与多聚腺苷作用信号. 生物化学与生物物理进展, 2001, **28**(4): 523−527
Zhang J, Shi X F, Liu C Q. Prog Biochem Biophys, 2001, **28**(4): 523−527

[30] 柳树群, 刘次全. mRNA 的序列、结构以及翻译速率与蛋白质结构的关系. 动物学研究, 1999, **20**(6): 457−461
Liu S Q, Liu C Q. Zoolog Res, 1999, **20**(6): 457−461

Influences of The mRNA Loop Structures on Protein Folding Rate*

LI Rui-Fang[1)**], LI Hong[2)], GUO Chun-Yang[1)], YANG Sa-Ru-La[1)]

([1)] *College of Physics and Electronic Information, Inner Mongolia Normal University, Hohhot* 010022, China;

[2)] *School of Physical Science and Technology, Inner Mongolia University, Hohhot* 010021, China)

Abstract The important factors in secondary structures of mRNA influencing on protein folding rates are found during our previous research, and there are kinds of complex loop structures in secondary structures of mRNA. Do these complex loop structures have important influences on protein folding rate? Do different loop structures have similar effects on protein folding rate? Based on this idea, a data set that contains both the information of internal loops, hairpin loops, bulge loops, multi-branch loops and protein folding rates was constructed. For each protein in the data set, the secondary structures of mRNA were predicted followed by calculations of the parameters of mRNA secondary structures, including the base content of each loop structure, the content of base pairs, and the base content of the single strand. Analyses of the relationship between the protein folding rates and each parameter of loop structures of mRNA reveal that the protein folding rate has a significant positive correlation with the content of each kind of the four loop structures, it means that the loop structures of mRNA act as a kind of influential factors for the protein folding rate. Given the proteins in the data set were classed into different folding types and different secondary structural types, the relationship analyses reveal that for proteins in different types, the effects of loop structures on protein folding rate are significantly different. This work will provide the theoretical basis for the future study of mRNA and protein folding rate.

Key words mRNA, internal loop, hairpin loop, bulge loop, multi-branch loop, protein folding rate

DOI: 10.16476/j.pibb.2017.0473

* This work was supported by grants from The National Natural Science Foundation of China (31260219), The Ph.D. Programs Foundation of Ministry of Education of China (20121501110006), The Natural Science Foundation of Inner Mongolia (2016MS0362) and the Research Program of Science and Technology at Universities of Inner Mongolia Autonomous Region(NJZY17043).

**Corresponding author.

Tel: 86-471-4392576, E-mail: liruifang@imnu.edu.cn

Received: December 26,2017 Accepted: April 9, 2018

附　录

Table S1　Basic information for the 59 two-state proteins

PDB ID	mRNA length	$\ln(k_f)$	$\widetilde{N}_{\text{int}}$	$\widetilde{N}_{\text{hair}}$	$\widetilde{N}_{\text{bul}}$	$\widetilde{N}_{\text{mul}}$	$\widetilde{N}_{\text{mat}}$	$\widetilde{N}_{\text{sc}}$	Structure class
1ARR	159	9.20	0.044	0.033	0.000	0.094	0.226	0.472	α
1BA5	159	5.90	0.017	0.027	0.016	0.069	0.258	0.484	α
1BDD	180	11.69	0.000	0.040	0.017	0.093	0.250	0.500	α
1EFX	177	8.19	0.042	0.025	0.006	0.040	0.294	0.401	α
1IDY	162	8.73	0.028	0.039	0.012	0.123	0.259	0.451	α
1IMQ	255	7.28	0.020	0.021	0.004	0.067	0.271	0.455	α
1LMB	240	10.4	0.014	0.025	0.004	0.067	0.292	0.408	α
1NTI	258	7.00	0.022	0.017	0.011	0.039	0.271	0.450	α
1PRB	141	12.9	0.041	0.043	0.050	0.000	0.262	0.447	α
1U5P	330	11.0	0.014	0.023	0.006	0.052	0.282	0.436	α
1VII	108	11.8	0.069	0.069	0.032	0.000	0.259	0.444	α
1YCC	309	9.62	0.014	0.023	0.008	0.033	0.285	0.395	α
256B	318	12.3	0.016	0.014	0.010	0.051	0.261	0.465	α
2PDD	123	9.80	0.049	0.045	0.011	0.089	0.333	0.333	α
1C9O	198	7.20	0.036	0.032	0.015	0.091	0.268	0.460	β
1E0L	111	10.6	0.039	0.068	0.009	0.153	0.243	0.514	β
1CSP	201	6.50	0.029	0.020	0.007	0.050	0.299	0.403	β
1E65	384	4.91	0.011	0.014	0.006	0.021	0.297	0.393	β
1FNF-10	282	5.50	0.012	0.024	0.018	0.074	0.291	0.418	β
1FNF-9	270	-0.90	0.012	0.020	0.009	0.113	0.270	0.430	β
1G6P	198	6.30	0.018	0.025	0.013	0.056	0.263	0.460	β
1JMQ	120	8.40	0.053	0.046	0.025	0.108	0.258	0.483	β
1JO8	174	2.50	0.022	0.034	0.017	0.069	0.310	0.379	β
1K0S	429	7.40	0.010	0.012	0.005	0.025	0.287	0.410	β
1LOP	492	6.60	0.009	0.013	0.003	0.041	0.276	0.447	β
1M9S	228	4.00	0.013	0.021	0.007	0.095	0.228	0.544	β
1MJC	207	5.30	0.030	0.023	0.006	0.072	0.285	0.406	β
1NYF	174	4.54	0.038	0.030	0.010	0.000	0.305	0.362	β
1PIN	102	9.40	0.036	0.103	0.013	0.029	0.284	0.431	β
1PSF	207	3.20	0.017	0.027	0.000	0.036	0.290	0.377	β
1SHG	171	1.10	0.023	0.032	0.006	0.070	0.263	0.462	β
1TEN	267	1.06	0.013	0.016	0.005	0.039	0.292	0.416	β
1C8C	192	6.95	0.021	0.035	0.005	0.234	0.161	0.646	β
1K8M	261	-0.71	0.021	0.025	0.005	0.034	0.284	0.360	β
1PNJ	252	-1.00	0.017	0.025	0.016	0.028	0.298	0.385	β
1PSE	207	1.17	0.017	0.027	0.000	0.036	0.290	0.377	β
1QTU	324	-0.36	0.017	0.012	0.003	0.038	0.290	0.420	β
1WIT	279	0.41	0.011	0.016	0.006	0.039	0.319	0.362	β
2AIT	222	4.21	0.020	0.021	0.008	0.052	0.302	0.387	β
1PKS	228	-1.06	0.021	0.026	0.007	0.059	0.250	0.500	β
1FMK	171	4.05	0.047	0.025	0.006	0.041	0.263	0.474	α-β
1K9Q	120	8.37	0.053	0.046	0.025	0.108	0.258	0.483	α-β
1AYE	225	6.90	0.016	0.027	0.004	0.055	0.289	0.409	α-β
1DIV-N	168	6.61	0.020	0.037	0.006	0.104	0.208	0.583	α-β
1FKF	321	1.60	0.015	0.016	0.005	0.056	0.302	0.396	α-β
1HDN	255	2.69	0.025	0.025	0.006	0.029	0.278	0.443	α-β
1N88	288	2.00	0.016	0.015	0.012	0.016	0.319	0.316	α-β
1O6X	213	6.80	0.015	0.028	0.005	0.042	0.286	0.347	α-β
1PGB-B	48	12.0	0.042	0.125	0.031	0.000	0.271	0.458	α-β
1RFA	234	8.40	0.017	0.026	0.013	0.043	0.274	0.444	α-β
1RIS	291	6.10	0.012	0.019	0.003	0.035	0.309	0.378	α-β
1SPR	309	8.70	0.019	0.016	0.003	0.045	0.285	0.427	α-β
1URN	288	4.60	0.013	0.017	0.010	0.104	0.270	0.458	α-β
2ACY	294	0.84	0.014	0.027	0.003	0.035	0.269	0.463	α-β
2CI2	192	5.80	0.022	0.021	0.007	0.026	0.313	0.354	α-β
2HQI	216	0.18	0.021	0.023	0.011	0.028	0.292	0.417	α-β
2PTL	180	4.10	0.022	0.043	0.008	0.058	0.256	0.489	α-β
2VIK	378	6.80	0.010	0.015	0.003	0.037	0.294	0.394	α-β
1DIY-C	257	3.30	0.010	0.020	0.007	0.047	0.226	0.370	α-β

Note: $\ln(k_f)$ is protein folding rate, $\widetilde{N}_{\text{int}}$, $\widetilde{N}_{\text{hair}}$, $\widetilde{N}_{\text{bul}}$ and $\widetilde{N}_{\text{mul}}$ is the content of internal loops, of hairpin loops, of bulge loops, and of multi-branch loops respectively, which were calculated with the equation (1), $\widetilde{N}_{\text{mat}}$ is the content of base pairs in mRNA secondary structures, which were calculated with the equation (2) $\widetilde{N}_{\text{sc}}$ is the base content of the single strand in mRNA secondary structures, which were calculated with the equation (3).

Continued

Table S2 Basic information for the 40 multi-state proteins

PDB ID	mRNA length	ln(k_f)	$\widetilde{N}_{int}$	$\widetilde{N}_{hair}$	$\widetilde{N}_{bul}$	$\widetilde{N}_{mul}$	$\widetilde{N}_{mat}$	$\widetilde{N}_{sc}$	Structure class
1A6N	453	1.10	0.008	0.013	0.003	0.027	0.283	0.435	α
1AYI	255	7.20	0.015	0.018	0.004	0.059	0.298	0.353	α
1CEI	255	5.80	0.016	0.018	0.004	0.055	0.306	0.384	α
1ENH	162	10.5	0.017	0.028	0.012	0.068	0.290	0.420	α
1UZC	207	8.68	0.032	0.025	0.000	0.046	0.242	0.444	α
2CRO	195	5.35	0.016	0.023	0.005	0.056	0.308	0.369	α
2ABD	258	7.86	0.022	0.017	0.011	0.039	0.271	0.450	α
1BEB	468	-2.20	0.009	0.014	0.003	0.017	0.295	0.374	β
1ADW	369	-7.60	0.017	0.013	0.003	0.039	0.274	0.453	β
1CBI	408	-3.20	0.008	0.014	0.003	0.024	0.304	0.392	β
1EAL	381	1.30	0.011	0.019	0.004	0.024	0.294	0.399	β
1HCD	354	-4.97	0.019	0.017	0.007	0.064	0.229	0.525	β
1HNG	285	1.80	0.013	0.020	0.008	0.036	0.295	0.411	β
1I1B	453	-4.01	0.008	0.012	0.005	0.033	0.289	0.422	β
1IFC	393	3.40	0.007	0.014	0.004	0.037	0.275	0.450	β
1JOO	447	0.30	0.009	0.017	0.002	0.035	0.253	0.494	β
1OPA	399	1.40	0.010	0.014	0.004	0.026	0.286	0.386	β
1TIT	267	3.60	0.013	0.022	0.007	0.041	0.326	0.348	β
1BNI	324	2.60	0.014	0.020	0.003	0.038	0.247	0.494	α-β
1BRS	267	3.40	0.013	0.018	0.008	0.036	0.270	0.466	α-β
1DK7	438	0.83	0.010	0.011	0.005	0.038	0.285	0.429	α-β
1GXT	264	4.40	0.019	0.019	0.004	0.087	0.277	0.443	α-β
1HEL	387	6.10	0.012	0.011	0.004	0.027	0.284	0.416	α-β
1HMK	363	2.79	0.010	0.013	0.004	0.045	0.289	0.419	α-β
1PHP-C	657	-3.50	0.006	0.007	0.002	0.018	0.309	0.382	α-β
1PHP-N	525	2.30	0.007	0.010	0.003	0.018	0.328	0.345	α-β
1QOP-A	807	-2.50	0.005	0.008	0.002	0.012	0.318	0.358	α-β
1QOP-B	1170	-6.90	0.003	0.004	0.002	0.009	0.312	0.374	α-β
1RA9	477	-3.20	0.009	0.013	0.002	0.022	0.321	0.358	α-β
1SCE	291	4.17	0.010	0.018	0.003	0.054	0.275	0.440	α-β
1QBU	228	7.30	0.023	0.022	0.014	0.026	0.307	0.333	α-β
2A5E	381	3.50	0.007	0.014	0.003	0.020	0.336	0.328	α-β
2BLM	780	-1.24	0.006	0.007	0.002	0.017	0.300	0.400	α-β
2LZM	492	4.10	0.009	0.009	0.004	0.029	0.285	0.429	α-β
2RN2	465	0.10	0.012	0.014	0.003	0.024	0.277	0.445	α-β
2VIK	378	11.9	0.010	0.015	0.003	0.037	0.294	0.394	α-β
3CHY	384	1.00	0.012	0.016	0.010	0.045	0.271	0.419	α-β
1AON	465	-1.50	0.008	0.010	0.005	0.029	0.288	0.424	α-β
1BTA	267	1.11	0.013	0.018	0.008	0.036	0.270	0.461	α-β
1B9C	672	-2.76	0.005	0.011	0.002	0.018	0.281	0.433	α-β

Note: ln(k_f) is protein folding rate, $\widetilde{N}_{int}$, $\widetilde{N}_{hair}$, $\widetilde{N}_{bul}$ and $\widetilde{N}_{mul}$ is the content of internal loops, of hairpin loops, of bulge loops, and of multi-branch loops respectively, which were calculated with the equation (1), $\widetilde{N}_{mat}$ is the content of base pairs in mRNA secondary structures, which were calculated with the equation (2) $\widetilde{N}_{sc}$ is the base content of the single strand in mRNA secondary structures, which were calculated with the equation (3).