基因序列内部特征与蛋白质对称结构的相关性分析

申小娟[1,2]　李光林[1,2]

[1]（ 中国科学院深圳先进技术研究院　深圳　518055 ）
[2]（ 中国科学院大学　北京　100049 ）

摘　要　蛋白质的翻译是一个非常复杂且至关重要的生命过程。翻译速率会沿着mRNA上发生改变借以调控伴随翻译的蛋白质折叠，并对蛋白质的最终构象产生重要影响。本文以TIM beta-alpha barrel折叠子中的两个不同物种HisA蛋白质为研究对象，初步分析了它们基因序列中局部密码子使用偏好、局部残基带电性分布、局部GC含量分布与蛋白质对称结构的相关性，探讨翻译过程中各种因素在蛋白质对称结构形成过程中的可能调控机制。结果表明，在两个不同物种的HisA蛋白质中，对称结构与密码子使用偏好、残基带电性以及GC含量都存在一定程度的相关性。

关键词　基因序列；对称结构；密码子使用偏好；蛋白质翻译

Analysis on the Association Between Intragenic Features and Symmetry in Protein Structures

SHEN Xiao-juan[1,2]　LI Guang-lin[1,2]

[1](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract　Protein translation is a remarkably complex and crucial process of life. Translation speed varies along mRNA to coordinate the co-translational protein folding, and such variations may have drastic effects on the final conformation of the protein encoded. In this paper, we choose HisA protein with TIM beta-alpha barrel fold from two different species to investigate the factors that may be involved in modulating translation process for the fomation of structure symmetry. The association between symmetry in protein structure and several intragenic features is explored, including local codon usage bias along codon sequence, local charge distribution of the encoded protein sequence, local GC content distribution along the nucleotide sequence. Our results show that for HisA proteins from two different species, symmetry in structure is correlated with codon usage bias, charge of the residues and GC content.

Keywords　gene sequence; structural symmetry; codon usage bias; protein translation

1　引　言

对称性普遍存在于蛋白质分子中且对蛋白质的稳定性及其生物功能有非常重要的作用[1-6]。据有关研究表明，常见的10种蛋白质折叠类型中有6种具有内部结构对称性[7,8]。进化的观点认为，蛋白质的结构对称性是基因复制和融合的结果，早期的蛋白质比现在的蛋白质要更小，较大的蛋白质是由小的结构单元组合而来[9,10]。

现在普遍认为细胞内许多蛋白质的折叠是在蛋白质翻译过程的同时就发生了，mRNA序列中包含了蛋白质正确折叠的信息，指导新生态链进行伴随翻译的蛋白质折叠[11,12]。核糖体内存在许多机制能够允许翻译速度根据蛋白质折叠过程进行调节，并对蛋白质的最终构象产生重要影响[13]。翻译速率的改变可能是由

基金项目：广东省低成本健康技术创新团队项目资助。
作者简介：申小娟，博士研究生，助理研究员，主要研究方向为生物物理与计算生物；李光林，博士，研究员，主要研究方向为生物医学信号处理、生物医学仪器、神经康复工程等，E-mail：gl.li@siat.ac.cn。

于局部mRNA的稳定性的改变或者是在mRNA翻译片段出现稀有密码子的原因[14]。新生肽链的氨基酸序列也可能会参与调控翻译速率，帮助蛋白质正确折叠[15,16]。

本文以TIM beta-alpha barrel折叠子两个不同物种中的HisA蛋白质为研究对象，初步探索翻译过程中各种因素在蛋白质对称结构形成过程中的可能调控机制。由于对称结构的普遍性以及新合成的肽链会发生伴随翻译的折叠，因此对称结构的形成除了基因复制和融合的原因以外，可能还存在其他保守模式来调控对称结构的翻译过程，协助对称蛋白质在体内有效地进行伴随翻译的折叠。我们应用改进的重现图方法分析HisA蛋白质的结构对称性[17]，确定对称子结构的信息。在此基础上，分别计算残基带电性分布、密码子使用偏好分布以及GC含量分布。研究发现，残基带电性、密码子使用偏好以及GC含量在选择的两个HisA蛋白质中出现了一种共同的变化特征，从而表明这三方面的因素可能参与调控翻译速率，帮助HisA蛋白质形成对称结构。

2 数据与方法

2.1 数据说明

HisA是一种组氨酸合成酶，由八个β-strand/α-helix重复单元连接而形成一个桶状结构(图1中A、D)。根据SCOP蛋白质结构分类数据库，HisA蛋白质有两个不同的物种：Streptomyces coelicolor（PDB id：1VZW）和Thermotoga maritima（PDB id：1QO2）。两个蛋白质的氨基酸序列相似度约为29%，其中蛋白质结构数据从RSCB中获取，基因数据从EXpasy中获取，基因组数据从genbank FTP中获取。

2.2 改进的重现图方法

简要介绍改进的重现图方法分析蛋白质内部结构对称性。对任意一条蛋白质，用一维矢量$S=x_1x_2x_3\cdots x_N$,来表示。只考虑主链上C_α原子，x_1代表该残基C_α原子的空间坐标，N表示蛋白质的长度。基于该矢量构建一个d维矢量：

$$
\begin{aligned}
X_1 &= x_1x_2\cdots x_d\\
X_2 &= x_2x_3\cdots x_{d+1}\\
&\cdots\cdots\\
X_i &= x_ix_{i+1}\cdots x_{i+d-1}\\
&\cdots\cdots\\
X_{N-d+1} &= x_{N-d+1}x_{N-d+2}\cdots x_N
\end{aligned}
\tag{1}
$$

i代表片段起始的残基位置，该组矢量对应S中所有长度为d的连续片段。计算对于x_i有多少片段与它相近。选取dRMSD来定义片段X_i和X_j的距离：

$$dRMSD(X_i, X_j)=\sqrt{\frac{1}{d(d-1)/2}\sum_{\substack{m=1\\n>m}}^{d}(r_{mn}^{i}-r_{mn}^{j})^2} \tag{2}$$

其中r^i_{mn}和r^j_{mn}分别是片段X_i和X_j中对应位置的第m个和第n个C_α之间的空间距离。选择一定的阈值由此构建改进的重现图，并进一步计算改进重现图上子矩阵的皮尔森关联系数。

2.3 局部密码子使用偏好分布

由于经典的密码子适应指数[18]（Codon Adaption Index，CAI）与序列片段长度没有内在的关系，所以我们选择CAI值来衡量局部密码子使用偏好。CAI的定义是基因序列中每个密码子相对同义密码子使用（RSCU）的几何平均值，除以该基因序列的最大CAI值[18]。

$$
\begin{aligned}
&CAI = CAI_{obs}/CAI_{\max}\\
&where\\
&CAI_{obs}=(\prod_{k=1}^{L}RSCU_k)^{1/L}\\
&CAI_{\max}=(\prod_{k=1}^{L}RSCU_{k\max})^{1/L}
\end{aligned}
\tag{3}
$$

其中$RSCU_k$和$RSCU_{max}$分别是同义密码子组中第K个密码子和最优密码子的RSCU值。RSCU值是某一密码子观测到的频率除以该组同义密码子中使用频率的期望。

$$RSCU_{ij}=\frac{x_{ij}}{\frac{1}{n}\sum_{j=1}^{n_i}x_{ij}} \tag{4}$$

其中x_{ij}是第i个氨基酸的第j个密码子出现的次数，n_i是第i个氨基酸同义密码子的数目。需要指出的是，在原始的定义中同义密码子使用值是以一组高表达的基因序列为参照来计算，我们以整个基因组的密码子使用值为参照。计算局部密码子使用偏好时，设置滑动窗口片段长度为20个密码子。

2.4 局部残基带电性分布

根据氨基酸带电性不同，将氨基酸分为三种不同带电类别，带正电的Arg、His和Lys,带负电的Asp和Glu，其余为电中性氨基酸。设定带正电的残基值为+1，带负电的残基值为-1，电中性残基值为0。设置滑动窗口长度为20个残基，计算局部残基带电性分布。

2.5 局部GC含量分布

设置窗口长度为40个核苷酸，计算局部GC含量分布（40个核苷酸长度约为mRNA上核糖体的移动步长）。

3　结　果

3.1　1VZW和1QO2结构对称性分析

图1为两个HisA蛋白质的结构对称性分析结果。其中图A、D为两个蛋白质的结构卡通图，我们分别用红色和绿色表示两个对称的子结构。图B为1VZW结构对称关联图，可以看到起始为86，长度为总长度一半的子结构与起始为第一个长度为序列长度一半的子结构具有很强的关联性。图C为1VZW的关联指数分布，在（86，0.82）处有一个明显的峰值，表示该位置起始的片段与起始为第一个位置的片段结构的皮尔森相关系数为0.82，蛋白质具有两对称性。图E为1Q02的结构对称关联图，可以看到在115的位置周围有较强的相关指数，图F上显示（115，0.83）处有一个明显峰值，蛋白质具有两对称结构。

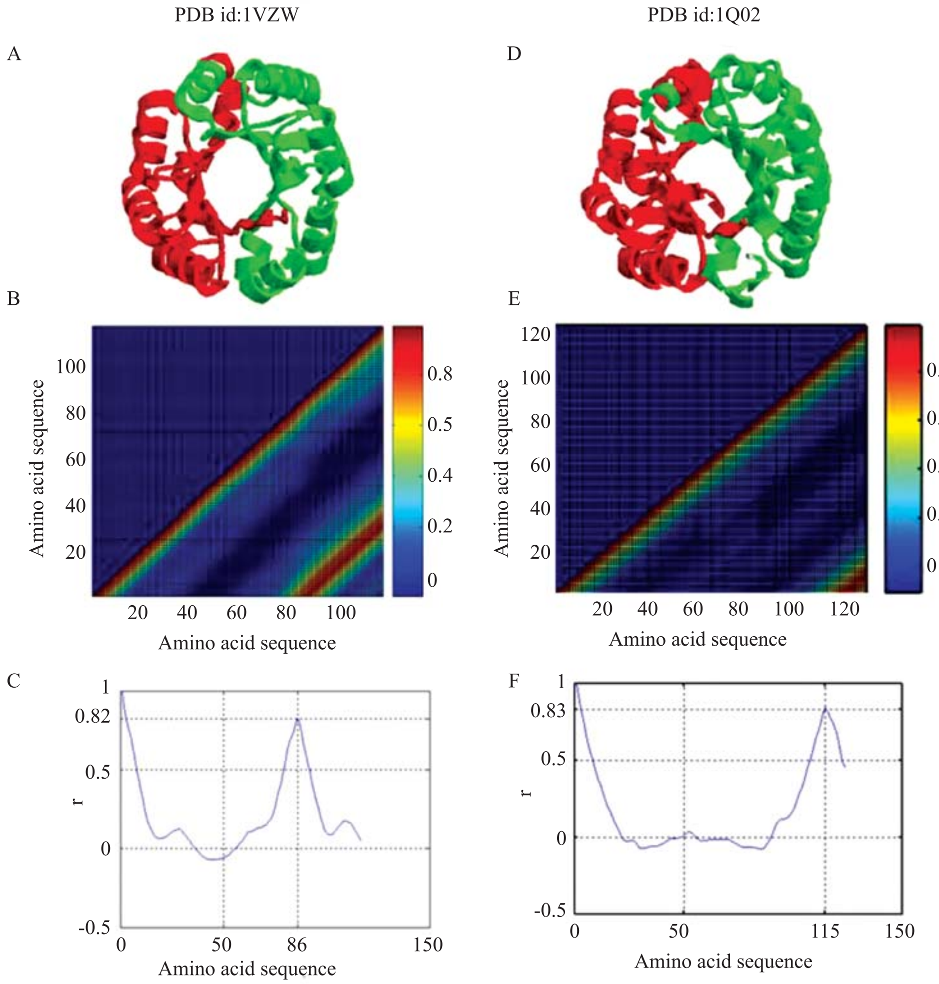

图1　蛋白质结构示意图、结构对称性关联图、关联指数分布图。图A、D分别为蛋白质1VZW以及1Q02的结构卡通示意图，对称的两部分结构分别用绿色标记出来。图B、E分别为两个蛋白质的结构对称关联图。关联图上与对角线平行的红色条纹表示以该位置起始，长度为总序列长度一半的子结构，与起始为第一个氨基酸同样长度的子结构具有很强的相似性。1VZW和1Q02蛋白质结构内部都探测到了两对称性。图C、F为关联指数分布图，表示序列内部长度为总序列长度一半的所有子结构与第一个片段结构的关联指数分布。

3.2　局部特征分布

图2分别是两个不同物种中HisA蛋白质的1VZW、1Q02的各种局部特征分布图。其中1Q02是Thermotoga maritima物种中的HisA蛋白质，我们曾经初步分析过该蛋白质的局部密码子使用偏好[19]，发现在密码子序列的中间对应与对称结构的连接区域出现CAI值的明显下降（图D）。1VZW是来自于Streptomyces coelicolor物种的HisA蛋白质，该蛋白质与1Q02的氨基酸序列相似度大约为29%，核苷酸序列的相似度约为50%。图A为1VZW的局部密码子使用偏好分布，红色箭头显示在密码子序列的中间部位同样出现了一处明显CAI值下降。结合图1中结构对称性分析结果，该下降的区域对应于对称结构的连接区。密码子偏好指数下降意味着该区域内出现了更多该物种的低频使用密

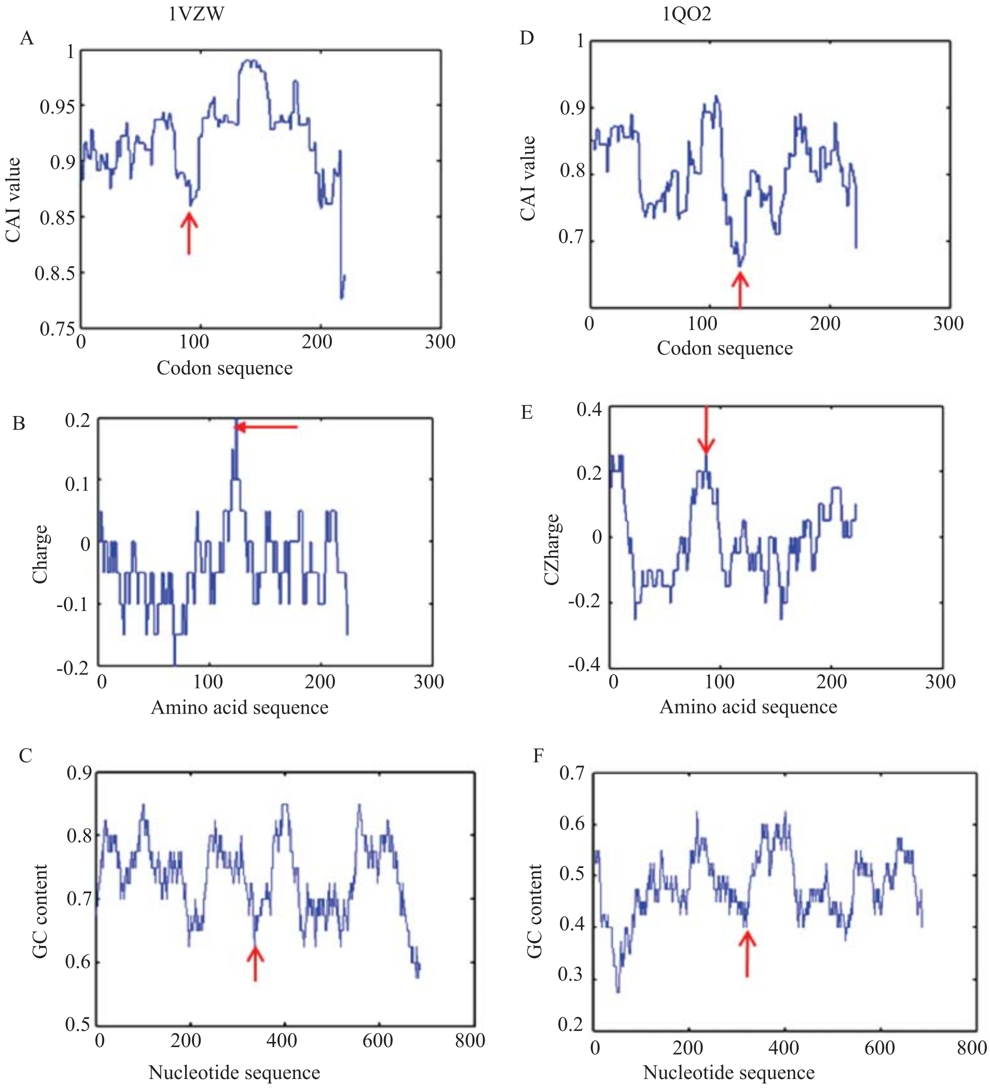

图2 密码子偏好分布图、残基带电性分布图、GC含量分布图。图A、D分别为蛋白质1VZW、1QO2蛋白质的密码子使用偏好CAI分布图。红色箭头显示该区域有一个明显的CAI值下降。B、E分别为两个蛋白质氨基酸序列内部残基带电性的分布图。红色箭头显示该部位序列片段的残基带电性明显升高。C、F分别为两个蛋白质基因序列内部GC含量的分布图。红色箭头显示该部位的GC含量明显下降。

码子。HisA蛋白质在两个不同的物种中密码子的使用拥有同样的特征。

图B和E分别为两个蛋白质局部残基的带电性分布。红色箭头标示出序列中残基带正电的峰值区域。在1VZW和1QO2两个蛋白质序列中，都能够看到在序列的中间部位出现一个明显的峰值，表示该区域内拥有更多的带正电残基。图C和F分别为两个蛋白质核苷酸序列中局部GC含量的分布。结果显示在两个蛋白质对称结构的连接区，局部GC含量也出现下降。

4 结论与展望

本文我们研究了两个不同物种中HisA蛋白质基因序列的内部特征与结构对称性的关系，发现局部密码子使用偏好、局部残基带电性、局部GC含量在蛋白质对称结构的连接区域都出现了一定程度的变化。首先，两个物种中HisA蛋白质的密码子使用偏好在连接区域都出现明显下降，说明稀有密码子可能参调控与翻译速率从而帮助HisA蛋白质对称结构的形成。其次，在1VZW和1QO2蛋白质序列中，对称结构连接区的残基带电性都呈现出升高，表明残基带电性也可能与对称结构的形成相关。最后，GC含量在这两个蛋白质的连接区域也都出现一定的下降，说明GC含量与对称结构也存在一定的关联。结果表明，蛋白质对称结构的形成除了基因复制和融合这一基本的假设以外，自然界可能还保留了其他一些保守的机制调控翻译过程从而帮助新生肽链有效地折叠到对称结构。这些理论是基于两个不同物种HisA蛋白质的分析结果，为了验

证这些理论的普遍性，我们还需要开展更进一步更广泛的研究。

致谢:感谢华中科技大学肖奕教授给予本论文的有益讨论和宝贵意见。

参 考 文 献

[1] Andrade M A, Perez-Iratxeta C, Ponting C P. Protein repeats: structures, functions, and evolution [J]. Journal of Structural Biology, 2001, 134(2-3): 117-131.

[2] Guerler A, Ernst-Walter K. Symmetric structures in the universe of protein folds [J]. Journal of Chemical Information and Modeling, 2009.

[3] Blaber M, Lee J. Designing proteins from simple motifs: opportunities in top-down symmetric deconstruction [J]. Current Opinion in Structural Biology, 2012.

[4] Blaber M, Lee J, Longo L. Emergence of symmetric protein architecture from a simple peptide motif: evolutionary models [J]. Cellular and Molecular Life Sciences, 2012: 1-8.

[5] Goodsell D S, Olson A J. Structural symmetry and protein function [J]. Annual Review of Biophysics and Biomolecular Structure, 2000, 29(1): 105-153.

[6] Guerler A, Wang C, Knapp E W. Symmetric structures in the universe of protein folds [J]. Journal of Chemical Information and Modeling, 2009, 49(9): 2147-2151.

[7] Marcotte E M, et al. A census of protein repeats [J]. Journal of Molecular Biology, 1999, 293(1): 151-160.

[8] Salem G M, et al. Correlation of observed fold frequency with the occurrence of local structural motifs [J]. Journal of Molecular Biology, 1999, 287(5): 969-981.

[9] McLachlan A. Repeating sequences and gene duplication in proteins [J]. Journal of Molecular Biology, 1972, 64(2): 417-437.

[10] McLachlan A. Analysis of gene duplication repeats in the myosin rod [J]. Journal of Molecular Biology, 1983, 169(1): 15.

[11] Cabrita L D, Dobson C M, Christodoulou J. Protein folding on the ribosome [J]. Current Opinion in Structural Biology, 2010, 20(1): 33-45.

[12] Hartl F U, Hayer-Hartl M. Converging concepts of protein folding in vitro and in vivo [J]. Nature structural & Molecular Biology, 2009, 16(6): 574-581.

[13] Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding [J]. Nature Structural & Molecular Biology, 2009. 16(3): 274-280.

[14] Tuller T, et al. Translation efficiency is determined by both codon bias and folding energy [J]. Proceedings of National Academy of Sciences, 2010, 107(8): 3645-50.

[15] Kramer G, et al. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins [J]. Nature Structural & Molecular Biology, 2009, 16(6): 589-97.

[16] Tuller T, et al., Composite effects of gene determinants on the translation speed and density of ribosomes [J]. Genome Biology, 2011, 12(11): R110.

[17] Chen H, Huang Y, Xiao Y. A simple method of identifying symmetric substructures of proteins [J]. Computational Biology and Chemistry, 2009, 33(1): 100-107.

[18] Sharp P M, Li W H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications [J]. Nucleic Acids Research, 1987, 15(3): 1281-1295.

[19] Shen X, Li G. Role for gene sequence, codon bias and mRNA folding energy in modulating structural symmetry of proteins [C] // IEEE Engineering in Medicine and Biology Society. 2013.