

# $\beta$ -螺旋桨折叠子基因内部特征与蛋白质结构的相关性分析

申小娟

邵阳学院信息科学与工程学院

DOI:10.12238/pe.v3i5.16616

**[摘要]** mRNA作为连接基因组与功能蛋白组的桥梁,它的序列特征不仅决定了蛋白质的翻译效率,而且还深刻影响着新生肽链的折叠过程。本文初步对五叶 $\beta$ -螺旋桨结构Tachylectin-2蛋白质的氨基酸序列以及基因序列的内部特征。研究表明,除了氨基酸序列存在与结构相应的五重复结构,它的基因序列上也存在多个特征与蛋白质的五叶 $\beta$ -螺旋桨结构相对应。密码子偏好、GC含量、氨基酸带电性以及mRNA局部折叠能量都出现了相应的局部连续上升或是连续下降。这些在不同特性中展现出的显著规律,共同构成了一个精细的生物学调控网络,影响翻译动力学和蛋白质折叠路径,并最终影响蛋白质的结构与功能。

**[关键词]** 基因序列; 螺旋桨结构; 密码子偏好; GC含量; mRNA局部折叠

**中图分类号:** U664.33 **文献标识码:** A

## Analysis on the association between intragenic features and symmetry in protein structures

Xiaojuan Shen

School of Information Science and Engineering, Shaoyang University

**[Abstract]** mRNA serves as a critical bridge between the genome and the functional proteome, with its sequence features influencing both translation efficiency and the folding of nascent polypeptides. Here, we performed a preliminary analysis of the amino acid and gene sequences of the five-bladed  $\beta$ -propeller protein Tachylectin-2. Results shows that beyond the same five repeat pattern in its protein sequence, the gene sequence also exhibits corresponding periodic features. Codon usage bias, GC content, amino acid charge distribution, and local mRNA folding energy all display coordinated local trends of continuous increase or decrease. These correlated patterns suggest a fine tuned regulatory network that modulates translation dynamics and co-translational folding, ultimately influencing the protein's structural organization and function.

**[Key words]** Gene sequence;  $\beta$ -propeller; Codon usage; GC content; local mRNA folding energy

## 引言

现在普遍认为细胞内蛋白质的三维结构的形成过程深受新生肽链翻译和折叠动力学的调控影响<sup>[1,2]</sup>。作为遗传信息的直接载体, mRNA不仅通过其开放阅读框(ORF)序列精确编码氨基酸序列,还通过隐含的“第二遗传密码”——如密码子使用偏好(codon usage bias)以及mRNA局部折叠能量(local folding energy)——精细调控翻译速率和共翻译折叠路径(co-translational folding)<sup>[3,4]</sup>。传统遗传学观点认为,同义密码子突变是“沉默”的,仅影响表达水平而非蛋白质结构;然而,近年来研究揭示,这些序列内部特征可诱导核糖体局部暂停(ribosome pausing),为多肽链提供时间窗口,避免错折(misfolding)并优化最终蛋白质构象。我们已经研究了左手螺旋管蛋白质(left-handed- $\beta$ -helix)的基因序列上存在高度保守的多层调控机制<sup>[5]</sup>。本文我们将以 $\beta$ -螺旋桨( $\beta$ -propeller)折叠子为对象,研究其基因

序列内部的特征,例如密码子使用偏好(codon usage bias)、GC含量(GC content)、氨基酸带电性(amino acid charge)以及mRNA局部折叠能量(local folding energy)与螺旋桨结构的相关性。

$\beta$ -螺旋桨结构是一种典型的折叠模块,由多个四股 $\beta$ -折叠片( $\beta$ -sheets)呈轮辐状排列形成,具有高度的稳定性和功能多样性<sup>[6]</sup>。 $\beta$ -螺旋桨折叠子的三级结构非常对称,具有4到8度的重复。这类结构以其对称的多叶重复单元(blades)广泛存在于信号转导、细胞黏附和免疫识别等功能蛋白中<sup>[7]</sup>。本文将以Tachylectin-2 (TL-2) (PDBid: 1TL2)为例,分析该蛋白基因序列及氨基酸序列的内部特征,包括氨基酸序列对称性、密码子偏好指数(codon adaptation index, CAI)、滑动窗口GC含量、氨基酸带电性分布以及局部mRNA二级结构自由能。我们假设,这些特征在基因序列中将呈现与五叶 $\beta$ -螺旋桨结构对应的局部连

续模式,如在叶边界处的密码子稀缺或GC上升,从而揭示一个精细的生物学调控网络。

## 1 数据与方法

### 1.1 数据说明

Tachylectin-2 (PDBid: 1TL2) 蛋白是一种源自马蹄蟹的凝集素,具有五叶 $\beta$ -螺旋桨核心结构<sup>[8]</sup>。它能够特异性结合糖分子,尤以N-乙酰脱氧葡萄糖胺(GlcNAc)/N-乙酰半乳糖胺(GalNAc)等糖型分子为目标。1TL2常被作为 $\beta$ -propeller折叠模型使用<sup>[9]</sup>。Smock et al. 在讨论对称蛋白进化de novo设计时,曾以tachylectin-2为典型的五叶螺旋桨结构例子。该蛋白质的氨基酸序列以及对应的基因序列从uniprot中获取,对应物种的基因组数据从Codon Usage Database上获得。

### 1.2 改进的重现图方法

氨基酸序列内部对称性可以利用改进的重现图方法进行计算。具体算法与相似性判定方法参见Shen (2016)中的相关描述。

### 1.3 局部密码子使用偏好分布

我们选用CAI值来衡量局部密码子使用偏好<sup>[10]</sup>。计算局部密码子使用偏好时,设置滑动窗口片段长度为40个密码子。

### 1.4 局部残基带电性分布

根据氨基酸侧链的电荷特性,可将氨基酸分为三类:带正电的氨基酸(Arg、His、Lys)、带负电的氨基酸(Asp、Glu)以及电中性的氨基酸(其余残基)。分别将带正电残基赋值为+1,带负电残基赋值为-1,电中性残基赋值为0。随后,采用长度为20个残基的滑动窗口,对蛋白质序列的局部电荷分布进行计算与分析。

### 1.5 局部GC含量分布

以30个核苷酸为滑动窗口长度,计算mRNA的局部GC含量分布。该窗口长度与核糖体在翻译过程中约40个核苷酸的移动步长相近,可用于表征局部序列组成变化。

### 1.6 局部mRNA折叠能量

本研究采用MATLAB中集成的RNAfold函数计算mRNA的折叠自由能。在计算中采用了局部折叠的方式来分析局部mRNA折叠能量的特征。我们设置窗口的长度为60个碱基,移动步长为1个碱基,用RNAfold计算每个窗口序列的折叠自由能,并获得mRNA上局部自由能的特征分布。

## 2 结果

### 2.1 蛋白质氨基酸序列对称性分析

图1 1TL2为蛋白质的结构及氨基酸序列对称性分析结果。其中图a为蛋白质的空间结构图。图c为1VZW结构对称关联图,可以看到氨基酸序列上有5个相似的片段,与蛋白质的5重对称结构正好对应。图b为1VZW的关联指数分布,箭头所指的5处有明显的峰值,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较高,蛋白质序列具有5重对称性,与结构的对称性正好对应。

### 2.2 局部特征分布

图2分别是蛋白质1TL2的各种局部特征分布图。图a蛋白质的密码子使用偏好CAI分布图。5个红色长方形框显示该区域有

明显的CAI值下降,表明这些区域内相较于其他区域有更多的使用稀有密码子。b为基因序列内部GC含量的分布图,5个红色长方形框显示该区域有明显的GC含量升高。更高的GC含量可能预示更稳定的局部mRNA结构。c为蛋白质氨基酸序列内部残基带电性的分布图,红色长方形框内显示该部位序列片段带负电的残基明显增多。d为mRNA局部折叠能量分布图,可以看到有5个明显的折叠能量下降的区域,更低的折叠能量预示更稳定的mRNA结构。

### 1TL2

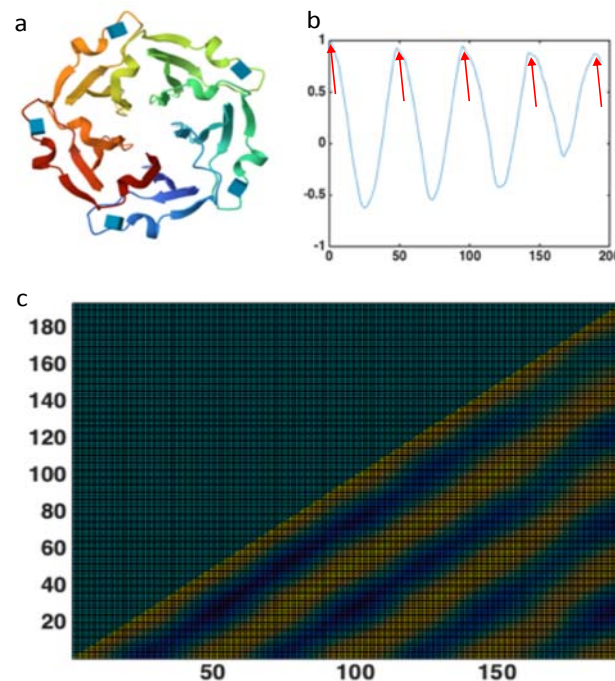


图1 蛋白质结构示意图、关联指数分布图、对称性关联图。图a、为蛋白质1TL2结构示意图,5对称的结构分别用不同的颜色标记出来。图b为关联指数分布图。图c为蛋白质的序列对称关联图。

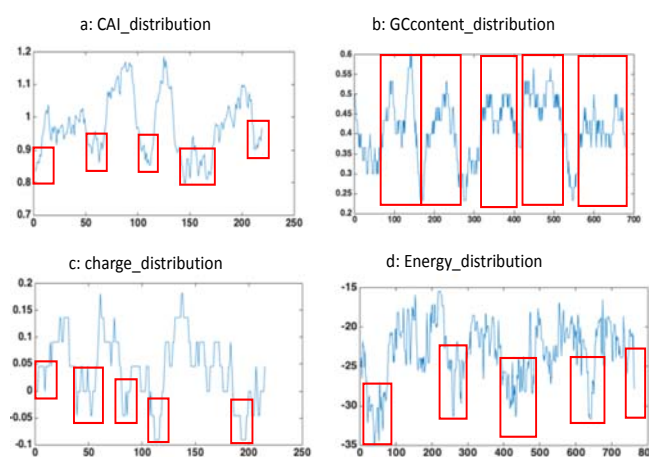


图2 密码子偏好分布图(a)、GC含量分布图(b)、残基带电性分布图(c)、局部mRNA折叠能量分布图(d)。

## 3 结论与展望

本文我们研究了 $\beta$ -螺旋桨折叠子中的Tachylectin-2(1TL2)蛋白质基因序列的内部特征与结构的关系,研究发现该蛋白质

的基因序列内部也存在与结构对称相应的保守信息。首先, 其氨基酸序列保留了对称信息, 同时, 局部密码子使用偏好、局部残基带电性、局部GC含量、以及mRNA局部折叠都出现了与结构5对称相对应的5重复结构: 5次局部密码子的下降, 5次GC含量的升高, 5次氨基酸序列的带电性下降, 5次局部能量的下降。这些特征说明1TL2螺旋桨蛋白质的基因序列里还保留了其他的保守机制, 它们之间可能形成一个协同网络, 驱动1TL2螺旋桨蛋白质结构在翻译过程中的精确组装, 并最终决定蛋白质的结构稳定性和功能特异性。

#### [致谢]

本研究得到湖南省自然科学基金省市联合基金项目(项目编号: 2022JJ50201)的资助。

#### [参考文献]

[1] Komar AA. A Code Within a Code: How Codons Fine-Tune Protein Folding in the Cell. *Biochemistry (Mosc)*. 2021 Aug; 86(8): 976–991.

[2] Liu Y, Yang Q, Zhao F. Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annual Review of Biochemistry*. 2021 Jun 20; 90:375–401. doi: 10.1146/annurev-biochem-071320-112701. Epub 2021 Jan 13. PMID: 33441035; PMCID: PMC8284178.

[3] Morales-Polanco, F., Lee, J.H., Barbosa, N.M. and Frydman, J., 2022. Cotranslational mechanisms of protein biogenesis and complex assembly in eukaryotes. *Annual Review of Biomedical Data Science*, 5(1), pp.67–94.

[4] Moss, M.J., Chamness, L.M. and Clark, P.L., 2024. The effects

of codon usage on protein structure and folding. *Annual Review of Biophysics*, 53.

[5] Shen, X., Huang, T., Wang, G. and Li, G., 2015. How the sequence of a gene specifies structural symmetry in proteins. *Plos one*, 10(12), p.e0144473.

[6] Fülöp, V. and Jones, D.T., 1999.  $\beta$  Propellers: structural rigidity and functional diversity. *Current Opinion in Structural Biology*, 9(6), pp.715–721.

[7] Pons, T., Gmez, R., Chinea, G. and Valencia, A., 2003. Beta-propellers: associated functions and their role in human diseases. *Current medicinal chemistry*, 10(6), pp.505–524.

[8] Beisel, H.G., Kawabata, S.I., Iwanaga, S., Huber, R. and Bode, W., 1999. Tachylectin - 2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab *Tachypleus tridentatus*. *The EMBO journal*.

[9] Smock, R.G., Yadid, I., Dym, O., Clarke, J. and Tawfik, D.S., 2016. De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell*, 164(3), pp.476–486.

[10] Shen, X., 2011. Conformation and sequence evidence for two-fold symmetry in left-handed beta-helix fold. *Journal of Theoretical Biology*, 285(1), pp.77–83.

#### 作者简介:

申小娟(1981—), 女, 汉族, 湖南省邵阳市人, 邵阳学院信息科学与工程学院, 研究生, 助理研究员, 计算生物。