

文章编号 :1009-038X(2001)06-0652-04

核酸序列是否包含更多蛋白质结构信息

李炜疆, 宋江宁

(江南大学 工业生物技术教育部重点实验室 江苏 无锡 214036)

摘 要 :氨基酸序列决定蛋白质高级结构是普遍接受的假设,也是蛋白质结构预测的理论基础.但近来不断有研究表明,同义密码子用语和蛋白质二级结构间存在着某些关联,这些结果可能对蛋白质折叠理论和基因工程有积极意义.本文对这一课题研究的最新进展情况进行了简要的分析与评述.

关键词 :同义密码子用语;蛋白质二级结构;统计关联;蛋白质折叠

中图分类号:Q 61

文献标识码:A

Do Nucleotide Sequences Contain More Information About Protein Structures

LI Wei-jiang, SONG Jiang-ning

(The Key Laboratory of Industrial Biotechnology under Ministry of Education, Southern Yangtze University, Wuxi 214036, China)

Abstract : It is generally accepted that amino acid sequences determine the spatial structures of proteins. Protein structure prediction methods are explicitly or implicitly based on this assumption. However, recent researches show that there do exist some specific correlations between the synonymous codon usage and protein secondary structure, which is meaningful to protein folding theory and genetic engineering. Recent advances on this subject are briefly reviewed.

Key words : synonymous codon usage; protein secondary structure; statistical correlation; protein folding

一般认为,蛋白质高级结构是由其一级结构即氨基酸序列所决定的,也就是说,氨基酸序列足以提供蛋白质折叠到其天然结构所需要的信息,即 Anfinsen 原理^[1].据此,人们倾向于认为多肽链合成之后能自发地折叠形成为其特定氨基酸序列所决定的三维空间结构,亦即蛋白质自组装原理(principle of protein self-assembly)^[2].几乎所有的蛋白质结构预测方法,无论其具体手段如何,事实上

都以此为基础.

氨基酸序列由 mRNA 序列编码,但由于遗传密码的简并性, mRNA 的编码序列与相应的氨基酸并非一一对应关系,而是多对一的关系.因此,核酸编码序列包含比氨基酸序列更多的信息.这些信息是否都与蛋白质无关,或者说,编码同一蛋白质的不同 mRNA 序列都能使其翻译产物——氨基酸序列折叠到同样的空间结构吗?对此问题,研究同义密

收稿日期 2001-10-22; 修订日期 2001-10-24.

基金项目 教育部工业生物技术重点实验室访问学者基金资助课题.

作者简介 李炜疆(1964-),男,陕西榆林人,理学博士,教授,博士生导师.

码子用语是否携带蛋白质二级结构信息是很好的切入点. 一方面, 蛋白质二级结构是三级结构的基础, 对二级结构的影响必然涉及整体空间结构; 另一方面, 二级结构(尤其是螺旋结构)通常在蛋白质折叠早期快速形成, 这是整个折叠过程最可能受翻译过程影响的阶段, 更高级结构的形成要慢得多, 如果同义密码子用语和蛋白质二级结构无关, 则几乎可以肯定也和更高级结构无关.

1 影响同义密码子用语的因素

同义密码子用语是非随机的, 受到多种因素的影响^[3] (1) 种属特异性^[4], 不同的生物有各自的同义密码子用语偏倚性, 即使是对于同一种生物, 同义密码子用语在不同基因中也有着显著的差异, 在细菌中, 即使是最常用的同义密码子, 其翻译准确性差别也很大^[5-9] (2) tRNA 丰度, 密码子用语随着 tRNA 丰度的变化而选择相应的同义密码子, 高表达基因偏好使用 tRNA 丰富的密码子或者能够与 tRNA 更好结合的密码子, Bulmer 推测这可能与密码子用语和 tRNA 丰度偶联演化相关^[10, 11] (3) 基因表达水平^[4, 12-14] (4) 密码子第三位碱基的 (G + C)%, 高效表达的大肠杆菌蛋白质的三联体密码子第三位更倾向于使用碱基 C (与 U 的使用频率相比)^[12] (5) 基因组中基因在细胞内的分布位置^[15]. Purvis 等人还假设同义密码子用语与蛋白质折叠位点相关, 在翻译暂停处引入翻译速度低的同义密码子, 可能影响蛋白质的折叠^[16].

蛋白质在核糖体上的合成过程中, 由于密码子与反密码子的特性以及细胞环境的不同, 编码同样氨基酸序列的不同 mRNA 翻译表达新生肽链的过程很不相同. 对大肠杆菌编码区不同位置上的同义密码子使用情况的研究表明, 多数氨基酸的密码子用语在翻译起始区变化显著, 只有少数氨基酸在翻译终止区有微弱变化^[17]. 此外, 同义密码子用语与氨基酸序列上下文间存在密切的关联, 尤其是氨基

酸对于 N 端紧邻密码子用语影响最为显著^[18]. 如果蛋白质折叠与翻译偶联 (co-translational folding), 核酸序列之间的差异就极有可能影响到蛋白质折叠的动力学过程, 因而就有可能对蛋白质空间结构产生影响. 近年来, 蛋白质结构数据库的空间结构数据不断增多, 使得对这个问题的进一步统计研究成为可能.

这方面的研究已有若干报导^[3, 9, 15, 19-21], 本文综述了主要相关文献的研究结果, 并比较了它们的研究方法和研究结果的差异.

2 基因序列 – 结构对应数据集

如果同义密码子的使用与蛋白质的结构有关, 那么唯一可能产生关联的环节就是核酸序列在核糖体上合成新生肽链的过程. 收集尽可能多足够准确的蛋白质结构与对应的核酸编码序列数据, 是研究的先决条件, 这涉及不同的大分子基础数据库, 蛋白质结构数据来自 PDB, 而核酸编码序列来自 EMBL 或 GenBank. 由于在蛋白质结构数据和相应的核酸编码序列间没有直接的交叉索引数据资料将它们联系起来, 使得数据收集工作比较繁琐并有一定的不确定性. 不同生物的蛋白质翻译机制有很大差异, 这很可能影响同义密码子与蛋白质的结构关联, 所以通常对不同物种的数据分别处理.

Adzhubei 等^[22, 23]曾试图构建一个比较完善的蛋白质结构与对应基因编码序列的数据库, 并在 Nucleic Acid Research 的数据库特刊号上两度撰文介绍, 但其设想一直未能很好实现. 各研究组的研究结果都基于各自建立的数据集^[3, 15, 19, 20], 其中以来源于人和大肠杆菌的蛋白质最多. 最近, 作者构建了一个高精度大肠杆菌蛋白质结构与基因序列对应数据集 EcoPDB^[24], 收录的蛋白质结构数据测量精度均好于 2.5 Å, 包含有 191 个 *E. coli* 基因及其相应的 PDB 蛋白质二级结构数据, 总残基数为 55 763. 有关数据集的情况见表 1.

表 1 基因序列 – 结构对应数据集

Tab. 1 The dataset of coding sequences and corresponding protein secondary structures				
数据集	<i>E. coli</i> 记录 (蛋白质/残基数目)	<i>B. subtilis</i> 记录 (蛋白质/残基数目)	<i>S. cerevisiae</i> 记录 (蛋白质/残基数目)	<i>Homo sapiens</i> 记录 (蛋白质/残基数目)
1	54/14 107	无	无	77/12 465
2	31/7 297	无	无	35/6 304
3	28/	28/	29/	79/
4	191/55 763	无	无	117/25 882

注: 数据集 1~4 分别为解涛、丁达夫^[19, 20], Oresic 和 Shalloway^[3], Gupta 等^[15], 以及作者构建的蛋白质结构与基因编码序列对应数据集. 其中 *E. coli* 记录来自 EcoPDB^[24], *Homo Sapiens* 记录来自 Adzhubei 等人构建的数据库 ISSD2.0^[23].

3 同义密码子与蛋白质二级结构 关联的统计检验

同义密码子用语与蛋白质二级结构是否存在关联,可以通过比较同义密码子的二级结构偏好性是否一致进行检验.如果同一氨基酸的不同密码子对二级结构的偏好有显著差异,则说明同义密码子与蛋白质二级结构有关,或说携带有二级结构信息.

解涛、丁达夫^[19,20]对所有同义密码子在不同的各种二级结构单元中的出现频率进行了比较,用 Z 分数作为统计检验量,发现大肠杆菌只有密码子 GGU 和 GGC 的 Z 分数明显超出了统计显著性的阈值,对哺乳类的数据统计发现,有 9 种氨基酸的同义密码子携带有蛋白质二级结构信息.据此,他们认为,真核生物的同义密码子用语要比原核生物携带更多的蛋白质结构信息.解涛、丁达夫的研究结论基本上与 Adzhubei 等^[21]的结论一致.

Oresic 和 Shalloway^[3]分析了大肠杆菌和人蛋白质的三维空间结构与相应的 mRNA 序列,采用联列表对同义密码子在不同二级结构单元中的出现概率进行比较.与解涛、丁达夫^[19,20]不同的是,他们还考虑了同义密码子对邻近氨基酸二级结构的影响.发现同义密码子和蛋白质结构存在种属特异性关联:人 Asp 的 GAU 密码子在 N 端处有明显的 α 螺旋倾向,而 *E. coli* Asn 的 AAC 密码子在 C 端则有明显的 α 折叠偏好.他们分析说,这种种属特异性关联可能反映了不同种类生物间的翻译机制差异,有助于确认那些对于蛋白质二级结构形成有重要意义的氨基酸残基,改进蛋白质二级结构预测算法,以及指导重组基因表达.

Gupta 等^[15]用 χ^2 检验分析比较编码蛋白质 α 螺旋和 β 折叠的同义密码子的出现频率,寻找同义密码子用语对 α 螺旋或 β 折叠的偏好性.他们发现在同义密码子和蛋白质二级结构单元之间并不存在普遍意义上的种属特异性关联.另外,他们通过分析高效表达基因的同义密码子用语在二级结构单元中的频率,没有发现象 Adzhubei 等^[21]提出的那样,两者之间存在普遍的关联.

4 讨论

同义密码子用语不可能是蛋白质折叠最重要

的影响因素,但并不能排除存在少数几种密码子用语对蛋白质的折叠产生重大作用的可能性.以上研究结果多少证实了这种可能性确实是存在的.尽管细节有所不同,但在总体上是一致的,即同义密码子与蛋白质二级结构存在较弱但是显著的关联.其中存在着某些争议之处,部分原因是数据量较小而导致的统计涨落.因为所研究的关联即使存在也是比较弱的,统计方法的不同并不导致统计结果本质性的差异.

原核生物和真核生物蛋白质的体内折叠机制不同,很多细菌蛋白质的折叠依赖于分子伴侣才能完成,而真核生物只有少数特殊的蛋白质需要分子伴侣协助折叠^[25-27],这反映了真核生物可能有其它机制来防止产生误折叠^[3],而通过同义密码子辅助调节则是一条可能且经济的途径.这可能对某些外源蛋白质表达的基因工程有重要的参考价值.

蛋白质二级结构本身就是氨基酸之间相互作用的结果,从这个意义上讲,单个氨基酸所包含的二级结构信息量很小,还必须考察密码子上下文的关系,而考察这种关系,最简单的研究对象就是二联体.密码子和蛋白质发生作用的纽带是蛋白质翻译过程, Irwin 等^[28]的研究表明,蛋白质翻译速率受密码子上下文的影响,而这种上下文效应与邻近的位于翻译核糖体表面 A、P 位点上的 tRNA 分子之间的融洽性有关,在翻译过程中,核糖体上 A、P 位点上有两个紧邻的密码子在作用,它们作用之间有一个衔接的过程,因而在此过程中两者极有可能发生相互作用.

由此我们注意到,前面介绍的统计分析研究的一个重大缺陷,即仅研究了单个同义密码子用语和蛋白质二级结构的关联,没有考察密码子二联体(codon dimmers)或密码子对(codon pairs)与蛋白质结构的关联,其主要原因是样本数据量规模的限制.作者在自己构建的更大规模数据集 EcoPDB^[24]的基础上,使用信息论方法对此进行更细致的检验,初步的研究结果是单个大肠杆菌密码子基本不携带二级结构信息,但是如果对于二肽和密码子二联体,情况则有很大的不同,密码子二联体携带的二级结构信息量要比二肽明显地大,这似乎表明,密码子上下文关联对于蛋白质结构有重要的影响,而外源基因表达往往不能形成正确的蛋白质结构就可能与此有关^[29].

参考文献:

- [1] ANFINSEN C.B. Principles that govern the folding of protein chains[J]. *Science*, 1973, 181: 223~230.

- [2] 静国忠. 基因及其分子生物学基础 [M]. 北京 : 北京大学出版社 , 1999.
- [3] ORESIC M , SHALLOWAY D. Specific correlation between relative synonymous codon usage and protein secondary structure [J]. **J Mol Biol** , 1998 , 281 : 31~48.
- [4] GRANTHAM R , GAUTIER C , GOUY M , *et al.* Codon catalog usage and the genome hypothesis [J]. **Nucleic Acids Res** , 1981 , 8 : 49~62.
- [5] GROSJEAN H , FIERIS W. Preferential codon usage in prokaryotic genes : the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes [J]. **Gene** , 1982 , 18 : 199~209.
- [6] PARKER J , JOHNSTON T C , BORGIA P T , *et al.* Codon usage and mistranslation. In vivo basal level misreading of the MS2 coat protein message [J]. **J Biol Chem** , 1983 , 258 : 10007~10012.
- [7] PRECUP J , PRKER J. Missense misreading of asparagines codons as a function of codon identity and context [J]. **J Biol Chem** , 1987 , 262 : 11351~11355.
- [8] MCPHERSON D T. Codon preference reflects mistranslational constraints : a proposal [J]. **Nucleic Acids Res** , 1988 , 16 : 4111~4120.
- [9] KARLIN S , MRAZEK J. What drives codon choices in human genes [J]. **J Mol Biol** , 1996 , 262 : 459~472.
- [10] BULMER M. Coevolution of codon usage and transfer RNA abundance [J]. **Nature** , 1987 , 325 : 728~730.
- [11] BULMER M. The selection-mutation-drift theory of synonymous codon usage [J]. **Genetics** , 1991 , 129 : 897~907.
- [12] GRANTHAM R , GAUTIER C , GOUY M , *et al.* Codon catalog usage is a genome strategy modulated for gene expressivity [J]. **Nucleic Acids Res** , 1987 , 9 : 43~74.
- [13] IKEMURA T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in protein genes [J]. **J Mol Biol** , 1981 , 146 : 1~21.
- [14] IKEMURA T. Codon usage and tRNA content in unicellular and multicellular organisms [J]. **Mol Biol Evol** , 1985 , 2 : 13~34.
- [15] GUPTA S K , MAJUMDAR S , BHATTACHARYA T K , *et al.* Studies on the synonymous codon usage and protein secondary structural units [J]. **Biochemical and Biophysical Research Communications** , 2000 , 269 : 692~696.
- [16] PURVIS I J , BETTANY A J E , SANTIAGO T C , *et al.* The efficiency of folding of some proteins is increased by controlled rates of translation in vivo-a hypothesis [J]. **J Mol Biol** , 1987 , 193 : 413~417.
- [17] 李炜疆. 同义密码子用语的位置依赖 [J]. 生物物理学报 , 2001 , 17 : 529~534.
- [18] 李炜疆. 同义密码子用语与氨基酸序列上下文的关联 [J]. 无锡轻工大学学报 , 2001 , 20 : 1~5.
- [19] 解涛 , 丁达夫. 同义密码子用语与蛋白质结构的关系 [J]. 生物物理学报 , 1998 , 14 : 296~302.
- [20] XIE TAO , DING DAFU. The relationship between synonymous codon usage and protein structure [J]. **FEBS Letters** , 1998 , 434 : 93~96.
- [21] ADZHUBEI I A , ADZHUBEI I A , KRASHENINNIKOV I A , *et al.* Non-random usage of " degenerate " codons is related to protein three-dimensional structure [J]. **FEBS Letters** , 1996 , 399 : 78~82.
- [22] ADZHUBEI I A , ADZHUBEI A A , NEIDLE S. An Integrated Sequence-Structure Database incorporating matching mRNA sequence , amino acid sequence and protein three-dimensional structure data [J]. **Nucleic Acids Res** , 1998 , 26 : 327~331.
- [23] ADZHUBEI I A , ADZHUBEI A A. ISSD Version 2.0 : taxonomic range extended [J], **Nucleic Acids Res** , 1999 , 27 : 268~271.
- [24] 李炜疆 , 宋江宁. EcoPDB : 高精度大肠杆菌蛋白质结构与对应基因序列数据集 [J]. 无锡轻工大学学报 , 2001(4) : 340~343.
- [25] HARTL F U. Molecular chaperones in cellular protein folding [J]. **Nature** , 1996 , 381 : 571~580.
- [26] KUBOTA H , HYNES G , WILLSON K. The chaperonin containing t-complex polypeptide 1 (tcp-1) : a multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol [J]. **Eur J Bio Chem** , 1995 , 230 : 3~16.
- [27] LEWIS S A , TIAN G L , VAINBERG I E , *et al.* Chaperonin-mediated folding of actin and tubulin [J]. **J Cell Biol** , 1996 , 132 : 1~4.
- [28] IRWIN B , HECK J D , HATFIELD G W. Codon pair utilization biases influence translational elongation step times [J]. **J Biol Chem** , 1995 , 270 : 22801~22806.
- [29] 宋江宁 , 李炜疆. 基因序列与蛋白质结构关联的再检验 [R]. 首届中国生物信息学大会 , 北京 , 2001.