

## A型流感病毒同义密码子的使用偏性对RNA二级结构的影响

李瑞芳\*, 于志芬, 黄俏, 郭春阳, 李雪

(内蒙古师范大学 物理与电子信息学院, 呼和浩特 010022)

**摘要:**以A型流感病毒为研究样本,分析了同义密码子使用偏性对RNA二级结构的影响,为进一步研究同义密码子存在的意义及A型流感病毒RNA特征提供一些理论依据。收集整理了NCBI中收录的全部A型流感病毒的核酸序列信息,计算了每一条核酸序列的RNA二级结构,计算出RNA环结构含量和茎结构含量及折叠自由能。在此基础上,计算了RNA二级结构的柔性。同时,计算了每一条核酸序列的相对同义密码子使用值。由此,建立了A型流感病毒RNA二级结构数据库。分析每条核酸序列的相对同义密码子使用值与RNA的环结构、茎结构及柔性之间的关系。分析结果表明,50%的氨基酸的相对同义密码子使用值与RNA茎结构含量和环结构含量显著相关;60%的氨基酸的相对同义密码子使用值与单位平均折叠自由能显著相关;50%的氨基酸的相对同义密码子使用值与RNA柔性显著相关。进一步分析发现,与茎结构含量和环结构含量都显著相关的密码子,它们的相对同义密码子使用值与两种结构含量的相关性截然相反,而且发现,在所选的参量中,RNA柔性相对同义密码子使用值显示出更好的相关性。结果证实,对于A型流感病毒,同义密码子的使用偏性对RNA二级结构存在很大的影响。

**关键词:**A型流感病毒;同义密码子使用偏性;RNA二级结构;RNA柔性;折叠自由能

**中图分类号:**R373.1<sup>+</sup>3 **文献标识码:**A **文章编号:**1000-8721(2016)06-0773-09

DOI:10.13242/j.cnki.bingduxuebao.003076

RNA扮演了基因表达的重要角色,它首先作为蛋白质合成的中间体被发现。在生物学功能上,mRNA是一种既可以负责遗传信息传递,又可以作为遗传信息载体的多功能分子,并以各种不同的方式行使这些功能。研究表明,mRNA分子形状是非常多变的,它作为遗传信息的临时携带者,不断与蛋白质之类的分子结合,且不具有独立而稳定的三级结构。而且mRNA分子的折叠二级结构比其三级结构更有利于行使作为信使的基本功能<sup>[1]</sup>。这使得对RNA分子二级结构的研究显得更为重要了。有关RNA二级结构的研究,人们在多方面已经作了许多尝试,并且得到了很多较成熟的预测RNA二级结构的方法<sup>[2,3]</sup>。而RNA序列对RNA二级结构的影响是多方面的,RNA不仅包含了编码蛋白的遗传语言,还包含了其它丰富的信息内容。这主要表现在密码子的“摆尾”现象,即密码子第三位点碱基使用的自由度问题。由于遗传密码的简并性,基因

序列包含了比其编码的氨基酸序列更多的信息。基因序列的附加自由度不仅可以减小突变危险性<sup>[4]</sup>,还参与基因表达调控<sup>[5,6]</sup>,但当基因序列最终转译为氨基酸序列后,额外的信息就消失了。生命过程是否利用了这些信息?这是值得深入研究的问题。

早期的研究认为密码子偏性与诸多因素有关,如:氨基酸组分<sup>[7]</sup>、翻译起始效应、G+C含量<sup>[8]</sup>、基因长度<sup>[9,10]</sup>、tRNA的丰度<sup>[11]</sup>、蛋白质的结构<sup>[12]</sup>等,同时,同义密码子的使用直接或者间接地影响着生物体内基因的转录,信使RNA的翻译,以及合成蛋白质的结构与活性,具有重要的生物学意义<sup>[11,13]</sup>。近年来,相关研究在进一步快速发展,研究范围也在不断的扩大。随着基因组学研究的继续深入和蛋白质组学的大规模兴起,密码子偏性问题必然会涉及更广泛的生物学领域。有关密码子使用偏性与mRNA的二级结构之间的相关性,研究者们曾以蚕丝蛋白,大肠杆菌等为样本进行了研究<sup>[14-16]</sup>,结果表明密码子偏好对mRNA二级结构有一定的影响,而且指出虽然影响不是很大,但它们之间的影响却是不能忽略的。我们认为同义密码子的使用偏性是影响RNA二级结构的形成和功能的重要因素。

自遗传密码破译以来,研究者通过对原核生物和真核生物的基因编码序列的转录和翻译功能的研究发现,同义密码子的使用与基因表达存在着密切的关系<sup>[17]</sup>。不但不同生物体使用的偏好密码子可

收稿日期:2015-11-09 修回日期:2016-09-28

基金项目:内蒙古自治区自然科学基金项目(项目编号:2016MS0362),题目:氨基酸特性及mRNA序列和结构对相应蛋白质折叠速率的影响

作者简介:李瑞芳(1974-),女,内蒙古自治区,副教授,主要从事量子光学及生物物理研究,Tel:15384815995,E-mail:lirui芳@imnu.edu.cn

\*通讯作者:李瑞芳(1974-),女,内蒙古自治区,副教授,主要从事量子光学及生物物理研究,Tel:15384815995,E-mail:lirui芳@imnu.edu.cn

能不同,甚至同一生物体内不同基因也有着不同的密码子使用倾向<sup>[18]</sup>,所以,为了避免所选样本物种差异对结果的影响,本文选择同一种病毒的所有亚型病毒 RNA 作为研究对象,我们注意到 A 型流感病毒(Influenza A virus, IAV)都是单股负链 RNA 病毒<sup>[19]</sup>,被分为 17 种 HA 亚型和 10 种 NA 亚型。目前对 A 型流感病毒宿主、分离、生活习性等方面已经有很多研究<sup>[20-24]</sup>,但是人类与 A 型流感病毒的战争还在继续。每年呈季节性流行的 H1N1 和 H3N2 A 型流感病毒可造成全球 25 万人~50 万人感染死亡<sup>[25]</sup>,历史上暴发过多次 A 型流感病毒引起的全球流感大流行,造成大规模人群感染死亡,近年来不断涌现的各种人感染禽流感病毒死亡事件也均是由 A 型流感病毒引发的,因此,系统研究 A 型流感病毒有助于加强人类对该病毒的防控能力。本文试图从这类病毒的 RNA 的序列和结构特性入手。寻找同义密码子的使用偏性对 RNA 二级结构的影响。期望通过本文的研究对 A 型流感病毒 RNA 序列和结构有更深入的了解。

## 材料与方法

**1 数据获得** 以 A 型流感病毒为研究对象,首先在 NCBI 数据库中整理出所有 IAV 核酸序列信息,得到 17 个 HA 亚型和 10 个 NA 亚型的全部流感病毒组合的核酸序列。然后,运用 RNAstructure 5.6 程序<sup>[26]</sup>预测计算所有核酸序列的 RNA 二级结构,计算出它们的茎结构含量、环结构含量,以及平均折叠自由能,以此为基础,计算折叠自由能标准偏差和 RNA 柔性值。将 IAV 病毒名称, RNA 茎结构含量,环结构含量及 RNA 柔性值分别列在附表 1 中。

**2 RNA 二级结构预测** 有关 RNA 二级结构预测的程序有很多种,本文使用的是 RNAstructure 5.6 程序。该软件应用了多种算法,包括最小自由能,计算配分函数和估计碱基配对概率,优化期望准确度,随机抽样等,是一个能够较精确地预测分析 RNA 二级结构的程序。对于每个病毒的每个蛋白 RNA 序列,本文用该程序计算它们的二级结构,得到它们的茎结构含量,环结构含量,折叠自由能以及 RNA 柔性值。

### 3 RNA 二级结构特征量提取

**3.1 RNA 的茎结构、环结构含量** RNA 的茎结构、环结构是组成 RNA 二级结构的基本单元,它们的含量必将会影响 RNA 的二级结构,因此对 RNA

的茎结构、环结构含量做如下定义:

$$\tilde{N}_{sp} = \frac{N_{sp}}{N} \quad (1)$$

$$\tilde{N}_L = \frac{N_L}{N} \quad (2)$$

式中,  $\tilde{N}_{sp}$  和  $\tilde{N}_L$  分别定义为某条 RNA 的茎结构含量和环结构含量,  $N_{sp}$ 、 $N_L$  分别为这条 RNA 中参与茎结构和环结构核苷酸的数量,对于  $N_{sp}$ ,定义为茎结构中其中一条链的结构含量,就是说只计算每一个茎结构中其中一条链的核苷酸数量(即参与总茎结构核苷酸数量的一半),  $N$  为这条 RNA 序列的长度(即包含核苷酸的个数)。

**3.2 RNA 单位平均折叠自由能** 折叠自由能通常被认为是影响 RNA 二级结构稳定性的重要因素,而且自由能越低结构就越稳定。本文使用 RNAstructure 5.6 程序计算 RNA 折叠自由能,然而在计算结果中折叠方式有多种,即对应的能量也就多种,同时为了避免序列长度对结果的影响,本文定义了 RNA 单位平均折叠自由能  $E^*$ :

$$E^* = \frac{\bar{E}}{N} \quad (3)$$

式中,  $E^*$  为单位平均折叠自由能(RNA 平均折叠自由能与序列长度的比值),  $\bar{E}$  为 RNA 所有可能折叠方式折叠自由能的平均值,  $N$  为 RNA 序列的长度。考虑到 20 种最可能的折叠方式已足够精确,我们的 RNAstructure 5.6 程序计算中设定最多折叠方式数为 20。例如, A 型流感病毒 H1N1 蛋白的 RNA 序列长度( $N$ )为 2 233bp,利用 RNAstructure 5.6 程序得到其可能的最小自由能值分别为 -597.8, -597.7, -597.4, -595.9, -595.8, -595.7, -595.7, -594.9, -594, -593.6, -593, -592.6, -592.2, -592, -591.7, -591.6, -591.4, -591, -590.9, -590.6。则平均能量  $\bar{E} = (-597.8 - 597.7 - 597.4 - 595.9 - 595.8 - 595.7 - 595.7 - 594.9 - 594 - 593.6 - 593 - 592.6 - 592.2 - 592 - 591.7 - 591.6 - 591.4 - 591 - 590.9 - 590.6) / 20 = -593.78$ 。单位平均折叠自由能  $E^* = -593.775 / 2233 = -0.27$ 。

**3.3 RNA 的柔性** 迄今为止,还没有描述 RNA 序列动力学柔性的合适参数。因此,仿照 DNA 序列柔性的描述方法<sup>[27,28]</sup>,引进 RNA 柔性的概念,在给定温度下,一条 RNA 序列通常存在多种可能的折叠方式,不同方式的折叠自由能之间存在差异,用序列折叠自由能的标准偏差来定义 RNA 的柔性。折叠自由能的差异越大, RNA 的柔性就越大。

RNAstructure 程序是预测 RNA 二级结构的

常用程序之一,使用该程序可计算 RNA 序列可能出现的二级结构及其折叠自由能。计算出每一条 RNA 序列折叠自由能的平均值  $\bar{E}$  和标准偏差  $Sd_E$

$$\bar{E} = \frac{\sum_{j=1}^n E_j}{n} \quad (4)$$

$$Sd_E = \sqrt{\frac{\sum_{j=1}^n (E_j - \bar{E})^2}{(n-1)}} \quad (5)$$

$$SF = 100 \times \frac{Sd_E}{|\bar{E}|} \quad (6)$$

式中,  $\bar{E}$  为 RNA 所有可能折叠方式折叠自由能的平均值,  $E_j$  ( $E_j \leq 0$ ) 为第  $j$  ( $j = 1, 2, \dots, n$ ) 种折叠结构对应的折叠自由能,  $n$  为可能折叠方式的数目,  $Sd_E$  反映了 RNA 结构的可变性。SF (specific flexibility) 是以 RNA 折叠自由能为基础的一个综合参量,这一参量有效地描述了 RNA 结构的可变性和稳定性,定义为 RNA 柔性<sup>[29]</sup>。

对于每个蛋白质对应的 RNA 序列,分别使用 RNAstructure 5.6 程序得到它们的茎结构和环结构的数量,以及可能的折叠方式的平均自由能,然后依据上述公式分别计算出每个蛋白质对应的 RNA 的茎结构含量,环结构含量,单位平均折叠自由能和 RNA 柔性值。

**4 相对同义密码子使用值** 本文用相对同义密码子使用值 (RSCU 值) 来描述密码子的偏性,它的计算方法如公式 (7)。由于该指标计算简单,而且除去了氨基酸的组成对密码子使用的影响,所以能比较直观的反映出同义密码子的使用偏好。

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}} \quad (7)$$

式中,  $RSCU_{ij}$  是编码第  $i$  个氨基酸的第  $j$  个密

码子的密码子使用值,  $X_{ij}$  是编码第  $i$  个氨基酸的第  $j$  个密码子的出现次数,  $n_i$  是编码第  $i$  个氨基酸的同义密码子数量 ( $n_i$  值为 1~6)<sup>[30]</sup>。本文运用了 Codon W 软件对密码子使用进行了分析,该软件可以对密码子使用进行多变量分析,还可以对密码子使用与氨基酸的相关性进行分析。

考虑到为 RSCU 的值只有在编码蛋白序列中计算才有意义,在计算 RSCU 时,去掉其中没有参与编码蛋白序列的碱基,发现虽然只有少数碱基没有参与编码蛋白序列,但是与用整体 RNA 序列计算出的 RSCU 结果差距甚大。所以,首先整理每条 A 型流感病毒 RNA 序列,去掉没有参与编码蛋白序列的碱基,然后通过使用 Codon W 软件,得到了 126 条 RNA 序列的相对同义密码子使用值。

## 结 果

对于全部的 A 型流感病毒,使用上述方法计算出相应 RNA 序列中包含的茎结构含量 ( $\tilde{N}_{sp}$ )、环结构含量 ( $\tilde{N}_L$ )、RNA 单位平均折叠自由能 ( $E^*$ ) 和 RNA 的柔性 (SF) 值。然后对这四种参量,分别分析了它们与相对同义密码子使用值 (RSCU 值) 之间的相关性。分析结果见表 1 至表 3。

### 1 A 型流感病毒相对同义密码子使用值对 RNA 二级结构环结构和茎结构的影响

对 A 型流感病毒的每条核酸序列,计算得到茎结构含量与环结构含量,并与 RSCU 值进行相关性分析,得到它们与 64 组密码子使用值的相关性,结果如表 1。表中只列出了与茎结构、环结构含量具有相关性的密码子。

表 1 相对同义密码子使用值与环结构含量及茎结构含量之间的相关性分析结果

Table 1 Linear regression between relative synonymous codon usage with the content of loop and stem structure

Amino	Codon	Loop structure		Stem structure		N	df
		P	R	P	R		
Leu	UUA	0.02	0.21*	0.02	-0.21*	126	124
Leu	UUG	0.0065	-0.24**	0.0065	0.24**	126	124
Ile	AUC	0.0207	-0.21*	0.0207	0.21*	126	124
Ile	AUA	2.0E-06	0.41***	2.0E-06	-0.41***	126	124
Val	GUU	0.0004	0.31***	0.0004	-0.31***	126	124
Val	GUG	0.0003	-0.31***	0.0003	0.31***	126	124
Thr	ACU	0.0244	-0.20*	0.0244	0.20*	126	124
Thr	ACC	0.0098	-0.23**	0.0098	0.23**	126	124
Thr	ACA	2.1E-07	0.44***	2.1E-07	-0.44***	126	124
Gln	CAA	2.9E-05	0.36***	2.9E-05	-0.36***	126	124

Amino	Codon	Loop structure		Stem structure		N	df
		P	R	P	R		
Gln	CAG	2.9E-05	-0.36***	2.9E-05	0.36***	126	124
Asp	GAU	0.0124	-0.22**	0.0124	0.22**	126	124
Asp	GAC	0.0124	0.22**	0.0124	-0.22**	126	124
Glu	GAA	0.0071	0.24**	0.0071	-0.24**	126	124
Glu	GAG	0.0071	-0.24**	0.0071	0.24**	126	124
Arg	CGU	0.0107	-0.23**	0.0107	0.23**	126	124
Arg	CGA	0.0131	-0.22**	0.0131	0.22**	126	124
Ser	AGU	0.0003	-0.31***	0.0003	0.31***	126	124
Ser	AGC	2.4E-07	0.44***	2.4E-07	-0.44***	126	124
Gly	GGU	0.0006	-0.30***	0.0006	0.30***	126	124
Gly	GGA	0.0033	0.26**	0.0033	-0.26**	126	124

Note: Two-tailed significance: \* :  $P \leq 0.05$ ; \*\* :  $P \leq 0.01$ ; \*\*\* :  $P \leq 0.001$ . R is the correlation coefficient, P is the significance level, N is the number of samples, and df is the degree of freedom.

对 RNA 二级结构的环结构和茎结构,均有 10 种氨基酸的 21 个密码子使用度与它们显著相关。相关密码子数占这 10 种氨基酸的所有密码子数的 56.76%。表 1 中,发现与环结构含量显著相关的密码子同时也与茎结构含量显著相关,且相关系数大小相同,正负相反。例如,同义密码子 CAG 和 CAA (编码氨基酸 Gln)的使用值与环结构含量和茎结构含量均极显著相关,相关系数都是 0.36 且相关性相反。说明密码子 CAG 在环结构中的使用对环结构

的形成有阻碍作用,而对茎结构的形成有促进作用。  
2 A 型流感病毒相对同义密码子使用值对 RNA 单位平均折叠自由能的影响

RNA 分子折叠自由能的极小化决定其结构,是衡量 RNA 二级结构稳定性的重要参量,能量越小结构越稳定。考虑到同义密码子的使用偏性可能是影响单位平均折叠自由能的重要因素,本文对每个病毒每个蛋白 RNA 序列的 RSCU 值与  $E^*$  进行相关性分析,结果如表 2。

表 2 相对同义密码子使用值与单位平均折叠自由能之间的相关性分析结果

Table 2 Linear regression between relative synonymous codon usage with average unit folding free energy

Amino	Codon	P	R	N	df
Phe	UUU	0.0001	0.33***	126	124
Phe	UUC	0.0001	-0.33***	126	124
Leu	UUG	3.6E-05	-0.36***	126	124
Leu	CUU	0.0021	0.27**	126	124
Ile	AUC	0.0085	-0.23**	126	124
Val	GUU	1.0E-06	0.42***	126	124
Val	GUG	0.0004	-0.31***	126	124
Ser	UCU	0.0077	0.24**	126	124
Ser	UCC	0.0051	0.25**	126	124
Ser	UCA	0.0008	-0.30***	126	124
Pro	CCC	0.0338	-0.19*	126	124
Pro	CCA	0.0033	-0.26**	126	124
Pro	CCG	8.8E-06	0.38***	126	124
Thr	ACA	0.0136	0.22*	126	124
Thr	ACG	0.0004	-0.31***	126	124
Gln	CAA	8.7E-08	0.45***	126	124
Gln	CAG	8.7E-08	-0.45***	126	124
Asn	AAU	0.0361	-0.19*	126	124
Asn	AAC	0.0361	0.19*	126	124
Asp	GAU	0.0471	0.18*	126	124
Asp	GAC	0.0471	-0.18*	126	124
Glu	GAA	7.0E-05	0.35***	126	124
Glu	GAG	7.0E-05	-0.35***	126	124
Cys	UGU	0.0070	0.24**	126	124
Cys	UGC	0.0070	-0.24**	126	124
Arg	CGG	0.0032	-0.26**	126	124
Ser	AGU	0.0224	-0.20*	126	124
Ser	AGC	0.0226	0.20*	126	124

Note: Two-tailed significance: \* :  $P \leq 0.05$ ; \*\* :  $P \leq 0.01$ ; \*\*\* :  $P \leq 0.001$ . R is the correlation coefficient, P is the significance level, N is the number of samples, and df is the degree of freedom.

从表 2 中可以看出,在本文分析的所有特征量中,与 RNA 单位平均折叠自由能相关的密码子最多。有 28 个密码子的使用与其显著相关,涉及显著的氨基酸有 12 种,这些密码子占编码相关氨基酸所有密码子数的 71.79%。当我们进一步比较表 1 和表 2 中的相关系数时,发现密码子使用偏性与单位平均折叠自由能的相关系数较高,说明密码子使用偏性对 RNA 折叠自由能有很重要的影响。

### 3 A 型流感病毒相对同义密码子使用值对 RNA 柔性(SF)的影响

RNA 柔性是以 RNA 折叠自由能为基础的一个综合参量,有效描述了 RNA 结构的可变性和稳定性。我们认为它是影响 RNA 功能的一个参量。所以,对 A 型流感病毒核酸序列的 RSCU 值与 RNA 柔性进行相关性分析,结果如表 3。

表 3 相对同义密码子使用值与 RNA 柔性的线性分析结果

Table 3 Linear regression between relative synonymous codon usage and specific flexibility of RNA

Codon	P	R	N	df
Phe	UUU	8.7E-05	0.34***	126
phe	UUC	8.7E-05	-0.34***	126
Ile	AUU	0.005 0	-0.25**	126
Ile	AUC	0.000 5	0.31***	126
Val	GUU	4.1E-05	0.36***	126
Val	GUG	8.8E-05	-0.34***	126
Ser	UCU	0.043 7	-0.18*	126
Ser	UCG	0.002 0	-0.27**	126
Thr	ACU	3.1E-05	0.36***	126
Thr	ACA	0.000 7	-0.30***	126
Ala	GCC	0.018 8	0.21*	126
Ala	GCG	0.026 1	-0.20*	126
Cys	UGU	0.000 2	0.32***	126
Cys	UGC	0.000 2	-0.32***	126
TER	UAG	1.8E-06	-0.41***	126
TER	UGA	0.004 2	0.25**	126
Arg	CGC	0.000 1	-0.33***	126
Ser	AGU	8.6E-07	0.42***	126
Ser	AGC	0.021 0	-0.21*	126
Gly	GGC	0.016 5	-0.21*	126

Note: Two-tailed significance: \* :  $P \leq 0.05$ ; \*\* :  $P \leq 0.01$ ; \*\*\* :  $P \leq 0.001$ . R is the correlation coefficient, P is the significance level, N is the number of samples and df is the degree of freedom.

对于 RNA 柔性(SF),有 10 种氨基酸的 20 个密码子使用度与 SF 显著相关。这些密码子占编码相关氨基酸所有密码子数的 57.14%。例如,表中同义密码子 UGU 和 UGC(编码 Cys)的使用与 SF 的相关系数均为 0.32,但密码子 UGU 的使用与 SF 呈正相关,而密码子 UGC 的使用与 SF 呈负相关,说明密码子 UGU 的使用能增强相应 RNA 的柔性,

而密码子 UGC 的使用则会减小相应 RNA 的柔性。而 RNA 柔性是一个能影响 RNA 结构的稳定性的参量,结果意味着同义密码子的使用偏性会影响到 RNA 结构的稳定性。

### 4 RNA 二级结构随相对同义密码子使用值变化的关系图

为了能够直观地看出茎结构含量( $\tilde{N}_{sp}$ )、环结构含量( $\tilde{N}_L$ )、单位平均折叠自由能以及 RNA 柔性随相对同义密码子使用值的变化关系,我们还作了这四个参量随相对同义密码子使用值的变化关系图(图 1、图 2、图 3)。由于涉及相关的密码子很多,我们以个别密码子为例作了相对同义密码子的使用值与 RNA 二级结构之间的关系图。

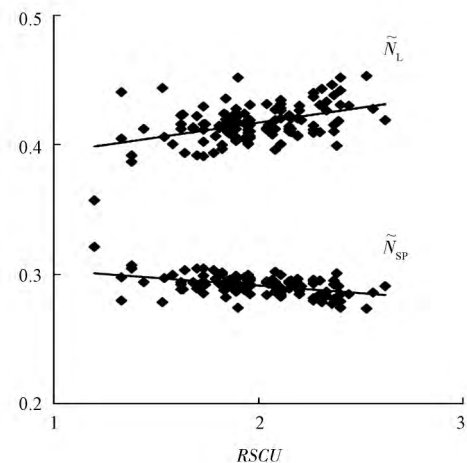


图 1 RNA 茎结构含量和环结构含量随密码子 ACA 使用值的变化关系

Figure 1 Correlation between the content of stem and loop structure of RNA and codon of usage of ACA

图 1 是以编码氨基酸 Thr 的密码子 ACA 为例绘制的关系图,很直观地表达出了茎环结构含量和环结构含量随密码子 ACA 使用值的变化趋势正好相反,且它们与相对同义密码子使用值相关性的强度很相近(斜率绝对值相近),图 2 和图 3 分别是以编码氨基酸 Glu 的密码子 GAA 和编码氨基酸 Ser 的密码子 AGU 为例绘制的关系图,说明单位平均折叠自由能与密码子 GAA 的使用呈正相关趋势,而 RNA 柔性随密码子 AGU 的使用呈负相关趋势,且在所选的参量中 RNA 柔性随相对同义密码子使用值的相关性最强(斜率最大)。其它涉及相关的密码子也可以得到类似的相关变化关系图,这些关系图很直观地说明相对同义密码子的使用偏性对 RNA 二级结构有很大影响。

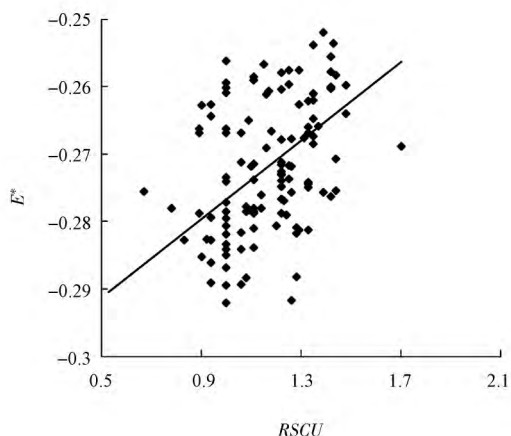


图 2 RNA 单位平均折叠自由能随密码子 GAA 使用值的变化关系

Figure 2 Correlation between the average unit folding free energy of RNA and codon of usage of GAA

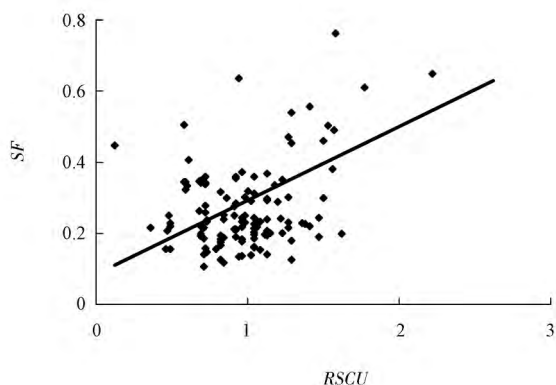


图 3 RNA 柔性随密码子 AGU 使用值的变化关系

Figure 3 Correlation between the flexibility of RNA and codon of usage of AGU

## 讨 论

RNA 分子是重要的细胞成分,其涉及许多基本的生物过程。了解它们的功能背后的机制就需要 RNA 三级结构知识。而目前直接对 RNA 进行三级结构的理论预测都不太顺利,因而通过预测二级结构的研究方法来获取其三级结构成为一个必然趋

势。RNA 能在结构上有高度相似性,不是由某段有稳定性序列所决定的,而是由它们在特有的二级结构中存有稳定性所决定的。因此, RNA 的这段特有的二级结构也决定了它们的功能特点。所以,为了更好得了解 RNA 的工作原理及功能多样性,须对 RNA 的结构进行剖析,这样既能为我们提供更多信息来进行生物分析,还能为蛋白质的结构分析及功能预测等提供参考信息<sup>[31,32]</sup>。

本文把 RNA 二级结构大致分为环结构和茎结构,并使用相对同义密码子使用值来衡量同义密码子的使用偏性。研究发现, A 型流感病毒的一些密码子的使用偏性对 RNA 的环结构和茎结构的影响都是极其明显的。相对同义密码子使用值与 RNA 的二级结构的茎、环结构含量显著相关,且与茎、环结构的相关性正好相反,涉及显著相关的氨基酸占总数的 50%。考虑到茎区是由氢键相连的配对碱基叠加而形成的,那么茎结构含量越大,氢键数量就越多,对应的自由能就越小,越有利于增强二级结构稳定性。而环区数量的增加会导致自由能的增长,会造成二级结构稳定性下降<sup>[33]</sup>。从而影响其功能。希望本文得到的密码子使用偏性与 RNA 的环结构和茎结构之间的相关性能为分子功能的研究和预测提供有价值参考。

在本文分析的所有特征量中,与 RNA 单位平均折叠自由能相关的密码子最多,涉及显著相关的氨基酸占总数的 60%,相关系数相对较高。而且,密码子使用偏性与 RNA 柔性相关性较显著,涉及显著相关的氨基酸占总数的 50%,且与其呈极显著相关的密码子使用占多数,说明密码子偏性可能是 RNA 柔性的一个起关键作用的影响因素。RNA 折叠自由能与 RNA 柔性是 RNA 二级结构中的两个综合参量,上述相关性证实了密码子使用偏性对 RNA 二级结构有很重要的影响。近年来,人们提出了基因改造方法<sup>[34,35]</sup>,其中一种有效的方法是通过遗传密码偏爱并结合基因内部二级结构调整来实现的。希望本文的研究结果能为通过基因改造而破坏病毒结构和功能提供参考。

附表 1 A 型流感病毒的核酸序列信息

Appendix Table 1 Information on the nucleic-acid sequence of all influenza-A viruses

No.	Name	$\tilde{N}_{sp}$	$\tilde{N}_L$	$E^*$	SF	No.	Name	$\tilde{N}_{sp}$	$\tilde{N}_L$	$E^*$	SF
1	H1N1	0.2956	0.4089	-0.27	0.41	64	H8N1	0.2898	0.4205	-0.28	0.31
2	H1N2	0.2900	0.4199	-0.28	0.25	65	H8N2	0.2989	0.4022	-0.28	0.34
3	H1N3	0.2790	0.4421	-0.25	0.19	66	H8N3	0.2863	0.4275	-0.29	0.24
4	H1N4	0.3020	0.3959	-0.28	0.30	67	H8N4	0.2957	0.4086	-0.27	0.18
5	H1N5	0.2867	0.4266	-0.26	0.32	68	H8N5	0.2939	0.4121	-0.26	0.35

No	Name	$\tilde{N}_{sp}$	$\tilde{N}_L$	$E^*$	SF	No	Name	$\tilde{N}_{sp}$	$\tilde{N}_L$	$E^*$	SF
6	H1N6	0.2883	0.4233	-0.28	0.16	69	H8N6	0.2799	0.4402	-0.26	0.15
7	H1N7	0.2923	0.4155	-0.26	0.21	70	H8N7	0.2742	0.4516	-0.25	0.49
8	H1N8	0.2865	0.4269	-0.27	0.19	71	H8N8	0.2963	0.4074	-0.28	0.20
9	H1N9	0.2769	0.4461	-0.26	0.26	72	H9N1	0.2741	0.4518	-0.24	1.08
10	H2N1	0.2943	0.4114	-0.28	0.19	73	H9N2	0.2947	0.4106	-0.27	0.64
11	H2N2	0.2918	0.4165	-0.26	0.18	74	H9N3	0.2940	0.4121	-0.29	0.30
12	H2N3	0.2952	0.4097	-0.28	0.24	75	H9N4	0.2809	0.4383	-0.26	0.25
13	H2N4	0.2879	0.4242	-0.26	0.33	76	H9N5	0.2935	0.4131	-0.28	0.24
14	H2N5	0.2921	0.4156	-0.28	0.23	77	H9N6	0.2930	0.4139	-0.28	0.45
15	H2N6	0.2917	0.4167	-0.26	0.13	78	H9N7	0.2843	0.4313	-0.27	0.22
16	H2N7	0.2938	0.4125	-0.27	0.23	79	H9N8	0.3001	0.3998	-0.28	0.19
17	H2N8	0.2920	0.4160	-0.27	0.24	80	H9N9	0.2940	0.4119	-0.28	0.19
18	H2N9	0.2905	0.4189	-0.29	0.20	81	H10N1	0.2782	0.4437	-0.26	0.20
19	H3N1	0.2849	0.4302	-0.26	0.20	82	H10N2	0.2828	0.4343	-0.26	0.34
20	H3N2	0.2885	0.4229	-0.26	0.24	83	H10N3	0.2949	0.4102	-0.28	0.21
21	H3N3	0.2934	0.4133	-0.26	0.25	84	H10N4	0.2966	0.4069	-0.27	0.16
22	H3N4	0.3030	0.3939	-0.27	0.29	85	H10N5	0.2996	0.4008	-0.28	0.23
23	H3N5	0.2734	0.4532	-0.27	0.16	86	H10N6	0.2898	0.4203	-0.26	0.26
24	H3N6	0.2797	0.4407	-0.24	1.13	87	H10N7	0.3043	0.3913	-0.27	0.23
25	H3N7	0.2943	0.4115	-0.26	0.16	88	H10N8	0.2915	0.4170	-0.27	0.14
26	H3N8	0.2981	0.4038	-0.27	0.86	89	H10N9	0.2929	0.4142	-0.27	0.16
27	H3N9	0.2951	0.4096	-0.28	0.23	90	H11N1	0.2909	0.4183	-0.28	0.21
28	H4N1	0.2937	0.4124	-0.27	0.23	91	H11N2	0.2975	0.4050	-0.28	0.35
29	H4N2	0.2861	0.4277	-0.27	0.34	92	H11N3	0.2845	0.4310	-0.28	0.16
30	H4N3	0.2869	0.4262	-0.26	0.22	93	H11N4	0.3042	0.3917	-0.31	0.45
31	H4N4	0.3016	0.3969	-0.29	0.37	94	H11N5	0.2828	0.4343	-0.26	0.34
32	H4N5	0.2906	0.4187	-0.27	0.17	95	H11N6	0.2998	0.4004	-0.28	0.15
33	H4N6	0.2949	0.4102	-0.26	0.30	96	H11N7	0.2923	0.4154	-0.28	0.20
34	H4N7	0.2815	0.4370	-0.27	0.22	97	H11N8	0.2935	0.4131	-0.28	0.14
35	H4N8	0.2943	0.4113	-0.27	0.32	98	H11N9	0.2979	0.4043	-0.27	0.32
36	H4N9	0.2834	0.4331	-0.27	0.11	99	H12N1	0.2900	0.4199	-0.27	0.13
37	H5N1	0.2998	0.4004	-0.28	0.47	100	H12N2	0.2845	0.4309	-0.28	0.24
38	H5N2	0.2884	0.4233	-0.27	0.61	101	H12N3	0.2966	0.4067	-0.28	0.23
39	H5N3	0.2978	0.4044	-0.27	0.54	102	H12N4	0.2879	0.4242	-0.28	0.25
40	H5N4	0.2933	0.4133	-0.28	0.18	103	H12N5	0.2987	0.4027	-0.28	0.21
41	H5N5	0.2925	0.4150	-0.27	0.23	104	H12N6	0.2897	0.4206	-0.29	0.30
42	H5N6	0.2997	0.4007	-0.27	0.21	105	H12N7	0.2905	0.4190	-0.27	0.29
43	H5N7	0.2909	0.4182	-0.26	0.24	106	H12N8	0.2938	0.4125	-0.27	0.19
44	H5N8	0.2954	0.4092	-0.26	0.36	107	H12N9	0.2786	0.4428	-0.26	0.29
45	H5N9	0.2938	0.4124	-0.28	0.34	108	H13N1	0.3066	0.3867	-0.28	0.50
46	H6N1	0.3213	0.3574	-0.32	0.90	109	H13N2	0.2973	0.4054	-0.27	0.19
47	H6N2	0.2921	0.4158	-0.25	0.20	110	H13N3	0.2847	0.4306	-0.26	0.24
48	H6N3	0.2891	0.4218	-0.28	0.28	111	H13N6	0.2921	0.4158	-0.26	0.14
49	H6N4	0.2919	0.4162	-0.29	0.19	112	H13N8	0.2842	0.4316	-0.27	0.38
50	H6N5	0.3005	0.3990	-0.27	0.18	113	H13N9	0.2906	0.4189	-0.26	0.28
51	H6N6	0.2938	0.4125	-0.27	0.37	114	H14N2	0.2884	0.4231	-0.29	0.36
52	H6N7	0.2850	0.4301	-0.26	0.51	115	H14N4	0.2969	0.4062	-0.27	0.46
53	H6N8	0.2943	0.4114	-0.28	0.29	116	H14N5	0.2853	0.4295	-0.26	0.22
54	H6N9	0.2858	0.4283	-0.28	0.22	117	H14N6	0.2942	0.4116	-0.28	0.20
55	H7N1	0.2960	0.4080	-0.28	0.65	118	H14N8	0.2821	0.4357	-0.26	0.22
56	H7N2	0.3042	0.3916	-0.28	0.76	119	H15N2	0.2935	0.4131	-0.27	0.22
57	H7N3	0.2888	0.4225	-0.28	0.36	120	H15N4	0.2901	0.4198	-0.27	0.13
58	H7N4	0.2906	0.4188	-0.29	0.34	121	H15N6	0.2927	0.4145	-0.26	0.17
59	H7N5	0.2966	0.4067	-0.28	0.21	122	H15N8	0.2918	0.4165	-0.27	0.14
60	H7N6	0.2882	0.4235	-0.27	0.12	123	H15N9	0.2927	0.4145	-0.27	0.23
61	H7N7	0.2868	0.4263	-0.26	0.36	124	H16N3	0.2969	0.4061	-0.28	0.20
62	H7N8	0.2973	0.4054	-0.29	0.35	125	H16N9	0.3033	0.3934	-0.29	0.56
63	H7N9	0.2953	0.4094	-0.28	0.22	126	H17N10	0.2879	0.4242	-0.25	0.30

Note:  $\tilde{N}_{sp}$  and  $\tilde{N}_L$  is the content of the stem structure and loop structure calculated using equation (1) and (2).  $E^*$  is average unit folding free energy calculated using equation (3). SF is the flexibility of RNA, calculated with equation (6).

## 参考文献:

- [1] Batey R T. Structures of regulatory elements in mRNAs [J]. *Curr Opin Struct Biol*, 2006, 16(3):299-306.
- [2] Mathews D H, Disney M D, Childs J L, Schroeder S J, Zuker M, Turner D H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure [J]. *Proc Natl Acad Sci USA*, 2004, 101(19):7287-7292.
- [3] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction [J]. *Nucleic Acids Res*, 2003, 31(13):3406-3415.
- [4] Luo L F. The degeneracy rule of genetic code [J]. *Orig Life Evol Biosph*, 1988, 18(1-2):65-70.
- [5] Bulmer M. The selection-mutation-drift theory of synonymous codon usage [J]. *Genetics*, 1991, 129(3):897-907.
- [6] Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity [J]. *Nucleic Acids Res*, 1981, 9(1):r43-r74.
- [7] Onofrio G D, Mouchiroud D, Aissani B, Gautier C, Bernardi G. Correlations between the compositional properties of human genes codon usage and amino acid composition of proteins [J]. *J Mol Evol*, 1991, 32(6):504-510.
- [8] Knight R D, Freeland S J, Landweber L F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes [J]. *Genome Biol*, 2001, 2(4): research0010.1-research0010.13.
- [9] Eyre-Walker A. Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy [J]. *Mol Biol Evol*, 1996, 13(6):864-872.
- [10] Marin A, Gonzalez F, Gutierrez G, Oliver J L. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli* [J]. *Nucleic Acids Res*, 1998, 26(19):4540-4540.
- [11] Buchan J R, Aucott L S, Stansfield I. tRNA properties help shape codon pair Preferences in open reading frames [J]. *Nucleic Acids Res*, 2006, 34(3):1015-1027.
- [12] Gupta S K, Majumdar S, Bhattacharya T K, Ghosh T C. Studies on the relationships between the synonymous codon usage and Protein secondary structural units [J]. *Biochem Biophys Res Commun*, 2000, 269(3):692-696.
- [13] 吴宪明, 吴松锋, 任大明. 密码子偏性的分析方法及相关研究进展 [J]. *遗传*, 2007, 29(4):420-426.
- [14] Martínez-Pérez F, Bendena W G, Chang B S W, Tobe S S. Influence of codon usage bias on FGLamide-alatostatin mRNA secondary structure [J]. *Peptides*, 2011, 32(3):509-517.
- [15] 王侃, 刘次全, 曹槐, 陈雪峰. 大肠杆菌 mRNA 编码区长度、形成二级结构倾向与密码子偏好性的关系 [J]. *微生物学报*, 2006, 46(6):895-899.
- [16] Zama M. Codon usage and secondary structure of mRNA [J]. *Nucleic Acids Symp Ser*, 1990, (22):93-94.
- [17] 朱平, 高雷, 徐振源. 基于拟氨基酸编码方法的同义密码子的偏好性与结合强度的相关性研究 [J]. *物理学报*, 2009, 6(58):4295-4300.
- [18] 石秀凡, 黄京飞, 梁宠荣, 柳树群, 谢君, 刘次全. 人类基因中同义密码子的偏好性与密码子-反密码子间的结合强度密切相关吗? [J]. *科学通报*, 2000, 45(23):2520-2525.
- [19] Neumann G, Noda T, Kawaoka Y. Emergence and pandemic potential of swine-origin H1N1 influenza virus [J]. *Nature*, 2009, 459(7249):931-939.
- [20] Olsen B, Munster V J, Wallensten A, Waldenstrom J, Osterhaus A. Global patterns of influenza a virus in wild birds [J]. *Science*, 2006, 312(5772):384-388.
- [21] Webster R G, Bean W J, Gorman O T, Gorman O T, Kawaoka Y. Evolution and ecology of influenza A viruses [J]. *Micro Rev*, 1992, 56(1):152-179.
- [22] Steinhauer D A, Skehel J J. Genetics of influenza viruses [J]. *Annu Rev Genet*, 2002, 36(2):305-332.
- [23] Li J, Yu X, Pu X, Xie L, Sun Y, Xiao H, Wang F, Din H, Wu Y, Liu D, Zhao G, Liu J, Pan J. Environmental connections of novel avian-origin H7N9 influenza virus infection and virus adaptation to the human [J]. *Sci China Life Sci*, 2013, 56(6):485-492.
- [24] Ilyushina N A, Chernyy E S, Korchagina E Y, Gambaryan A S, Henry S M, Bovin N V. Labeling of influenza viruses with synthetic fluorescent and biotin-labeled lipids [J]. *Viol Sin*, 2014, 29(4):199-210.
- [25] Ng A K, Wang J H, Shaw P C. Structure and sequence analysis of influenza A virus nucleoprotein [J]. *Sci China C Life Sci*, 2009, 52(5):439-449.
- [26] Mathews D H, Disney M D, Childs J L, Schroeder S J, Zuker M, Turner D H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure [J]. *Proc Natl Acad Sci USA*, 2004, 101(19):7287-7292.
- [27] Tsai L, Luo L F. A statistical mechanical model for predicting B-DNA curvature and flexibility [J]. *J Theor Biol*, 2000, 207(2):177-194.



- [28] Luo L F. Theoretic-physical Approach to Molecular Biology. Shanghai Science and Technology Publisher, 2004: 408-428.
- [29] Chen W, Luo L F, Zhang L R. The organization of nucleosomes around splice sites [J]. Nucleic Acids Res, 2010, 38(9):2788-2798.
- [30] 于志芬, 李瑞芳. 同义密码子的使用偏性对蛋白质折叠速率的影响[J]. 生物物理学报, 2013, 8(29):603-613.
- [31] 李誌, 刘元宁, 张浩, 沈延杰, 李杰, 霍洪亮. 基于动态规划二级结构的比较[J]. 吉林大学学报工学版, 2011, 2(41):250-253.
- [32] 刘哲言, 屈武斌, 张成岗. 用于寡核苷酸二级结构预测的热力学数据库研究进展[J]. 生物信息学, 2014, 3(12):196-205.
- [33] 廖波, 王天明. RNA 二级结构的最小自由能算法[J]. 生物数学学报, 2003, 18(3): 364-368.
- [34] 夏亚穆, 李晨晨. 环糊精葡萄糖基转移酶的基因改造与高效表达[J]. 中国生物工程杂志, 2015, 35(2): 105-110.
- [35] 曹瑞兵, 徐学清, 周斌, 陈德胜, 陈溥言. 猪  $\alpha 1$ -干扰素的基因改造与高效原核表达[J]. 生物工程学报, 2004, 20(2):291-294.

## Influence of Synonymous Codon Bias on the RNA Secondary Structure in Influenza-A Viruses

LI Ruifang\*, YU Zhifen, HUANG Qiao, GUO Chunyang, LI Xue

(College of Physics and Electronic Information, Inner Mongolia Normal University, Hohhot 010022, China)

**Abstract:** Influenza-A viruses were selected as samples. Then, the relationship between synonymous codon bias and mRNA secondary structures was analyzed. Our data will provide a theoretical basis for further research on the synonymous codon and RNA properties of the influenza-A virus. Information on the nucleic acid sequence of all influenza-A viruses in the National Center for Biotechnology Information was collected. The secondary structure of each nucleic-acid sequence was predicted. Then, the loop structure, stem structure, and free replicating energy of RNA were calculated. Based on this information, the specific flexibility of RNA was worked out. Simultaneously, the synonymous codon bias in each nucleic-acid sequence was counted. A database on the RNA secondary structure of the influenza-A virus was set up. Then, the relationship between the synonymous codon bias and the content of loop structure, stem structure, and flexibility were analyzed. Codon usage of 50% of amino acids was correlated significantly with the content of stem structure or the content of loop structure. Also, 60% of amino acids were correlated significantly with average unit folding free energy. In addition, codon usage of 50% of amino acids was correlated significantly with the specific flexibility of RNA. For codons that were correlated significantly with stem structure and loop structure, the correlation between their usage with two types of structural content was completely contrary. Also, the correlation between specific flexibility with synonymous codon bias was better than that observed for other parameters. These results suggest that synonymous codon usage bias has important influences on RNA secondary structure.

**Key words:** Synonymous codon bias; RNA secondary structure; Influenza-A virus; RNA specific flexibility; Folding free energy

\* Corresponding author: LI Ruifang, E-mail: lirui.fang@imnu.edu.cn