Hans Journal of Biomedicine 生物医学, 2023, 13(4), 368-377
Published Online October 2023 in Hans. https://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2023.134043

Hans汉斯

编码疾病相关的固有无序蛋白同义密码子使用偏好性的研究

孙灿壮，冯永娥*

内蒙古农业大学理学院，内蒙古 呼和浩特

收稿日期：2023年8月19日；录用日期：2023年9月28日；发布日期：2023年10月17日

摘　要

固有无序蛋白(简称IDPs)在生理条件下不具有稳定的空间结构，但是在生物体内发挥重要的生物学功能，除此之外，尤为重要的是它们与人类许多重大疾病密切相关。因此，研究疾病相关的IDPs，可以进一步了解IDPs与一系列疾病的病理学上的关系，这为蛋白质的药物设计和疾病治疗提供新思路。研究表明：密码子的使用偏好与固有无序蛋白的无序程度存在一定的关联性。本文研究了编码疾病相关的固有无序蛋白(IDPs)中同义密码子和GC含量的使用偏好，以及编码这些蛋白质的有序/无序区域的密码子中四个核苷酸在密码子的三个位点的分布差异。结果表明，这些蛋白中同义密码子的使用、GC含量均存在显著差异，另外四个核苷酸在其密码子三个位点的分布在编码3类疾病相关的IDPs有序和无序区中均存在偏好性。这些结果为后期研究IDPs提供了很好的参考信息。

关键词

固有无序蛋白，同义密码子，密码子相对使用度，GC含量，方差分析

Synonymous Codon Usage Bias in Nucleic Acids Encoding Disease-Associated Intrinsically Disordered Proteins

Canzhuang Sun, Yong'e Feng*

College of Science, Inner Mongolia Agriculture University, Hohhot Inner Mongolia

Received: Aug. 19th, 2023; accepted: Sep. 28th, 2023; published: Oct. 17th, 2023

*通讯作者。

文章引用: 孙灿壮, 冯永娥. 编码疾病相关的固有无序蛋白同义密码子使用偏好性的研究[J]. 生物医学, 2023, 13(4): 368-377. DOI: 10.12677/hjbm.2023.134043

Abstract

Despite the fact that intrinsically disordered proteins (IDPs) lack stable spatial structures under physiological conditions, they perform critical biological functions in organisms. Moreover, intrinsically disordered proteins (IDPs) are associated with many major human diseases. Therefore, the systematic study of disease-associated intrinsically disordered proteins (IDPs) can further understand the pathological relationship between intrinsically disordered proteins (IDPs) and a series of diseases, providing a new idea for protein drug design and disease treatment. Some studies have shown a specific correlation between codon usage bias and intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs). In this paper, we analyzed the use bias of synonymous codons and GC content in nucleic acids encoding the disease-associated intrinsically disordered proteins (IDPs). Moreover, we studied the distribution of four nucleotides at three sites of the codon in encoding the ordered/disordered regions of these proteins. The results showed that the use of synonymous codons, the content of GC, and the distribution of four nucleotides at three sites of the codon are biased in the encoding of ordered regions and disordered regions of disease-associated intrinsically disordered proteins (IDPs). The results can provide a reference for further study of intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs).

Keywords

Intrinsically Disordered Proteins, Synonymous Codons, GC Content, Relative Synonymous Codon Usage, Analysis of Variance

1. 引言

一直以来，蛋白质结构生物学的观点是：蛋白质序列必须折叠形成稳定的空间结构才能正常发挥功能，即“序列－结构－功能”的范式研究路线，固有无序蛋白(Intrinsically Disordered Proteins，简称 IDPs) [1]的出现打破了这种范式。固有无序蛋白，它们在天然状态下不具有一个稳定的三维结构，但是依然能够行使重要的生物学功能。这类蛋白的无序化程度可发生于不同水平，可以是整个蛋白质序列无序，也可以是蛋白质序列中的几个残基、长环区域或蛋白质的末端，或结构域的一部分无序，这些无序片段，被称为固有无序区(Intrinsically Disordered Regions，简称 IDRs)，而这个蛋白质序列中其他有构象的区域被称为有序区(ORs)。研究表明，固有无序蛋白在真核生物中含量较丰富，而且有一些功能是固有无序蛋白独立发挥的，较原核生物而言，在真核生物中 IDPs 的增加可能是对付真核生物中复杂的细胞调控[2]。固有无序蛋白结构松散、灵活性高、在生理条件下虽然不具有稳定的二级或三级结构，但是在生物体内参与大量的细胞信号转导、DNA 代谢过程、mRNA 可变剪切、还有翻译后修饰等重要功能[3] [4] [5] [6]。固有无序蛋白除了在生物体内发挥上述重要的功能之外，尤为重要的是它们与人类许多重大疾病密切相关，如阿尔茨海默病、帕金森疾病、艾滋病等这些举世闻名的恶疾都与固有无序蛋白或其无序区(IDRs)密切相关[7] [8]。NUPR1 是一种多功能的固有无序蛋白，它通过控制胰腺癌细胞的迁移、侵袭和粘附，来参与胰腺导管癌的发生。P53 蛋白是人体中与肿瘤密切相关的固有无序蛋白，在细胞信号转导网络中

起着关键作用，它的功能丧失可增加细胞通过狭窄孔的迁移率，而在肿瘤中常发现 P53 的突变型 TP53，P53 发生突变后其抑制途径可能会失效，进而引发肿瘤。同时 P53 家族在生物体内的糖酵解，糖异生，有氧呼吸和细胞凋亡等多种代谢途径中发挥着重要的作用[9] [10]。故研究疾病相关的固有无序蛋白(IDPs)或 IDRs 具有重大的生物学意义和药物学价值。

一般而言，氨基酸在固有无序蛋白(IDPs)中的无序区(IDRs)和有序区域的分布存在偏较大偏差，无序区(IDRs)富含极性的，带电的，亲水性的残基[11]。与此相一致的是，最近的研究表明，编码蛋白质的核酸在编码无序区(IDRs)和编码有序区域(ORs)的序列之间表现出密码子偏差[12]。由于无序区域较为松散，自由度较高，因此无序区域比有序区域更容易发生翻译错误。另一些研究表明：无序区(IDRs)与密码子选择有关，其中次优密码子使用减慢了翻译速度[13] [14]。此外，研究表明：在真核生物中编码有序区域的基因序列相比无序区(IDRs)具有较高的 GC 含量[15]。另外，一些固有无序蛋白中无序区(IDRs)由较高的 GC 含量编码[15] [16]。无序区(IDRs)与 GC 含量之间的关系归因于有序区域(ORs)和无序区(IDRs)之间氨基酸使用的偏差[16]。基于上述分析，在本文，我们首先构建了一个与 3 类疾病相关的固有无序蛋白(IDPs)的数据库。然后，在这些蛋白中，我们研究了同义密码子的使用偏好，GC 含量，以及编码这些蛋白的密码子中三个位点的四个核苷酸的分布，结果发现：同义密码子、GC 含量在 3 类疾病相关的 IDPs 中有序和无序区均存在不同程度的使用偏好性；编码 3 类疾病相关的 IDPs 的密码子在有序和无序区 3 个位点碱基的组成上也存在一定的差异，这些结果为后期研究 IDPs 提供了有益的信息。

2. 材料与方法

2.1. 资料库

鉴于实验中测定的固有无序蛋白(IDPs)数量有限，本项目从 UniProt 数据库[17]下载人类全部蛋白(物种号：9606；蛋白质 ID：UP000005640)，并通过注释信息挑出与癌症(cancer)、心血管疾病(cardiovascular)、神经退化性疾病(neurodegenerative)相关的蛋白，共 20386 个蛋白的数据集。将数据集中的蛋白先与 DisProt 数据库[18]中的蛋白序列进行比对，若已经被 DisProt 数据库收录，则该蛋白的固有无序信息以 DisProt 数据库为准；对于未被 DisProt 数据库收录的蛋白序列，利用 MobiDB3.0 [19]中 MobiDB-lite 预测算法进行固有无序蛋白预测。综合上述得到的固有无序蛋白，然后用 CD-hit 软件去除同源性 > 30%的序列，最终得到疾病相关的固有无序蛋白数据库(disease-related-disordered proteins)，简称 DDP 数据库，然后通过 EMBL 库[20]找到对应的 mRNA 序列(注释：没有 mRNA 序列的就舍掉对应的 IDPs 蛋白)，这样最终得到与癌症相关的 IDPs 有 7166 个 mRNA 序列，与心血管疾病相关的 IDPs 有 4092 个 mRNA 序列，与神经退行性疾病相关的 IDPs 有 4341 个 mRNA 序列，作为我们研究的数据集。

2.2. 特征及方法

2.2.1. 密码子使用频率

我们在建立的数据库中分别统计了 3 种疾病相关的固有无序蛋白无序序列对应 mRNA 序列中 64 种密码子出现的频次。最终，一条固有无序蛋白无序序列表示成了 64 维的特征向量：

$$P_{codon} = \left[P_1, P_2, \cdots, P_{64}\right] \tag{1}$$

$$P_i = \frac{n_i}{\sum n_i} \ \left(i = 1, 2, 3, \cdots, 64\right) \tag{2}$$

其中，n_i 是第 i 种密码子在 mRNA 序列上出现的频次。

2.2.2. GC 含量

GC 含量为：[(G + C 的总数量)/(A + T + C + G 的总数量)] * 100%。GC 含量被用于分类学，一般基因内 GC 含量高于基因组，外显子高于内含子[21] [22] [23] [24]。本文对 3 类疾病相关的固有无序蛋白对应 mRNA 序列中，分别统计了有序区和无序区中的 GC 含量。

2.2.3. 方差分析

方差分析(简称 ANOVA)也称“变异数分析”或“*F* 检验”，一类常用的统计方法，常用在两组以及两组以上的样本均数差别的显著性检验[25] [26] [27] [28] [29]。ANOVA 计算多组样本均数的显著性差异可以用公式(3)来表示：

$$\mathrm{MS_T} = \mathrm{MS_B} + \mathrm{MS_W} \tag{3}$$

其中，$\mathrm{MS_T}$ 代表总均方，$\mathrm{MS_B}$ 代表组间均方，$\mathrm{MS_W}$ 代表组内均方。

我们用组间均方和组内均方的比值作为统计值，即 *F* 值，来防止各个组样本数目不同而带来的一系列的影响，*F* 值可以用公式(3)表示：

$$\text{F-value} = \mathrm{MS_B}/\mathrm{MS_W} \tag{4}$$

通过公式(4)可以看出方差分析是用组间均方与组内均方的比值(即 *F* 值)，通常 *F* 值越大，*P* 值越小。一般 *P* 值小于 0.05，说明在统计学上，各组样本彼此间有差异；若 *P* 值小于 0.001，说明各组样本之间的差异在统计学上有显著性差异。

我们用方差分析统计 64 种密码子使用频率，以及 GC 含量在有序区和无序区是否存在显著差异性。

2.2.4. 密码子相对使用度

我们定义第 *i* 个氨基酸的第 *j* 个密码子的相对同义密码子使用度 *RSCU* [30]，

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}} \tag{5}$$

其中 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的数量。对 59 个同义密码子(不包括 3 个终止密码子 TAG、TGG、TGA 和仅由一个密码子编码的蛋氨酸 ATG 及色氨酸 TGG 密码子)在有序和无序区的使用偏好性进行评估。若密码子使用无偏好性，则 *RSCU* 值为 1；如果该密码子比其他同义密码子使用更频繁，则其 *RSCU* 值大于 1；反之亦然。

3. 结果与讨论

3.1. 3 类疾病相关的固有无序蛋白中有序和无序区氨基酸的分布

利用公式(1)~(2)，我们分别统计了 64 个密码子在 3 类疾病相关的固有无序蛋白中有序区和无序区出现的频率，并用公式(3)~(4)方差分析统计了 64 个密码子出现在有序/无序区是否有显著的差异。结果表明：编码癌症相关的固有无序蛋白中有 58 个密码子在有序和无序区是存在显著差异的；编码心血管疾病相关的固有无序蛋白中有 57 个密码子在有序和无序区是存在显著差异的；编码神经退行性疾病相关的固有无序蛋白中有 56 个密码子在有序和无序区是存在显著差异的。这些结果表明：编码这 3 类疾病相关的固有无序蛋白的密码子在有序区和无序区均存在显著差异，鉴于固有无序蛋白与很多疾病密切相关，下一步可利用这些差异显著的密码子作为特征来识别固有无序蛋白及其无序区。

3.2. 3 类疾病相关的固有无序蛋白中有序和无序区 GC 含量的统计

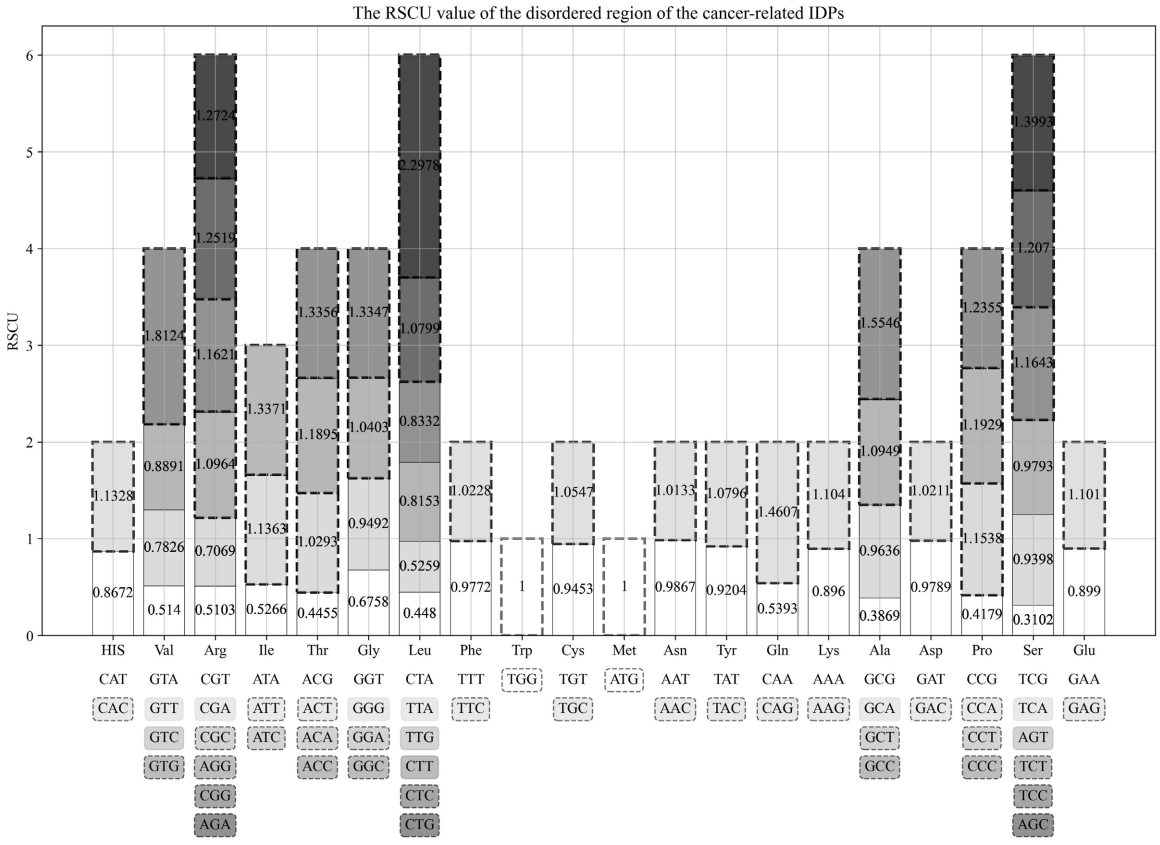
我们分别统计了 3 类疾病相关的固有无序蛋白中有序区和无序区的 GC 含量，并用公式(3)~(4)方差分析进一步检验 GC 含量在有序和无序区是否有显著的差异。GC 含量影响 DNA 结构的稳定性，GC 含量与蛋白质无序含量有关[21] [22] [23]。本文研究结果表明：GC 含量在这 3 类疾病相关的固有无序蛋白有序和无序区均存在显著差异，这验证了 GC 含量与无序区有一定的关联[21] [22] [23]，下一步也可以将 GC 含量作为参数用于识别固有无序蛋白的无序区。

3.3. 3 类疾病相关的固有无序蛋白中有序和无序区同义密码子使用偏好

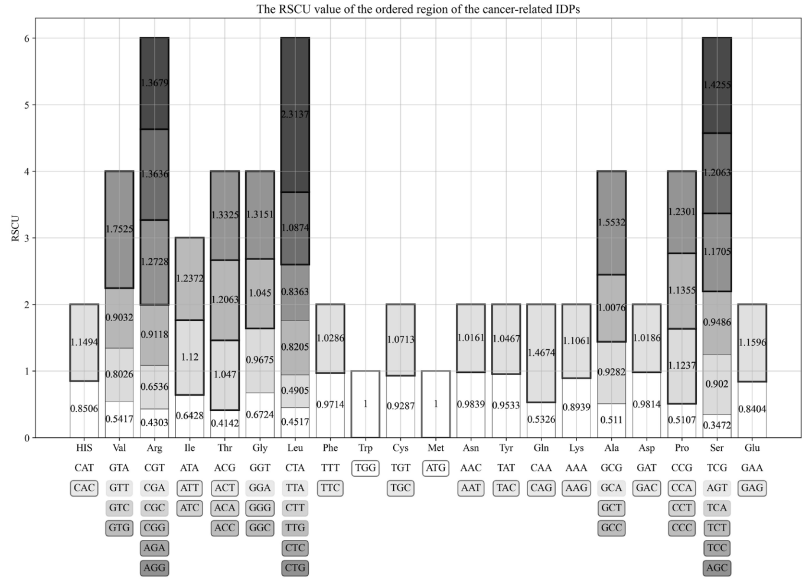
利用公式(5)，我们分别统计了 59 个同义密码子在 3 类疾病相关的固有无序蛋白中有序区和无序区的相对使用频率，结果发现：癌症相关的 IDPS 中共有 28 个密码子的 RSCU 大于 1，其中有序和无序区同义密码子存在差异的有 4 个密码子，CGC，AAC 在有序区出现频率偏低；在无序区出现频率偏高；AAT，GGG 在无序区出现频率偏低；在有序区出现频率偏高。

心血管疾病相关的 IDPs 中有 27 个密码子的 RSCU 大于 1，其中有序和无序区同义密码子存在差异的有 3 个密码子，GCT 在有序区出现频率偏低；在无序区出现频率偏高；ACT，GGG 在无序区出现频率偏低；在有序区出现频率偏高。

神经退行性疾病相关的 IDPs 中有 28 个密码子的 RSCU 大于 1，其中有序和无序区同义密码子存在差异的有 4 个密码子，CGC，AAC 在有序区出现频率偏低；在无序区出现频率偏高；AAT，GGG 在无序区出现频率偏低；在有序区出现频率偏高。详细结果如下图 1~3 所示。

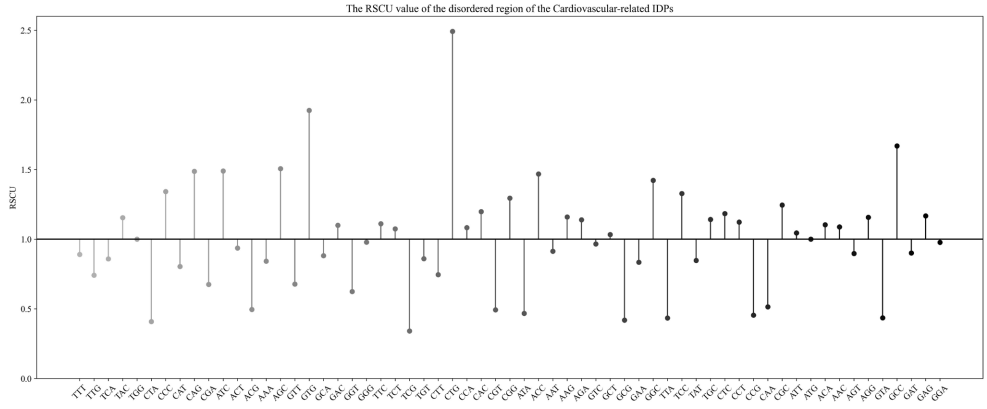


(a) Disordered regions

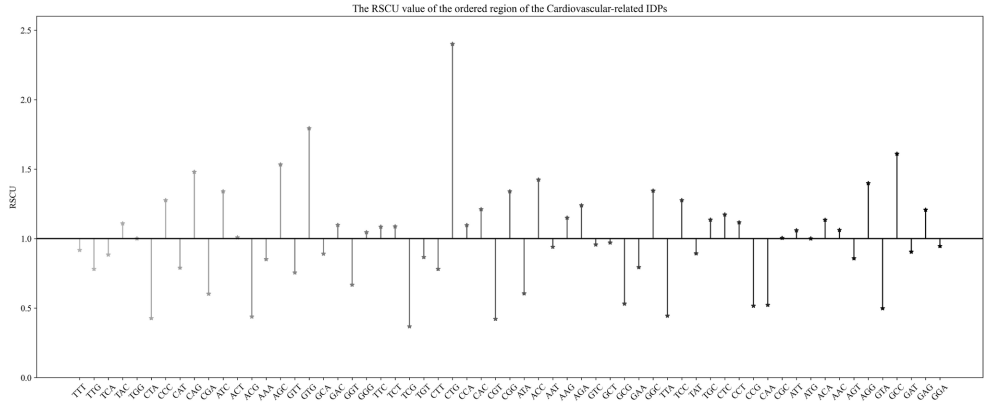


(b) Ordered regions

Figure 1. RSCU value of Cancer-related IDPs
图 1. 与癌症相关的固有无序蛋白同义密码子的 RSCU 值

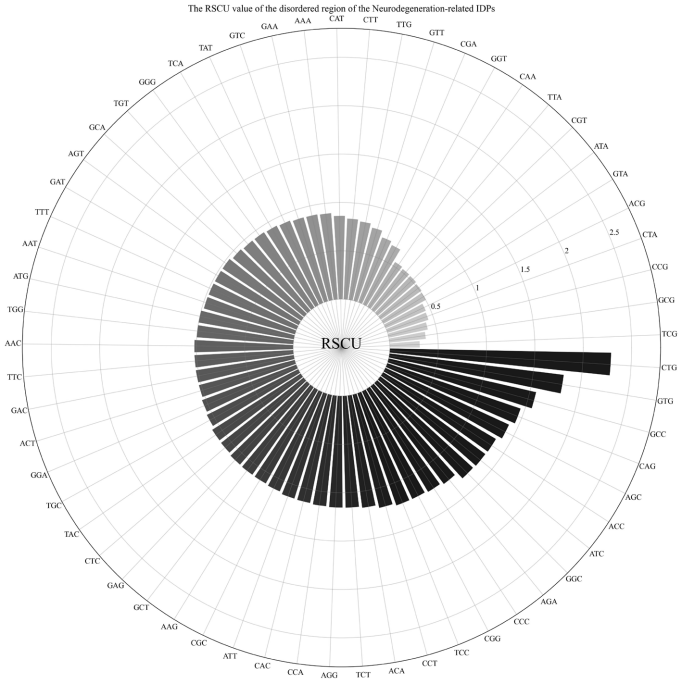


(a) Disordered regions

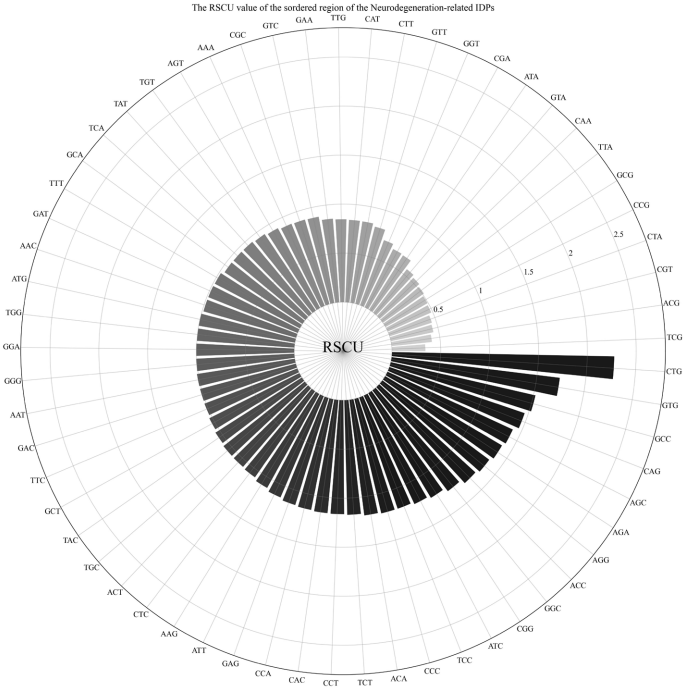


(b) Ordered regions

Figure 2. RSCU value of Cardiovascular-related IDPs
图 2. 与心血管疾病相关的固有无序蛋白同义密码子的 RSCU 值



(a) Disordered regions



(b) Ordered regions

Figure 3. RSCU value of Neurodegeneration-related IDPs
图 3. 与神经退行性相关的固有无序蛋白同义密码子的 RSCU 值

以上结果表明，在编码 3 类疾病相关的 IDPs 中同义密码子的使用均存在不同程度的偏好性，而且对应编码有序区和无序区密码子的分布也是有不同程度的差异，这和前期研究有序区和无序区氨基酸分布有差异的结果[12]是相一致的。

3.4. 编码 3 类疾病相关的固有无序蛋白密码子 3 个位点上碱基的分布

以上的研究是针对同义密码子使用偏好性的分析，接下我们对编码 3 类疾病相关的固有无序蛋白的密码子 3 个位点上碱基的组成，以及有序和无序区密码子每个位点上 4 碱基的出现频率是否存在差异进行统计，结果列在表 1~3。从表中可见：在编码癌症相关的固有无序蛋白有序和无序区中第一位点，碱基 A，G，C，T 均有显著差异；第二位点上碱基 G，C，T 有显著差异；第三位点碱基 A，C，T 有显著差异。在编码心血管疾病相关的固有无序蛋白有序和无序区中第一位点碱基 G，C，T 有显著差异，第二位点碱基 A，G，C，T 有显著差异；第三位点碱基 A，C 有显著差异。在编码神经退行性疾病相关的固有无序蛋白有序和无序区中第一位点碱基 A，G，C，T 有显著差异；第二位点碱基 G，C，T 有显著差异；第三位点碱基 A，C，T 有显著差异。可见，编码 3 类疾病相关的 IDPs 的密码子各个位点上碱基的组成上也是有显著差异的。这些结果为下一步从 DNA 角度分析固有无序蛋白与疾病的关联提供了有价值的信息。

Table 1. Four base content in three-site of encoding Cancer-related-IDPs codons
表 1. 编码癌症相关的固有无序蛋白的密码子三个位点 4 碱基的含量

Base	Disordered regions			ordered regions		
	First	second	third	First	second	third
A	26.11%	31.61%	22.77%	27.19%	32.32%	20.06%
G	34.09%	21.76%	27.56%	31.36%	17.96%	28.54%
C	26.75%	33.23%	27.74%	24.32%	22.00%	28.30%
T	13.05%	13.4%	21.93%	17.14%	27.72%	23.10%

Table 2. Four base content in three-site of encoding Neurodegeneration-related-IDPs codons
表 2. 编码神经退行性疾病相关的固有无序蛋白的密码子三个位点 4 碱基的含量

Base	Disordered regions			ordered regions		
	First	second	third	First	second	third
A	32.3%	39.12%	28.58%	34.36%	40.65%	24.98%
G	41.12%	26.22%	32.67%	40.35%	23.18%	36.47%
C	30.65	37.83%	31.52%	32.31%	29.32%	38.37%
T	26.54%	27.84%	45.61%	25.27%	40.76%	33.97%

Table 3. Four base content in three-site of encoding Cardiovascular-related IDPs codons
表 3. 编码心血管疾病相关的固有无序蛋白的密码子三个位点 4 碱基的含量

Base	Disordered regions			ordered regions		
	First	second	third	First	second	third
A	32.86%	39.50%	27.65%	32.70%	40.96%	23.75%
G	40.80%	26.09%	33.11%	39.77%	23.26%	36.97%
C	30.28%	37.02%	32.70%	31.37%	28.35%	40.27%
T	26.74%	29.11%	44.15%	25.84%	42.27%	31.89%

4. 结论

本文基于 Uniprot 库以及 disprot、mobidb 数据库，获得人类 3 类疾病相关的固有无序蛋白数据集。从其 mRNA 序列角度分析，结果发现：1) 3 类疾病相关的固有无序蛋白中 64 个密码子出现频率在有序和无序区存在差异，其中在癌症相关的固有无序蛋白中有 58 个密码子在有序和无序区出现是存在显著差异的；在心血管疾病相关的固有无序蛋白中有 57 个密码子在有序和无序区出现是存在显著差异的；在神经退行性疾病相关的固有无序蛋白中有 56 个密码子在有序和无序区出现是存在显著差异的；2) 3 类疾病相关的固有无序蛋白中 GC 含量在有序和无序区均存在显著差异；3) 同义密码子在 3 类疾病相关的 IDPs 中有序和无序区均存在不同程度的使用偏好性；4) 编码 3 类疾病相关的 IDPs 的密码子在有序和无序区 3 个位点碱基的组成上也存在一定的差异。这些结果为下一步识别固有无序蛋白，以及固有无序蛋白与疾病的关联提供了很好的参考信息。

致　谢

感谢匿名的评审专家给出的宝贵意见。

基金项目

本文由国家自然科学基金项目(62262050)和国家自然科学基金专项项目(62141204)资助下完成。

参考文献

[1] Bo, H., *et al*. (2009) Predicting Intrinsic Disorder in Proteins: An Overview. *Cell Research*, **19**, 929-949. https://doi.org/10.1038/cr.2009.87

[2] Munsky, B., Neuert, G. and van Oudenaarden, A. (2012) Using Gene Expression Noise to Understand Gene Regulation. *Science*, **336**, 183-187. https://doi.org/10.1126/science.1216379

[3] Csizmok, V., Follis, A.V., Kriwacki, R.W. and Forman-Kay, J.D. (2016) Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chemical Reviews*, **116**, 6424-6462. https://doi.org/10.1021/acs.chemrev.5b00548

[4] Wright, P.E. and Dyson, H.J. (2015) Intrinsically Disordered Proteins in Cellular Signaling and Regulation. *Nature Reviews Molecular Cell Biology*, **16**, 18-29. https://doi.org/10.1038/nrm3920

[5] Binolfi, A., Limatola, A., Verzini, S., Kosten, J., Theillet, F.X., Rose, H.M., Bekei, B., Stuiver, M., Rossum, M.V. and Selenko, P. (2016) Intracellular Repair of Oxidation-Damaged α-Synuclein Fails to Target C-Terminal Modification Sites. *Nature Communications*, **7**, Article No. 10251. https://doi.org/10.1038/ncomms10251

[6] Fung, H.Y.J., Birol, M. and Rhoades, E. (2018) IDPs in Macromolecular Complexes: The Roles of Multivalent Interactions in Diverse Assemblie. *Current Opinion in Structural Biology*, **49**, 36-43. https://doi.org/10.1016/j.sbi.2017.12.007

[7] Babu, M.M. (2016) The Contribution of Intrinsically Disordered Regions to Protein Function, Cellular Complexity, and Human Disease. *Biochemical Society Transactions*, **44**, 1185-1200. https://doi.org/10.1042/BST20160172

[8] Babu, M.M., van der Lee, R. and de Groot, N.S. (2011) Intrinsically Disordered Proteins: Regulation and Disease. *Current Opinion in Structural Biology*, **21**, 432-440. https://doi.org/10.1016/j.sbi.2011.03.011

[9] Kastenhuber, E.R. and Lowe, S.W. (2017) Putting p53 in Context. *Cell*, **170**, 1062-1078. https://doi.org/10.1016/j.cell.2017.08.028

[10] Bykov, V.J., Eriksson, S.E. and Bianchi, J. (2018) Targeting Mutant p53 for Efficient Cancer Therapy. *Nature Reviews Cancer*, **18**, 89-102. https://doi.org/10.1038/nrc.2017.109

[11] Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Sequence Complexity of Disordered Protein. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **42**, 38-48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3

[12] Oldfield, C.J., Peng, Z., Uversky, V.N. and Kurgan, L. (2020) Codon Selection Reduces GC Content Bias in Nucleic Acids Encoding for Intrinsically Disordered Proteins. *Cellular and Molecular Life Sciences*, **77**, 149-160. https://doi.org/10.1007/s00018-019-03166-6

[13] Zhou, M., Wang, T., Fu, J., Xiao, G. and Liu, Y. (2015) Nonoptimal Codon Usage Influences Protein Structure in In-

trinsically Disordered Regions. *Molecular Microbiology*, **97**, 974-987. https://doi.org/10.1111/mmi.13079

[14] Homma, K., Noguchi, T. and Fukuchi, S. (2016) Codon Usage Is Less Optimized in Eukaryotic Gene Segments Encoding Intrinsically Disordered Regions than in Those Encoding Structural Domains. *Nucleic Acids Research*, **44**, 10051-10061. https://doi.org/10.1093/nar/gkw899

[15] Peng, Z., Uversky, V.N. and Kurgan, L. (2016) Genes Encoding Intrinsic Disorder in Eukaryota Have High GC Content. *Intrinsically Disordered Proteins*, **4**, e1262225. https://doi.org/10.1080/21690707.2016.1262225

[16] Basile, W., Sachenkova, O., Light, S. and Elofsson, A. (2017) High GC Content Causes Orphan Proteins to Be Intrinsically Disordered. *PLOS Computational Biology*, **13**, e1005375. https://doi.org/10.1371/journal.pcbi.1005375

[17] Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R. and Bursteinas, B. (2021) UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480-D489.

[18] Hatos, A., Hajdu-Soltész, B., Monzon, A.M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., *et al*. (2019) DisProt: Intrinsic Protein Disorder Annotation in 2020. *Nucleic Acids Research*, **48**, D269-D276. https://doi.org/10.1093/nar/gkz975

[19] Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A.M., Parisi, G., Schad, E., Sormanni, P., Tompa, P., Vendruscolo, M., Vranken, W.F. and Tosatto, S.C.E. (2018) MobiDB 3.0: More Annotations for Intrinsic Disorder, Conformational Diversity and Interactions in Proteins. *Nucleic Acids Research*, **46**, D471-D476. https://doi.org/10.1093/nar/gkx1071

[20] Madeira, F., Park, Y.M., Lee, J., *et al*. (2019) The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Research*, **47**, W636-W641. https://doi.org/10.1093/nar/gkz268

[21] Oldfeld, C.J., Peng, Z.L., Uversky, V.N. and Kurgan, L. (2020) Codon Selection Reduces GC Content Bias in Nucleic Acids Encoding for Intrinsically Disordered Proteins. *Cellular and Molecular Life Sciences*, **77**, 149-160. https://doi.org/10.1007/s00018-019-03166-6

[22] Panda, A., Podder, S., Chakraborty, S. and Ghosh, T.C. (2014) GC-Made Protein Disorder Sheds New Light on Vertebrate Evolution. *Genomics*, **104**, 530-537.

[23] Bi, K., Lu, Z.H., Ge, Q.Y. and Gu, W.J. (2022) Extended XOR Algorithm with Biotechnology Constraints for Data Security in DNA Storage. *Current Bioinformatics*, **17**, 401-410. https://doi.org/10.2174/1574893617666220314114732

[24] Sauvat, A., *et al*. (2021) High-Throughput Label-Free Detection of DNA-to-RNA Transcription Inhibition Using Brightfield Microscopy and Deep Neural Networks. *Computers in Biology and Medicine*, **133**, Article ID: 104371. https://doi.org/10.1016/j.compbiomed.2021.104371

[25] Kou, G.S. and Feng, Y.E. (2015) Identify Five Kinds of Simple Super Secondary Structures with Quadratic Discriminant Algorithm Based on the Chemical Shifts. *Journal of Theoretical Biology*, **380**, 392-398. https://doi.org/10.1016/j.jtbi.2015.06.006

[26] Li, J., *et al*. (2021) Comprehensive Analysis Reveals GPRIN1 Is a Potential Biomarker for Non-Small Cell Lung Cancer. *Current Bioinformatics*, **16**, 130-138. https://doi.org/10.2174/1574893615999200530201333

[27] Prakosa, A., Southworth, M.K., Avari Silva, J.N., Silva, J.R. and Trayanova, N.A. (2021) Impact of Augmented-Reality Improvement in Ablation Catheter Navigation as Assessed by Virtual-Heart Simulations of Ventricular Tachycardia Ablation. *Computers in Biology and Medicine*, **133**, Article ID: 104366. https://doi.org/10.1016/j.compbiomed.2021.104366

[28] Tang, H., Zhao, Y.W., Zou, P., Zhang, C.M., Chen, R., Huang, P. and Lin, H. (2018) HBPred: A Tool to Identify Growth Hormone-Binding Proteins. *International Journal of Biological Sciences*, **14**, 957-964. https://doi.org/10.7150/ijbs.24174

[29] Dao, F.Y., Lv, H., Fullwood, M.J. and Lin, H. (2022) Accurate Identification of DNA Replication Origin by Fusing Epigenomics and Chromatin Interaction Information. *Research*, **2022**, Article ID: 9780293. https://doi.org/10.34133/2022/9780293

[30] Sharp, P.M. and Li, W. (1986) An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms. *Journal of Molecular Evolution*, **24**, 28-38. https://doi.org/10.1007/BF02099948