# 特征说明

#### 蛋白序列+核酸序列模型

Table 数据集统计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | UniProt | Protein seq | Na seq | Complex |
| Total | 767 | 3010 | 4200 | 8553 |
| Wt | 767 | 767 | 4159 | 4843 |
| Wt-SDNA | 421 | 421 | 1457 | 1682 |
| Wt-DDNA | 409 | 409 | 1758 | 2058 |
| Wt-SRNA | 224 | 224 | 856 | 1001 |
| Wt-DRNA | 33 | 33 | 88 | 102 |
| Mut | 346 | 2243 | 885 | 3710 |
| Mut-SDNA | 175 | 933 | 316 | 1258 |
| Mut-DDNA | 163 | 947 | 342 | 1322 |
| Mut-SRNA | 107 | 807 | 215 | 1070 |
| Mut-DRNA | 11 | 59 | 12 | 60 |

#### 蛋白序列特征

基于残基的蛋白序列特征。Protein Length = 100 🡪 100\*N matrix，N表示特征维度;

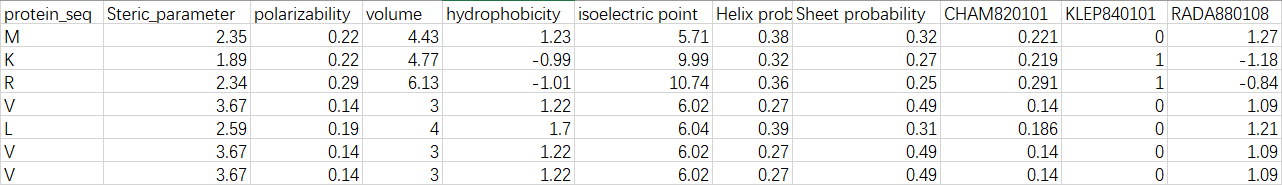
##### 结合相关理化性质

* 理由：

**某些类型的氨基酸更加倾向于和某种类型的配体相互作用**，综述通过统计学分析,发现AAindex中的polarizability (AAindex CHAM820101) 和 DNA结合的蛋白残基相关度高，charge (AAindex KLEP840101) 和RNA结合的蛋白残基相关度高，polarity (AAindex RADA880108) 和蛋白结合的蛋白残基相关度高。**用AAindex1全部，AAindex3结合模型框架考虑使用。（AAindex网址：**[**https://www.genome.jp/aaindex/**](https://www.genome.jp/aaindex/) **，AAindex1 下载网址：** [**https://www.genome.jp/ftp/db/community/aaindex/aaindex1**](https://www.genome.jp/ftp/db/community/aaindex/aaindex1) **）**

此外，目前iDRNA-ITF（工具预测蛋白序列上的结合位点）还利用了氨基酸的7种和结合位点显著相关的理化性质。

* 输入：蛋白序列；
* 输出：每个残基**10维特征;**
* 计算时间：< 1 s；



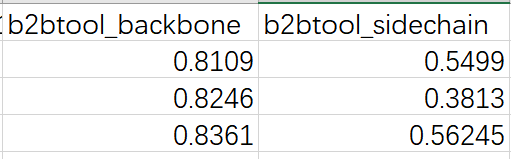
* 参考文献

Zhang, J., Ma, Z., & Kurgan, L. (2019). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Briefings in bioinformatics, 20(4), 1250–1268. <https://doi.org/10.1093/bib/bbx168>

Wang, N., Yan, K., Zhang, J., & Liu, B. (2022). iDRNA-ITF: identifying DNA- and RNA-binding residues in proteins based on induction and transfer framework. Briefings in bioinformatics, 23(4), bbac236. <https://doi.org/10.1093/bib/bbac236>

##### 刚柔性

* 理由：蛋白质表面残基的移动性与其功能密切相关。因此，识别极其刚性或柔性的表面残基有助于识别蛋白表面功能重要的残基。
* 工具：b2btools
* 输入：蛋白序列；
* 输出：每个残基**2维特征;**
* 计算时间：< 1 min;；



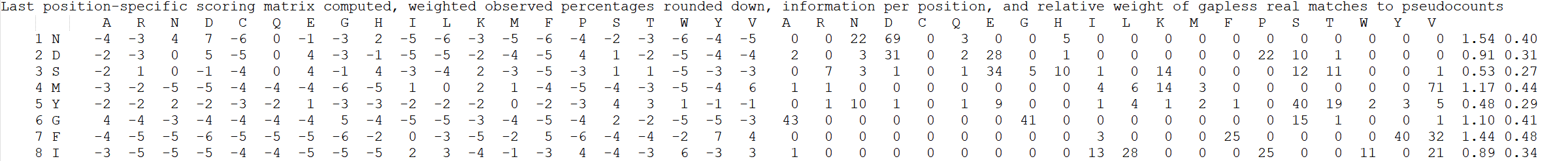
值越高，刚性的可能性越高。

* 参考文献

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., & Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. Nucleic acids research, 42(Web Server issue), W264–W270. <https://doi.org/10.1093/nar/gku270>

##### 保守性

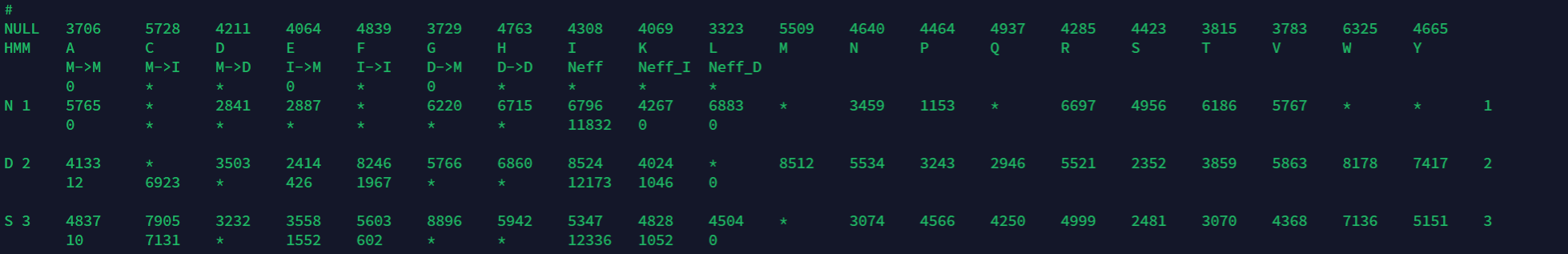
* 理由：核酸结合残基在进化过程中是保守的，核酸通常与蛋白质表面的蛋白质相互作用。
* 工具：PROVEAN、PSI-BLAST、HHblits
* 输入：蛋白序列
* 输出：每个残基的保守性指标,每个残基共**输出71维特征**。PROVEAN(1),PSSM(40), HMM(30)
  + PROVEAN:将每个氨基酸突变为ALA，检索NR库（NR00-NR05），得到PROVEAN Score;
  + PSI-BLAST：检索UniRef90库（参考GraphSite）；前20列表示PSSM，后20列表示PSFM。PSSM表示20种氨基酸的保守性，**每个位置上的数值表示对应氨基酸在该位置上的保守程度，数值越高表示该位置上的氨基酸更为保守**。PSFM表示每个位置上氨基酸的出现频率。



* + - Norm Method

Each element x in the PSSM is normalized to the range [0, 1]

* + HHblits：检索uniclust30库（参考GraphBind）. The HMM matrix consists of 20 columns of observed frequencies for 20 amino acids in homologous sequences, seven columns of transition frequencies and three columns of local diversities. \*号表示这个位置对应的是一个gap（缺失）状态



* + - Norm Method

Convert each score to the range [0, 1], \* was replace as 9999;

* 计算时间：PROVEAN (~ 10 min,300AA)，PSI-BLAST & HHblits (~ 30 min,300AA, 5个并行)
* 参考文献

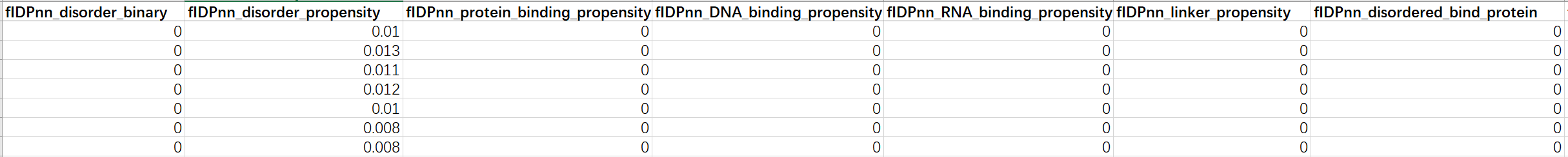
Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics (Oxford, England), 31(16), 2745–2747. <https://doi.org/10.1093/bioinformatics/btv195>

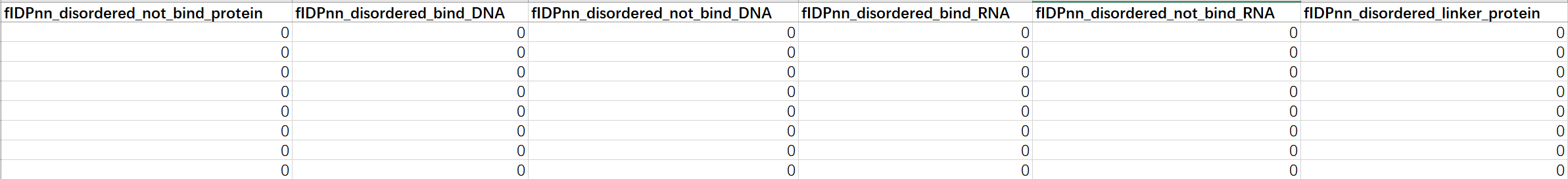
Bhagwat M, Aravind L. PSI-BLAST Tutorial. In: Bergman NH, editor. Comparative Genomics: Volumes 1 and 2. Totowa (NJ): Humana Press; 2007. Chapter 10. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK2590/>

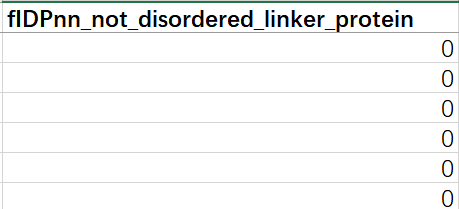
Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9(2), 173–175. <https://doi.org/10.1038/nmeth.1818>

##### Amino acid Disorder

* 理由：Intrinsic disorder amino acid在蛋白-蛋白、蛋白-核酸结合中富集。
* 工具：flDPnn
* 输入：蛋白序列；只拿取DNA和RNA的；
* 输出：每个氨基酸的disorder的二分类和可能性;**共14维度；**







# The disordered flexible linkers (DFLs) are disordered regions that serve as linkers or spacers between domains in multi-domain proteins and between structured constituents in domains

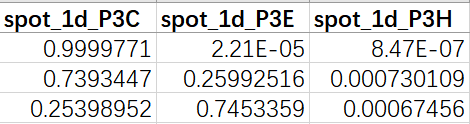
Disorder linker 用于连接不同的蛋白质结构域。

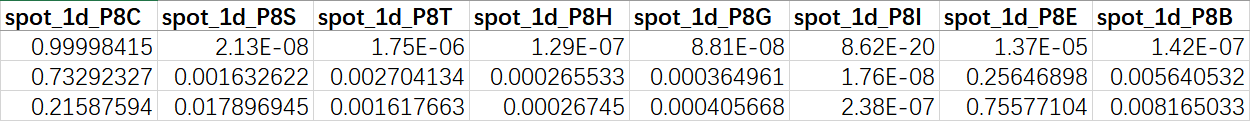
* 计算时间：蛋白序列链长 529，耗时：~1min;
* 参考文献：

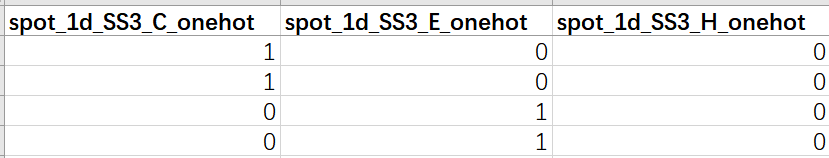
Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., & Kurgan, L. (2021). flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nature communications, 12(1), 4438. <https://doi.org/10.1038/s41467-021-24773-7>

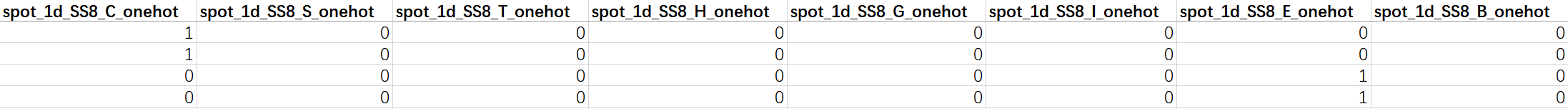
##### 二级结构

* 理由：蛋白质的二级结构可以提供功能区域，如蛋白质的活性位点或结合位点。特定的二级结构元素（如螺旋或折叠）可以形成特定的功能区域，从而实现蛋白质的特定功能，如酶活性、配体结合等。
* 工具：SPOT-1D-Single
* 输入：蛋白序列
* 输出：每个氨基酸不同二级结构类型的binary prediction 和 propensity；**共22维;**







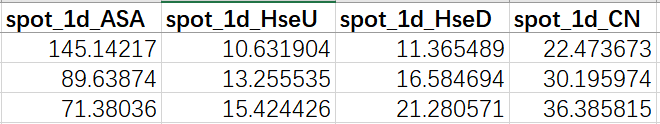


* 计算时间：蛋白序列链长 529，耗时：< 1min;
* 参考文献：

Singh, J., Litfin, T., Paliwal, K., Singh, J., Hanumanthappa, A. K., & Zhou, Y. (2021). SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. Bioinformatics (Oxford, England), 37(20), 3464–3472. <https://doi.org/10.1093/bioinformatics/btab316>

##### 溶剂可及性

* 理由：蛋白核酸结合多发生在蛋白质表面。
* 工具：SPOT-1D-Single
* 输入：蛋白序列
* 输出：溶液可及表面积、半球暴露程度（HSE-up、HSE-down）和 残基接触数量（CN）；**共4维。**
  + 半球暴露程度（HSE-up、HSE-down）：以残基CA原子为球心，画一个半径13A的球体。通过残基的CA-CB向量将球体分为两半，统计上下半球中的CA原子数量。
  + 残基接触数量（CN）：残基CA原子，半径13 A内的CA原子数量;
* 计算时间：蛋白序列链长 529，耗时：< 1min;

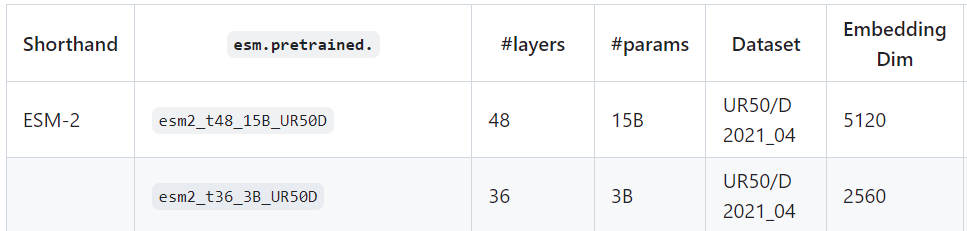


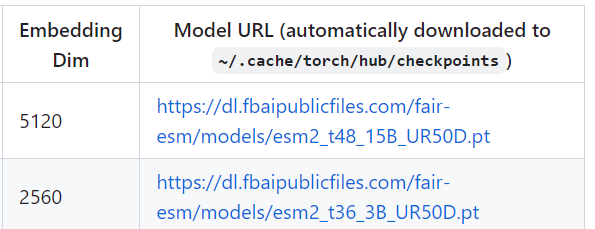
* 参考文献：

Singh, J., Litfin, T., Paliwal, K., Singh, J., Hanumanthappa, A. K., & Zhou, Y. (2021). SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. Bioinformatics (Oxford, England), 37(20), 3464–3472. <https://doi.org/10.1093/bioinformatics/btab316>

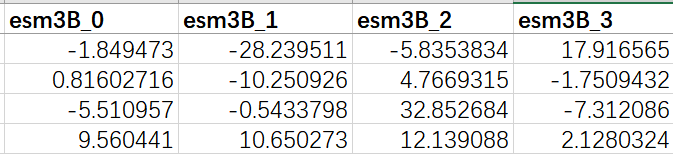
##### ESM

* 工具：ESM2-3B





* 输入：蛋白序列
* 输出：每个残基的特征，每个残基有 2560 维度;



* 计算时间： < 1 min （300 AA）
* 参考文献：

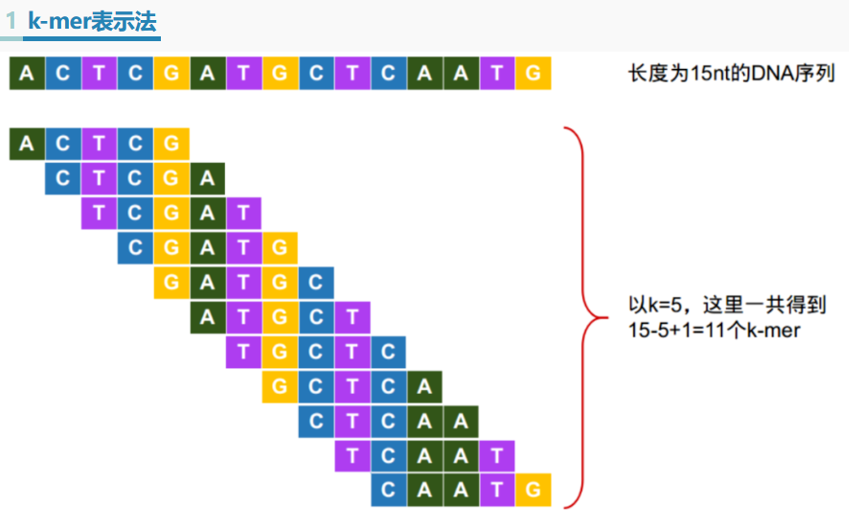
Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences of the United States of America, 118(15), e2016239118. <https://doi.org/10.1073/pnas.2016239118>

#### 核酸序列特征

DNA 理化特征

**DNA embedding特征**

* DNABERT
  + 模型：DNABERT3、DNABERT4、DNABERT5
  + 输入：DNA序列依据滑动窗口大小，步长为1，进行划分。划分kmer的示例如下：



* + 输出：**768维度/kmer**；
  + 计算时间：~ 1 min；
  + 参考文献：

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics (Oxford, England), 37(15), 2112–2120. <https://doi.org/10.1093/bioinformatics/btab083>

* dna2vec
  + 模型：dna2vec3、dna2vec4、dna2vec5
  + 输入：DNA序列依据滑动窗口大小，步长为1，进行划分。
  + 输出：**100维度/kmer**；
  + 计算时间：~ 1 min；
* EnhancerBERT
  + 模型：EnhancerBERT3、EnhancerBERT4、EnhancerBERT5
  + 输入：DNA序列依据滑动窗口大小，步长为1，进行划分。
  + 输出：**768维度/kmer**；
  + 计算时间：~ 1 min；

**单独用，平均和最大池化都试一下。**

RNA embedding特征

* SpliceBERT
* 输出：每个核苷酸输出**512维特征**;
* 计算时间：~ 1 min；
* 参考文献：

Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. doi: <https://doi.org/10.1101/2023.01.31.526427>

**RNA同时使用DNA的embedding特征，将U 改成 T.**

### 蛋白结构+核酸序列模型

Table 数据集统计

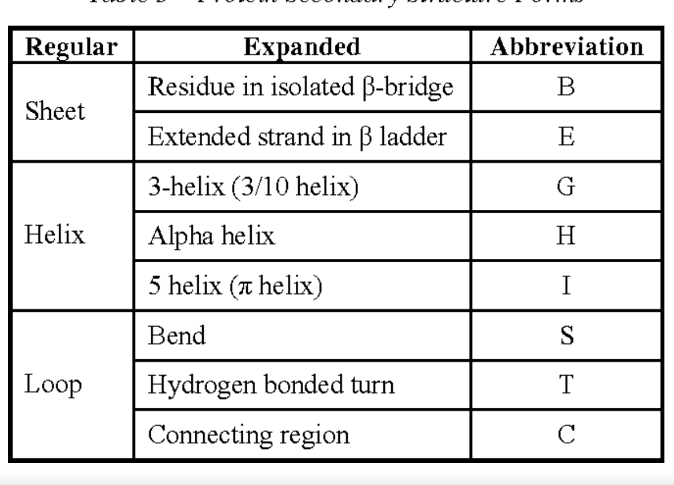
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | UniProt | Protein seq | Na seq | Complex |
| Total | 767 | 3010 | 4200 | 8553 |
| Total-AF2 | 745/(757) | 2951 | 4123 | 8416 |
| pLDDT-median > 70 | 537 | 2204 | 3252 | 6363 |
| Wt-DNA | 391 | 1420 | 2561 | 4695 |
| Wt-RNA | 172 | 820 | 691 | 1668 |
| pLDDT-median <= 70 | 208 | 747 | 1023 | 2053 |
| Wt-DNA | 162 | 562 | 770 | 1540 |
| Wt-RNA | 53 | 224 | 253 | 513 |

#### 蛋白结构残基特征

##### 二级结构和溶液可及表面积

我们输入蛋白单体结构，使用DSSP，计算出：（4维特征/残基）

* + 残基的八种二级结构;
  + 残基在蛋白单体结构中的溶表，并使用（PMID: 4023714）中提供的残基的标准溶表，来计算残基的相对溶表；



**问题**：编码方式需要确定：one-hot 还是 数字编码；

##### 结构特性

1. 波动性：我们输入蛋白单体结构，使用FlexPred，计算出每个残基的波动性；（1维特征/残基）
2. 动态特征：我们输入蛋白单体结构，使用ProDy，构建蛋白质结构的弹性网络,**计算得到每个残基的机械刚度, 扰动响应, 扰动响应效应, 扰动响应灵敏度；**（4维特征/残基）
3. 残基对蛋白折叠的影响：我们输入蛋白单体结构，使用FoldX的ALAScan模块，将结构中的每个残基突变为ALA，取得预测的ΔΔGfold**；**（1维特征/残基）
4. 残基对结构的贡献：FoldX的SequenceDetail模块，我们输入蛋白单体结构，计算每一个残基对于整体结构的能量贡献以及残基的键角。分别计算了残基的Phi、Psi，以及对整个结构的总能量贡献、主链氢键贡献，侧链氢键贡献，范德华能量贡献，静电相互作用，埋藏极性基团罚分，疏水基团能量，范德华冲突罚分，侧链熵，主链熵,生成顺式肽键的能量，范德华扭转冲突的贡献，对螺旋偶极子的静电能量贡献，侧链接触比，主链接触比等残基能量项; （31维特征/残基）