



多模态数据融合的 蛋白质功能预测

南京农业大学 人工智能学院

汇报人：朱一亨

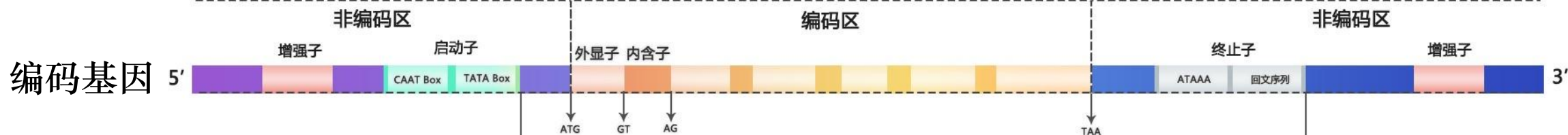
2024年06月20日

1 研究背景



非编码基因

编码基因



转录



mRNA

翻译



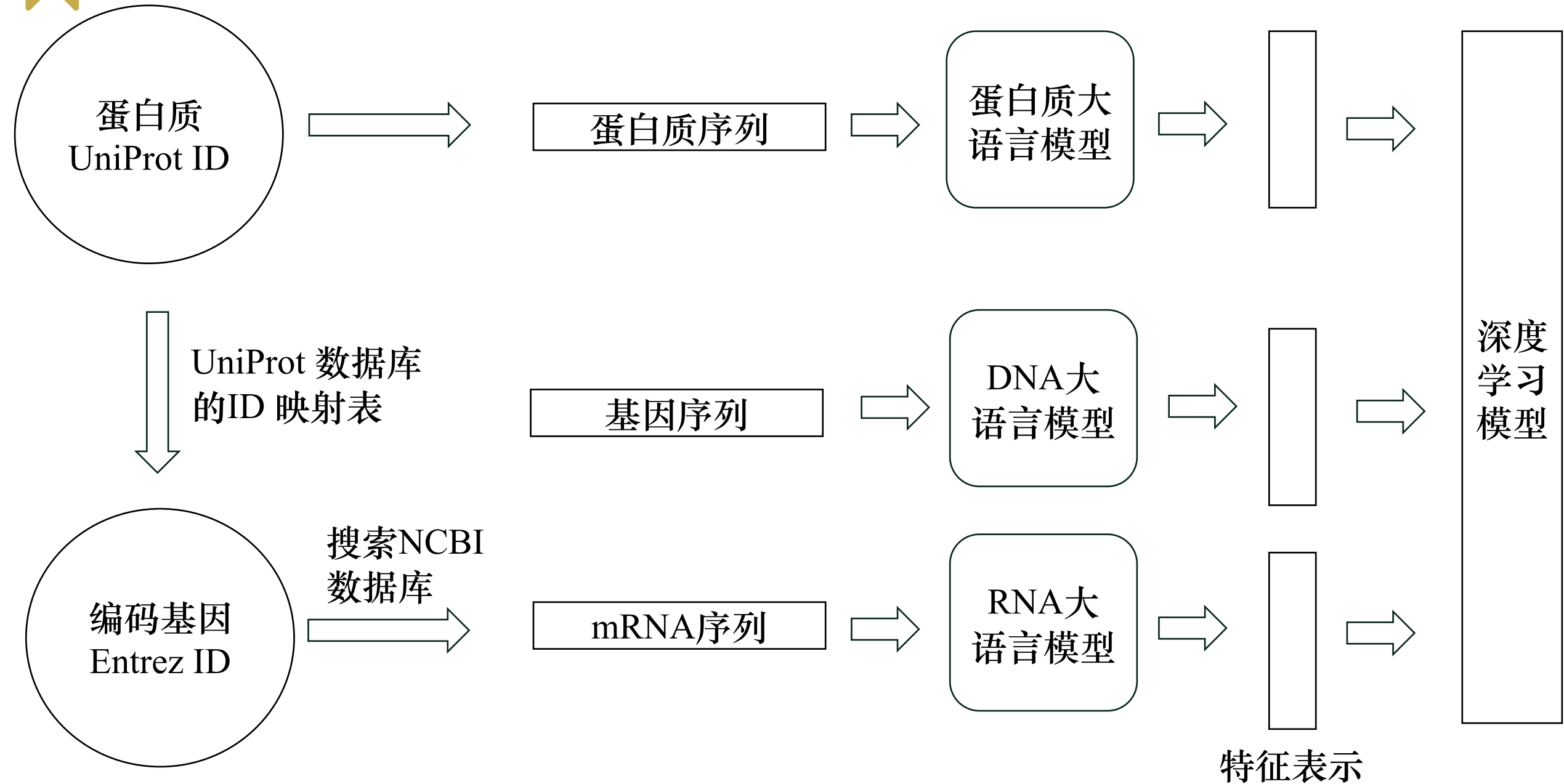
蛋白质(肽链)

生物中心法则

2 现有方法的不足和挑战

- 现有的蛋白质功能预测方法只注重从蛋白质自身挖掘知识，忽略了编码基因和mRNA中的知识。
- 基因组数据：基因序列（DNA序列）和表达数据
- 转录组数据：mRNA序列

3 初步的方法和设想



4 数据集构建

- (1) 从蛋白质功能注释数据库GOA中，下载全部的蛋白质，共计134778个。
- (2) 部分蛋白质，在UniProt中找不到对应的编码基因记录，故丢弃。剩余100947个蛋白质。
- (3) Training dataset: 81657 proteins, before 2021-12-31
- (4) Validation dataset: 588 proteins, 2022-01-01 between 2022-06-30
- (5) Test dataset: 1723 proteins, 2022-07-01 between 2023-06-30
- (6) CD-HIT: (sequence identity<30%, training, validation, test datasets)

初步的实验结果

Dataset	Method	F_{\max}			AUPR		
		MF	BP	CC	MF	BP	CC
Validation Dataset (588 proteins)	DIOMAND	0.609	0.370	0.499	0.321	0.207	0.279
	BLAST	0.631	0.363	0.520	0.434	0.236	0.360
	Naive	0.432	0.256	0.449	0.244	0.155	0.331
	ESM2	0.689	0.411	0.589	0.657	0.292	0.563
	DNABERT2	0.461	0.319	0.520	0.357	0.219	0.468
	ESM2 + DNABERT2	0.681	0.427	0.615	0.651	0.310	0.606
Test Dataset (1723 proteins)	DIOMAND	0.632	0.340	0.474	0.386	0.193	0.249
	BLAST	0.630	0.331	0.490	0.445	0.207	0.323
	Naïve	0.367	0.218	0.389	0.196	0.112	0.257
	ESM2	0.665	0.396	0.552	0.627	0.266	0.521
	DNABERT2	0.444	0.356	0.483	0.339	0.266	0.427
	ESM2 + DNABERT2	0.677	0.433	0.585	0.661	0.324	0.562

目前存在的问题

➤ 部分基因序列过长，不好处理。

<10000: 61945

10000-20000: 8954

20000-30000: 5322

30000-40000: 3432

40000-50000: 2562

50000-60000: 1872

60000-70000: 1403

70000-80000: 1075

80000-90000: 961

>90000: 7112

Average: 28668

Go to [reference sequence details](#)

Primary Assembly ⌵

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

Tools | Tracks | Download | ?

7 M | 7,500 K | 8 M

RBFOX1

Gene: RBFOX1
Name: RNA binding fox-1 homolog 1
Location: 5,239,721..7,713,340
Length: 2,473,620 nt

[Positional Info]

NC_000016.10 position: 6,923,633
Gene position: 1,683,913

Links & Tools
GeneID: [54715 \(RBFOX1\)](#)
HGNC: [18222](#)
MIM: [605104](#)

BLAST nr: [NC_000016.10 \(5,239,721..7,713,340\)](#)
BLAST to Genome: [NC_000016.10 \(5,239,721..7,713,340\)](#)
FASTA record: [NC_000016.10 \(5,239,721..7,713,340\)](#)
GenBank record: [NC_000016.10 \(5,239,721..7,713,340\)](#)

SELECTED TRACKS

- GTR
- Nucleotide
- OMIM
- Probe
- Protein
- PubChem Compound
- PubChem Substance
- ed
- ed (GeneR)
- ed (OMIM)
- ed(nucleot
- q Proteins
- q RNAs
- qGene
- d gene-spe

下一步的计划

- 测试mRNA序列的性能
- 将DNABERT2替换成其他语言模型，如NC-Transformer。
- 在蛋白质通道加入AF2预测的结构，采用图注意力网络，使模型更复杂。

数据统计

- 目前的数据集共计83968 proteins (Dataset I)

Training dataset: 81657 proteins

Validation dataset: 588 proteins

Test dataset: 1723 proteins

- 只有62683proteins在NCBI中找到mRNA sequences (Dataset II)


Training dataset: 60994 proteins

Validation dataset: 453proteins

Test dataset: 1236 proteins

Dataset I	Method	F _{max}			AUPR		
		MF	BP	CC	MF	BP	CC
Validation Dataset (588 proteins)	DIOMAND	0.609	0.370	0.499	0.321	0.207	0.279
	BLAST	0.631	0.363	0.520	0.434	0.236	0.360
	Naive	0.432	0.256	0.449	0.244	0.155	0.331
	ESM2	0.689	0.411	0.589	0.657	0.292	0.563
	DNABERT2 (Gene Sequence)	0.461	0.319	0.520	0.357	0.219	0.468
	NC-Transformer (Gene Sequence)	0.467	0.301	0.543	0.299	0.184	0.454
	ESM2 + DNABERT2 (Gene Sequence)	0.681	0.427	0.615	0.651	0.310	0.606
	ESM2 + DNABERT2 (mRNA Sequence)	0.678	0.407	0.590	0.655	0.294	0.579
	ESM2 + NC-Transformer (Gene Sequence)	0.693	0.421	0.597	0.665	0.310	0.576
Test Dataset (1723 proteins)	DIOMAND	0.632	0.340	0.474	0.386	0.193	0.249
	BLAST	0.630	0.331	0.490	0.445	0.207	0.323
	Naïve	0.367	0.218	0.389	0.196	0.112	0.257
	ESM2	0.665	0.396	0.552	0.627	0.266	0.521
	DNABERT2 (Gene Sequence)	0.444	0.356	0.483	0.339	0.266	0.427
	NC-Transformer (Gene Sequence)	0.438	0.370	0.493	0.329	0.271	0.396
	ESM2 + DNABERT2	0.677	0.433	0.585	0.661	0.324	0.562
	ESM2 + DNABERT2 (mRNA Sequence)	0.671	0.420	0.579	0.622	0.300	0.544
	ESM2 + NC-Transformer (Gene Sequence)	0.670	0.433	0.584	0.645	0.329	0.540

Dataset II	Method	F_{\max}			AUPR		
		MF	BP	CC	MF	BP	CC
Validation Dataset (453 proteins)	ESM2	0.698	0.390	0.596	0.667	0.269	0.580
	DNABERT2 (Gene Sequence)	0.504	0.317	0.543	0.387	0.218	0.523
	DNABERT2 (mRNA Sequence)	0.531	0.297	0.518	0.411	0.198	0.492
	ESM2 + DNABERT2 (Gene Sequence)	0.691	0.405	0.618	0.670	0.298	0.601
	ESM2 + DNABERT2 (mRNA Sequence)	0.687	0.393	0.594	0.661	0.272	0.586
Test Dataset (1236 proteins)	ESM2	0.659	0.392	0.550	0.621	0.258	0.525
	DNABERT2 (Gene Sequence)	0.464	0.364	0.489	0.341	0.271	0.450
	DNABERT2 (mRNA Sequence)	0.470	0.330	0.484	0.347	0.225	0.439
	ESM2 + DNABERT2 (Gene Sequence)	0.677	0.434	0.588	0.637	0.326	0.564
	ESM2 + DNABERT2 (mRNA Sequence)	0.667	0.414	0.568	0.620	0.293	0.545



➤ mRNA不起作用的原因：mRNA中只有外显子序列，它们直接通过密码子表翻译成氨基酸序列，因此mRNA序列中的大部分知识和蛋白质序列中的知识重复。

>NC_004327.3:c489287-488579 Plasmodium falciparum 3D7 genome assembly, chromosome:

6

ATGGTAATAAAAAGAAGAAAAAGAAAAAACAACACTTATAAATATATAAATAATTTATATATATA
AATCAGCGTTATTATAAAATGAAAATAAATAAATATATAAATATATAAATATATATAATGAATAGA
TTAATAAAAATAAGAAAATCAAAGAAATATATACCCCTTTAATATTTTAGTTAACATATATATATAT
ATATATATATATATATATATGAATATACATTAGTTAAAATATTTTCCTTGATTATGTTTATATTTATG
AATTTTTTTTTTTTTTTTTTTGAAGGGATCAATTAACGTTTTAGATTAACAAGACTTGGAAAATGCAG
AAGGCAAATAGGCCTGTACCCATTGGTATAGATTAAGAAAGATACAAAATAAGGTAAGGCTTAAAC
AAAAATATATTTATATATATGTATGTCAATGAATATATGCTATGCTATGAATAAAAAGAAAAATTATAA
AAAACGAGTACTTTTTATTTTATGCAAATAACTTAATGTTGTTATAATATATAACATGGATTTGTTG
TATTTGTATATAATATTATATTATATTATATTATATTATATTATATTATATTATAATATGTTTTT
TTTTTTTTTTAATTTTTTTTTTTTATTAGATATAATACAAAAGAAGACACTGGAGAAGAACCAAATTA
GGATTATAA

← Gene sequence

>XM_002808677.1 Plasmodium falciparum 3D7 60S ribosomal protein L39
(PF3D7_0611700), partial mRNA

ATGGGATCAATTAACGTTTTAGATTAACAAGACTTGGAAAATGCAGAAGGCAAATAGGCCTGTAC
CCCATTGGTATAGATTAAGAAAGATACAAAATAAGATATAATACAAAAGAAGACACTGGAGAAGAAC
CAAATTAGGATTATAA

← mRNA sequence
(cDNA)

>C0H4H3

MGSIKRFRLKQRLGKCRQRNPVPHWYRLKKDTKIRYNTKRRHWRRTKLGL

← Protein sequence

		密码子的第二位				
		T	C	A	G	
密码子的第一位 (5'端)	T	TTT: Phe F	TCT: Ser S	TAT: Tyr Y	TGT: Cys C	T
	TTC: Phe F	TCC: Ser S	TAC: Tyr Y	TGC: Cys C	C	
	TTC: LeI L	TCA: Ser S	TAA: Ter *	TGA: Ter *	A	
	TTC: LeI L	TCG: Ser S	TAG: Ter *	TGG: Trp W	G	
C	CTT: LeI L	CCT: Pro P	CAT: His H	CGT: Arg R	T	
	CTC: LeI L	CCC: Pro P	CAC: His H	CGA: Arg R	C	
	CTA: LeI L	CCA: Pro P	CAA: Gln Q	CGC: Arg R	A	
	CTG: LeI L	CCG: Pro P	CAG: Gln Q	CGG: Arg R	G	
A	ATT: Ile I	ACT: Thr T	AAT: Asn N	AGT: Ser S	T	
	ATC: Ile I	ACC: Thr T	AAC: Asn N	AGC: Ser S	C	
	ATA: Ile I	ACA: Thr T	AAA: Lys K	AGA: Arg R	A	
	ATG: Met M	ACG: Thr T	AAG: Lys K	AGG: Arg R	G	
G	GTT: Val V	GCT: Ala A	GAT: Asp D	GGT: Gly G	T	
	GTC: Val V	GCC: Ala A	GAC: Asp D	GGC: Gly G	C	
	GTA: Val V	GCA: Ala A	GAA: GlT E	GGA: Gly G	A	
	GTG: Val V	GCG: Ala A	GAG: GlT E	GGG: Gly G	G	

密码子的第三位 (3'端)