

Enhancing RNA Motif Representation with Contrastive Learning and Language Models for Sequence-Structure Analysis

Vinay Chaudhari¹, Md. Sharear Saon², Grace Fu³, Brent M. Znosko², Jie Hou^{1*}



¹Department of Computer Science, Saint Louis University, St. Louis, MO, 63112, ²Department of Chemistry,

Saint Louis University, St. Louis, MO, 63112, and ³Parkway South High School, Manchester, MO, 63021, USA

Introduction

RNA secondary structural motifs, such as stems, loops, ar st pr IN ac ch re tra Tł re R st

Dataset

Bioinformatics +

SAINT LOUIS UNIVERSITY...

Computational Biology

RNA Cossmos

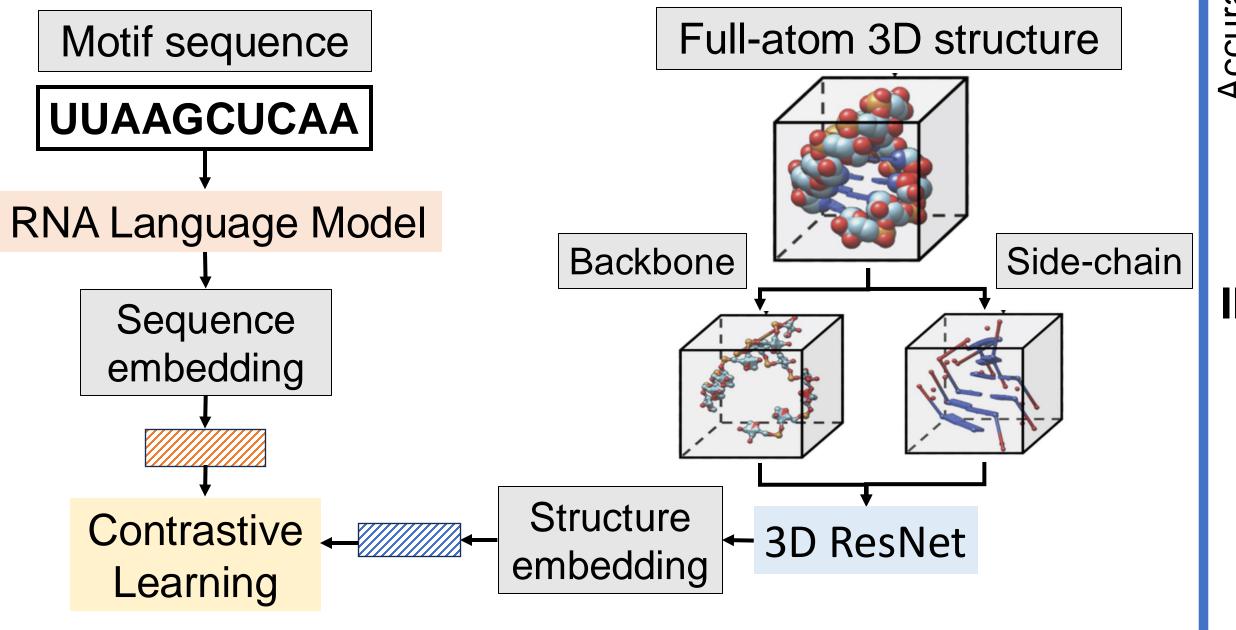
Statistics and Visualization of Motifs in CoSSMos² Database

and bulges, are fundamental units of RNA folding and	Туре	Motif	All PDB	X-ray	NMR	cryo-EM					
structure. These motifs influence critical biological		1x1	5605	1390	465	3750	1 x 1	2 x 2	3 x 3	4 x 4	5 x 5
processes, including gene regulation and molecular	Symmetric	2x2	3291	1374	130	1787				<u> </u>	
interactions. Despite extensive RNA sequence data,	Internal Loops	3x3 4x4	3078 114	1169 21	186 30	1723 63	ade le				
		5x5	341	169	0	172					Rac
		1x2	4298	1519	275	2504	1 x 2	1 x 3	1 x 4	1 x 5	2 x 5
challenging due to the scarcity of experimentally		1x3	3770	1806	121	1843				(hand)	
resolved RNA structures and the limitations of	-	1x4 1x5	2339 458	1044 158	85	1210 300	2 x 3		3 x 4	3 x 5	4 x 5
traditional tools to identify homologous RNA motifs.		2x3	4 <u>3</u> 8 3409	1459	89	1861					
This work focuses on enhancing RNA motif		2x4	1413	554	39	820		2 x 4			
representation by integrating contrastive learning and		2x5	585	143	30	412			R		
RNA language models. By combining sequence and		3x4	1645	497	19	1129					
structure embeddings, we aim to improve motif		3x5 4x5	115 325	74 106	0 9	41 210					
clustering, classification, and sequence-structure		hairpin3	3274	1440	243	1591	Hairpin3	Hairpin4	Hairpin5	Hairpin6	Hairpin7
	Hairpin Loops	hairpin4	42973	18228	2048	22697					
alignment.		hairpin5	17836	7033	801	10002					
Methods		hairpin6 hairpin7	13890 11561	5512 5039	688 318	7690 6204					
A. Sequence-Based Learning		bulge1	37048	14568	1318	21162	Bulge1	Bulge2	Bulge3	Bulge4	Bulge5
		•			240		Duigei	Duigez	Duiges	Duige4	Buiges
	Dulas	bulge2	13258	5565	310	7383					
RNA motif sequences are processed using the RNA	Bulge Loops	bulge3	3474	1876	224	1354					
RNA motif sequences are processed using the RNA language model ¹ for embedding extraction:	Bulge Loops	bulge3 bulge4	3474 851	1876 205	224 79	1354 567					
RNA motif sequences are processed using the RNA language model ¹ for embedding extraction: Method 1: Default embeddings from RNA-FM.		bulge3 bulge4 bulge5	3474 851 318	1876 205 151	224 79 84	1354 567 83					
RNA motif sequences are processed using the RNA language model ¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM,		bulge3 bulge4	3474 851	1876 205	224 79 84	1354 567 83 96558					
RNA motif sequences are processed using the RNA language model ¹ for embedding extraction: Method 1: Default embeddings from RNA-FM.	Loops	bulge3 bulge4 bulge5 Total	3474 851 318 175269	1876 205 151 71100	224 79 84 7591	1354 567 83 96558 Res	Sults				
RNA motif sequences are processed using the RNA language model ¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification.	Loops I. Visualiza	bulge3 bulge4 bulge5 Total	3474 851 318 175269 NA mo	1876 205 151 71100	224 79 84 7591	1354 567 83 96558 Res ngs for	Sults three diff		-	-	
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning 	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total	3474 851 318 175269 NA mo bedding	1876 205 151 71100	224 79 84 7591 beddi Motif	1354 567 83 96558 Res ngs for Sequenc	Sults three diff ce Embeddin	g	Motif Str	ucture Embe	edding
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D 	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total	3474 851 318 175269 NA mo bedding	1876 205 151 71100	224 79 84 7591 beddi Motif (Fine	1354 567 83 96558 Res ngs for Sequence tune with	Sults three diff	g	Motif Str	-	edding
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: 	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total	3474 851 318 175269 NA mo bedding	1876 205 151 71100	224 79 84 7591 beddi Motif (Fine	1354 567 83 96558 Res ngs for Sequence tune with	Sults three diff ce Embeddin	g	Motif Str	ucture Embe	edding
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total	3474 851 318 175269 NA mo bedding	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine	1354 567 83 96558 Res ngs for Sequence tune with	Sults three diff classification	g	Motif Str (Finetune	ucture Embe	edding ication)
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class classification (25 motif types)	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total tion of R quence Emb aw RNA-FM	3474 851 318 175269 NA mo bedding	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine	1354 567 83 96558 Res ngs for Sequence tune with	Sults three diff classification	g	Motif Str (Finetune	with classif	edding
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class classification (25 motif types) Method 2: Apply contrastive learning directly to 	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total tion of R quence Emb aw RNA-FM	3474 851 318 175269 NA mo bedding	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine 1x5 2x4 3x3 bul	1354 567 83 96558 Res ngs for Sequence tune with 2x5 3x4 4x5 5x5 5x5	Sults Sults three diff classification x ⁵ 4x4 2x2 hairpin 2x2 1x3	g n) 1x2 bulge3	Motif Str (Finetune	with classif	edding ication) bulge3 1x2 hairpin7
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class classification (25 motif types) Method 2: Apply contrastive learning directly to structural data for motif representation. 	Loops I. Visualiza Motif Sed	bulge3 bulge4 bulge5 Total Ition of R quence Emb aw RNA-FM	3474 851 318 175269 NA mo bedding) 2x2 1x5 1x bulc	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine 1x5 2x4 3x3 bul	1354 567 83 96558 Res ngs for Sequence tune with 2x5 3x4 4x5 5x5	x ⁵ 4x4	g 1x2 bulge3 hairpin3	Motif Str (Finetune	with classif	edding ication) bulge3 1x2 hairpin7
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class classification (25 motif types) Method 2: Apply contrastive learning directly to structural data for motif representation. C. Contrastive Learning 	Loops I. Visualiza Motif Sea (R	bulge3 bulge4 bulge5 Total Ition of R quence Emb aw RNA-FM	3474 851 318 175269 NA mo bedding) 2x2 1x5 1x 1x5 1x bulc airpin7 x5 hairpir	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine 1x5 2x4 3x3 bul	1354 567 83 96558 Res ngs for Sequence tune with 2x5 3x4 4x5 5x5 5x5	Sults three diff c Embeddin classification x ⁵ 4x4 x ⁵ 4x4 x ⁵ 4x4	g n) 1x2 bulge3	Motif Str (Finetune 2x5 3x4 hairpin6	with classif	edding ication) bulge3 1x2 hairpin7
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class classification (25 motif types) Method 2: Apply contrastive learning directly to structural data for motif representation. 	Loops I. Visualiza Motif Sea (R	bulge3 bulge4 bulge5 Total Ition of R quence Emb aw RNA-FM	3474 851 318 175269 NA mo bedding) 2x2 1x5 1x 1x5 1x bulc airpin7 x5 hairpir	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine 1x5 2x4 3x3 bul	1354 567 83 96558 Res ngs for Sequence tune with 2x5 3x4 4x5 5x5 5x5	Sults three diff c Embeddin classification x ⁵ 4x4 x ⁵ 4x4 x ⁵ 4x4	g 1x2 bulge3 hairpin3	Motif Str (Finetune 2x5 3x4 hairpin6 bulge2 1x1 hairpin	bulge1 hairpin	edding ication) bulge3 1x2 hairpin7
 RNA motif sequences are processed using the RNA language model¹ for embedding extraction: Method 1: Default embeddings from RNA-FM. Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification. B. Structure-Based Learning RNA motif 3D structures are encoded using a 3D ResNet model: Method 1: Train 3D ResNet for multi-class classification (25 motif types) Method 2: Apply contrastive learning directly to structural data for motif representation. C. Contrastive Learning 	Loops I. Visualiza Motif Sea (R	bulge3 bulge4 bulge5 Total Ition of R quence Emb aw RNA-FM	3474 851 318 175269 NA mo bedding) 2x2 1x5 1x 1x5 1x bulc airpin7 x5 hairpir	1876 205 151 71100 tif emb	224 79 84 7591 beddi Motif (Fine 1x5 2x4 3x3 bul	1354 567 83 96558 Res ngs for Sequence tune with 2x5 3x4 4x5 5x5 ge5 sbulge4 x3	Sults three diff c Embeddin classification x ⁵ 4x4 x ⁵ 4x4 x ⁵ 4x4	g 1x2 bulge3 hairpin3	Motif Str (Finetune 2x5 3x4 hairpin6	bulge1 hairpin bulge1 hairpin hairpin4	edding ication) bulge3 1x2 hairpin7

II. Evaluation of Motif classification using sequence model and structure model

- 2) capture relationships between motifs' sequence features and their structural characteristics.
- **D. Framework of Contrastive Learning for RNA** secondary structural motifs

RNA motifs.



References

- 1. Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., ... & Li, Y. (2022). Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*. 2. Richardson, K. E., Kirkpatrick, C. C., & Znosko, B. M. (2020). RNA CoSSMos
- 2.0: an improved searchable database of secondary structure motifs in RNA threedimensional structures. Database, 2020, baz153.

