When will RNA get its AlphaFold Moment?

Bohdan Schneider, et al.

Stephen Owesney



Overview

RNA Refresher

Motivation

RNA Background

RNA 3D Structure Prediction

RNA-Puzzles / CASP-RNA

Challenges

Conclusions / Future Work



• Messenger RNA (mRNA)



• Messenger RNA (mRNA)



- Messenger RNA (mRNA)
- Transfer RNA (tRNA)



- Messenger RNA (mRNA)
- Transfer RNA (tRNA)
- Ribosomal RNA (rRNA)



ribosomal RNA (rRNA)

- Messenger RNA (mRNA)
- Transfer RNA (tRNA)
- Ribosomal RNA (rRNA)



TYPES OF RNA



TYPES OF RNA

	RNAs involved in protein synthesis												•				
	Туре			•	Abbr.		Function			Distribution + Dof +					Regulatory RNAs		
	Messenger RNA		n	mRNA		Codes for protein			All	Туре	٠	Abbr.	٠	Function			
	Ribosomal RNA			ri	rRNA Tr		Translat	Translation		All			aRNA,		Transcriptional attenuation / mRNA de		
	Signal recognition particle RNA			RNA 7	7SL RNA or SRP RNA		Membrane integrati		ion	All	Antisense RNA		asRNA		stabilisation / Translation block		
Transfer RNA			tF	tRNA Trans		Translat	anslation		All	Cis-natural antisense		cie NAT		Gene regulation			
Transfer-messenger RNA		RNA	tr	tmRNA		Rescuing stalled ribo		bosome	s Bad	transcript		CI3-NAT					
F	1	RNA	s involv	ved ir	n post-t	transcription	nal modif	ication o	r DNA repl	ication		CRISPR RNA		crRNA		Resistance to parasites, by targeting the	
Туре		•	Abbr.	٠		F	unction		•		Distrib	Long noncoding RN	A	IncRNA		Regulation of gene transcription, epigen	
Small nuclear RNA snRNA				Splicing and other functions				Eukary	yotes a	MicroRNA		miRNA Gene regulation		Gene regulation			
Small nucleolar RNA			noRNA		Nucleotide modification of RN			NAs	As Eukary		yotes a	Piwi-interacting RN/	Piwi-interacting RNA			Transposon defense, maybe other function	
SmY RNA		S	SmY		mRNA trans-splicing			Ner		Nema	todes	Small interfering RNA		siRNA		Gene regulation	
Small Cajal body-specific RNA		s	caRNA		Type of snoRNA: Nucleotide modification of RNAs				Parasitic RNAs								
		+		_			Тур	oe ∳	Functio	on ♦				Distrib	ution	•	
Guide RNA		9	JRNA mRNA		mRNA	nucleotide	Retrotransposon Self-		Eukaryotes and some bacteri		ia						
Ribonuclease P		F	RNase P tRN		tRNA m	naturation	propa		propagati								
Ribonuclease MRP		RNase MRP		rRNA m	naturation	Viral genome carrie		carrier	viruses, many satellite viruse		es, positive-sense RNA viruses, negative-sense RNA es and reverse transcribing viruses			NA viruses, negative-sense RNA cribing viruses			
Y RNA		RNA		RNA pr	ocessing,	Viroid		Self- propagating		Infected plants							
Telomerase RNA C	omponent	t T	TERC Telome		Telome	re synthe		Self-		_							
Spliced Leader RNA		S	SL RNA		mRNA trans-spli		Satellite	Satellite RNA propagatii		Infected cells							

Alde.	• Name •	Family a	Securiphen •	-
march da	man and ng KAA			
method.	non messanger RNA	-	agreenym of ref. 200	
-JEAA	small PDSA		Appendix provide and the second	
summittee.	small nen messanger RSG.	-	Approxyment of Apple 201	
UPDA.	Internalise HPark	NACCOUNTS IN		
-SEAA	redución POLA.	-	agreen you all 1974.	
mPDGA.	mensuringer RAA	-		
paRNA.	protein cooling Mb2A.	-	Apren production of a second	
-99.65	classes at PDA	- Aller		
				-
		and the second s		-
	the second second second second	Company of		-
1.00.000	A large tailored discound 824	CL IN LTOP		-
Sec. 10	maked a second line second a			-
RNA	annoxiating PDGA	REPORTED IN		
pRNA.	processiles PDQA.	NUMBER OF STREET	Aprendym of AbRC RNA	
42.855	ECE POND.	NUMBER OF STREET		
and real	·	REFECCEPTING	Aprendym of CE RNA	
aros.	arthurna RbsA	-	agreeness of an RAS	_
an POst.	arthurne MIGA			-
-anim POGA	arthurme miles PDOL	-		-
sin hitz	site rational and terms in amortal		Natural architecture instruction transmitted from the same generation boost an Unit General Instruction Operator DNA	
			situand and form perfect pairs	
serPOID.	CRIMER ROA			
Inscrimture A	Interval Software of State			
CRIEPE	CREEKS Car RNA	multiple		
00.000	DNA damage process With		83A but adjustes 25A days or reserve	
	DBB induced small Khole.			
shifthin	disable stranded RNA			
and a	makeness small interfering			
- POLA	POLA.	-		
maRDate.	mainteenfluitee 90% A			
Sharer	grade Khah	×		
In street	International interlation			
-	International second in the			
hers/POLA	POLO.	-	Approxyment from PAC	
			agranges for parent PAL (in the strict server, it may induction	
N-POQA	Interregenerate madeue PDUI.	-	mailear KNN interacipit. Ital da roll emiliap as cylioplaseria anteraci	
<u> </u>			Recently Realized With Indian Section 1	-
10.00	POLO. Interdomental	-	Introduction, by resultativing targetest cr870A materialies.	
Brock State	lang-integrats non-sailing PDO.			
Brand Parks	long ner coding MNA			
-	minute R545.	- tele		
		familian .		-
	milashanikal KNA presencing rilamadrater	-	April 1997 B. Allows 1982	
	mitual antiseme-sheet			
	 Interfering Khih. 		Station & Station	
	national antitionne school.		Endogenous MNA regulators, Schemens 21 and 24 of in	
	Interfering WAR		Inamaripis which are heller possessed into 1955.	
Cupit	and the share an array with the	warmon at	Recall exercising RNA which is induced in response to	
8555			and deliver stress in Bennelintie and	
piPELS.	plast interacting PDLN.	and the later	Zenali KNLAs that have MNA protein completers with pixel sectories in animal cells.	
10225.0	CEDE 2 Interdenting MINA			-
			Read WA had is incident in the WA interference.	-
-and PD-QA	Perpendi announcement viPOQA		pull-way, is type of piPDA.	
RNING	mitachenikid KNA presenting	NAMES OF A DESCRIPTION OF		
	riarmateurur			-
	The rest of the second se	and the second s	A 120 II CONTRACTOR BEACH MENT	-
the state	small saint MMA			-
- APRIL	small syleplasmic RNA			
	small considered SNA			
-		-	Read RNA that is and safed by Sig-S in Bachariable and	
		and the second s	sheing glessen phosphale sinne.	
-sheeter.	short hatyan RNA.			
-JROLA	small interliering PDVA.			-
31,955	spliced braker 80ch	And in the		
-			Zenall maleur RNAs lowed in same species of remainder	
		and the second s	warms, thought in the involvent in ePDA interaceplicing	
number.	small maileater RNA	- the		
		and the local division of the local division		
	small mailear 824A	(and)an		
-	small mailear chanadels			
	penderine.			
2010	and Z university states in the			
Instructure.	menanding 874A			
230 104	 signal recognition particle PDGA. 	CLEOCEM		
surgery a	single situation PDGA			
-SPOLA	small temporal PDOA	·		
SandPOrds.	hare, sailing s265pt.	a		
Increase.	Internal or mensuring or 105.5.	REPORTED IN	Residential PDA endersate with shall PDA the and menoamous PDA the opportunity.	
		-		
- 40.4	Li spilleronemai PibiA.	(and last		
WEAR	mod RNA		Approxyment of a Real	
with the	model MDSA	REPORTER IN		
210 MAA	X inacitive specific internation			
1.000	X 894	SPOOT ST	Comparents of the Roll Colomating states particle and comparents for DNA multiplier through interactions with	
			sitesenativ and initiation postern.	
8405	natural antiseme transmistr		Natural arthurne transmightermatical within a set that	
_			Name transmitty is surgitized with the allow RNA transmitty is	
	vice la RM	-	intervented single stated of metalenger ROLA	
	sectable PRof.	-	Restarts of a science second of	
	mallingg, single situation PDA		which form a RT mu2NA complete that anits as an inti-	1-10
			sheet artificain	

RNA abbreviations (res)

sitten and her PDM

Other RNAs

Type •	Abbr. •	Function •	Distribution •	Ref. •
Vault RNA	vRNA, vtRNA	Expulsion of xenobiotics (conjectured)		[24]
GlycoRNA	-	Unknown		



Motivation

• RNA function tied to structure



Motivation

- RNA function tied to structure
- Advancements in science



Motivation

- RNA function tied to structure
- Advancements in science
- Advancements in medicine





Nucleotides break down into three key elements

• Aromatic nitrogenous bases



- Aromatic nitrogenous bases
- Ribose sugar rings



- Aromatic nitrogenous bases
- Ribose sugar rings
- Phosphate groups



- Aromatic nitrogenous bases
- Ribose sugar rings
- Phosphate groups



- Aromatic nitrogenous bases
- Ribose sugar rings
- Phosphate groups











RNA is structurally determined by interactions

• Base pairing hydrogen bonds



- Base pairing hydrogen bonds
- Sugar-base stacking



- Base pairing hydrogen bonds
- Sugar-base stacking
- Phosphate-base hydrogen bonds



- Base pairing hydrogen bonds
- Sugar-base stacking
- Phosphate-base hydrogen bonds
- Water-formed hydrogen bonds



- Base pairing hydrogen bonds
- Sugar-base stacking
- Phosphate-base hydrogen bonds
- Water-formed hydrogen bonds





RNA 3D Structure Prediction



RNA 3D Structure Prediction

Early Attempts

- First began in 1960
- "in silico"



RNA 3D Structure Prediction

Early Attempts

- First began in 1960
- "in silico"
- First interactive tool 1998



Ab initio

- FARFAR
- iFoldRNA
- NAST
- SimRNA
- Vfold



Homology Modeling

- RNABuilder
- ModelRNA



Fragment-based Assembly

- MC-Fold/MC-Sym
- Assemble
- RNAComposer
- 3dRNA



Deep Learning

- Townshend et al.
- End-to-end prediction



RNA-Puzzles: Evaluating RNA Structure Predictions

- Launched in 2010
- 38 challenges
- 2 projects



RNA-Puzzles: Evaluating RNA Structure Predictions

- Compared against reference structure
- Several similarity metrics
 - Interaction Network Fidelity (INF)
 - INF-WC
 - INF-NWC
 - INF-stacking



12 Years of Challenges

- INF-WC: 0.75 1.0
- INF-NWC: ~ 0



CASP Recap

- CASP launched (1994)
- Fully automatic webserver prediction (2000)
- AlphaFold enters (2020)



Figure 3. Numbers of RNA and protein structure predictions made in RNA-Puzzles and CASP competitions. The solid lines represent the numbers of groups competing in CASP and RNA-Puzzles; the dashed lines are for the number of protein/RNA targets. From 2010 to 2021, RNAs were predicted only in RNA-Puzzles and in 2022, CASP included also RNA targets, which is responsible for the recent spike in targets and groups involved in 3D RNA structure prediction.



CASP-RNA

- RNA Puzzles web server (2015)
- Deep learning models (2022)
- CASP-RNA (2022)
 - 42 groups
 - 12 RNA sequences
 - 18 teams used DL models
 - DeepFoldRNA
 - RhoFold
 - trRosettaRNA
 - OpenComplexRNA
- Top 4 were non-ML





Table 1. Numbers of all PDB-released structures (*) and residues in X-ray and cryo-EM structures (**) with high resolution ($\leq 2.0 \text{ Å}$) over decades. In the first column, amino acids are abbreviated as AAs, and nucleotides as nts

	≤1980	1981-1990	1991-2000	2001-2010	2011-2022	Total	% of the total
Proteins (*)	78	634	12 121	43 205	108 677	164 715	91.57
AAs ≤2.0 Å (**)	5050	45 236	1 609 401	11 390 238	28 513 777	41 563 702	99.78
RNA (*)	2	23	306	1392	4488	6211	3.45
RNA nts ≤2.0 Å (**)	0	0	1270	5974	26 921	34 165	0.08
DNA (*)	1	91	1061	2009	5800	8962	4.98
DNA nts ≤2.0 Å	0	238	5430	15 730	38 107	59 505	0.14

Complexity of RNA Architecture

- A-form antiparallel double helix is the easiest
- RNA junctions are complex and challenging
- Scarce data for non-Watson-Crick base pairs
- Inter-/Intra-molecular interactions



Limited Sequence and Alignment Data



Figure 5. Rfam versus Pfam alignments compared based on (A) a number of sequences, (B) a number of columns and (C) the average pairwise percent identity for each family. The points on the plots indicate the mean, and the vertical bars indicate the standard deviation.

1) Rfam construction is expensive



Figure 5. Rfam versus Pfam alignments compared based on (A) a number of sequences, (B) a number of columns and (C) the average pairwise percent identity for each family. The points on the plots indicate the mean, and the vertical bars indicate the standard deviation.

Rfam construction is expensive
 RNA alignments are smaller



Figure 5. Rfam versus Pfam alignments compared based on (A) a number of sequences, (B) a number of columns and (C) the average pairwise percent identity for each family. The points on the plots indicate the mean, and the vertical bars indicate the standard deviation.

- 1) Rfam construction is expensive
- 2) RNA alignments are smaller
- 3) Rfam has global biases



Figure 6. Counts of Rfam families, seed sequences, full sequences and structures for all Rfam families organized by Rfam RNA type.

- 1) Rfam construction is expensive
- 2) RNA alignments are smaller
- 3) Rfam has global biases
- 4) Inconsistency with NWC base pairs



- 1) Rfam construction is expensive
- 2) RNA alignments are smaller
- 3) Rfam has global biases
- 4) Inconsistency with NWC base pairs
- 5) Secondary structure representation



- 1) Rfam construction is expensive
- 2) RNA alignments are smaller
- 3) Rfam has global biases
- 4) Inconsistency with NWC base pairs
- 5) Secondary structure representation
- 6) Lack of pseudoknots





1) Rfam construction is expensive

2) RNA alignments are smaller

3) Rfam has global biases

4) Inconsistency with NWC base pairs

5) Secondary structure representation

6) Lack of pseudoknots

The Future RNA Prediction Landscape

Speculations

- AlphaFold took 50 years
- AlphaFold trained on 170,000 proteins
- Pfam has 19,632 alignments
- Rfam will have 19,000 in 70 years

Improvements

- RNA motifs and global structures
- More research on molecular interactions
- Improved benchmarking
- Improved MSA
- Consistent annotation of RNA families
- Incorporation of new technologies



Final Call to Action

DDD

NNNNN

Thank You!

Questions?