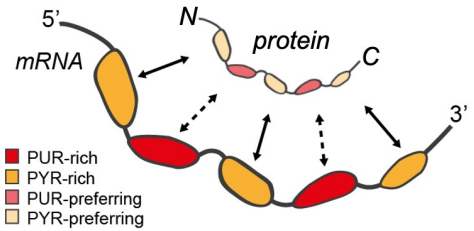


Towards physicochemical bioinformatics

Bojan Žagrović
Max Perutz Labs & University of Vienna

Computational Chemistry Days
IRB Zagreb, May 9th 2026

MAX PERUTZ LABS VIENNA

universität wien

Biomolecular sequences analysis: a foundation of modern molecular biology

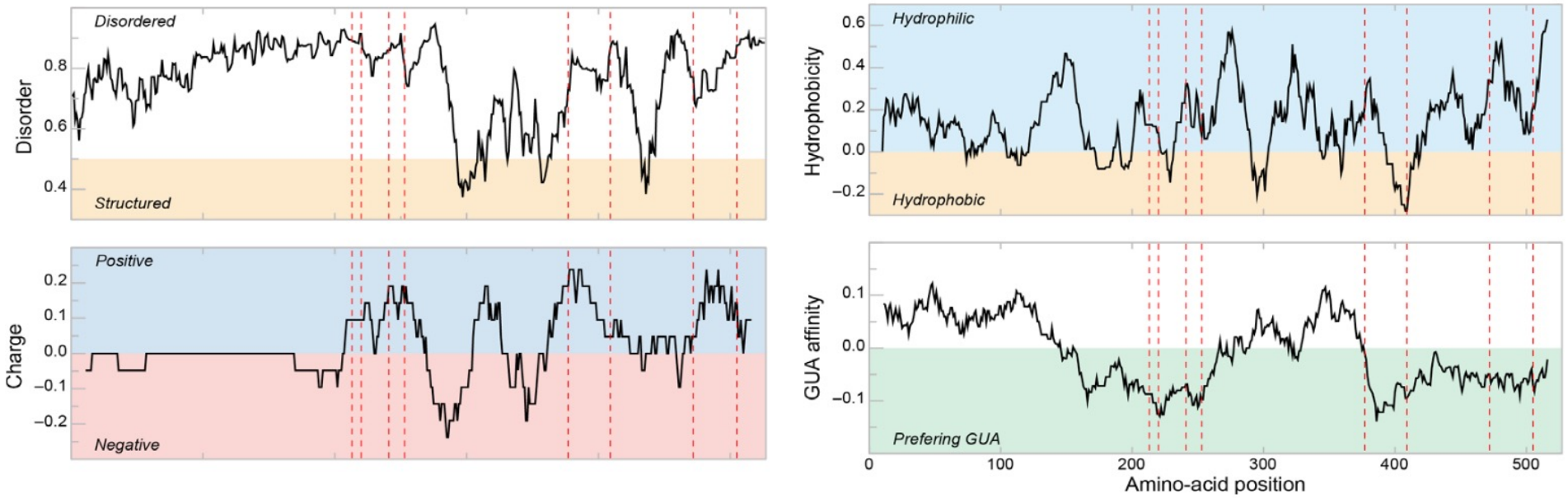
FUS (human)

MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSYGQSQNTGYGTQSTPQGYGS
TGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSSQSSSYGQPQSGSYSQQPSYGGQQQSYGQQQSYNPPQGY
GQQNQYNSSSGGGGGGGGGGNYGQDQSSMSSGGGSGGGYGNQDQSGGGGSGGYGQQDRGGRGRGGSGGGGG
GGGGGYNRSSGGYEPRGRGGGRGGRGGMGGSDRGGFNKFGGPRDQGSRHDSEQDNSDNNTIFVQGLGENVTIESV
ADYFKQIGIIKTNKKTGQPMINLYTDRETGKLKGEATVSFDDPPSAKAAIDWFDGKEFSGNPIKVSFATRRADFNRGGGNG
RGGRGRGGPMGRGGYGGGGSGGGGRGGFPSGGGGGGGQQRAGDWKCPNPTCENMNFSWRNECNQCKAPKPDG
PGGGPGGSHMGGNYGDDRRGGRGGYDRGGYRGRGGDRGGFRGGRGGGDRGGFGPGKMDSRGEHRQDRRERPY

Biomolecular sequences analysis: a foundation of modern molecular biology

FUS (human)

MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSYGQSQNTGYGTQSTPQGYGS
TGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSSQSSSYGQPQSGSYSQQPSYGGQQQSYGQQQSYNPPQGY
GQQNQYNSSSGGGGGGGGGGNYGQDQSSMSSGGGSGGGYGNQDQSGGGGSGGYGQQDRGGRGRGGSGGGGG
GGGGGYNRSSGGYEPRGRGGGRGGRGGMGGSDRGGFNKFGGPRDQGSRHDSEQDNSDNNTIFVQGLGENVTIESV
ADYFKQIGIIKTNKKTGQPMINLYTDRETGKLKGEATVSFDDPPSAKAAIDWFDGKEFSGNPIKVSFATRRADFNRGGGNG
RGGRGRGGPMGRGGYGGGGSGGGGRGGFPSGGGGGGGQQRAGDWKCPNPTCENMNFSWRNECNQCKAPKPDG
PGGGPGGSHMGGNYGDDRRGGRGGYDRGGYRGRGGDRGGFRGGRGGGDRGGFGPGKMDSRGEHRQDRRERPY



understanding biomolecular sequences as physicochemical objects

Example 1

Frameshifts: major impact on reading of genetic message

mRNA

```
atggggaaat tgatcaggat ggggccgcaa gagaggtggt tactccggac
aaagcggctt cattggagtc gcctcctctt cttactggga atgttgatca
tcggttctac ttatcagcac cttaggagac cccggggcct ttcctcattg
```

protein

wt ^{1}MGKLIRMGPQERWLLRTKRLHWSRLLFLLGMLIIGSTYQHLRRPRGLSSL^{50}

+1 ^{1}WGN-SGWGRKRGGYSGQSGFIGVASSSYWEC-SSVLLISTLGDPGAFPHC^{50}

-1 ^{1}GEIDQDGAAREVVTPDKAASLESPPLLTGNVDHRFYLSAP-ETPGPFLIV^{50}

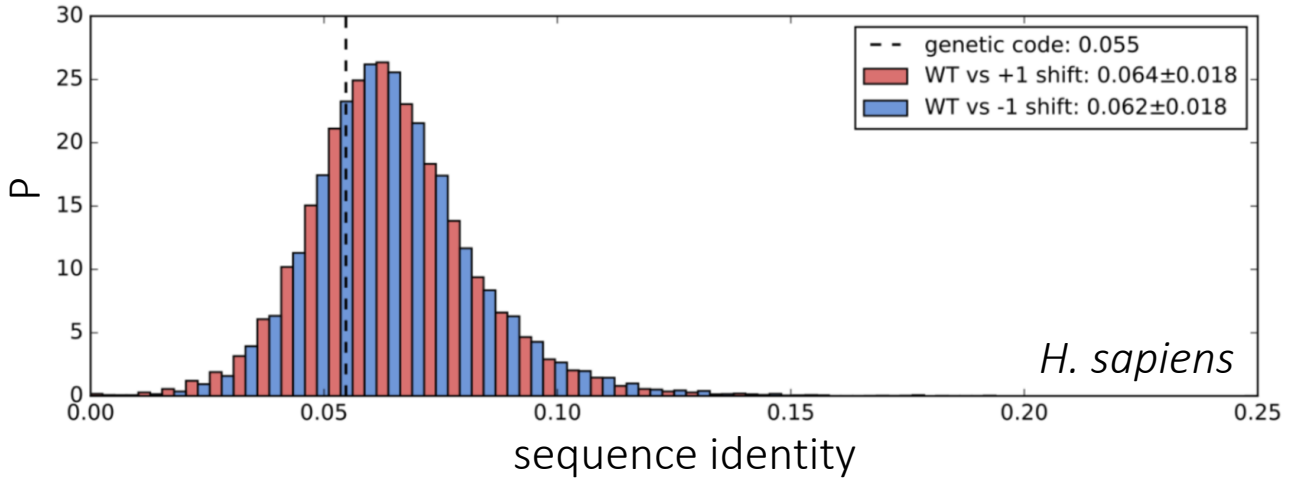
Frameshifts: major impact on reading of genetic message

mRNA

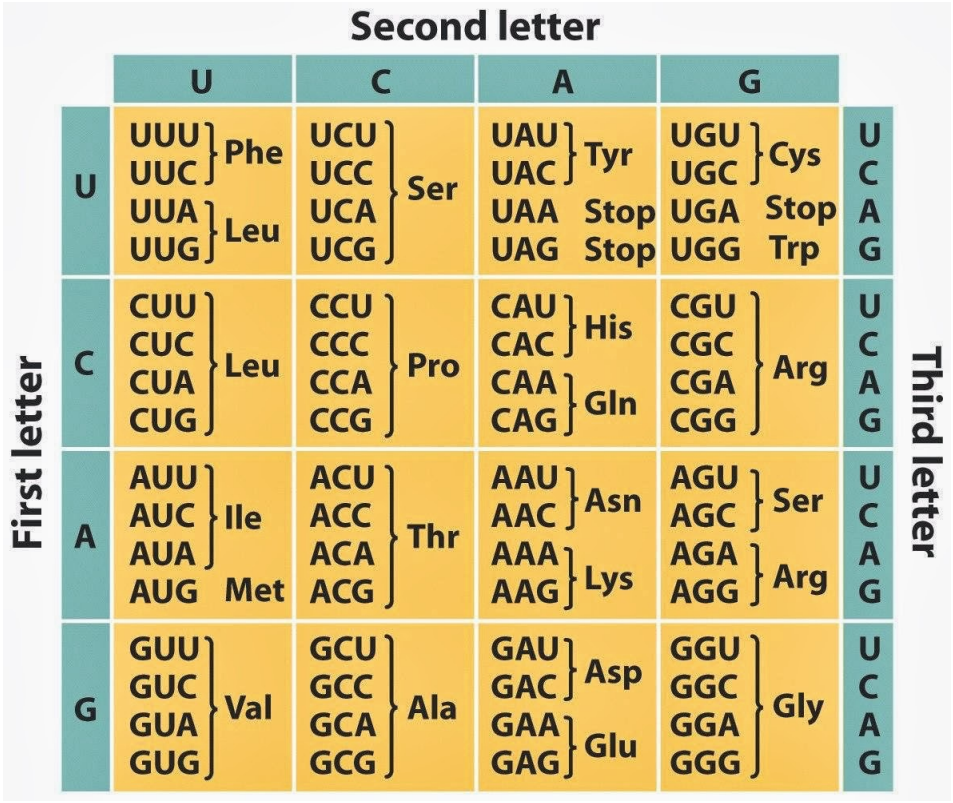
```
atggggaaat tgatcaggat ggggccgcaa gagaggtggt tactccggac
aaagcggctt cattggagtc gcctcctctt cttactggga atgttgatca
tcggttctac ttatcagcac cttaggagac cccggggcct ttcctcattg
```

protein

```
wt  1MGKLIRMGPQERWLLRTKRLHWSRLLFLLGMLIIGSTYQHLRRPRGLSSL50
+1  1WGN-SGWGRKRGGYSGQSGFIGVASSSYWEC-SSVLLISTLGDPGAFPHC50
-1  1GEIDQDGAARE VVTPDKAASLESPPLLTGNVDHRFYLSAP-ETPGPFLIV50
```



How related are amino acids encoded by frameshifted codons?

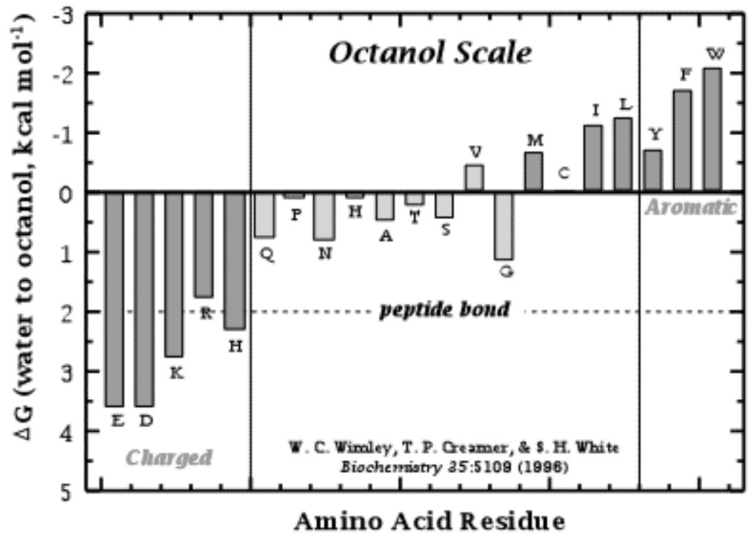


First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; UAA: Stop; UAG: Stop	UGU, UGC: Cys; UGA: Stop; UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile; AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

e.g. **Phe** → Phe, Leu, Ile, Val, Ser

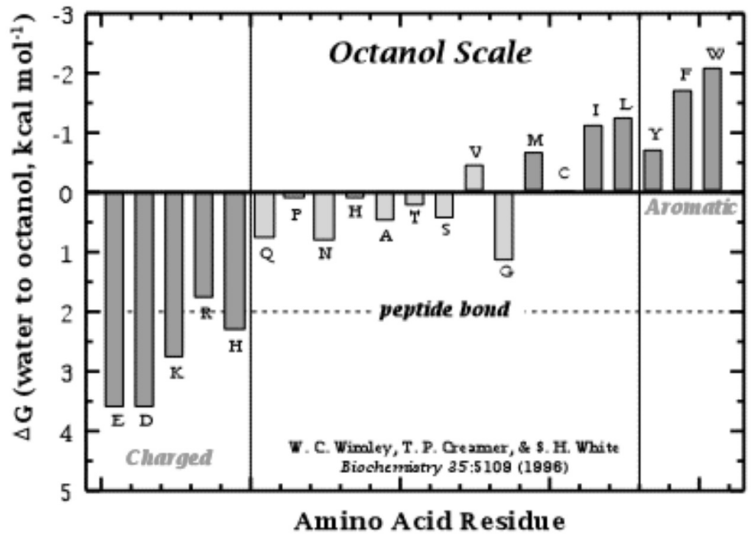
>600 different amino-acid properties (Kawashima et al., **NAR**, D202, 2008)

e.g.



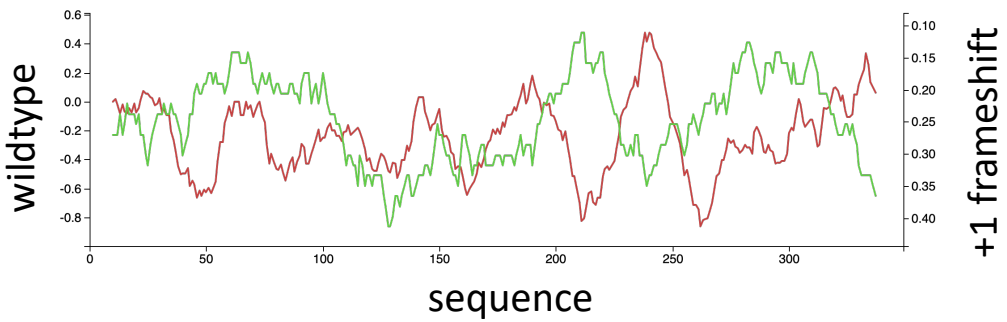
>600 different amino-acid properties (Kawashima et al., **NAR**, D202, 2008)

e.g.

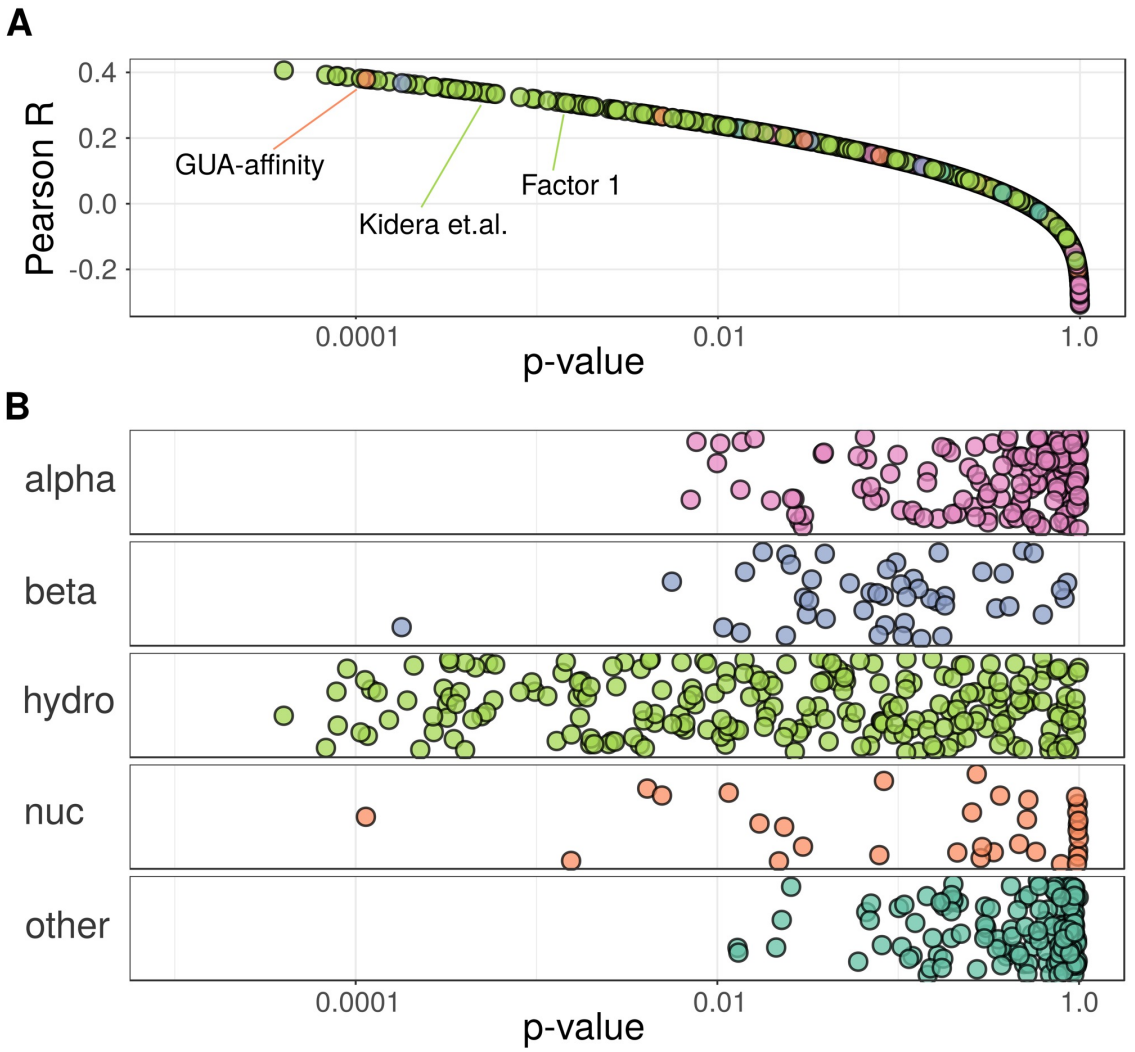


1. genetic code

2. sequence profiles

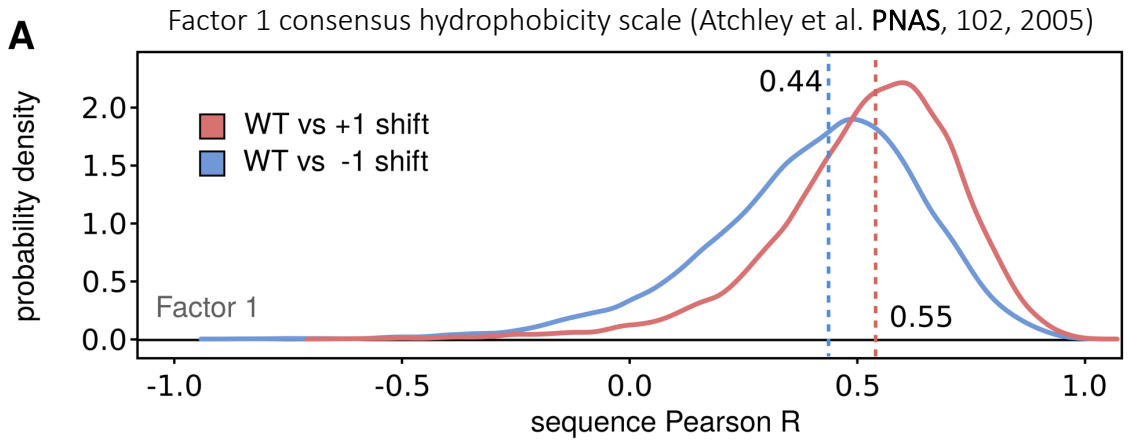


Impact of frameshifting at the level of genetic code

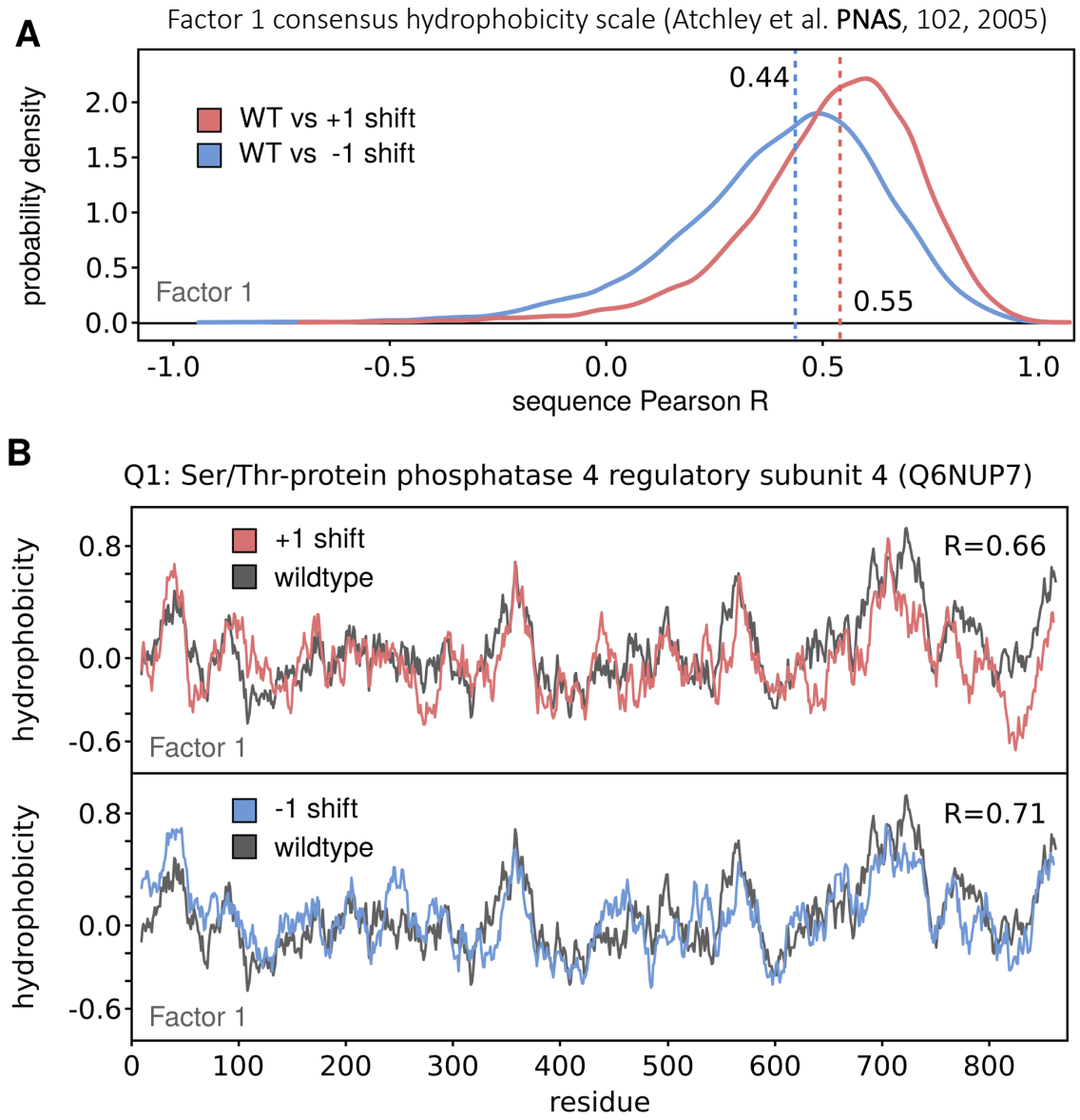


Hydrophobicity is significantly preserved upon frameshifting at the genetic code level

Sequence hydrophobicity profiles are robust against frameshifting

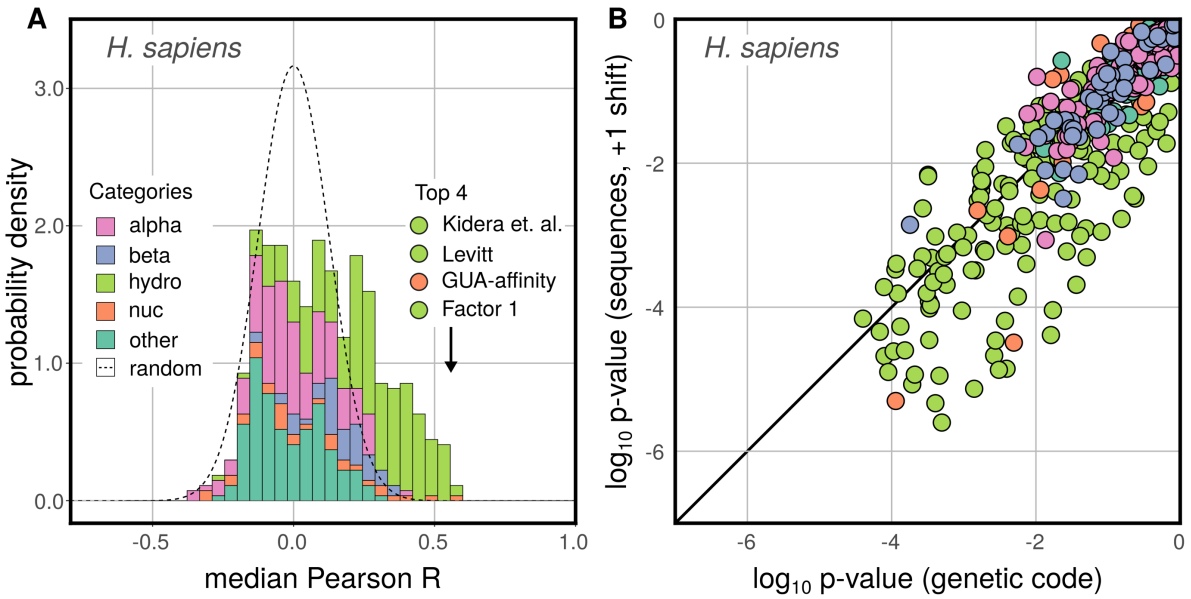


Sequence hydrophobicity profiles are robust against frameshifting

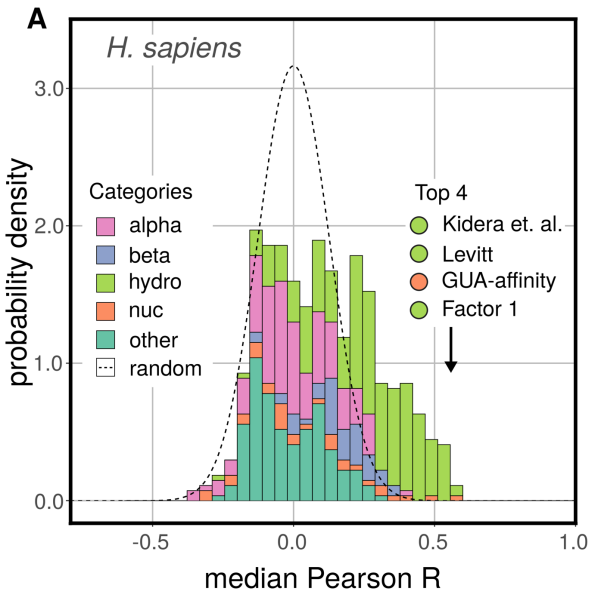


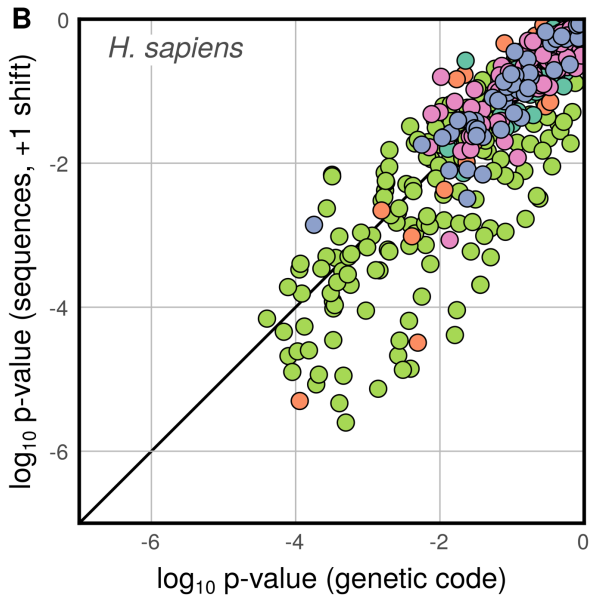
1st quartile

Sequence hydrophobicity profiles are robust against frameshifting



Sequence hydrophobicity profiles are robust against frameshifting





top 25%

C

GO enrichment for Factor 1

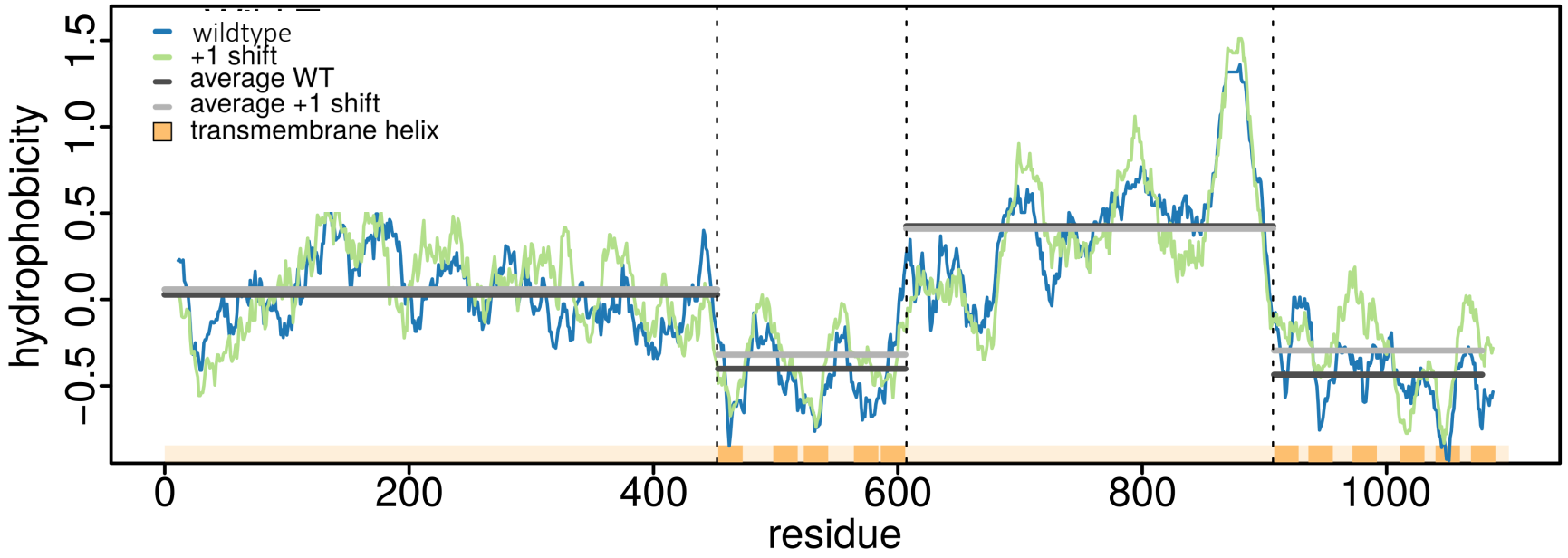
+1 shift	p-value
integral component of membrane	1.1×10^{-26}
keratin filament	2.2×10^{-4}
endoplasmatic reticulum	2.8×10^{-3}
nucleolus	3.8×10^{-3}

-1 shift	
nucleolus	2.0×10^{-9}
nucleoplasm	1.7×10^{-7}
ribonucleoprotein complex	5.8×10^{-5}
chromosome	1.4×10^{-4}
intracellular non-membrane organelle	2.2×10^{-3}

An example of a membrane protein

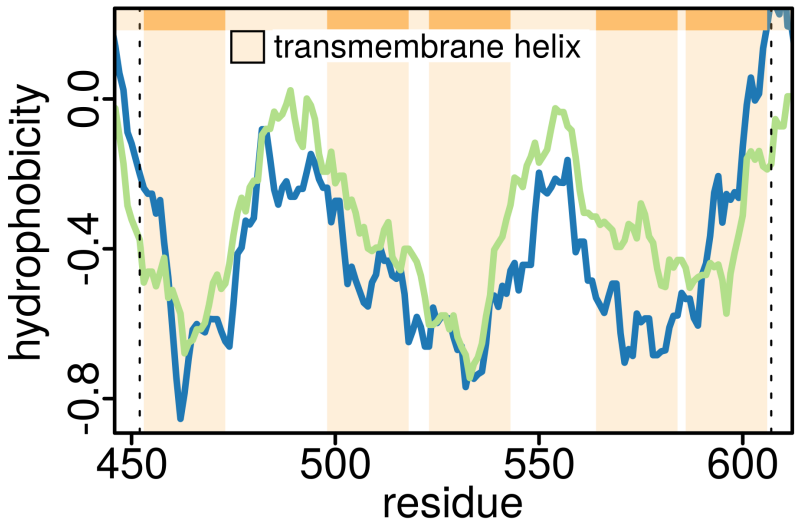
sodium/potassium/calcium exchanger 1 (O60721)

```
wt     1 MGKLIRMGPQERWLLRTKRLHWSRLLFLLGMLIIGSTYQHLRRPRGLSSL 50
+1     1 WGN-SGWGRKRGGYSGQSGFIGVASSSYWEC-SSVLLISTLGDPGAFPHC 50
                                 .....
```

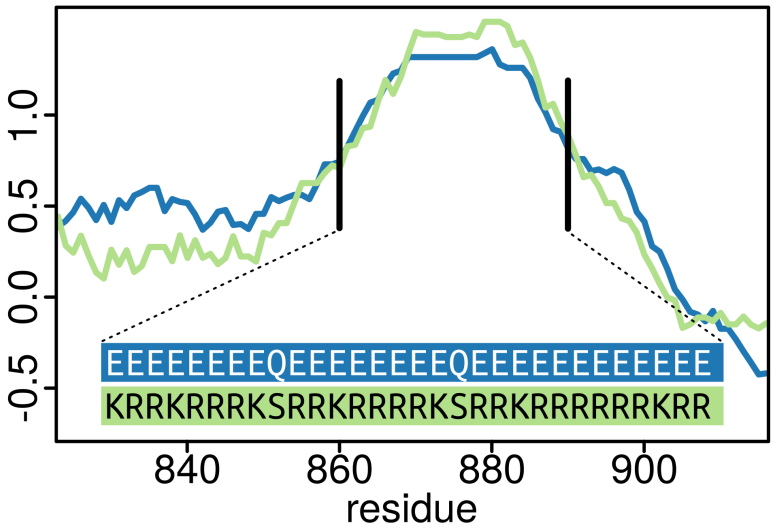


Factor 1 consensus hydrophobicity scale (Atchley et al. **PNAS**, 102, 2005)

sodium/potassium/calcium exchanger 1 (O60721)

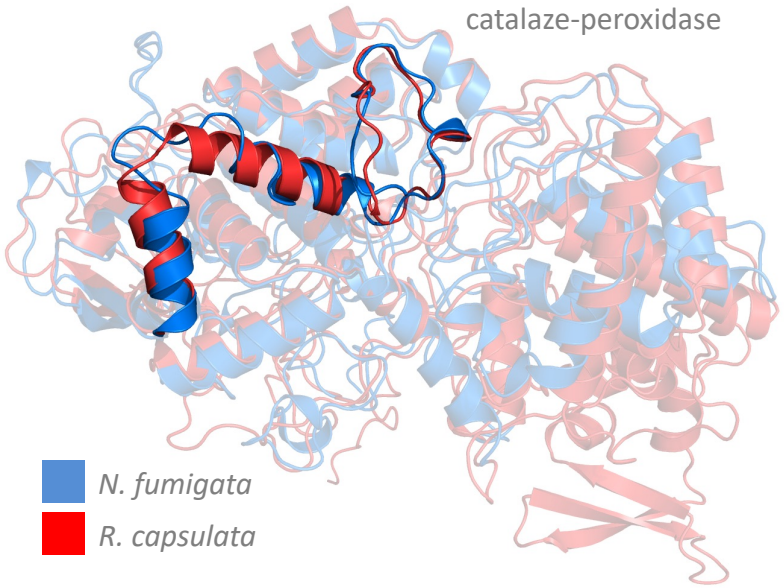


retained molecular topology



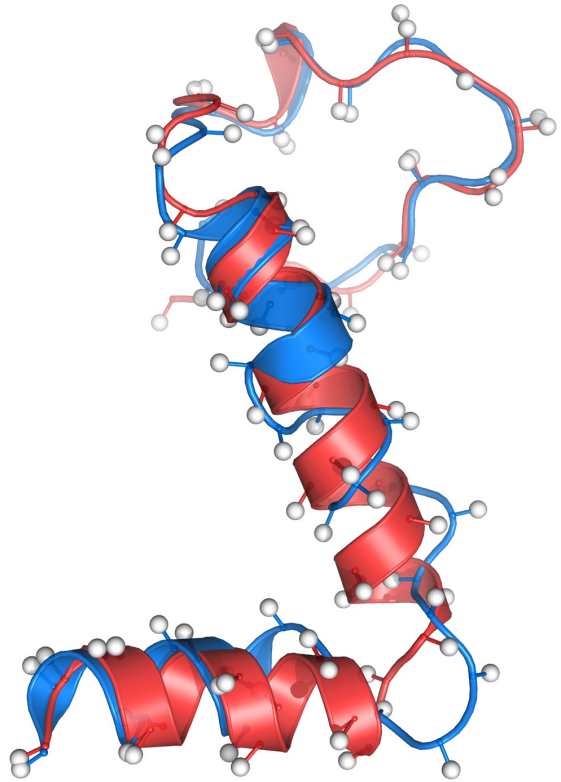
retained hydrophobicity, but charge inversion

Frameshifted regions could be structurally related




```
PAYEKIARRFLEHPDQFADAFARAWFKLTHRDMGPRARYLGPEVPSEVLI
HHHHHHHHHHHHH-HHHHHHHHHHHHHHHHHHTTS-SGGG-BSTT--SS--G
HHHHHHHHHH--SS--TTHHHHSHHHHHHHHT---GGG--STT--S-SSS
PSTMRSVRSSWPIRPPSTTLSRAPGSSCCIATWGRRRATSAPMCPPRIWS
```

sequence identity = 12%

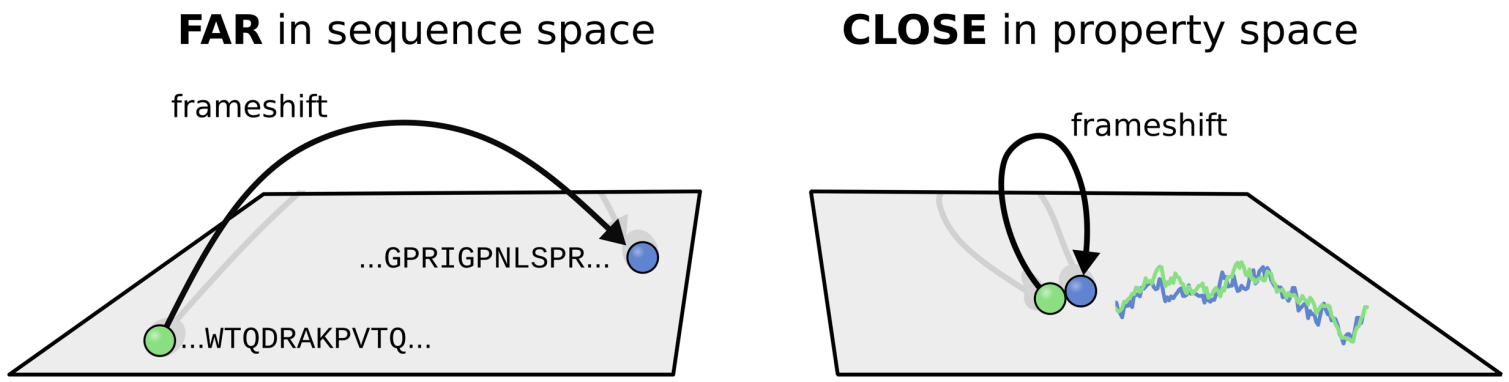


backbone RMSD = 2.8 Å

Marlene Adlhart, Daniel Braun, Lena Reitinger & Bojan Žagrović (unpublished)

Conclusion 1

- different protein sequence properties e.g. hydrophobicity, structural disorder and nucleobase affinity are significantly robust against frameshifting
- implications: overprinted genes; natural frameshifts; stop-codon reassignment; gain-of-function
- a potentially powerful way for evolution of novel sequences from optimized starting points



Lukas Bartonek*, Daniel Braun* & Bojan Zagrovic, **PNAS**, 117, 2020

Example 2

Stereochemical hypothesis of the origin of the genetic code

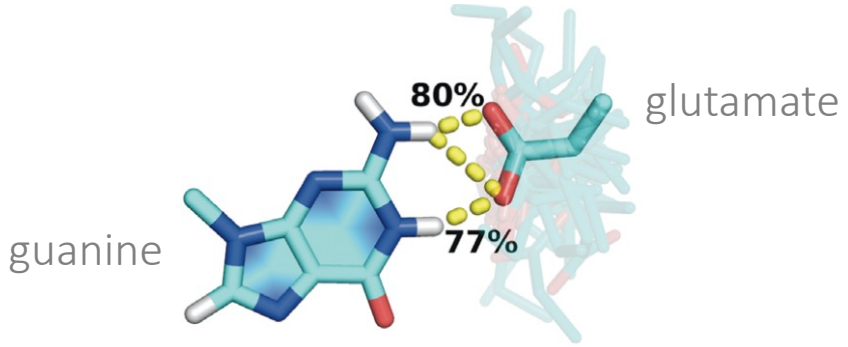
Gamow, Woese, Yarus and others …

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

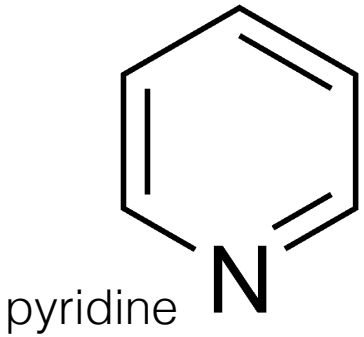
universal genetic code

⟺

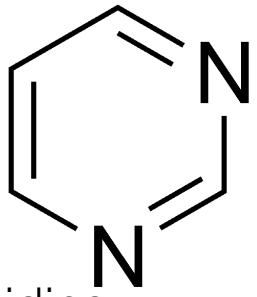
nucleotide/amino-acid affinities



What is the affinity of amino acids for pyrimidines?

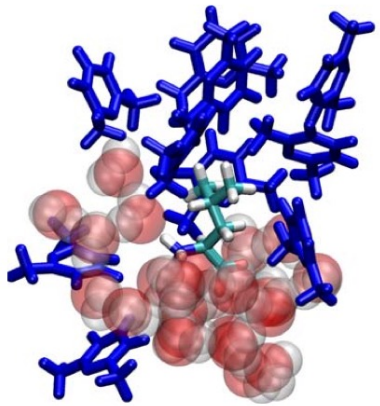


pyridine

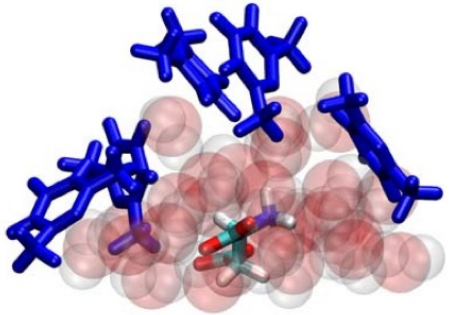


pyrimidine

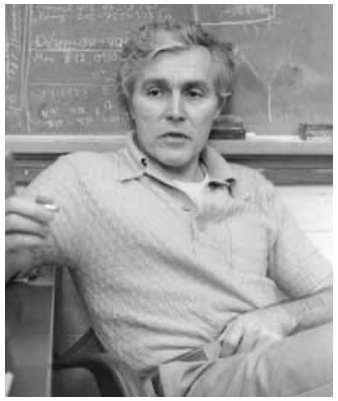
$$\Delta\Delta G \sim -\log(C_{DMP}/C_{water})$$



LEU in DMP:H_2O



ASP in DMP:H_2O

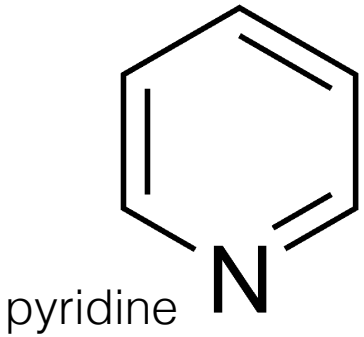


Carl Woese

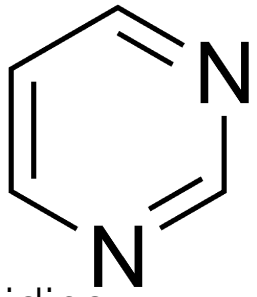
*DMP: 2,6-dimethylpyridine

Woese CR et al., **PNAS**, 55, 1966
Mathew DC & Luthey-Schulten Z, **J Mol Evol**, 66, 2008

What is the affinity of amino acids for pyrimidines?

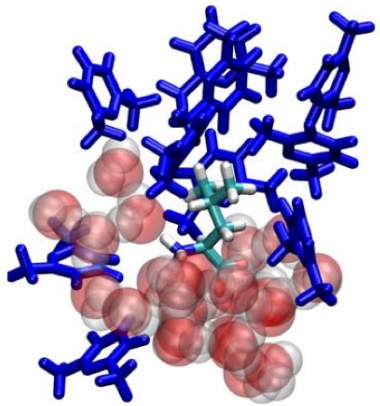


pyridine

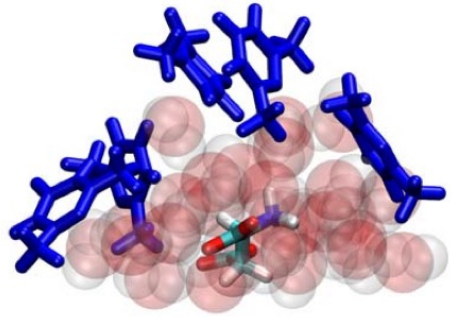


pyrimidine

$$\Delta\Delta G \sim -\log(C_{DMP}/C_{water})$$

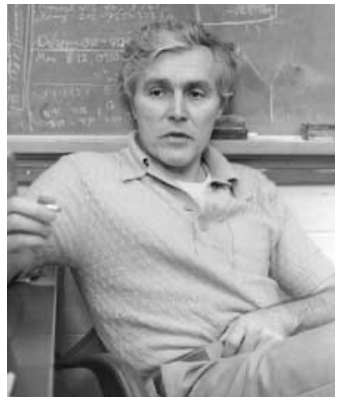


LEU in DMP:H_2O



ASP in DMP:H_2O

*DMP: 2,6-dimethylpyridine



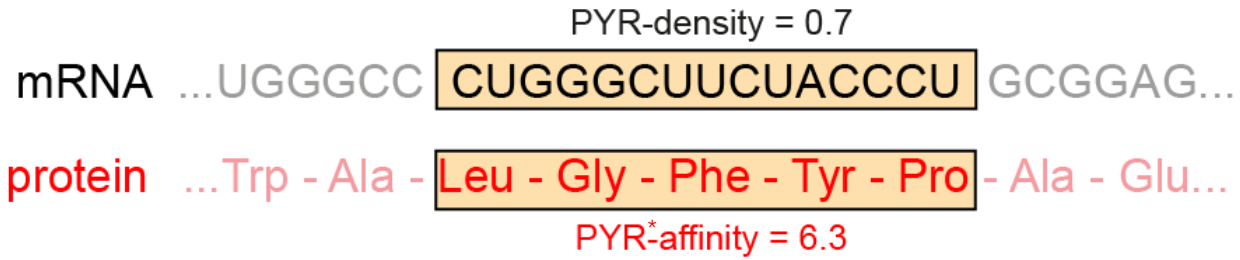
Carl Woese

C	L	F	W	I	M	P	V	T	A	S	Y	H	R	Q	G	N	K	D	E
4.3	4.4	4.5	4.9	5.0	5.0	6.1	6.2	6.2	6.5	7.5	7.7	7.9	8.6	8.9	9.0	9.6	10.2	12.2	13.6

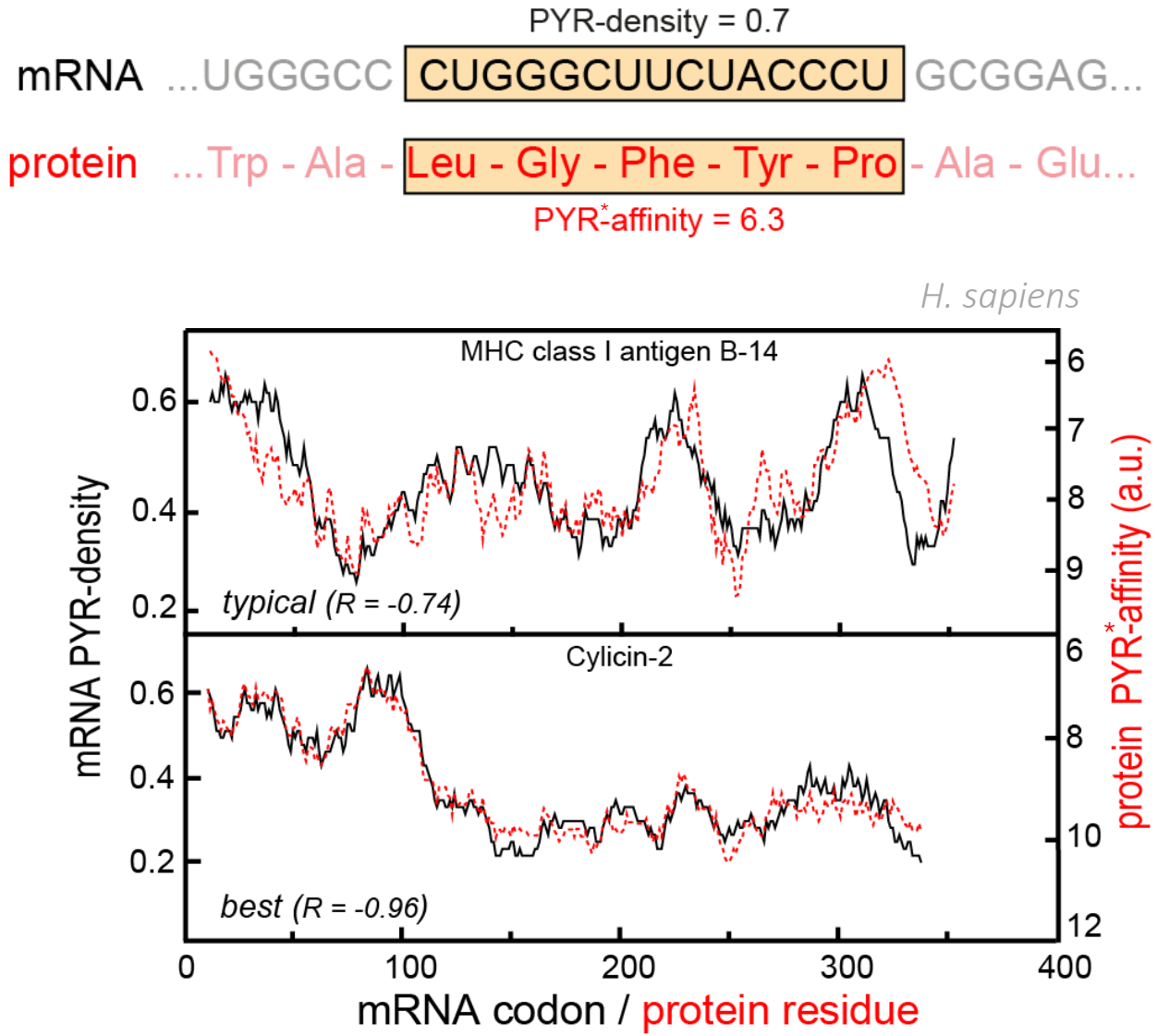
a. u.

Woese CR et al., **PNAS**, 55, 1966

Mathew DC & Luthey-Schulten Z, **J Mol Evol**, 66, 2008



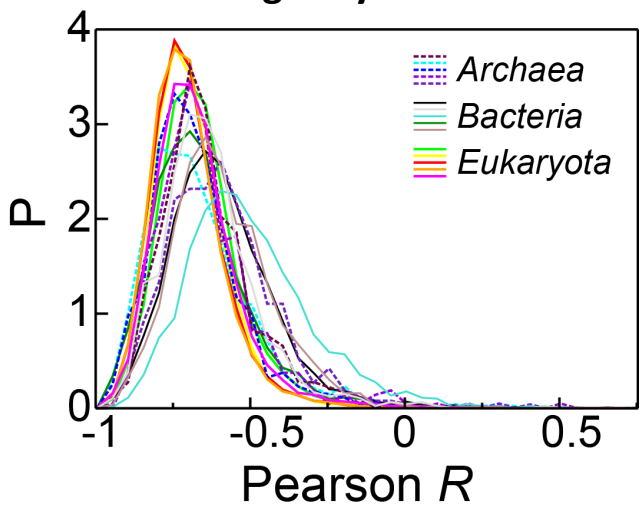
Hlevnjak et al. **NAR**, 40, 2012; Polyansky et al. **NAR**, 2013; 2012; Zagrovic et al. **FEBS Letters**, 592, 2018; Zagrovic et al. **Ann. Rev. Biophys.** 52, 2023



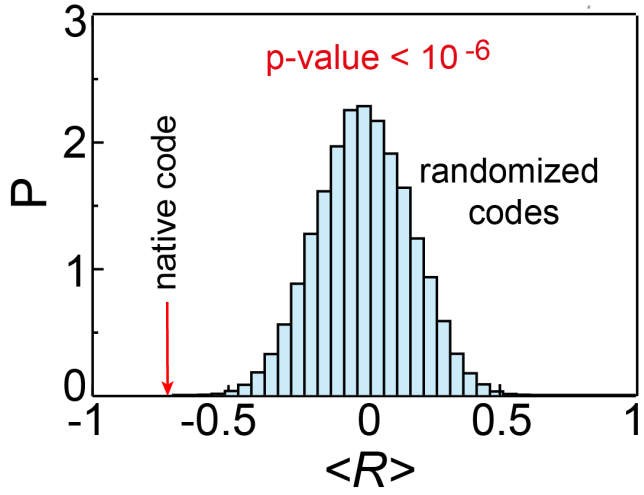
Hlevnjak et al. **NAR**, 40, 2012; Polyansky et al. **NAR**, 2013; 2012; Zagrovic et al. **FEBS Letters**, 592, 2018; Zagrovic et al. **Ann. Rev. Biophys.** 52, 2023

A robust and reproducible result of general validity

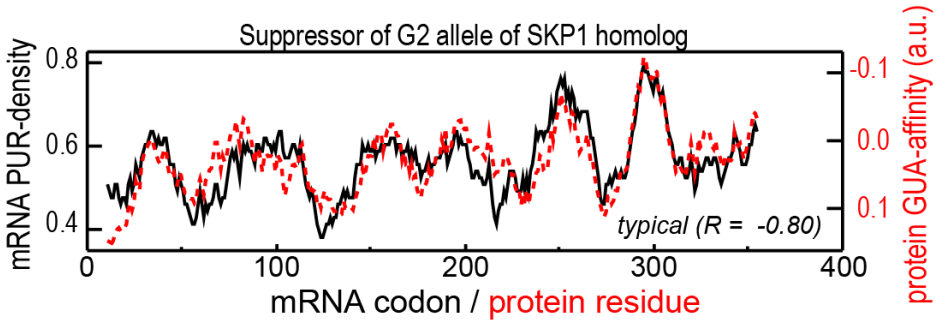
biologically universal



statistically significant



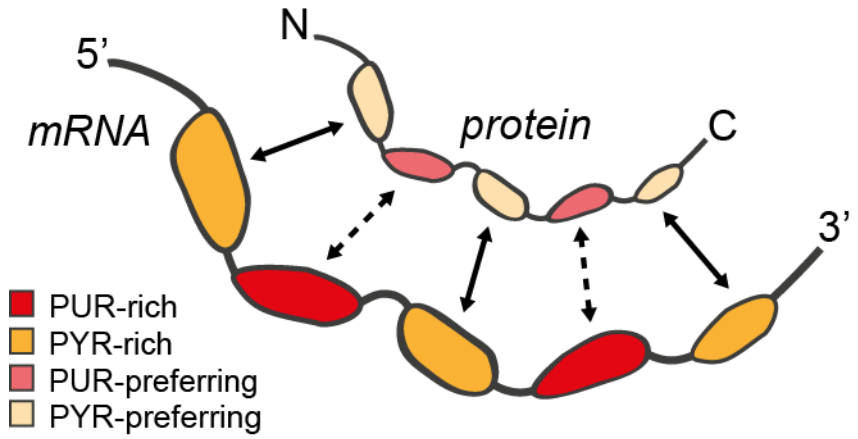
confirmed with multiple affinity scales



Hlevnjak, Polyansky & Zagrovic, **NAR**, 40, 2012; Polyansky & Zagrovic, **NAR,** 41, 2013; de Ruiter & Zagrovic, **NAR**, 43, 2015; Hlevnjak & Zagrovic, **NAR**, 43, 2015; Bartonek & Zagrovic, **PLOS CB**, 13, 2017; Zagrovic, Bartonek & Polyansky, **FEBS Letters**, 592, 2018; Bartonek, Braun & Zagrovic, **PNAS**, 117, 2020; Kapral, Farnhammer, Zhao, Lu & Zagrovic, **NAR**, 50, 2022; Zagrovic, Adlhart & Kapral, **Ann Rev Biophys**, 52, 2023; Adlhart, Hoffmann, Polyansky & Zagrovic, **PNAS**, 122, 2025

CENTRAL HYPOTHESIS

mRNA CDS and their autogenous proteins are **complementary to each other and bind in a co-aligned fashion**, especially if unstructured, reflecting the driving forces behind the origin of the genetic code. Complementarity is negatively regulated by the mRNA adenine content.



Comparison with experiment

- crosslinking and immunoprecipitation (CLIP-seq) informs on RNA binding targets of individual proteins
- 341 different RNA-binding proteins (RBPs) tested so far

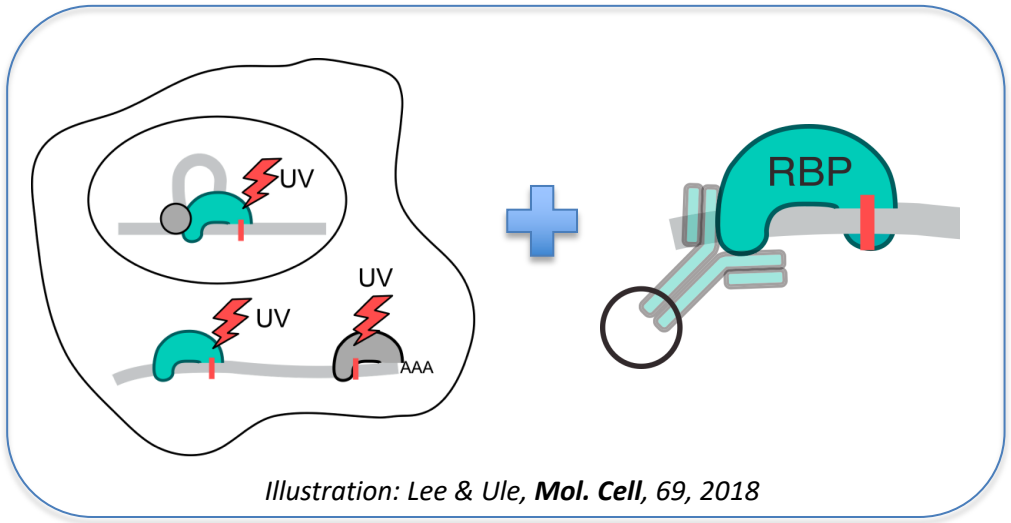


Illustration: Lee & Ule, ***Mol. Cell****, 69, 2018*

POSTAR3 Modules Submit Help About Contact

RBP binding sites of TAF15

Export data to CSV file Search:

Target gene symbol	Target gene ID	Target gene type (processed_transcript)	Protocol	Target gene exp. level	Binding site records
DLEU1	ENSG00000176124	processed_transcript	eCLIP	display	154
CASC19	ENSG00000254166	processed_transcript	eCLIP	display	107
SNHG1	ENSG00000255717	processed_transcript	PAR-CLIP,PARalyzer	display	98
GAS5	ENSG00000234741	processed_transcript	PAR-CLIP,Piranha_0.01	display	58
SNHG14	ENSG00000224078	processed_transcript	HITS-CLIP,Piranha_0.01	display	53
PROX1-AS1	ENSG00000230461	processed_transcript	PAR-CLIP,Piranha_0.01	display	38
GAS5	ENSG00000234741	processed_transcript	PAR-CLIP,PARalyzer	display	36
SNHG1	ENSG00000255717	processed_transcript	PAR-CLIP,Piranha_0.01	display	35
LRRC75A-AS1	ENSG00000175061	processed_transcript	PAR-CLIP,PARalyzer	display	34
DLEU1	ENSG00000176124	processed_transcript	PAR-CLIP,PARalyzer	display	25

Showing 1 to 10 of 143 entries (filtered from 13,176 total entries) Previous 1 2 3 4 5 ... 15 Next

Jump to page View

Comparison with experiment

- crosslinking and immunoprecipitation (CLIP-seq) informs on RNA binding targets of individual proteins
- 341 different RNA-binding proteins (RBPs) tested so far

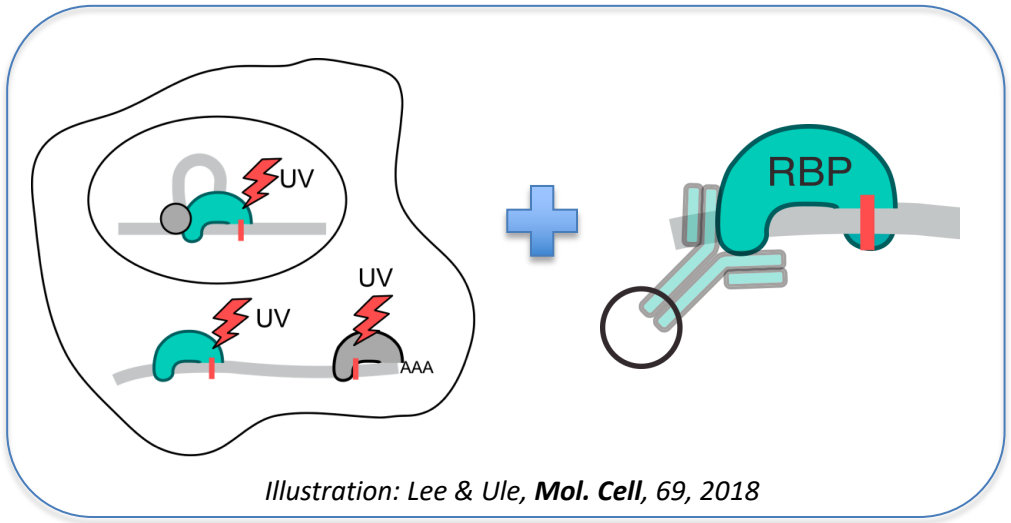


*Illustration: Lee & Ule, **Mol. Cell**, 69, 2018*

POSTAR3 Modules Submit Help About Contact

RBP binding sites of TAF15

Export data to CSV file Search:

Target gene symbol	Target gene ID	Target gene type (processed_transcript)	Protocol	Target gene exp. level	Binding site records
DLEU1	ENSG00000176124	processed_transcript	eCLIP	display	154
CASC19	ENSG00000254166	processed_transcript	eCLIP	display	107
SNHG1	ENSG00000255717	processed_transcript	PAR-CLIP,PARalyzer	display	98
GAS5	ENSG00000234741	processed_transcript	PAR-CLIP,Piranha_0.01	display	58
SNHG14	ENSG00000224078	processed_transcript	HITS-CLIP,Piranha_0.01	display	53
PROX1-AS1	ENSG00000230461	processed_transcript	PAR-CLIP,Piranha_0.01	display	38
GAS5	ENSG00000234741	processed_transcript	PAR-CLIP,PARalyzer	display	36
SNHG1	ENSG00000255717	processed_transcript	PAR-CLIP,Piranha_0.01	display	35
LRRC75A-AS1	ENSG00000175061	processed_transcript	PAR-CLIP,PARalyzer	display	34
DLEU1	ENSG00000176124	processed_transcript	PAR-CLIP,PARalyzer	display	25

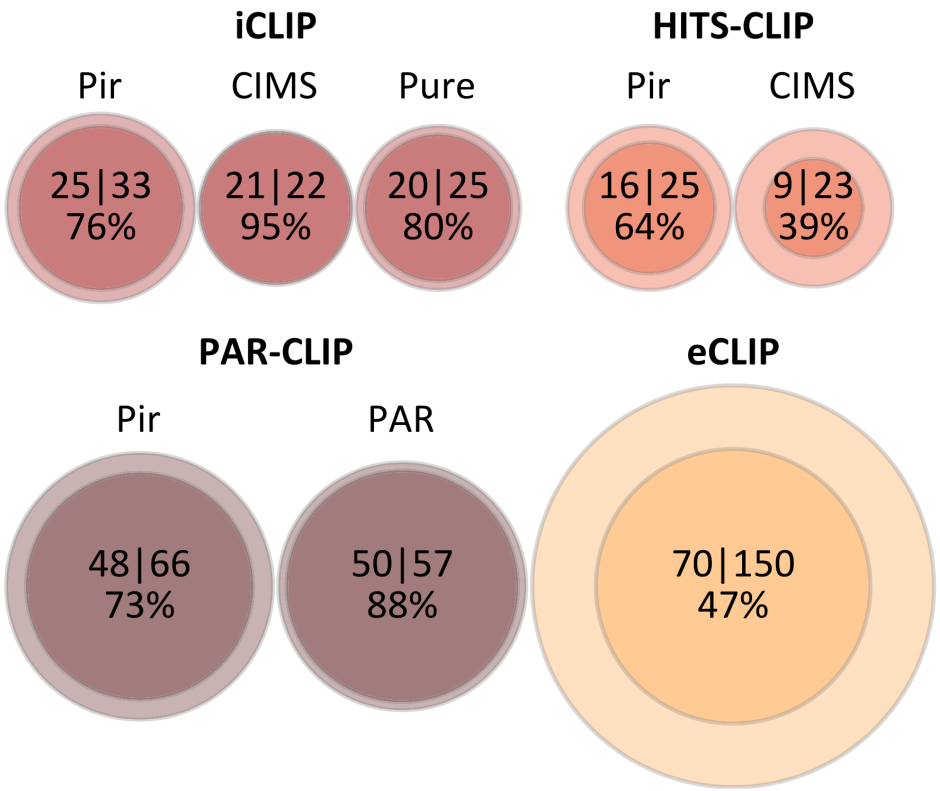
Showing 1 to 10 of 143 entries (filtered from 13,176 total entries) Previous 1 2 3 4 5 ... 15 Next

Jump to page View

Two key predictions:

1. autogenous mRNA/protein interactions are enriched over background
2. autogenous interactions occur preferentially in the coding sequence (CDS)

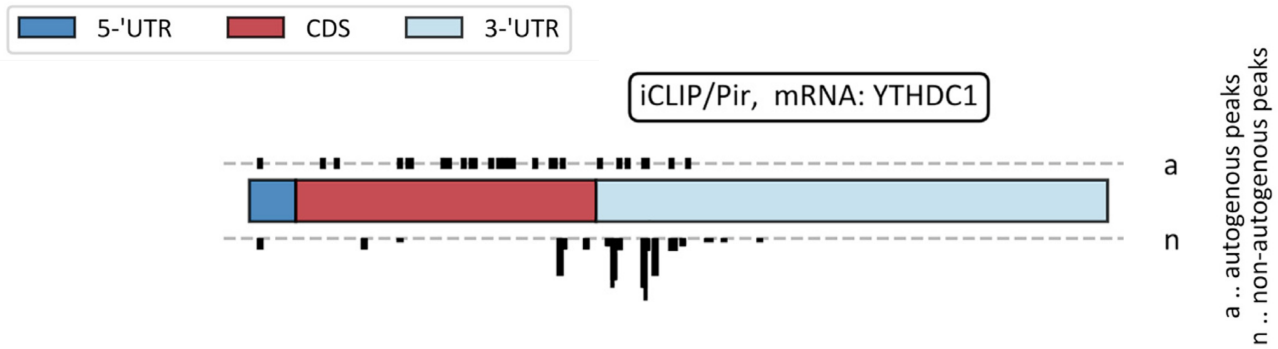
Autogenous mRNA-protein interactions are seen frequently in CLIP-seq experiments

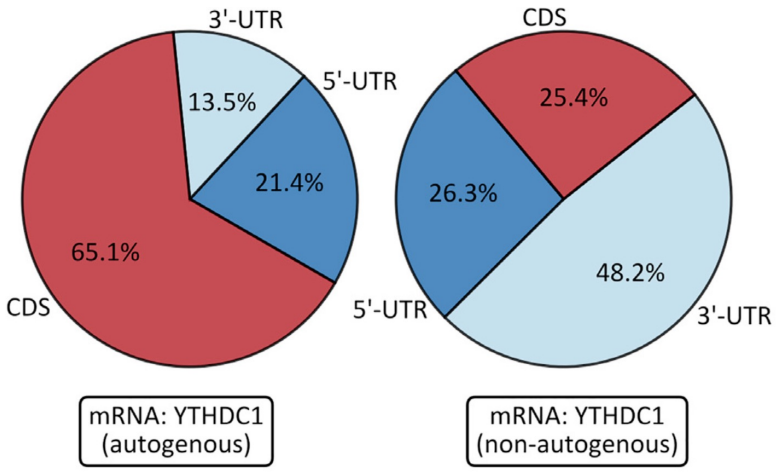


auto binding:
230/341 (67%)

peak callers: Pir- Piranha, PAR – Paralyzer, CIMS, CLIPPER

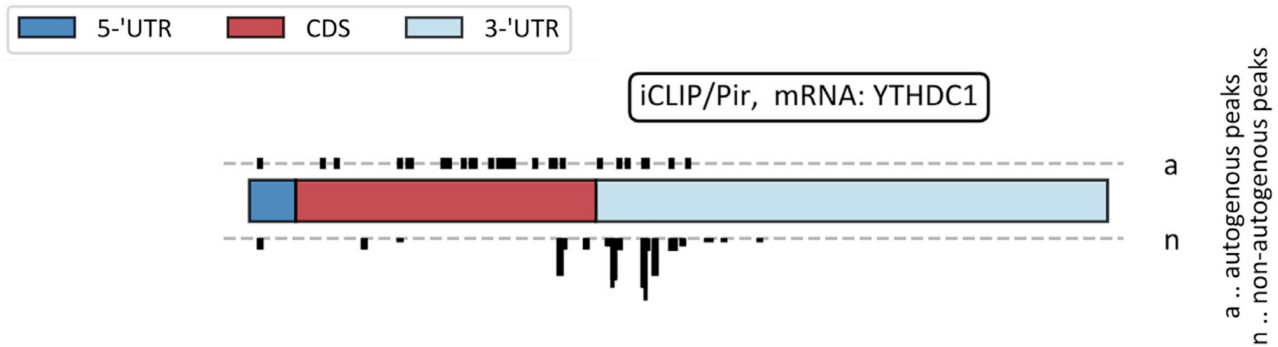
Autogenous interactions occur preferentially in the CDS

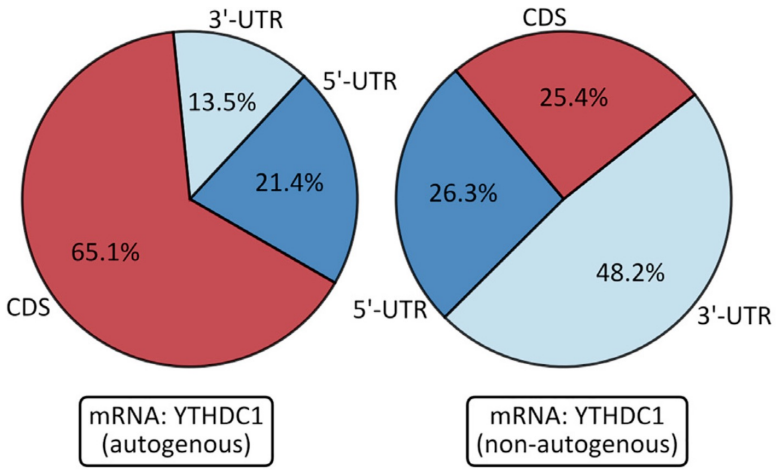


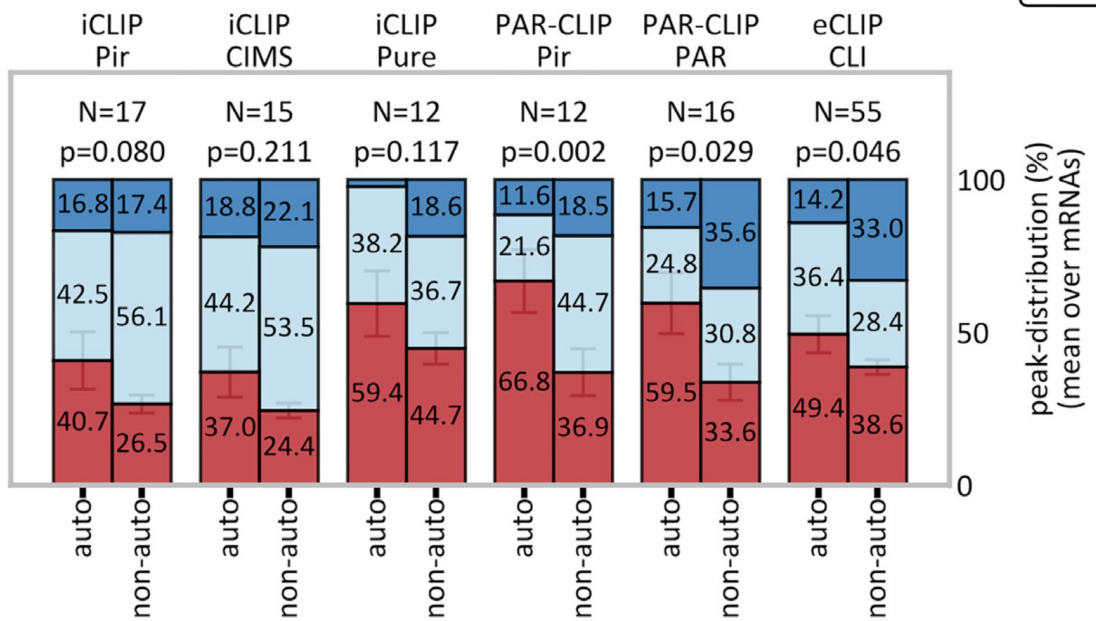


Kapral TH, Farnhammer F, Zhao W, Lu ZJ & Zagrovic B, **Nucleic Acids Research**, 50, 2022

Autogenous interactions occur preferentially in the CDS





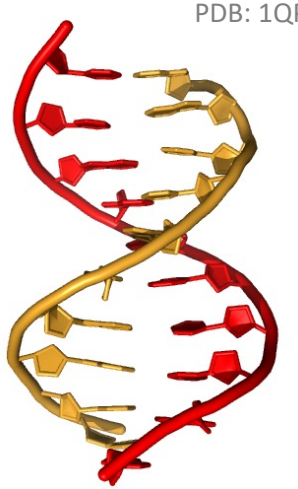


Kapral TH, Farnhammer F, Zhao W, Lu ZJ & Zagrovic B, **Nucleic Acids Research**, 50, 2022

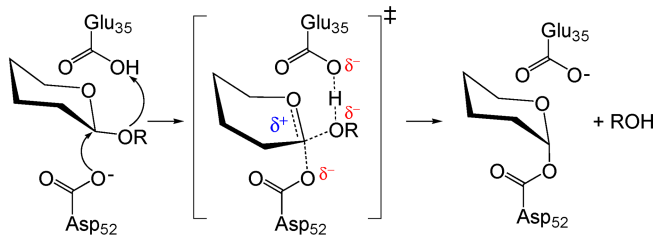
Physicochemical complementarity

one of the most powerful paradigms in biology

PDB: 1QPH



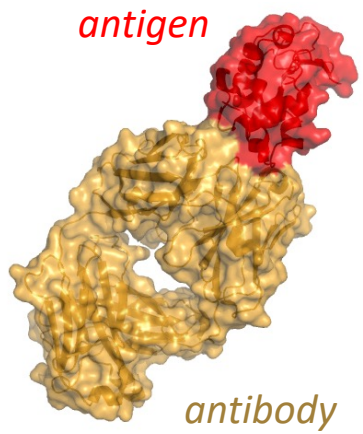
DNA replication



Kraut et al. PLOSB, 4, 2006

enzymatic catalysis

PDB: 1WBU

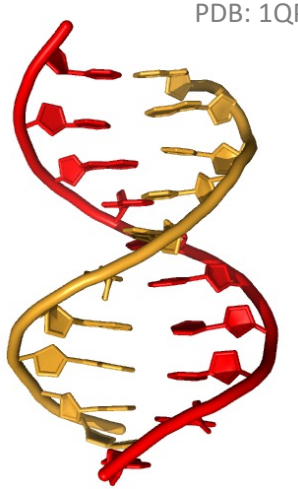


immune response

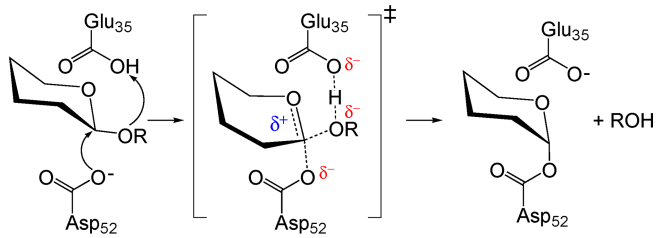
Physicochemical complementarity

one of the most powerful paradigms in biology

PDB: 1QPH



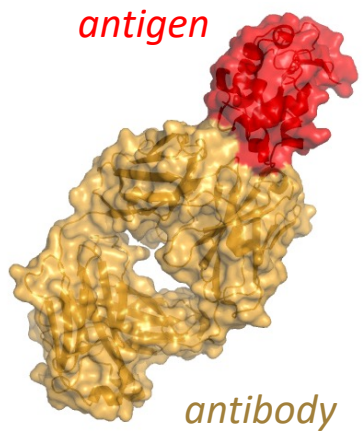
DNA replication



Kraut et al. PLOSB, 4, 2006

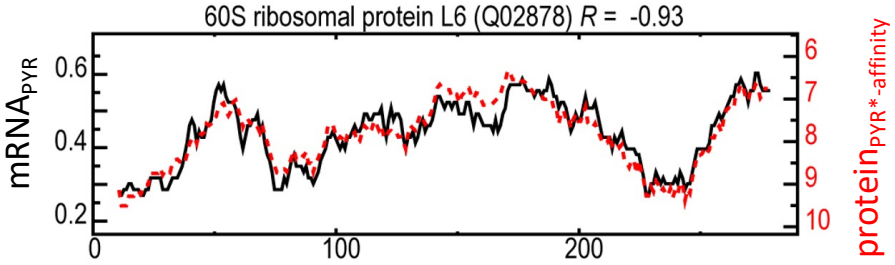
enzymatic catalysis

PDB: 1WBU



immune response

a novel complementarity?



Conclusion 2: coding and binding as two faces of the same coin?

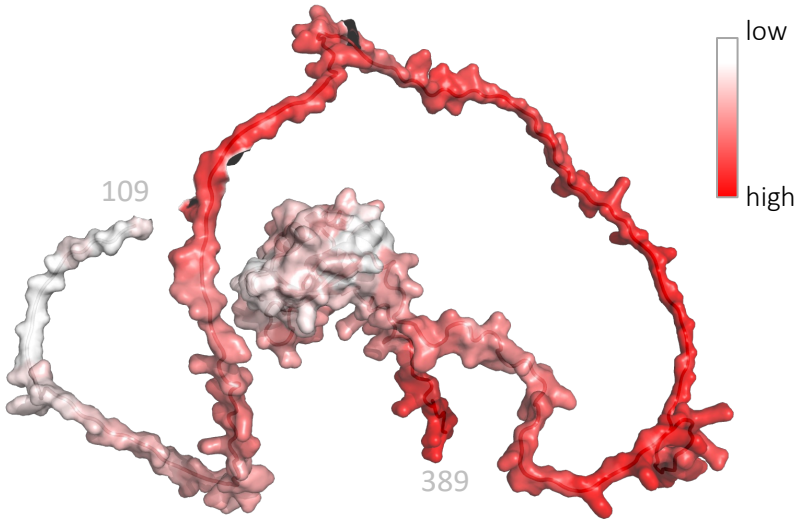
- evidence of complementary interactions between mRNAs and own proteins in unstructured state
- proteins interact with RNAs that are compositionally similar to their own mRNA and *vice versa*
- genetic code as a key for understanding large-structure of the cellular RNA-protein interactome

OPEN CHALLENGES

geometry; ΔGs; functional significance; experimental testing

Towards physicochemical bioinformatics

- **Analyzing biomolecular sequences from a physicochemical perspective**: a rich source of unexpected patterns
- **Particularly relevant in an unstructured context:** sequences as physicochemical objects
- **Advances required when it comes to:**
  - robust algorithms for sequence profile alignment
  - new measures for profile comparison
  - development of fast methods for detection of historical frameshifts
  - experimental testing of complementarity hypothesis



FUS GUA-affinity

Thanks

Present members

- **Thomas Kapral**
- **Anton A. Polyansky**
- Fran Miočić Stošić
- Felix Fischer
- Hannah Dydziul
- Christian Stelmach

Former members

- **Marlene Adlhart**
- **Mario Hlevnjak**
- **Theres Friesacher**
- **Daniel Hoffmann**
- **Anita de Ruiter**
- **Florian Poetsch**
- **Matea Hajnic**
- **Megan Hoogmoed**
- **Fiona Farnhammer**

& other members of the lab

Collaborators

- Renee Schroeder, Max Perutz Labs & University of Vienna
- Dea Slade, Max Perutz Labs & University of Vienna
- Thomas Leonard, Max Perutz Labs & University of Vienna
- John D. Sutherland, LMB Cambridge
- Zoya Ignatova, University of Hamburg
- Markus Zweckstetter, Max Planck Göttingen
- John Zhi Lu, Tsinghua University, Beijing
- Vanja Nagy, Medical University of Vienna

Funding

erc
FWF
VIENNA SCIENTIFIC CLUSTER
VolkswagenStiftung