UNIVERSITY OF WARSAW
MIM
FACULTY
OF MATHEMATICS, INFORMATICS
AND MECHANICS

UNIWERSYTET WARSZAWSKI

Architecture of large projects in bioinformatics (ADP)

Lecture 12

Łukasz P. Kozłowski

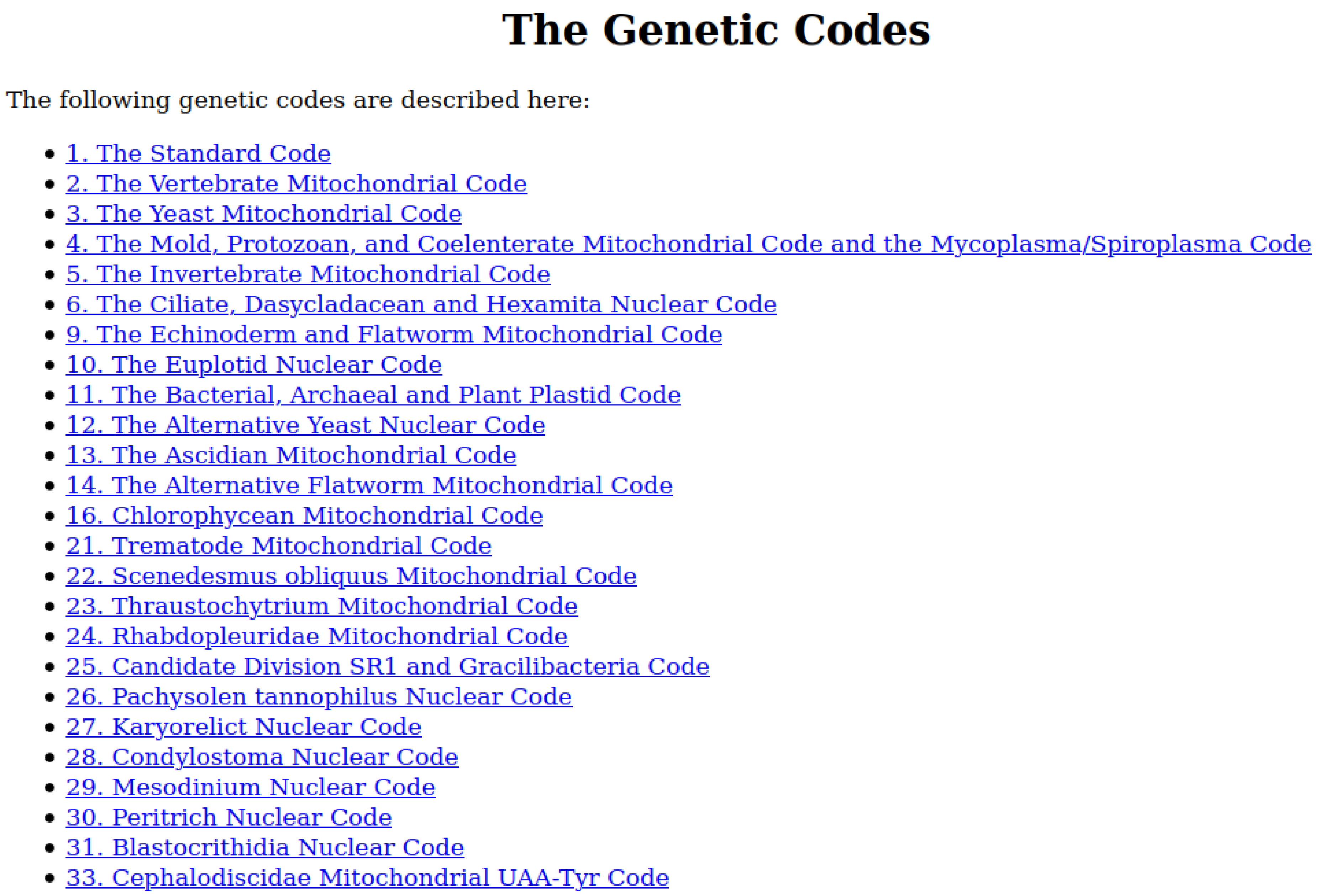
lukaskoz@mimuw.edu.pl

Warsaw, 2025

Large scale bioinformatics projects (some examples)

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

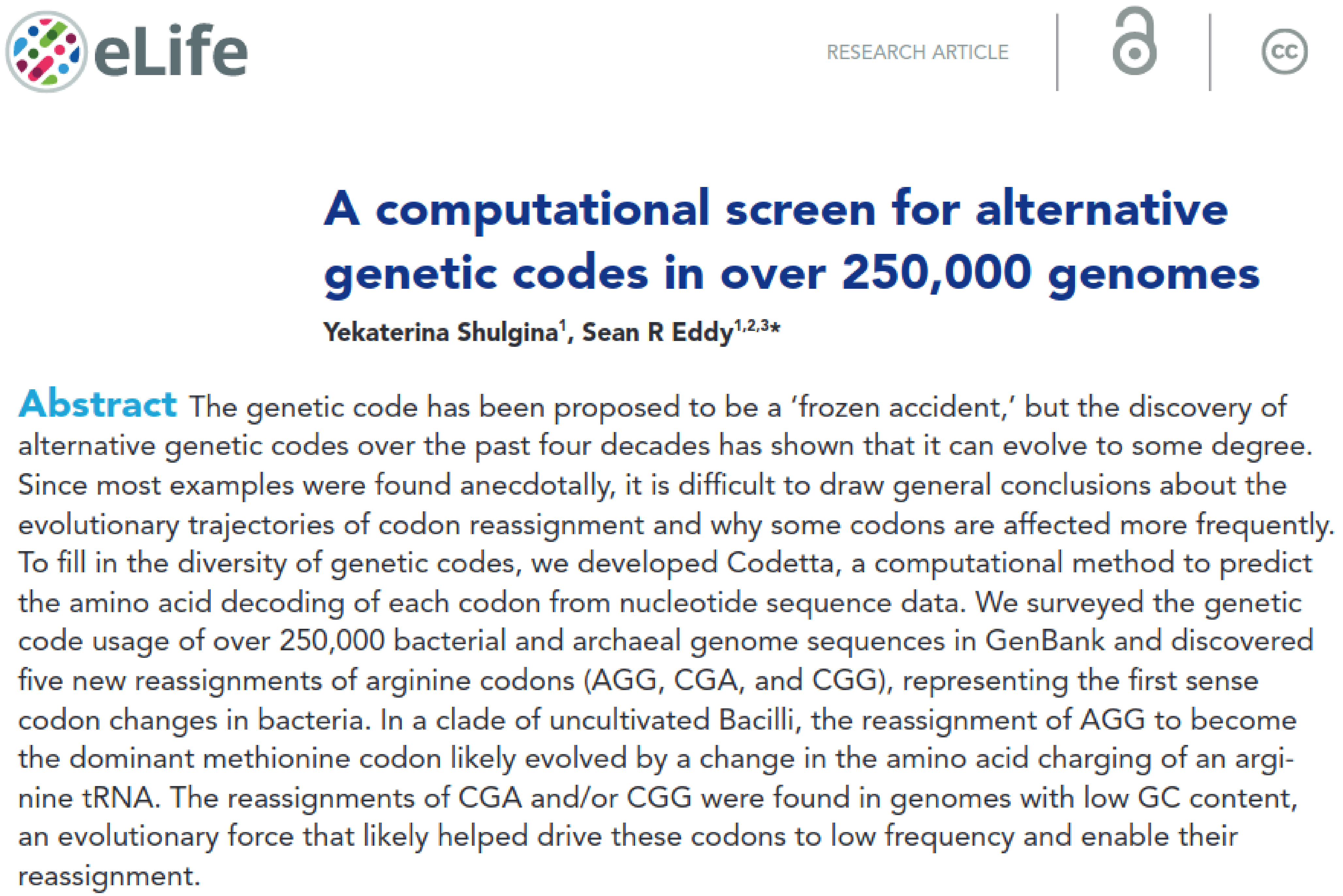
https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi



The Genetic Codes

The following genetic codes are described here:

- 1. The Standard Code
- 2. The Vertebrate Mitochondrial Code
- 3. The Yeast Mitochondrial Code
- 4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
- 5. The Invertebrate Mitochondrial Code
- 6. The Ciliate, Dasycladacean and Hexamita Nuclear Code
- 9. The Echinoderm and Flatworm Mitochondrial Code
- 10. The Euplotid Nuclear Code
- 11. The Bacterial, Archaeal and Plant Plastid Code
- 12. The Alternative Yeast Nuclear Code
- 13. The Ascidian Mitochondrial Code
- 14. The Alternative Flatworm Mitochondrial Code
- 16. Chlorophycean Mitochondrial Code
- 21. Trematode Mitochondrial Code
- 22. Scenedesmus obliquus Mitochondrial Code
- 23. Thraustochytrium Mitochondrial Code
- 24. Rhabdopleuridae Mitochondrial Code
- 25. Candidate Division SR1 and Gracilibacteria Code
- 26. Pachysolen tannophilus Nuclear Code
- 27. Karyorelict Nuclear Code
- 28. Condylostoma Nuclear Code
- 29. Mesodinium Nuclear Code
- 30. Peritrich Nuclear Code
- 31. Blastocrithidia Nuclear Code
- 33. Cephalodiscidae Mitochondrial UAA-Tyr Code



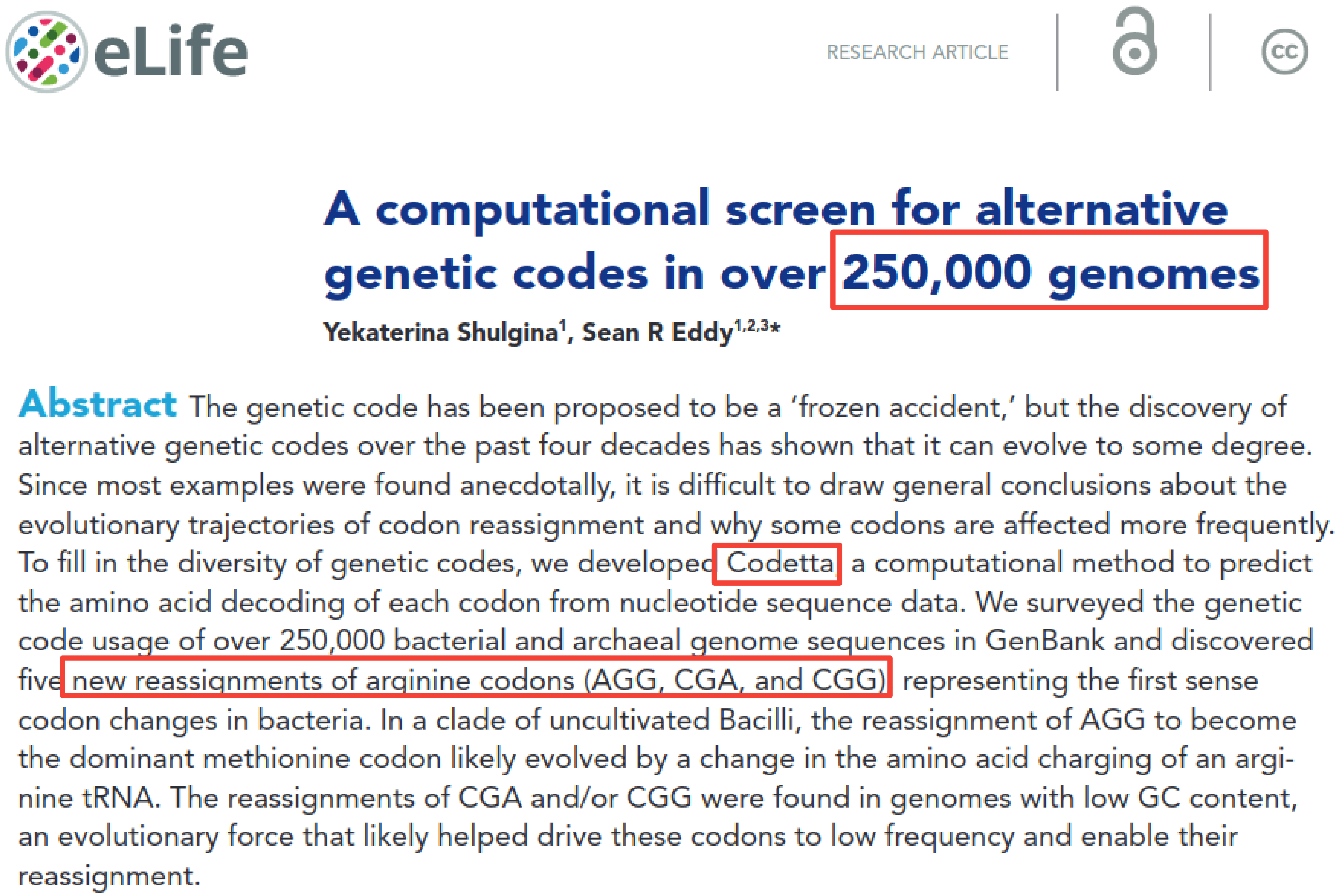
eLife

RESEARCH ARTICLE

A computational screen for alternative genetic codes in over 250,000 genomes

Yekaterina Shulgina[1], Sean R Eddy[1,2,3]*

Abstract The genetic code has been proposed to be a 'frozen accident,' but the discovery of alternative genetic codes over the past four decades has shown that it can evolve to some degree. Since most examples were found anecdotally, it is difficult to draw general conclusions about the evolutionary trajectories of codon reassignment and why some codons are affected more frequently. To fill in the diversity of genetic codes, we developed Codetta, a computational method to predict the amino acid decoding of each codon from nucleotide sequence data. We surveyed the genetic code usage of over 250,000 bacterial and archaeal genome sequences in GenBank and discovered five new reassignments of arginine codons (AGG, CGA, and CGG), representing the first sense codon changes in bacteria. In a clade of uncultivated Bacilli, the reassignment of AGG to become the dominant methionine codon likely evolved by a change in the amino acid charging of an arginine tRNA. The reassignments of CGA and/or CGG were found in genomes with low GC content, an evolutionary force that likely helped drive these codons to low frequency and enable their reassignment.



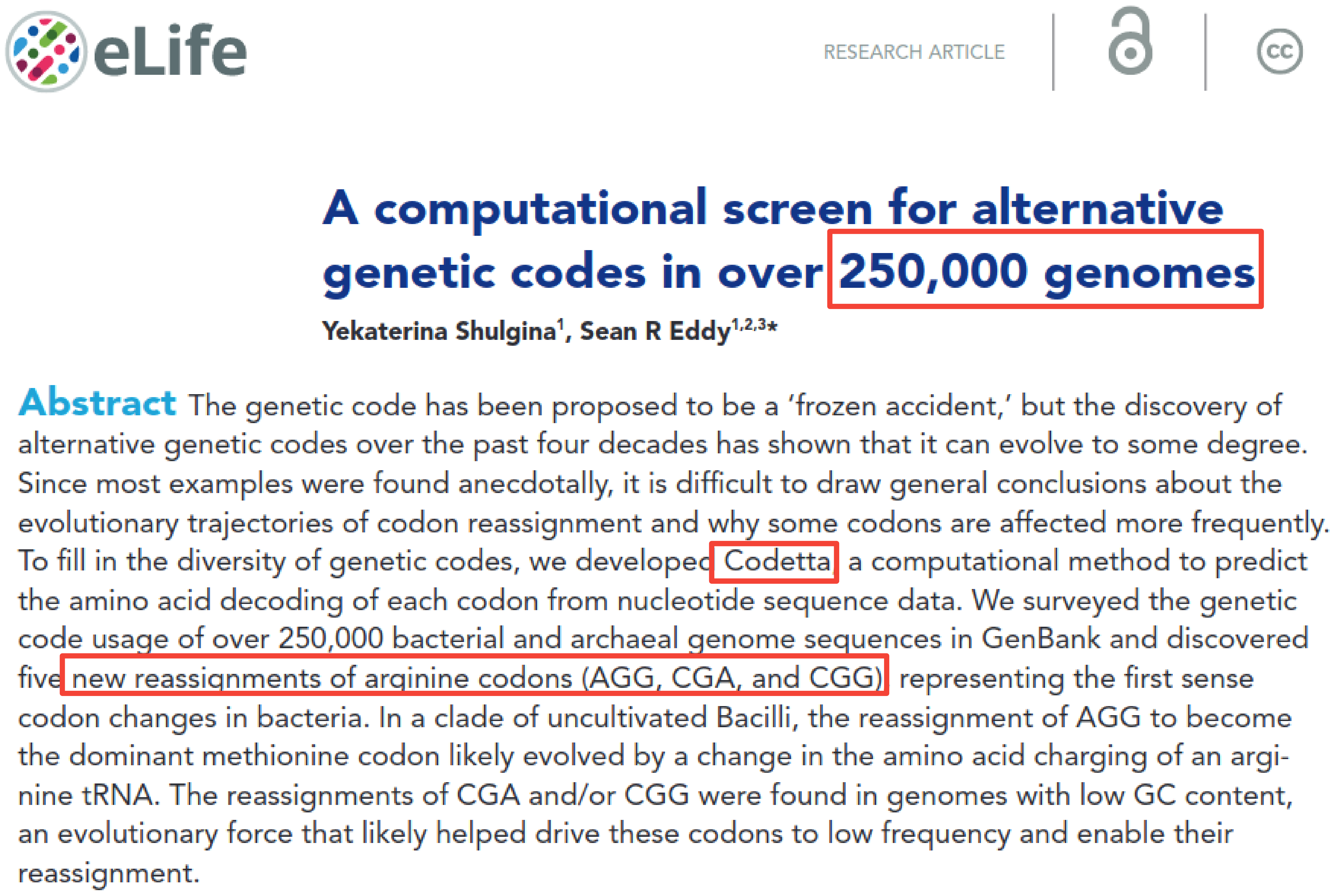
eLife

RESEARCH ARTICLE

A computational screen for alternative genetic codes in over 250,000 genomes

Yekaterina Shulgina[1], Sean R Eddy[1,2,3]*

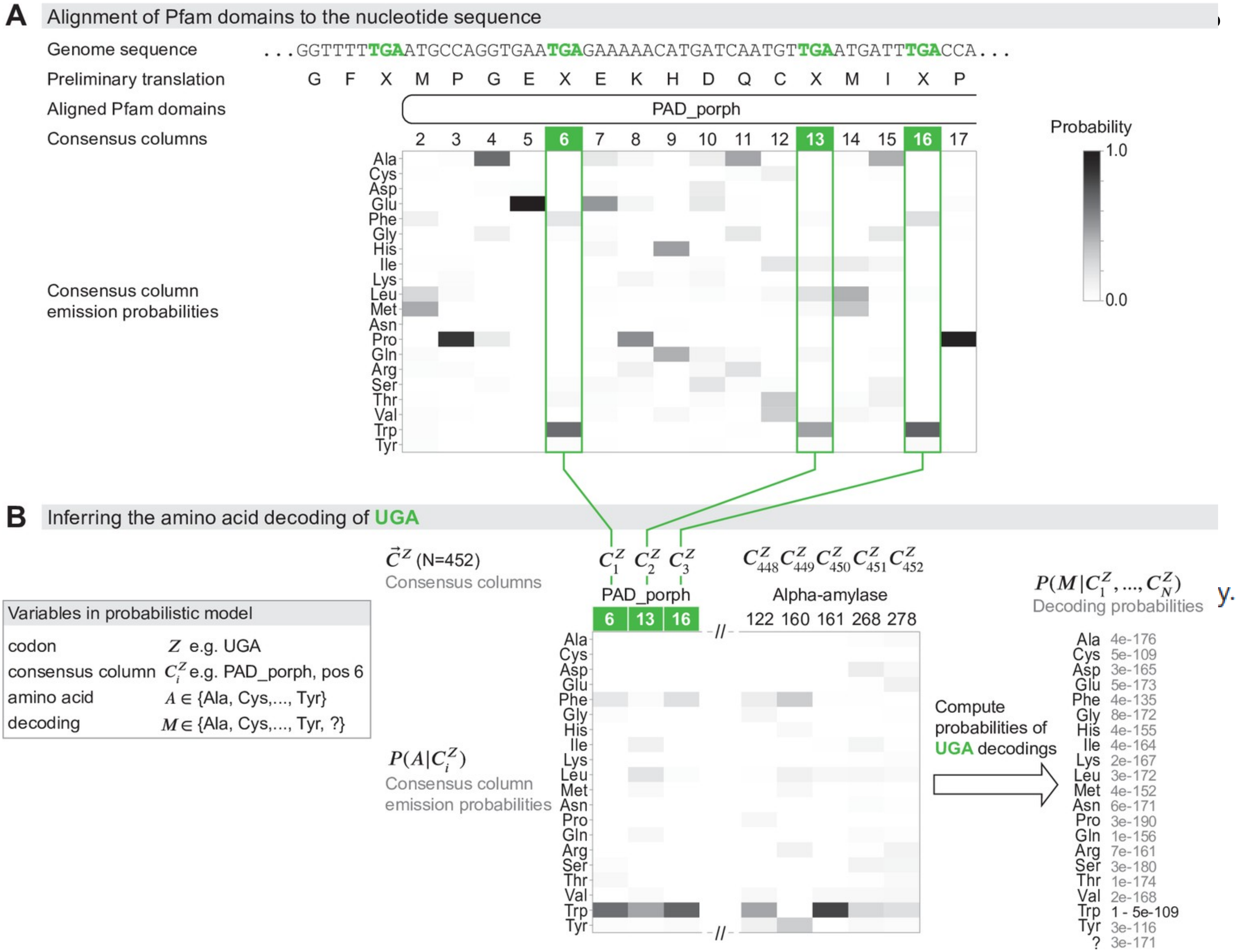
Abstract The genetic code has been proposed to be a 'frozen accident,' but the discovery of alternative genetic codes over the past four decades has shown that it can evolve to some degree. Since most examples were found anecdotally, it is difficult to draw general conclusions about the evolutionary trajectories of codon reassignment and why some codons are affected more frequently. To fill in the diversity of genetic codes, we developed Codetta, a computational method to predict the amino acid decoding of each codon from nucleotide sequence data. We surveyed the genetic code usage of over 250,000 bacterial and archaeal genome sequences in GenBank and discovered five new reassignments of arginine codons (AGG, CGA, and CGG), representing the first sense codon changes in bacteria. In a clade of uncultivated Bacilli, the reassignment of AGG to become the dominant methionine codon likely evolved by a change in the amino acid charging of an arginine tRNA. The reassignments of CGA and/or CGG were found in genomes with low GC content, an evolutionary force that likely helped drive these codons to low frequency and enable their reassignment.



A computational screen for alternative genetic codes in over 250,000 genomes

Yekaterina Shulgina[1], Sean R Eddy[1,2,3]*

Abstract The genetic code has been proposed to be a 'frozen accident,' but the discovery of alternative genetic codes over the past four decades has shown that it can evolve to some degree. Since most examples were found anecdotally, it is difficult to draw general conclusions about the evolutionary trajectories of codon reassignment and why some codons are affected more frequently. To fill in the diversity of genetic codes, we developed Codetta, a computational method to predict the amino acid decoding of each codon from nucleotide sequence data. We surveyed the genetic code usage of over 250,000 bacterial and archaeal genome sequences in GenBank and discovered five new reassignments of arginine codons (AGG, CGA, and CGG), representing the first sense codon changes in bacteria. In a clade of uncultivated Bacilli, the reassignment of AGG to become the dominant methionine codon likely evolved by a change in the amino acid charging of an arginine tRNA. The reassignments of CGA and/or CGG were found in genomes with low GC content, an evolutionary force that likely helped drive these codons to low frequency and enable their reassignment.

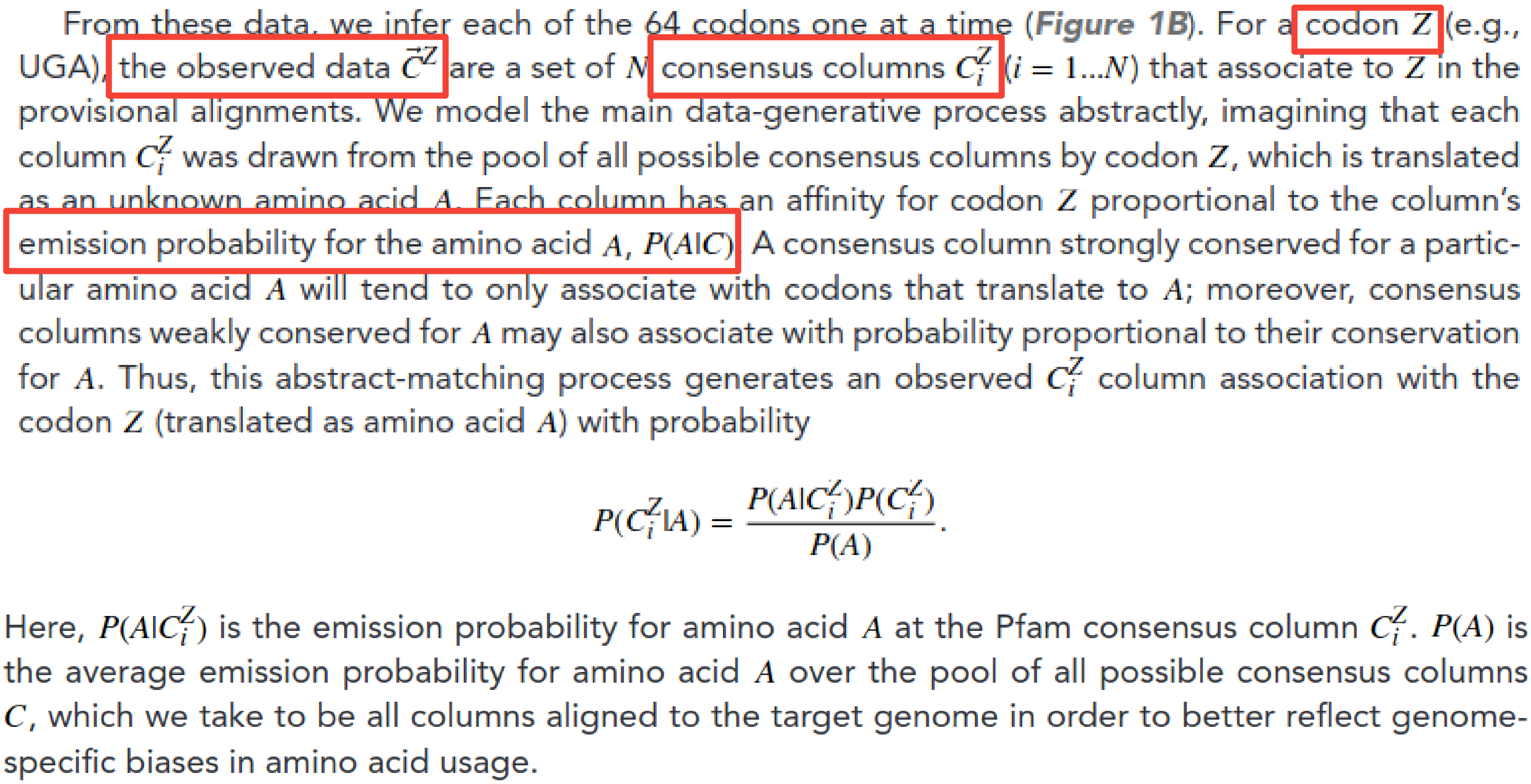

A
Alignment of Pfam domains to the nucleotide sequence
Genome sequence
...GGTTTTTGAATGCCAGGTGAATGAGAAAAACATGATCAATGTTGAATGATTTGACCA...
Preliminary translation
G F X M P G E X E K H D Q C X M I X P
Aligned Pfam domains
PAD_porph
Consensus columns
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Consensus column emission probabilities
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Probability
1.0
0.0
B
Inferring the amino acid decoding of UGA
$\vec{C}^Z$ (N=452)
Consensus columns
C_1^Z C_2^Z C_3^Z
C_{448}^Z C_{449}^Z C_{450}^Z C_{451}^Z C_{452}^Z
PAD_porph
6 13 16
Alpha-amylase
122 160 161 268 278
Variables in probabilistic model
codon Z e.g. UGA
consensus column C_i^Z e.g. PAD_porph, pos 6
amino acid $A \in$ {Ala, Cys,..., Tyr}
decoding $M \in$ {Ala, Cys,..., Tyr, ?}
$P(A|C_i^Z)$
Consensus column emission probabilities
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Compute probabilities of UGA decodings
$P(M|C_1^Z, ..., C_N^Z)$
Decoding probabilities
Ala 4e-176
Cys 5e-109
Asp 3e-165
Glu 5e-173
Phe 4e-135
Gly 8e-172
His 4e-155
Ile 4e-164
Lys 2e-167
Leu 3e-172
Met 4e-152
Asn 6e-171
Pro 3e-190
Gln 1e-156
Arg 7e-161
Ser 3e-180
Thr 1e-174
Val 2e-168
Trp 1 - 5e-109
Tyr 3e-116
? 3e-171

y.

From these data, we infer each of the 64 codons one at a time (***Figure 1B***). For a codon Z (e.g., UGA), the observed data $\vec{C}^Z$ are a set of N consensus columns C_i^Z ($i = 1...N$) that associate to Z in the provisional alignments. We model the main data-generative process abstractly, imagining that each column C_i^Z was drawn from the pool of all possible consensus columns by codon Z, which is translated as an unknown amino acid A. Each column has an affinity for codon Z proportional to the column's emission probability for the amino acid A, $P(A|C)$. A consensus column strongly conserved for a particular amino acid A will tend to only associate with codons that translate to A; moreover, consensus columns weakly conserved for A may also associate with probability proportional to their conservation for A. Thus, this abstract-matching process generates an observed C_i^Z column association with the codon Z (translated as amino acid A) with probability

$$P(C_i^Z|\mathrm{A}) = \frac{P(A|C_i^Z)P(C_i^Z)}{P(A)}.$$

Here, $P(A|C_i^Z)$ is the emission probability for amino acid A at the Pfam consensus column C_i^Z. $P(A)$ is the average emission probability for amino acid A over the pool of all possible consensus columns C, which we take to be all columns aligned to the target genome in order to better reflect genome-specific biases in amino acid usage.



From these data, we infer each of the 64 codons one at a time (***Figure 1B***). For a codon Z (e.g., UGA), the observed data $\vec{C}^Z$ are a set of N consensus columns C_i^Z ($i = 1...N$) that associate to Z in the provisional alignments. We model the main data-generative process abstractly, imagining that each column C_i^Z was drawn from the pool of all possible consensus columns by codon Z, which is translated as an unknown amino acid A. Each column has an affinity for codon Z proportional to the column's emission probability for the amino acid A, $P(A|C)$. A consensus column strongly conserved for a particular amino acid A will tend to only associate with codons that translate to A; moreover, consensus columns weakly conserved for A may also associate with probability proportional to their conservation for A. Thus, this abstract-matching process generates an observed C_i^Z column association with the codon Z (translated as amino acid A) with probability

$$P(C_i^Z|A) = \frac{P(A|C_i^Z)P(C_i^Z)}{P(A)}.$$

Here, $P(A|C_i^Z)$ is the emission probability for amino acid A at the Pfam consensus column C_i^Z. $P(A)$ is the average emission probability for amino acid A over the pool of all possible consensus columns C, which we take to be all columns aligned to the target genome in order to better reflect genome-specific biases in amino acid usage.

Given the data $\vec{C}^Z$ and this abstract generative model, we infer the most likely decoding M for codon Z out of 21 possibilities $M \in \{\mathrm{Ala}, \mathrm{Cys}, ..., \mathrm{Tyr}, ?\}$ (***Figure 1B***). The $M = ?$ model of nonspecific translation draws columns randomly and serves to catch codons that do not encode a specific amino acid, such as stop codons and ambiguously translated codons. For a given decoding M, the probability of the observed columns $\vec{C}^Z$ is then

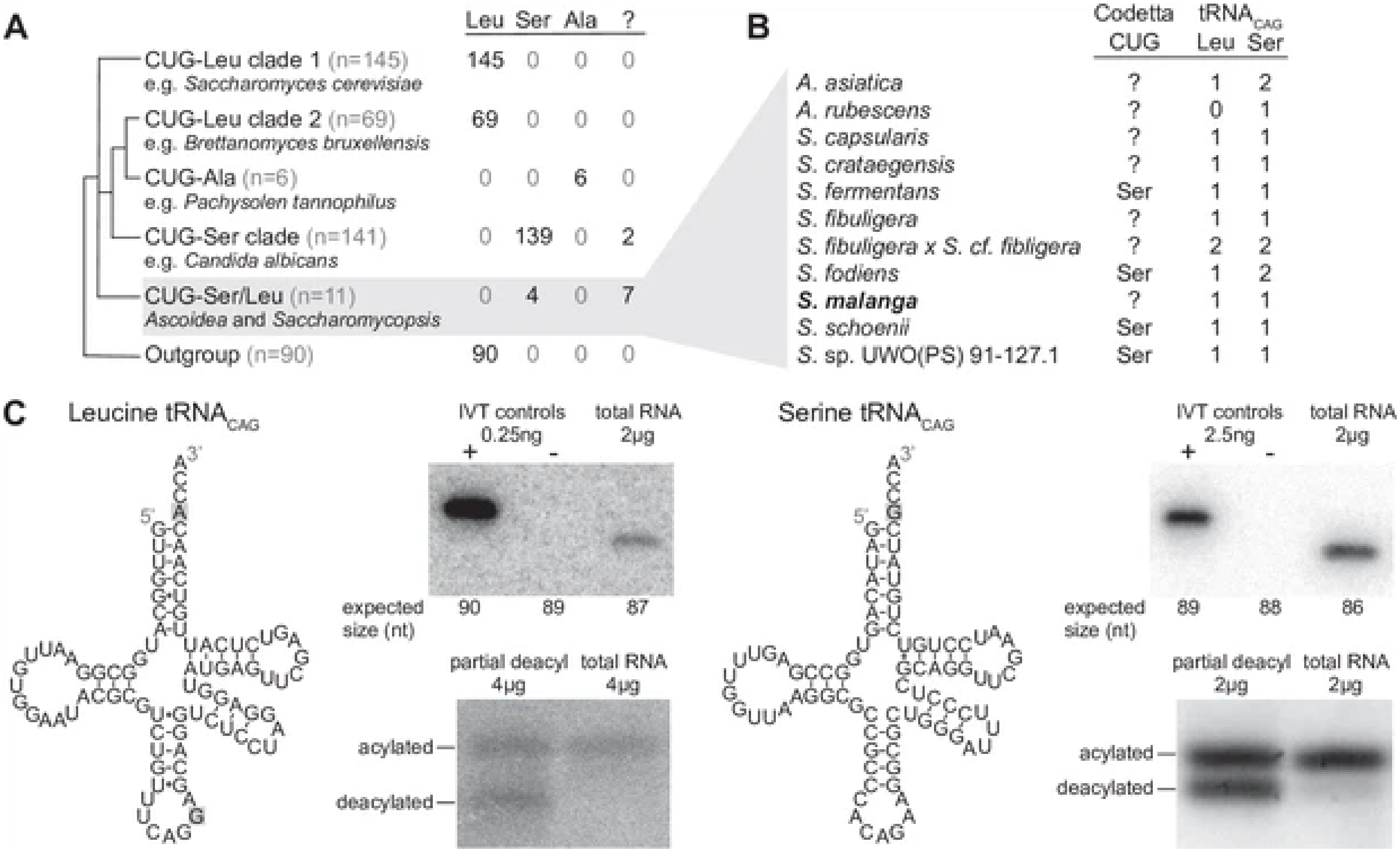
$$P(\vec{C}^Z|M) = \begin{cases} \prod_{i=1}^{N} \frac{P(A=M|C_i^Z)P(C_i^Z)}{P(A=M)} & \text{if } M \in \{\mathrm{Ala}, \mathrm{Cys}, ..., \mathrm{Tyr}\} \\ \prod_{i=1}^{N} P(C_i^Z) & \text{if } M = ? \end{cases}$$

Setting the prior probability of each decoding, $P(M)$, to be uniform, we compute the probability of the decoding M as

$$P(M|\vec{C}^Z) = \frac{P(\vec{C}^Z|M)}{\Sigma_{M'} P(\vec{C}^Z|M')}$$

Genetic code prediction of 462 yeast species confirms known distributions of CUG reassignment

Genetic code prediction of 462 yeast species confirms known distributions of CUG reassignment



A

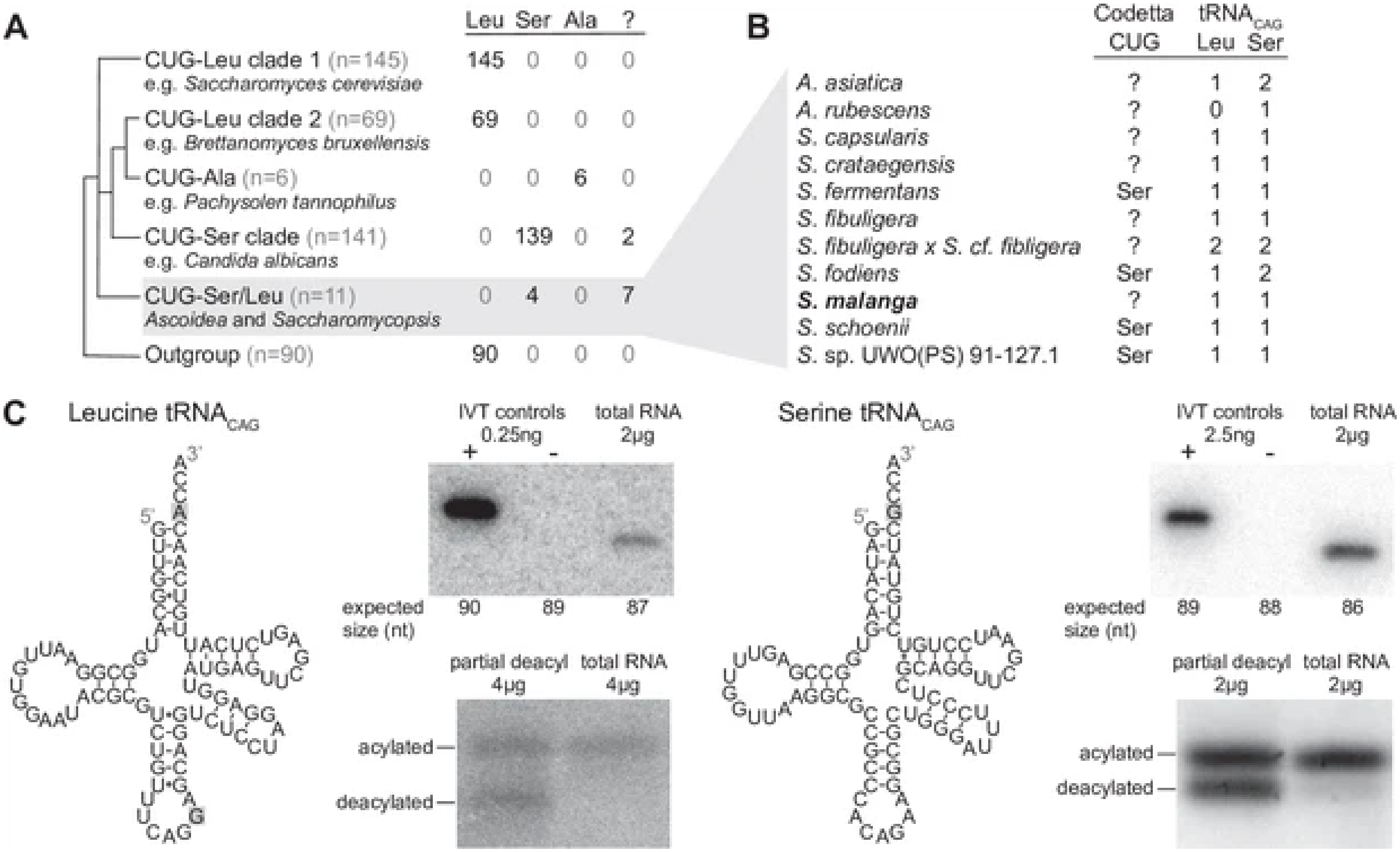
	Leu	Ser	Ala	?
CUG-Leu clade 1 (n=145) e.g. *Saccharomyces cerevisiae*	145	0	0	0
CUG-Leu clade 2 (n=69) e.g. *Brettanomyces bruxellensis*	69	0	0	0
CUG-Ala (n=6) e.g. *Pachysolen tannophilus*	0	0	6	0
CUG-Ser clade (n=141) e.g. *Candida albicans*	0	139	0	2
CUG-Ser/Leu (n=11) *Ascoidea* and *Saccharomycopsis*	0	4	0	7
Outgroup (n=90)	90	0	0	0

B

	Codetta	$tRNA_{CAG}$	
	CUG	Leu	Ser
A. asiatica	?	1	2
A. rubescens	?	0	1
S. capsularis	?	1	1
S. crataegensis	?	1	1
S. fermentans	Ser	1	1
S. fibuligera	?	1	1
S. fibuligera x S. cf. fibligera	?	2	2
S. fodiens	Ser	1	2
S. malanga	?	1	1
S. schoenii	Ser	1	1
S. sp. UWO(PS) 91-127.1	Ser	1	1

Genetic code prediction of 462 yeast species confirms known distributions of CUG reassignment

tRNACAG genes were identified using tRNAscan-SE 2.0



A

	Leu	Ser	Ala	?
CUG-Leu clade 1 (n=145) e.g. *Saccharomyces cerevisiae*	145	0	0	0
CUG-Leu clade 2 (n=69) e.g. *Brettanomyces bruxellensis*	69	0	0	0
CUG-Ala (n=6) e.g. *Pachysolen tannophilus*	0	0	6	0
CUG-Ser clade (n=141) e.g. *Candida albicans*	0	139	0	2
CUG-Ser/Leu (n=11) *Ascoidea* and *Saccharomycopsis*	0	4	0	7
Outgroup (n=90)	90	0	0	0

B

	Codetta	$tRNA_{CAG}$	
	CUG	Leu	Ser
A. asiatica	?	1	2
A. rubescens	?	0	1
S. capsularis	?	1	1
S. crataegensis	?	1	1
S. fermentans	Ser	1	1
S. fibuligera	?	1	1
S. fibuligera x S. cf. fibligera	?	2	2
S. fodiens	Ser	1	2
S. malanga	?	1	1
S. schoenii	Ser	1	1
S. sp. UWO(PS) 91-127.1	Ser	1	1

Table 1. A summary of all bacterial clades previously known to use a codon reassignment.

For each clade, the NCBI taxonomic IDs (taxids) shown most closely correspond to the known phylogenetic distribution from the literature. For each codon reassignment, we show the number of sequenced species analyzed by Codetta and how many were inferred to use the expected amino acid or had no inferred amino acid. None of the analyzed species belonging to reassigned clades were predicted to use an unexpected amino acid at the reassigned codon. [1] ***Bové, 1993***, [2] ***Volokhov et al., 2007***, [3] ***McCutcheon et al., 2009***, [4] ***Bennett and Moran, 2013***, [5] ***McCutcheon and Moran, 2010***, [6] ***Salem et al., 2017***, [7] ***Rinke et al., 2013***, and [8] ***Campbell et al., 2013***.

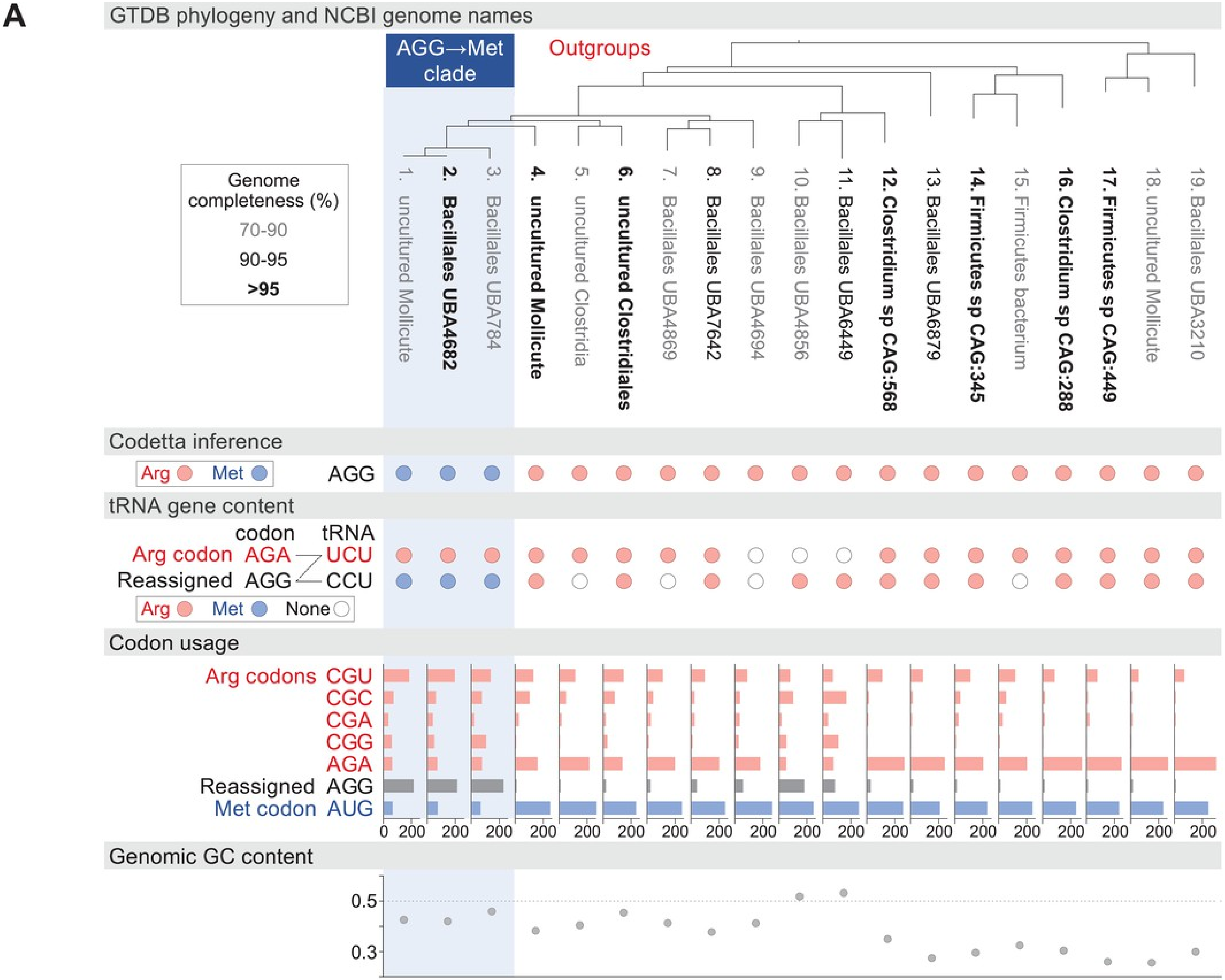
					Reassigned codon	
Phylogenetic distribution	**NCBI taxids**	**Reference**	**N species**	**Codon reassignment**	**Expected amino acid**	**Uninferred ('?')**
Entomoplasmatales and Mycoplasmatales	186328, 264638, 2085	[1, 2]	199	UGA Stop→W	191	8
Hodgkinia cicadicola	573658	[3]	1	UGA Stop→W	1	0
Nasuia deltocephalinicola	1160784	[4]	1	UGA Stop→W	1	0
Zinderia insecticola	884215	[5]	1	UGA Stop→W	1	0
Stammera capleta	2608262	[6]	1	UGA Stop→W	1	0
Gracilibacteria	363464	[7]	15	UGA Stop→G	13	2
Absconditabacteria	221235	[8]	6	UGA Stop→G	6	0

Table 2. A summary of codon inferences from the bacterial and archaeal genomes analyzed by Codetta, dereplicated to one assembly per species.

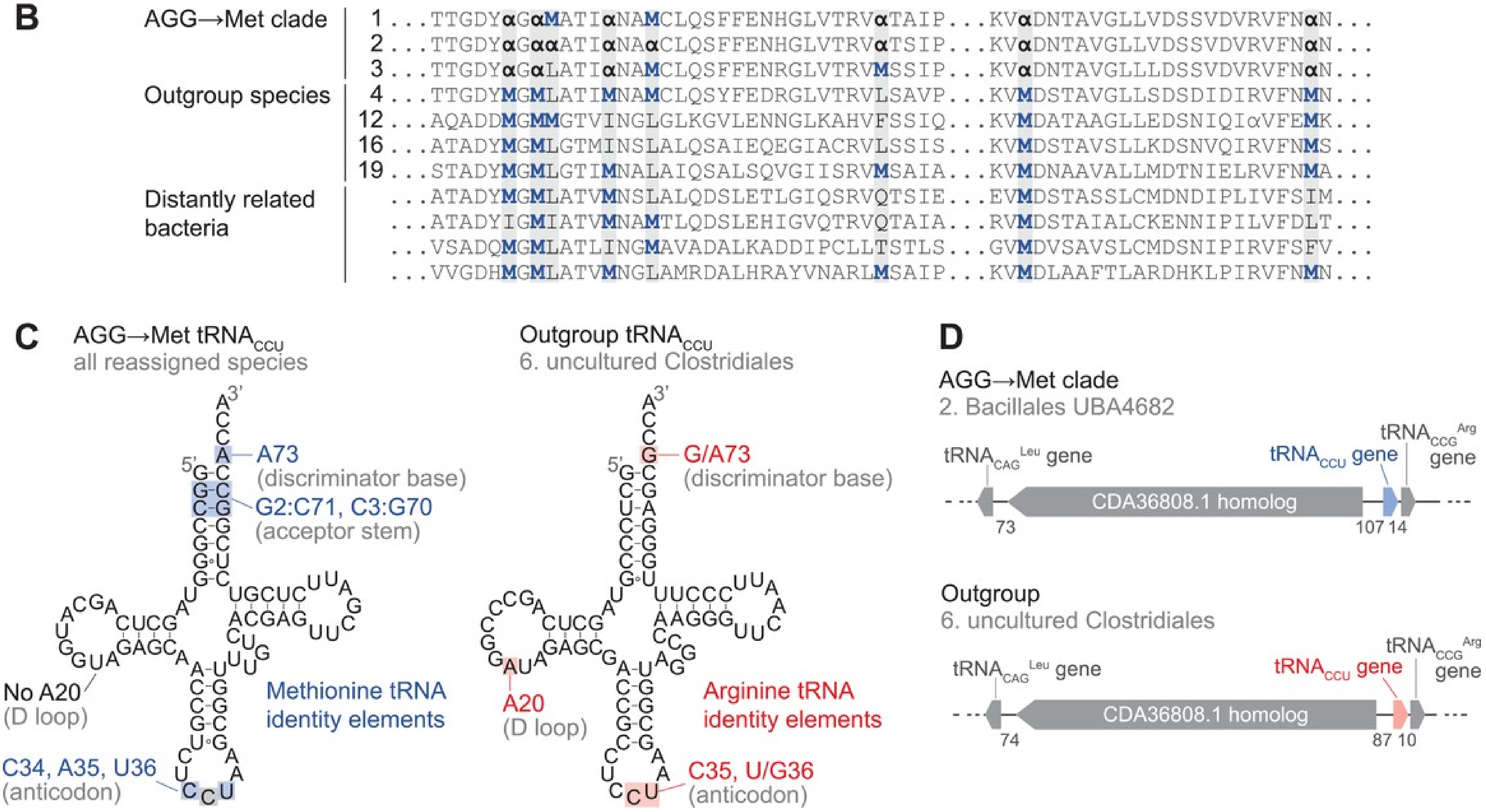
The Codetta inference for each codon is compared against a genetic code annotation derived by layering the known bacterial genetic codes in ***Table 1*** over the NCBI taxonomy. Reassigned stop codons are included with sense codons. Values can be calculated from ***Supplementary file 1***.

		Bacteria		**Archaea**	
		46,384 species		2309 species	
Sense	Total (N codons × N species)	2,829,648		140,849	
	Expected amino acid	2,823,497	99.78%	140,631	99.85%
	Other amino acid	612	0.02%	0	0.00%
	Uninferred ('?')	5539	0.20%	218	0.15%
Stop	Total (N codons × N species)	138,928		6927	
	Amino acid	290	0.21%	9	0.13%
	Uninferred ('?')	138,638	99.79%	6918	99.87%

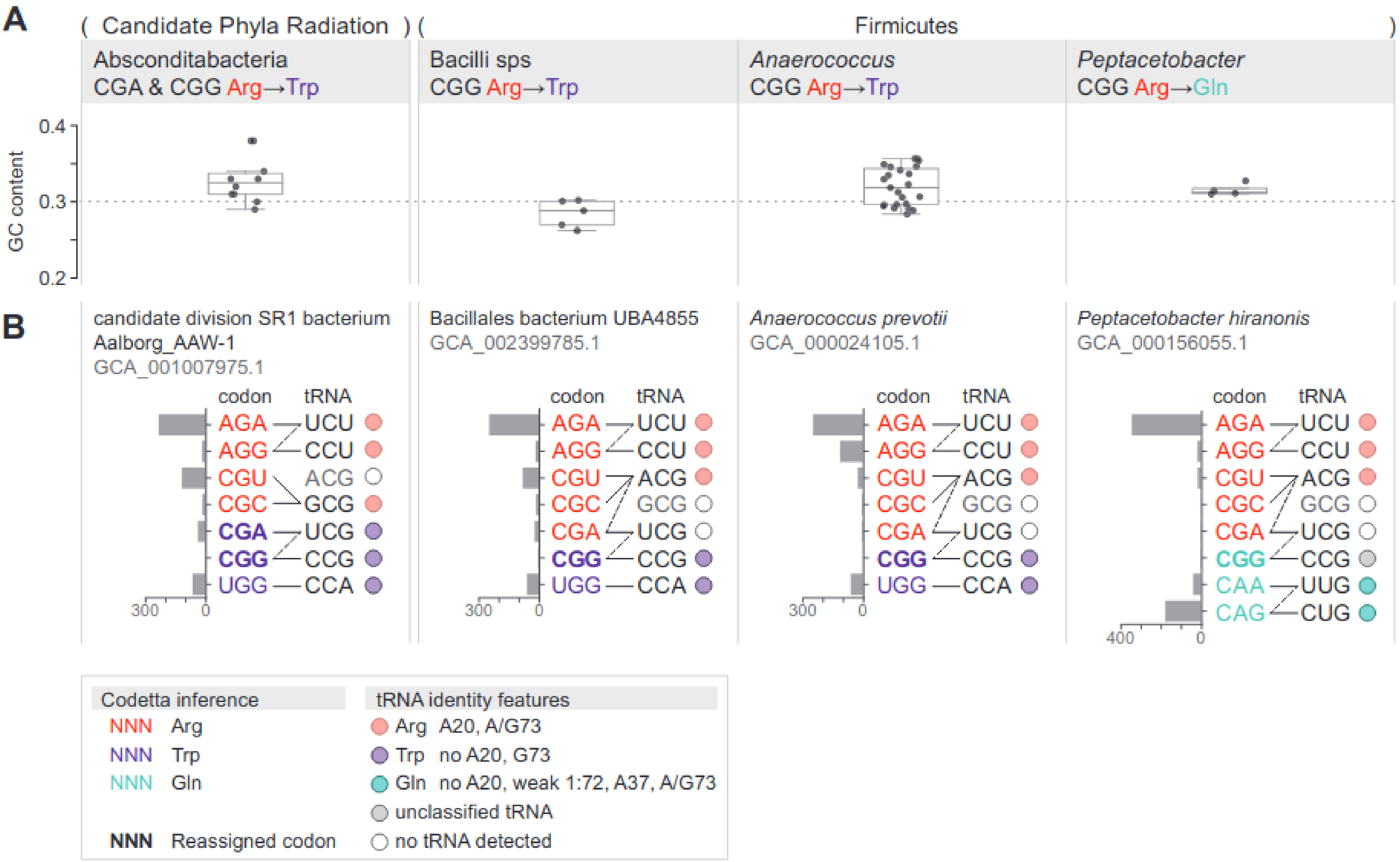
Reassignment of the canonical arginine codon AGG to methionine in a clade of uncultivated Bacilli



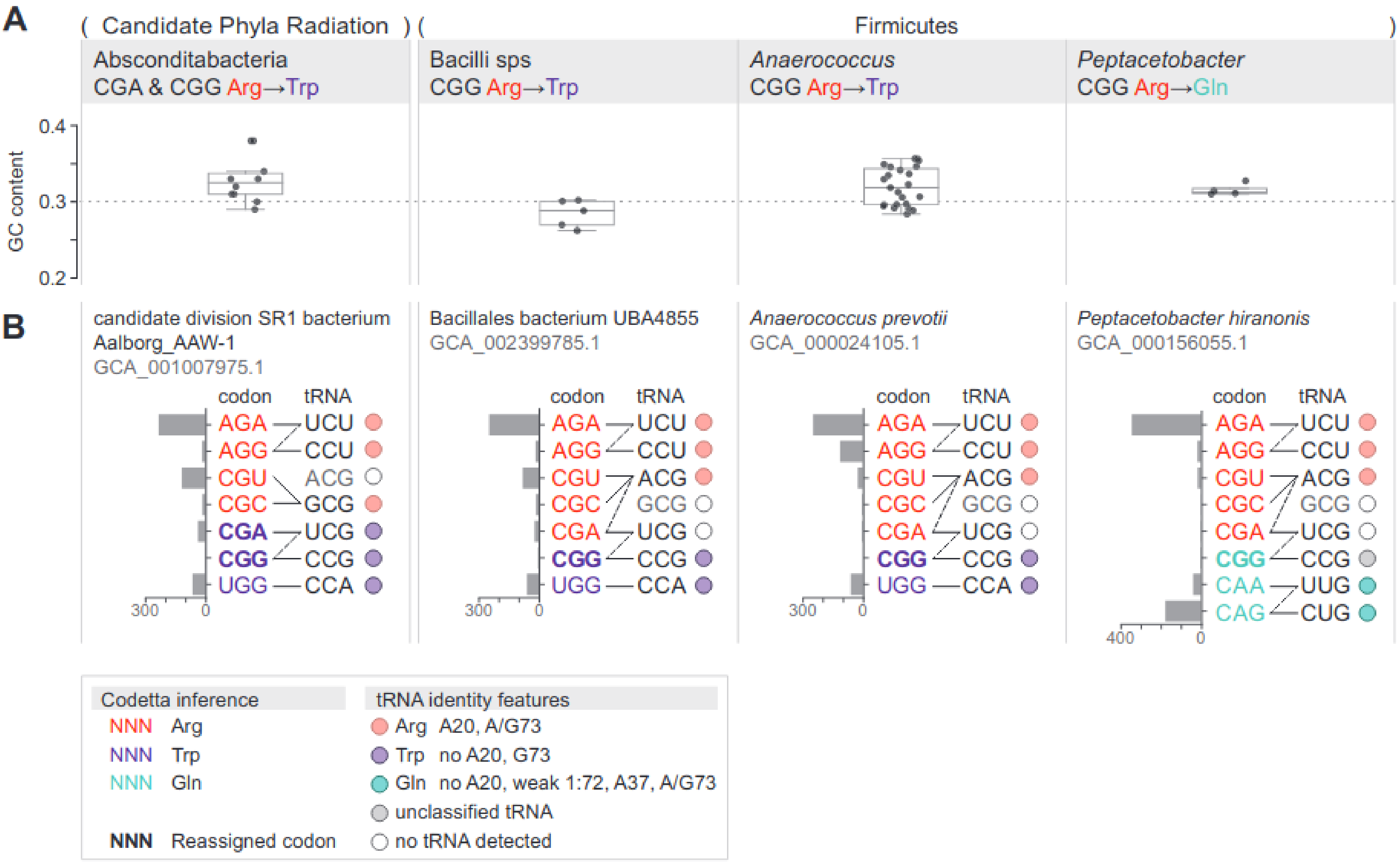
Reassignment of the canonical arginine codon AGG to methionine in a clade of uncultivated Bacilli


B
AGG→Met clade
Outgroup species
Distantly related bacteria
1 ...TTGDYαGαMATIαNAMCLQSFFENHGLVTRVαTAIP...KVαDNTAVGLLVDSSVDVRVFNαN...
2 ...TTGDYαGααATIαNAαCLQSFFENHGLVTRVαTSIP...KVαDNTAVGLLVDSSVDVRVFNαN...
3 ...TTGDYαGαLATIαNAMCLQSFFENRGLVTRVMSSIP...KVαDNTAVGLLLDSSVDVRVFNαN...
4 ...TTGDYMGMLATIMNAMCLQSYFEDRGLVTRVLSAVP...KVMDSTAVGLLSDSDIDIRVFNMN...
12 ...AQADDMGMMGTVINGLGLKGVLENNGLKAHVFSSIQ...KVMDATAAGLLEDSNIQIαVFEMK...
16 ...ATADYMGMLGTMINSLALQSAIEQEGIACRVLSSIS...KVMDSTAVSLLKDSNVQIRVFNMS...
19 ...STADYMGMLGTIMNALAIQSALSQVGIISRVMSAIA...KVMDNAAVALLMDTNIELRVFNMA...
...ATADYMGMLATVMNSLALQDSLETLGIQSRVQTSIE...EVMDSTASSLCMDNDIPLIVFSIM...
...ATADYIGMIATVMNAMTLQDSLEHIGVQTRVQTAIA...RVMDSTAIALCKENNIPILVFDLT...
...VSADQMGMLATLINGMAVADALKADDIPCLLTSTLS...GVMDVSAVSLCMDSNIPIRVFSFV...
...VVGDHMGMLATVMNGLAMRDALHRAYVNARLMSAIP...KVMDLAAFTLARDHKLPIRVFNMN...
C
AGG→Met tRNA$_{CCU}$
all reassigned species
A73
(discriminator base)
G2:C71, C3:G70
(acceptor stem)
No A20
(D loop)
Methionine tRNA
identity elements
C34, A35, U36
(anticodon)
Outgroup tRNA$_{CCU}$
6. uncultured Clostridiales
G/A73
(discriminator base)
A20
(D loop)
Arginine tRNA
identity elements
C35, U/G36
(anticodon)
D
AGG→Met clade
2. Bacillales UBA4682
tRNA$_{CAG}^{Leu}$ gene
tRNA$_{CCU}$ gene
tRNA$_{CCG}^{Arg}$ gene
CDA36808.1 homolog
73
107 14
Outgroup
6. uncultured Clostridiales
tRNA$_{CAG}^{Leu}$ gene
tRNA$_{CCU}$ gene
tRNA$_{CCG}^{Arg}$ gene
CDA36808.1 homolog
74
87 10

Summary of GC content, codon usage, and tRNA genes of four CGA and/or CGG reassignments.

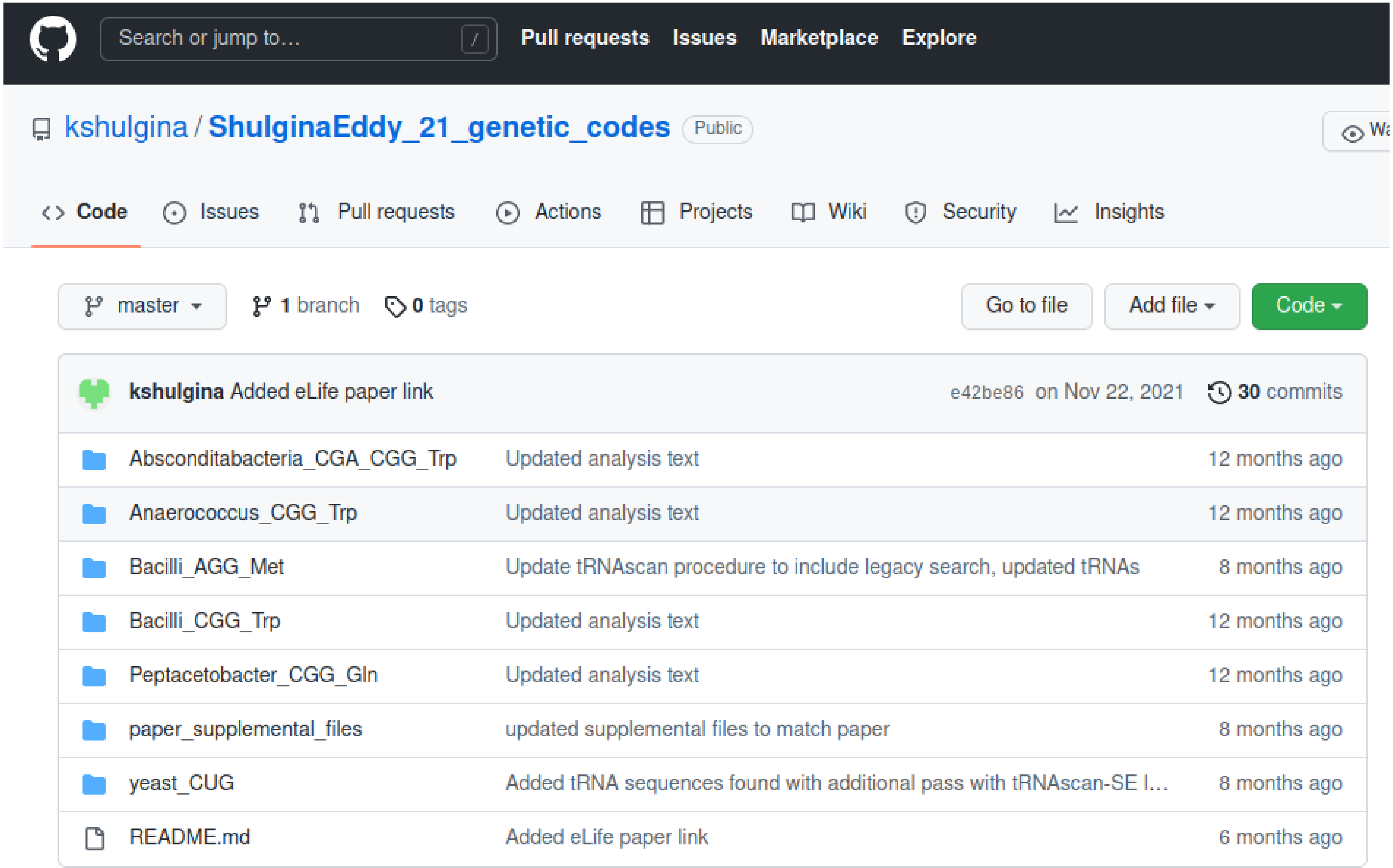


Summary of GC content, codon usage, and tRNA genes of four CGA and/or CGG reassignments.



Reassignments of arginine codons CGA and CGG occur in clades with low genomic GC content

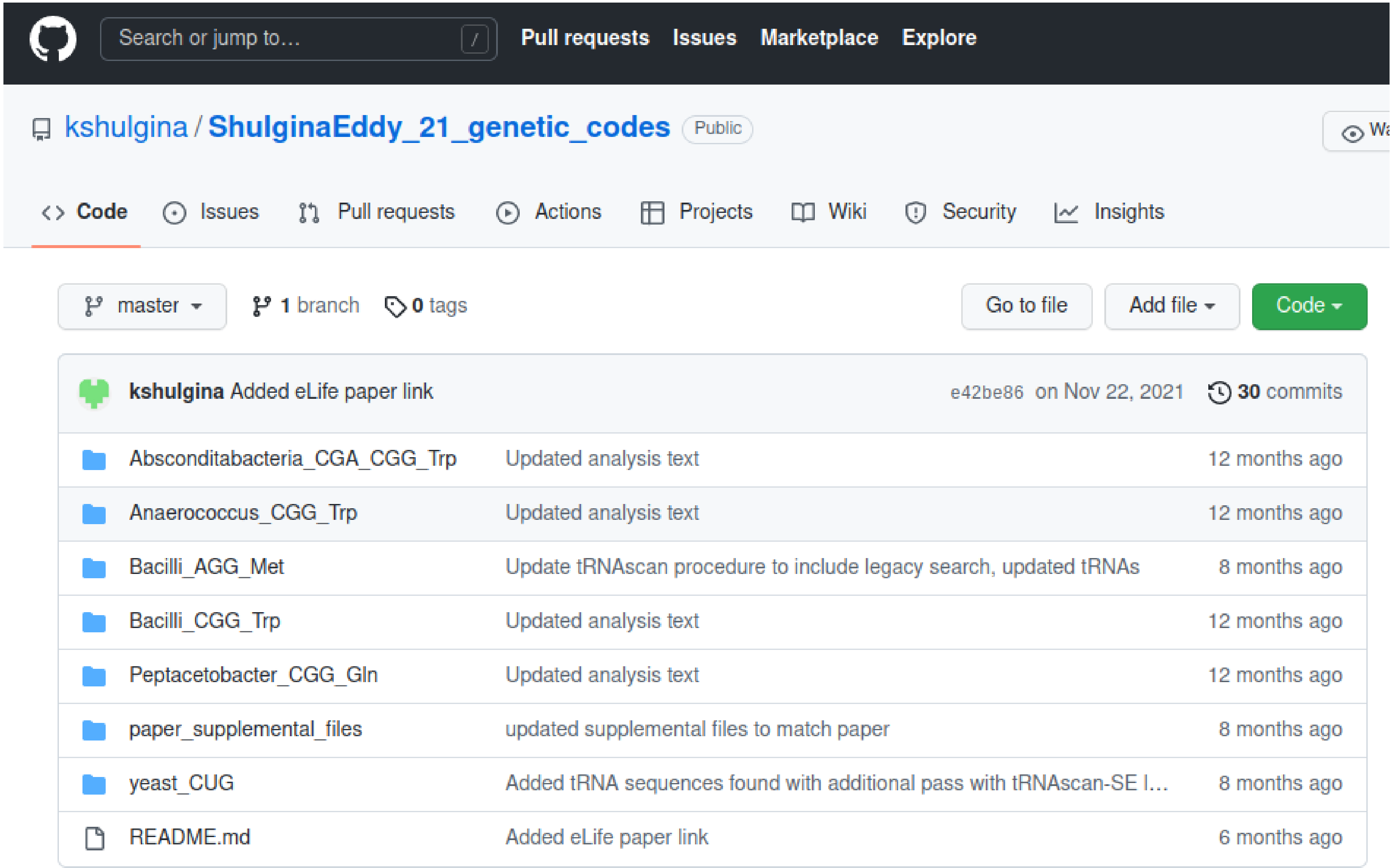
The computational requirements are dominated by the hmmscan step, which takes about an hour on a single CPU core for an ~12 Maa six-frame translation of a typical 6 Mb bacterial genome. We ran different genomes in parallel on a 30,000 core computing resource, the Harvard Cannon cluster. We implemented this method as Codetta v1.0, a Python 3 program that can be found at https://github.com/kshulgina/codetta/releases/tag/v1.0, (copy archived at swh:1:rev:4f5f31a33beed19bc3e10745154705ad002273df, ***Yekaterina, 2021***).



kshulgina Added eLife paper link e42be86 on Nov 22, 2021 **30** commits

Absconditabacteria_CGA_CGG_Trp	Updated analysis text	12 months ago
Anaerococcus_CGG_Trp	Updated analysis text	12 months ago
Bacilli_AGG_Met	Update tRNAscan procedure to include legacy search, updated tRNAs	8 months ago
Bacilli_CGG_Trp	Updated analysis text	12 months ago
Peptacetobacter_CGG_Gln	Updated analysis text	12 months ago
paper_supplemental_files	updated supplemental files to match paper	8 months ago
yeast_CUG	Added tRNA sequences found with additional pass with tRNAscan-SE l…	8 months ago
README.md	Added eLife paper link	6 months ago

The computational requirements are dominated by the hmmscan step, which takes about an hour on a single CPU core for an ~12 Maa six-frame translation of a typical 6 Mb bacterial genome. We ran different genomes in parallel on a 30,000 core computing resource, the Harvard Cannon cluster. We implemented this method as Codetta v1.0, a Python 3 program that can be found at https://github.com/kshulgina/codetta/releases/tag/v1.0, (copy archived at swh:1:rev:4f5f31a33beed19bc3e10745154705ad002273df, *Yekaterina, 2021*).

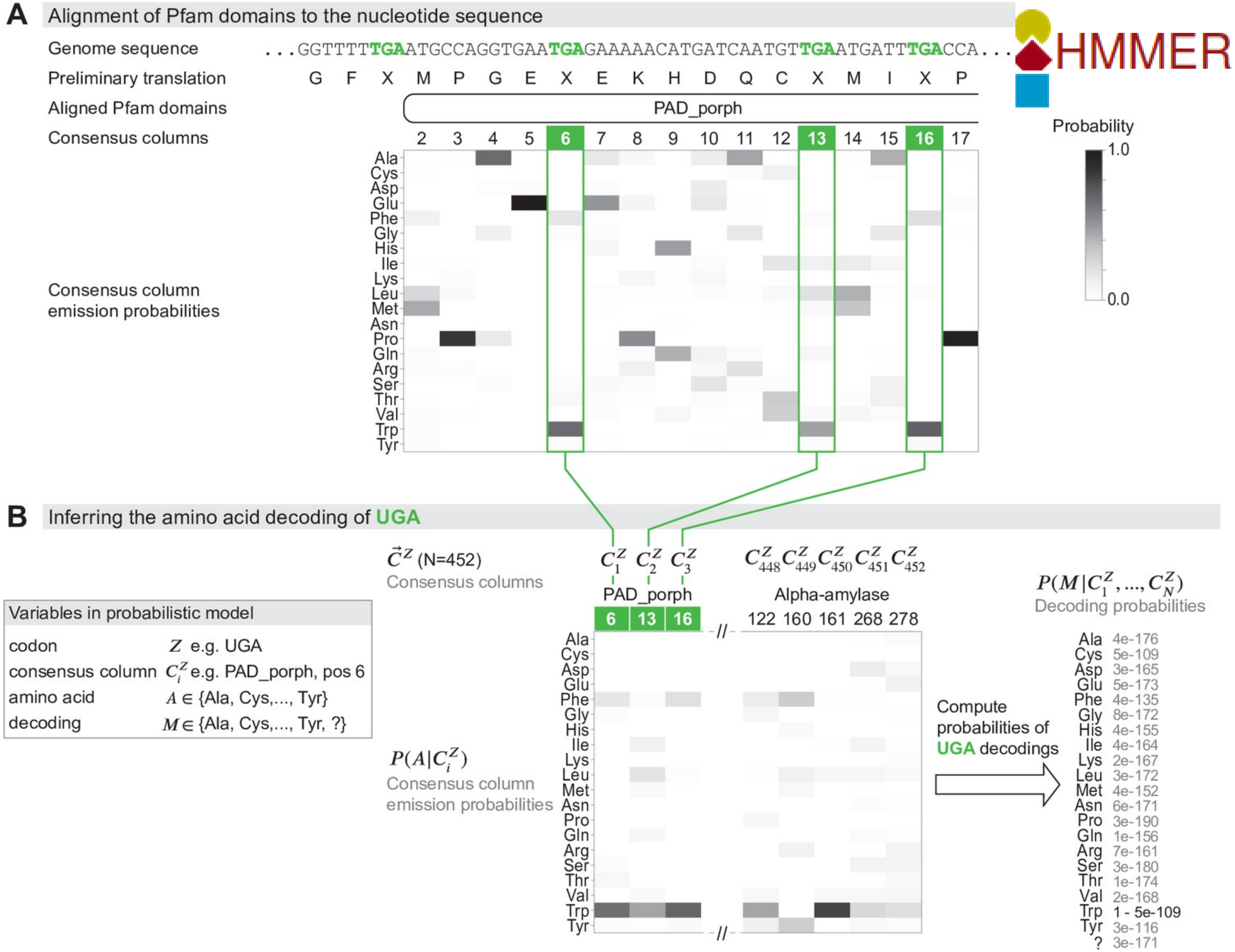


kshulgina / **ShulginaEddy_21_genetic_codes**

kshulgina Added eLife paper link e42be86 on Nov 22, 2021 **30** commits

Absconditabacteria_CGA_CGG_Trp	Updated analysis text	12 months ago
Anaerococcus_CGG_Trp	Updated analysis text	12 months ago
Bacilli_AGG_Met	Update tRNAscan procedure to include legacy search, updated tRNAs	8 months ago
Bacilli_CGG_Trp	Updated analysis text	12 months ago
Peptacetobacter_CGG_Gln	Updated analysis text	12 months ago
paper_supplemental_files	updated supplemental files to match paper	8 months ago
yeast_CUG	Added tRNA sequences found with additional pass with tRNAscan-SE l…	8 months ago
README.md	Added eLife paper link	6 months ago

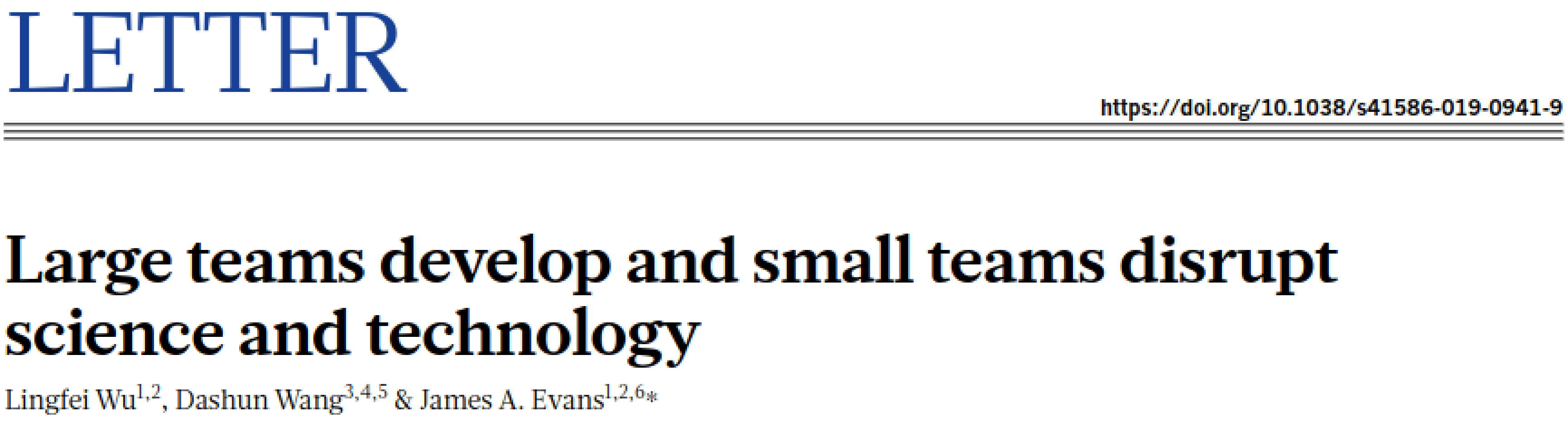
The computational requirements are dominated by the hmmscan step, which takes about an hour on a single CPU core for an ~12 Maa six-frame translation of a typical 6 Mb bacterial genome. We ran different genomes in parallel on a 30,000 core computing resource, the Harvard Cannon cluster. We implemented this method as Codetta v1.0, a Python 3 program that can be found at https://github.com/kshulgina/codetta/releases/tag/v1.0, (copy archived at swh:1:rev:4f5f31a33beed19b-c3e10745154705ad002273df, *Yekaterina, 2021*).


A
Alignment of Pfam domains to the nucleotide sequence
Genome sequence
...GGTTTTTGAATGCCAGGTGAATGAGAAAAACATGATCAATGTTGAATGATTTGACCA...
Preliminary translation
G F X M P G E X E K H D Q C X M I X P
Aligned Pfam domains
PAD_porph
Consensus columns
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Consensus column emission probabilities
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Probability
1.0
0.0
B
Inferring the amino acid decoding of UGA
$\vec{C}^Z$ (N=452)
Consensus columns
C_1^Z C_2^Z C_3^Z
$C_{448}^Z C_{449}^Z C_{450}^Z C_{451}^Z C_{452}^Z$
PAD_porph
Alpha-amylase
6 13 16
122 160 161 268 278
$P(A|C_i^Z)$
Consensus column emission probabilities
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Compute probabilities of UGA decodings
$P(M|C_1^Z,...,C_N^Z)$
Decoding probabilities

Variables in probabilistic model	
codon	Z e.g. UGA
consensus column	C_i^Z e.g. PAD_porph, pos 6
amino acid	$A \in$ {Ala, Cys,..., Tyr}
decoding	$M \in$ {Ala, Cys,..., Tyr, ?}

Ala	4e-176
Cys	5e-109
Asp	3e-165
Glu	5e-173
Phe	4e-135
Gly	8e-172
His	4e-155
Ile	4e-164
Lys	2e-167
Leu	3e-172
Met	4e-152
Asn	6e-171
Pro	3e-190
Gln	1e-156
Arg	7e-161
Ser	3e-180
Thr	1e-174
Val	2e-168
Trp	1 - 5e-109
Tyr	3e-116
?	3e-171

Other (large) bioinformatics projects

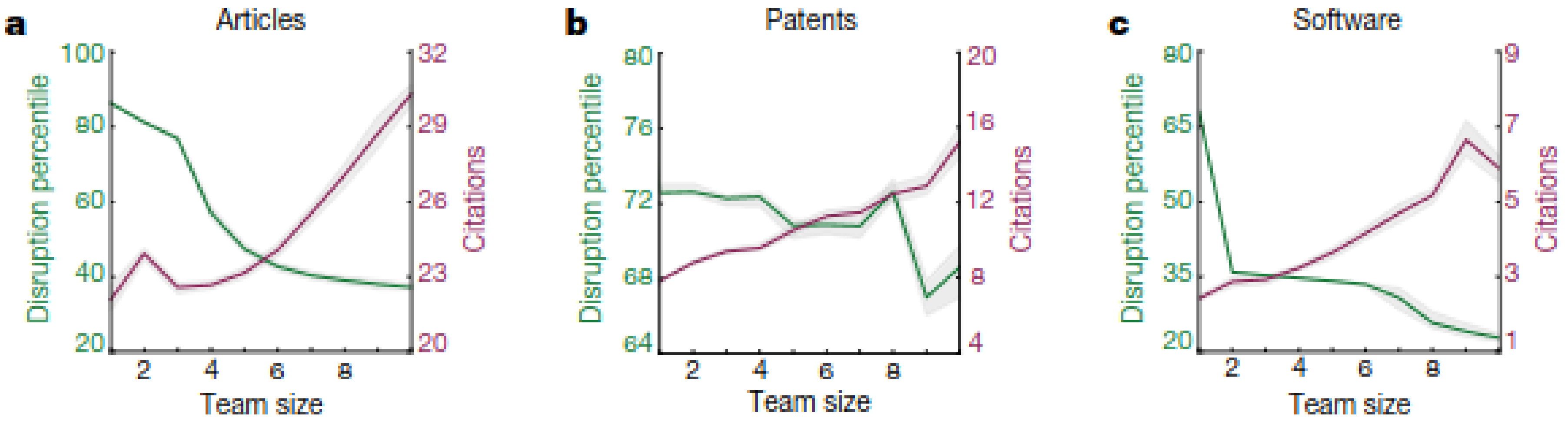


LETTER

https://doi.org/10.1038/s41586-019-0941-9

Large teams develop and small teams disrupt science and technology

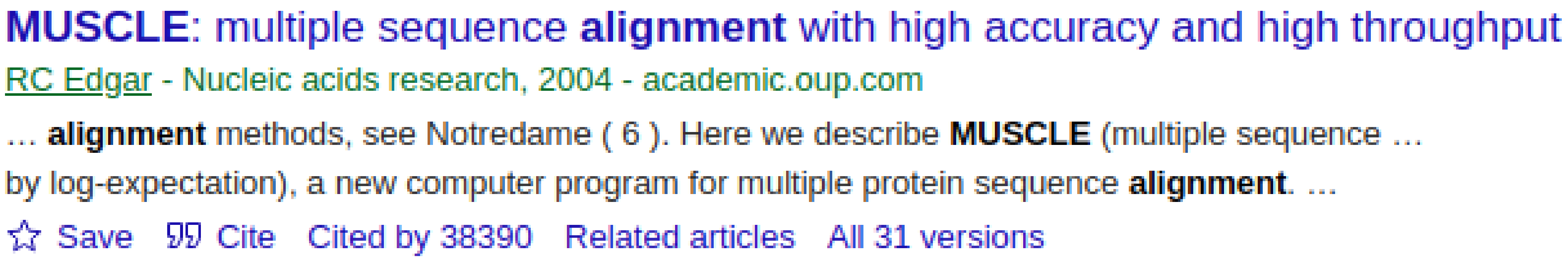
Lingfei Wu[1,2], Dashun Wang[3,4,5] & James A. Evans[1,2,6]*



Alex Bateman

Pfam

Rfam



MUSCLE: multiple sequence **alignment** with high accuracy and high throughput

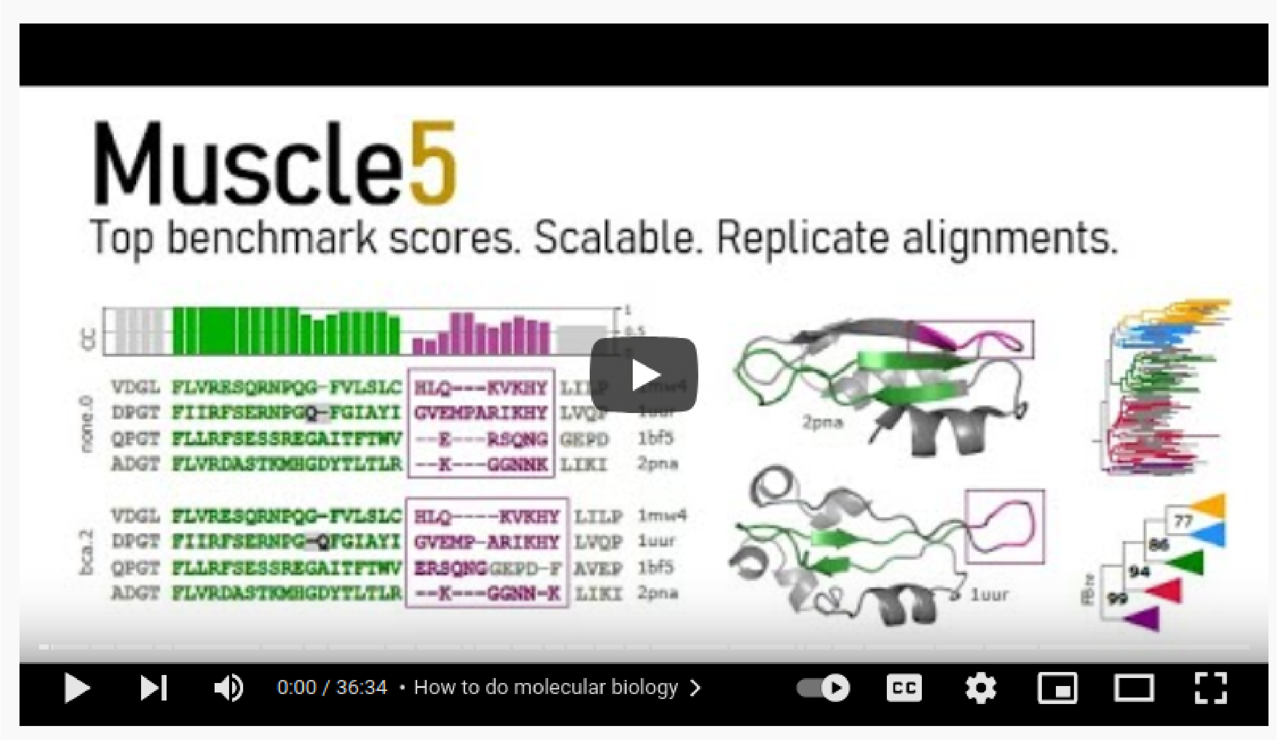
RC Edgar - Nucleic acids research, 2004 - academic.oup.com

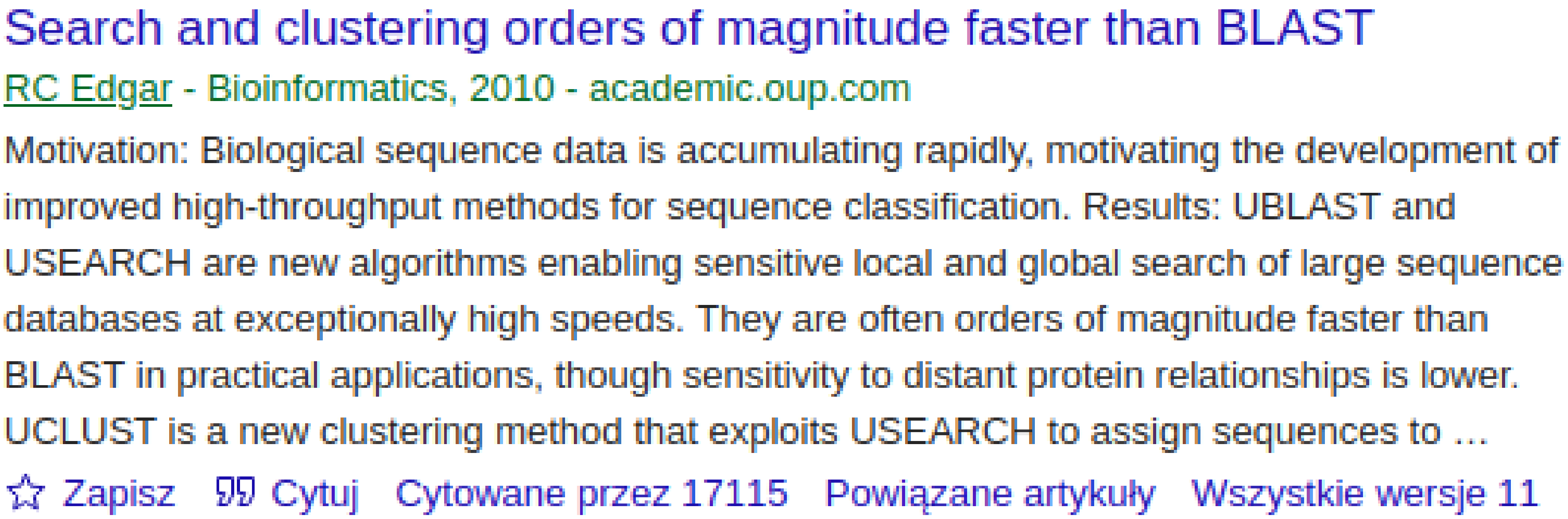
… **alignment** methods, see Notredame (6). Here we describe **MUSCLE** (multiple sequence … by log-expectation), a new computer program for multiple protein sequence **alignment**. …

☆ Save 99 Cite Cited by 38390 Related articles All 31 versions

YouTube video:

https://www.youtube.com/watch?v=2HmjHStpu7I


Muscle5
Top benchmark scores. Scalable. Replicate alignments.
cc
none.0
bca.2
2pna
1uur
0:00 / 36:34 • How to do molecular biology



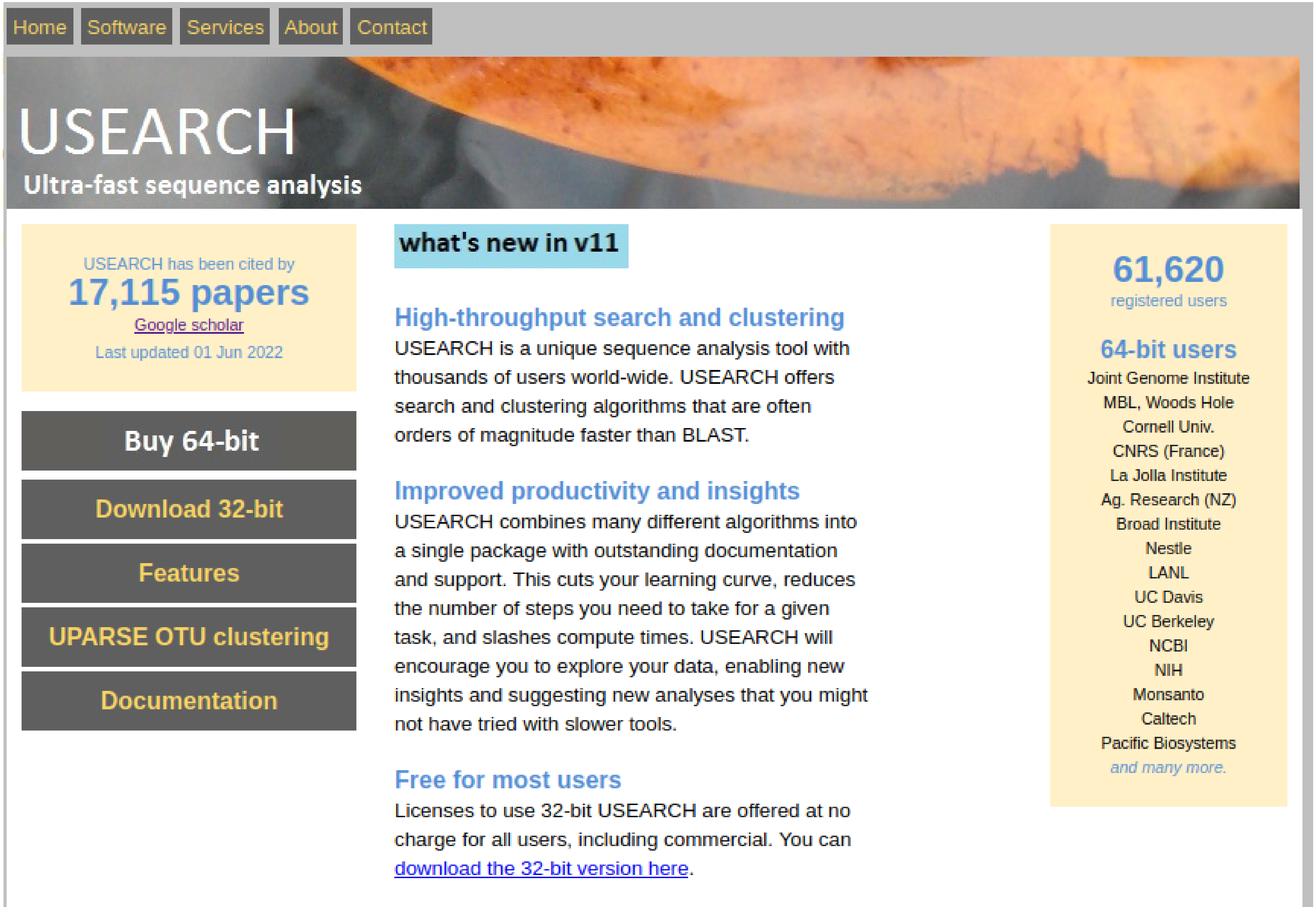
Search and clustering orders of magnitude faster than BLAST

RC Edgar - Bioinformatics, 2010 - academic.oup.com

Motivation: Biological sequence data is accumulating rapidly, motivating the development of improved high-throughput methods for sequence classification. Results: UBLAST and USEARCH are new algorithms enabling sensitive local and global search of large sequence databases at exceptionally high speeds. They are often orders of magnitude faster than BLAST in practical applications, though sensitivity to distant protein relationships is lower. UCLUST is a new clustering method that exploits USEARCH to assign sequences to …

Search and clustering orders of magnitude faster than BLAST

RC Edgar - Bioinformatics, 2010 - academic.oup.com



USEARCH

Ultra-fast sequence analysis

USEARCH has been cited by

17,115 papers

Google scholar

Last updated 01 Jun 2022

what's new in v11

High-throughput search and clustering

USEARCH is a unique sequence analysis tool with thousands of users world-wide. USEARCH offers search and clustering algorithms that are often orders of magnitude faster than BLAST.

Improved productivity and insights

USEARCH combines many different algorithms into a single package with outstanding documentation and support. This cuts your learning curve, reduces the number of steps you need to take for a given task, and slashes compute times. USEARCH will encourage you to explore your data, enabling new insights and suggesting new analyses that you might not have tried with slower tools.

Free for most users

Licenses to use 32-bit USEARCH are offered at no charge for all users, including commercial. You can download the 32-bit version here.

61,620

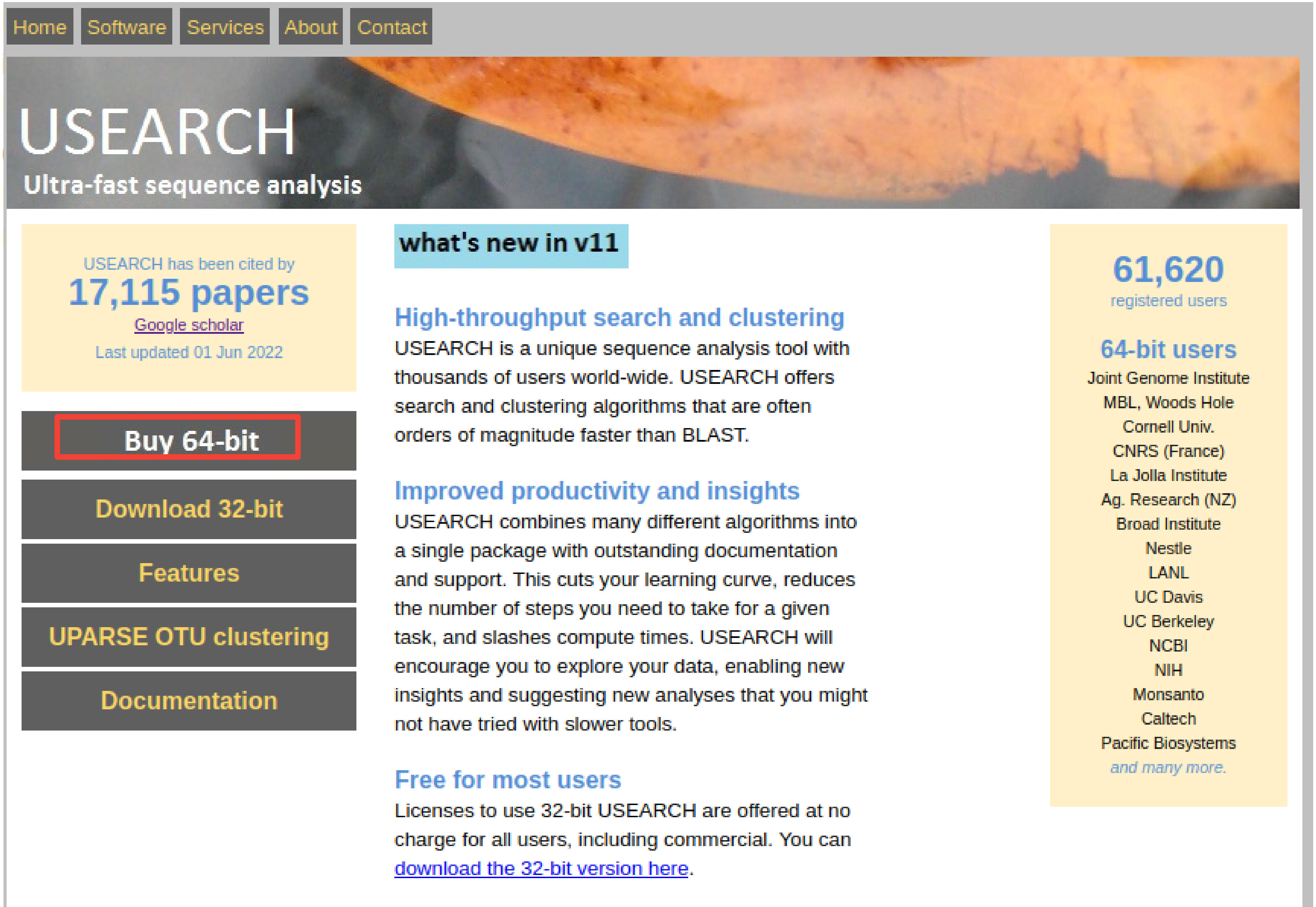
registered users

64-bit users

Joint Genome Institute
MBL, Woods Hole
Cornell Univ.
CNRS (France)
La Jolla Institute
Ag. Research (NZ)
Broad Institute
Nestle
LANL
UC Davis
UC Berkeley
NCBI
NIH
Monsanto
Caltech
Pacific Biosystems
and many more.

Search and clustering orders of magnitude faster than BLAST

RC Edgar - Bioinformatics, 2010 - academic.oup.com



USEARCH has been cited by

17,115 papers

Google scholar

Last updated 01 Jun 2022

what's new in v11

High-throughput search and clustering

USEARCH is a unique sequence analysis tool with thousands of users world-wide. USEARCH offers search and clustering algorithms that are often orders of magnitude faster than BLAST.

Improved productivity and insights

USEARCH combines many different algorithms into a single package with outstanding documentation and support. This cuts your learning curve, reduces the number of steps you need to take for a given task, and slashes compute times. USEARCH will encourage you to explore your data, enabling new insights and suggesting new analyses that you might not have tried with slower tools.

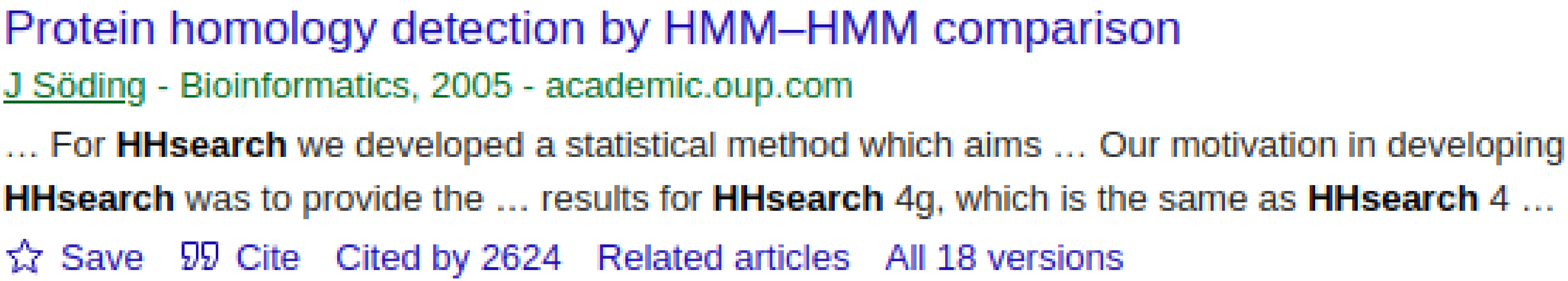
Free for most users

Licenses to use 32-bit USEARCH are offered at no charge for all users, including commercial. You can download the 32-bit version here.

61,620
registered users

64-bit users

Joint Genome Institute
MBL, Woods Hole
Cornell Univ.
CNRS (France)
La Jolla Institute
Ag. Research (NZ)
Broad Institute
Nestle
LANL
UC Davis
UC Berkeley
NCBI
NIH
Monsanto
Caltech
Pacific Biosystems
and many more.

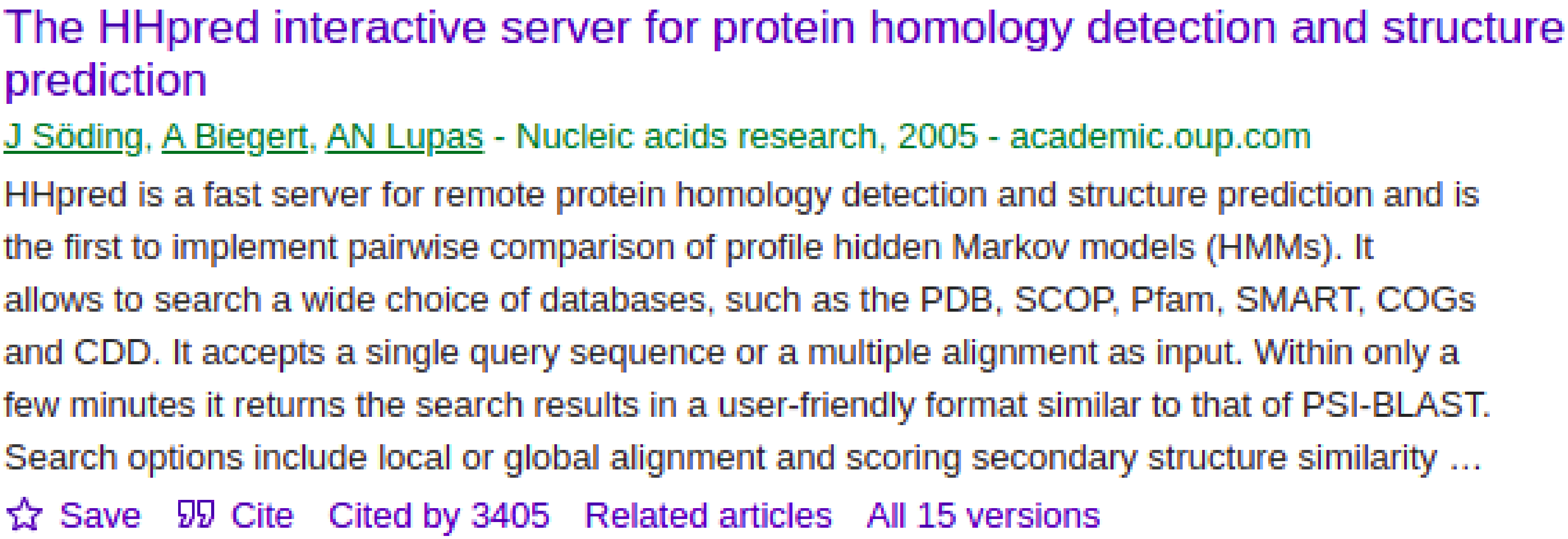


Protein homology detection by HMM–HMM comparison

J Söding - Bioinformatics, 2005 - academic.oup.com

… For **HHsearch** we developed a statistical method which aims … Our motivation in developing **HHsearch** was to provide the … results for **HHsearch** 4g, which is the same as **HHsearch** 4 …

☆ Save ❞ Cite Cited by 2624 Related articles All 18 versions

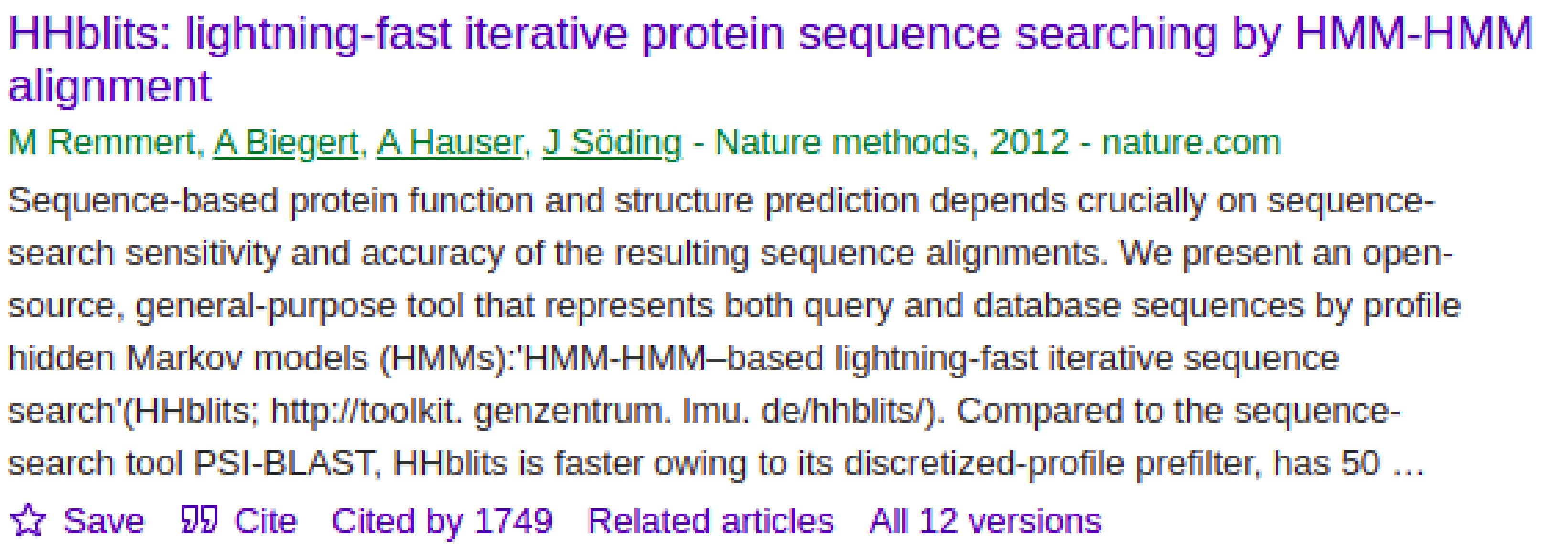


The HHpred interactive server for protein homology detection and structure prediction

J Söding, A Biegert, AN Lupas - Nucleic acids research, 2005 - academic.oup.com

HHpred is a fast server for remote protein homology detection and structure prediction and is the first to implement pairwise comparison of profile hidden Markov models (HMMs). It allows to search a wide choice of databases, such as the PDB, SCOP, Pfam, SMART, COGs and CDD. It accepts a single query sequence or a multiple alignment as input. Within only a few minutes it returns the search results in a user-friendly format similar to that of PSI-BLAST. Search options include local or global alignment and scoring secondary structure similarity …

☆ Save ❞ Cite Cited by 3405 Related articles All 15 versions



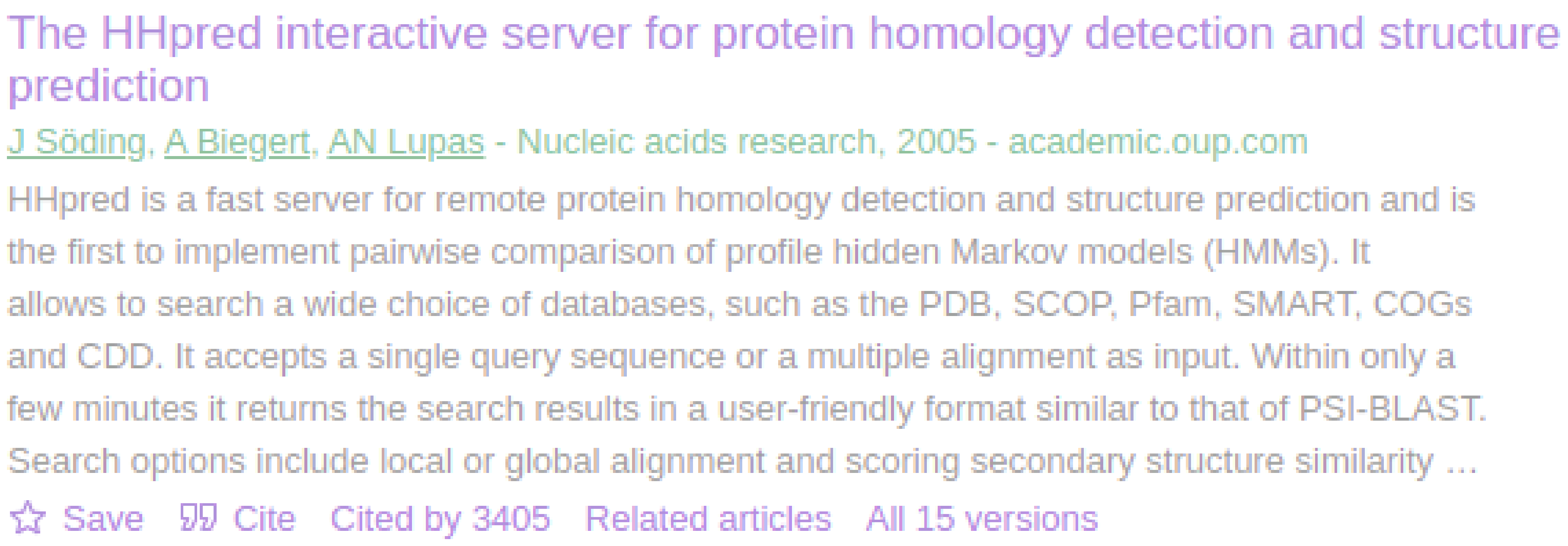
HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment

M Remmert, A Biegert, A Hauser, J Söding - Nature methods, 2012 - nature.com

Sequence-based protein function and structure prediction depends crucially on sequence-search sensitivity and accuracy of the resulting sequence alignments. We present an open-source, general-purpose tool that represents both query and database sequences by profile hidden Markov models (HMMs):'HMM-HMM–based lightning-fast iterative sequence search'(HHblits; http://toolkit. genzentrum. lmu. de/hhblits/). Compared to the sequence-search tool PSI-BLAST, HHblits is faster owing to its discretized-profile prefilter, has 50 …

n in developing

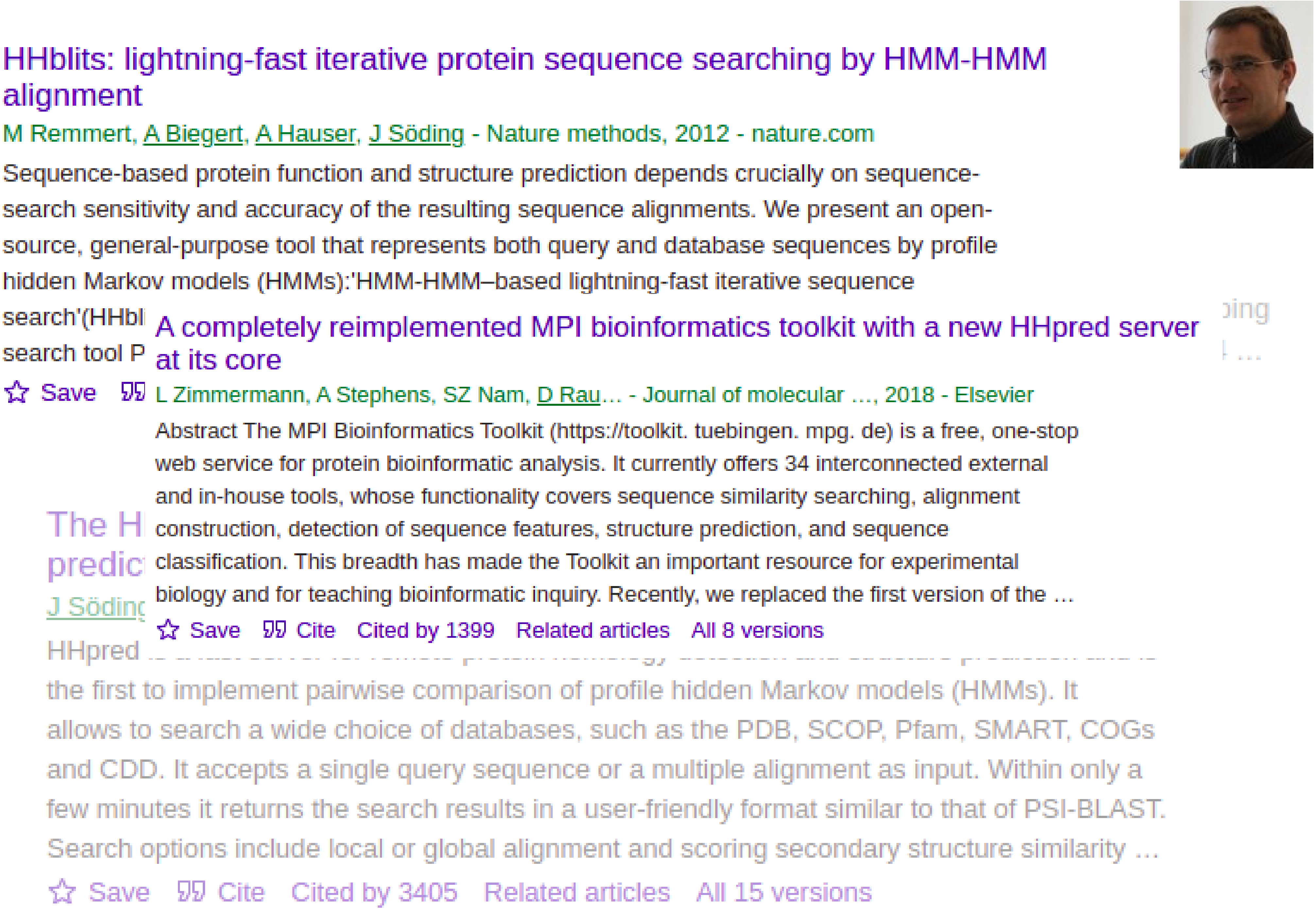
HHsearch 4 …



The HHpred interactive server for protein homology detection and structure prediction

J Söding, A Biegert, AN Lupas - Nucleic acids research, 2005 - academic.oup.com

HHpred is a fast server for remote protein homology detection and structure prediction and is the first to implement pairwise comparison of profile hidden Markov models (HMMs). It allows to search a wide choice of databases, such as the PDB, SCOP, Pfam, SMART, COGs and CDD. It accepts a single query sequence or a multiple alignment as input. Within only a few minutes it returns the search results in a user-friendly format similar to that of PSI-BLAST. Search options include local or global alignment and scoring secondary structure similarity …



HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment

M Remmert, A Biegert, A Hauser, J Söding - Nature methods, 2012 - nature.com

Sequence-based protein function and structure prediction depends crucially on sequence-search sensitivity and accuracy of the resulting sequence alignments. We present an open-source, general-purpose tool that represents both query and database sequences by profile hidden Markov models (HMMs):'HMM-HMM–based lightning-fast iterative sequence search'(HHbl … ɔing search tool P … l …

☆ Save 99

A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core

L Zimmermann, A Stephens, SZ Nam, D Rau… - Journal of molecular …, 2018 - Elsevier

Abstract The MPI Bioinformatics Toolkit (https://toolkit. tuebingen. mpg. de) is a free, one-stop web service for protein bioinformatic analysis. It currently offers 34 interconnected external and in-house tools, whose functionality covers sequence similarity searching, alignment construction, detection of sequence features, structure prediction, and sequence classification. This breadth has made the Toolkit an important resource for experimental biology and for teaching bioinformatic inquiry. Recently, we replaced the first version of the …

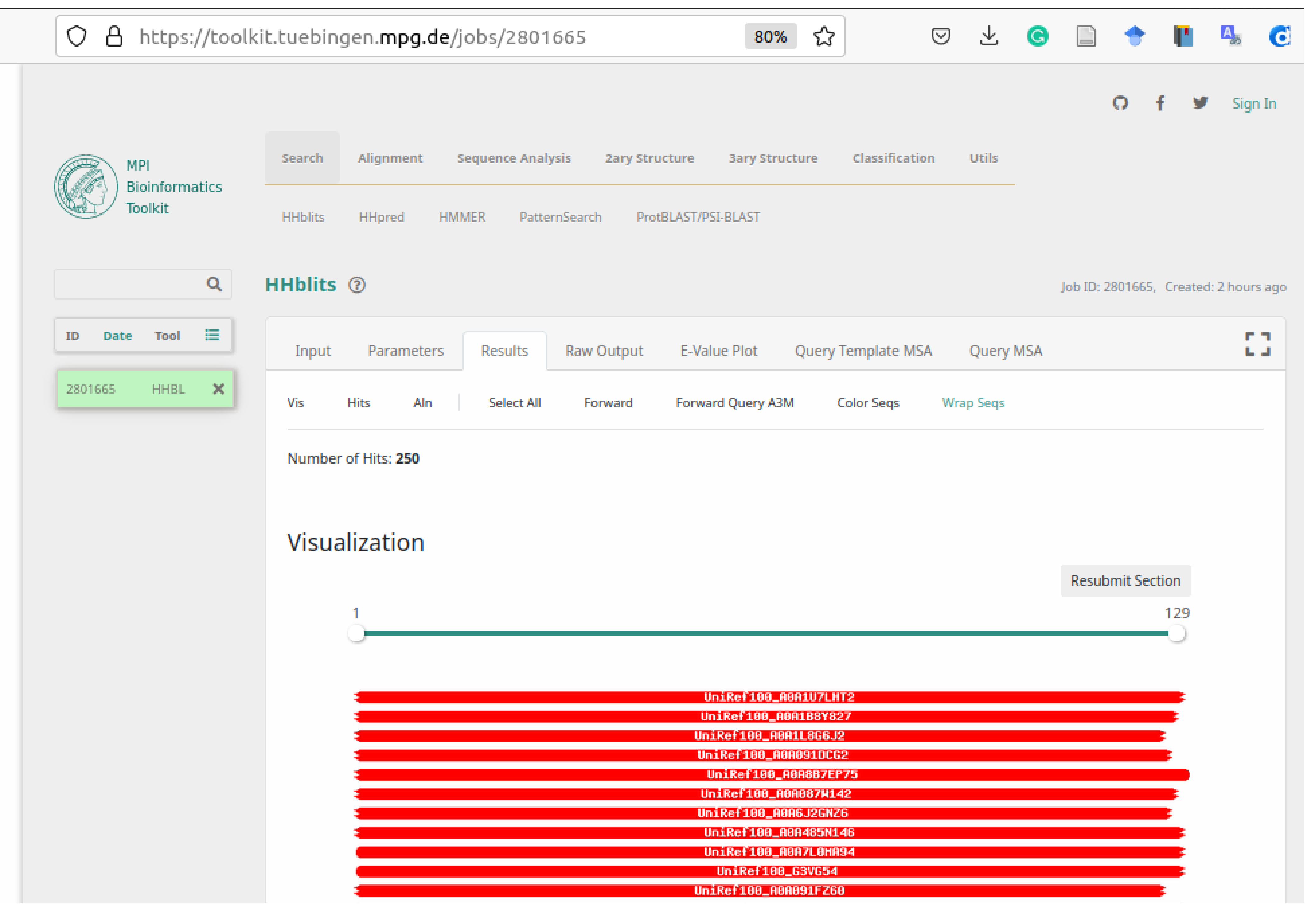
☆ Save 99 Cite Cited by 1399 Related articles All 8 versions

The H… predic…

J Söding…

HHpred … the first to implement pairwise comparison of profile hidden Markov models (HMMs). It allows to search a wide choice of databases, such as the PDB, SCOP, Pfam, SMART, COGs and CDD. It accepts a single query sequence or a multiple alignment as input. Within only a few minutes it returns the search results in a user-friendly format similar to that of PSI-BLAST. Search options include local or global alignment and scoring secondary structure similarity …

☆ Save 99 Cite Cited by 3405 Related articles All 15 versions



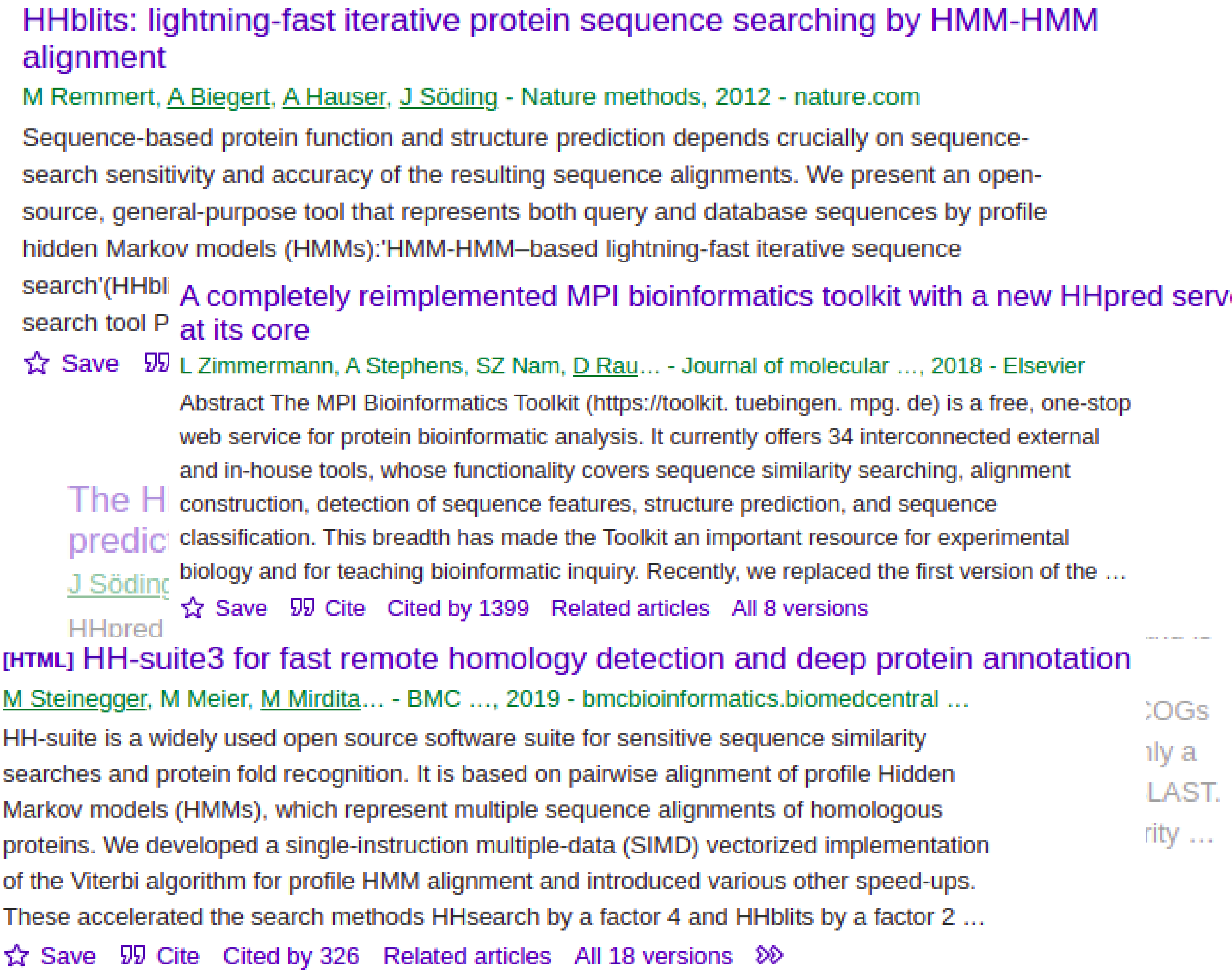
HHblits

Job ID: 2801665, Created: 2 hours ago

Number of Hits: **250**

Visualization

UniRef100_A0A1U7LHT2
UniRef100_A0A1B8Y827
UniRef100_A0A1L8G6J2
UniRef100_A0A091DCG2
UniRef100_A0A8B7EP75
UniRef100_A0A087W142
UniRef100_A0A6J2GNZ6
UniRef100_A0A485N146
UniRef100_A0A7L0MA94
UniRef100_G3VG54
UniRef100_A0A091FZ60



HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment

M Remmert, A Biegert, A Hauser, J Söding - Nature methods, 2012 - nature.com

Sequence-based protein function and structure prediction depends crucially on sequence-search sensitivity and accuracy of the resulting sequence alignments. We present an open-source, general-purpose tool that represents both query and database sequences by profile hidden Markov models (HMMs):'HMM-HMM–based lightning-fast iterative sequence search'(HHbl … oing

search tool P … l …

☆ Save 99

A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core

L Zimmermann, A Stephens, SZ Nam, D Rau… - Journal of molecular …, 2018 - Elsevier

Abstract The MPI Bioinformatics Toolkit (https://toolkit. tuebingen. mpg. de) is a free, one-stop web service for protein bioinformatic analysis. It currently offers 34 interconnected external and in-house tools, whose functionality covers sequence similarity searching, alignment construction, detection of sequence features, structure prediction, and sequence classification. This breadth has made the Toolkit an important resource for experimental biology and for teaching bioinformatic inquiry. Recently, we replaced the first version of the …

☆ Save 99 Cite Cited by 1399 Related articles All 8 versions

The H

predic

J Söding

HHpred

[HTML] HH-suite3 for fast remote homology detection and deep protein annotation

M Steinegger, M Meier, M Mirdita… - BMC …, 2019 - bmcbioinformatics.biomedcentral …

HH-suite is a widely used open source software suite for sensitive sequence similarity searches and protein fold recognition. It is based on pairwise alignment of profile Hidden Markov models (HMMs), which represent multiple sequence alignments of homologous proteins. We developed a single-instruction multiple-data (SIMD) vectorized implementation of the Viterbi algorithm for profile HMM alignment and introduced various other speed-ups. These accelerated the search methods HHsearch by a factor 4 and HHblits by a factor 2 …

☆ Save 99 Cite Cited by 326 Related articles All 18 versions »

OGs

ly a

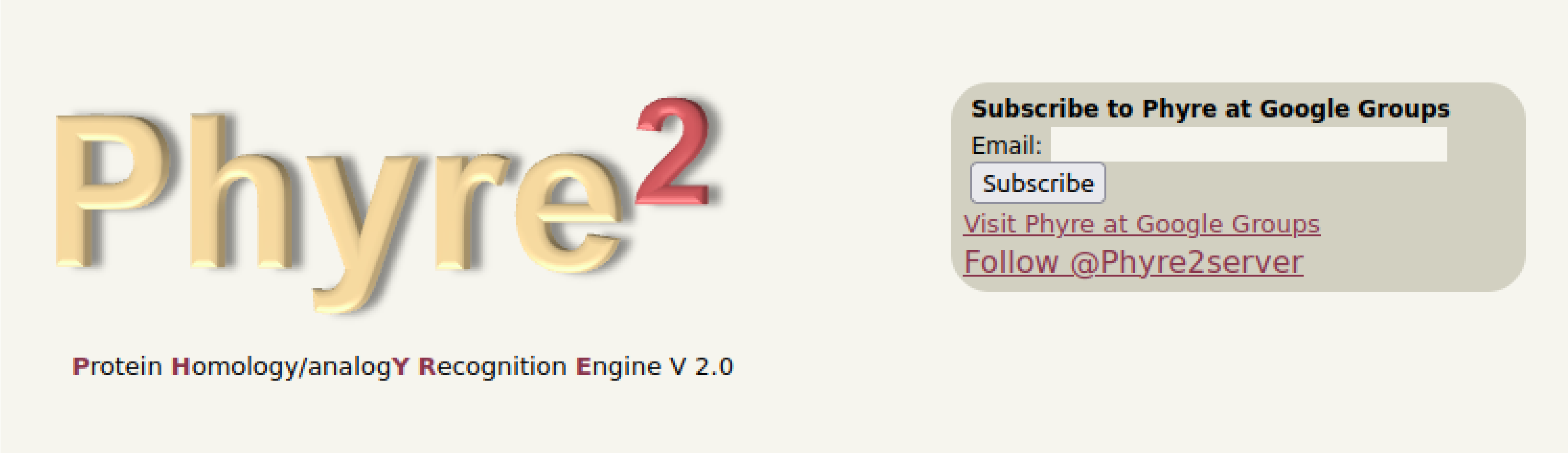
LAST.

rity …

The Phyre2 web portal for protein modeling, prediction and analysis LA Kelley, S Mezulis, CM Yates, MN Wass, MJE Sternberg Nature protocols 10 (6), 845-858	7627	2015
Protein structure prediction on the Web: a case study using the Phyre server LA Kelley, MJE Sternberg Nature protocols 4 (3), 363-371	4983	2009

https://www.youtube.com/watch?v=Adm8JQZMmj4&t=1s

https://www.youtube.com/watch?v=XoYHTF6XSY0



Protein Homology/analogY Recognition Engine V 2.0

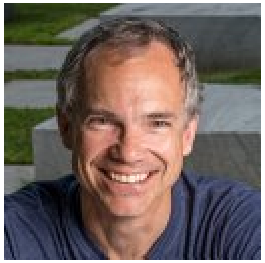
Subscribe to Phyre at Google Groups

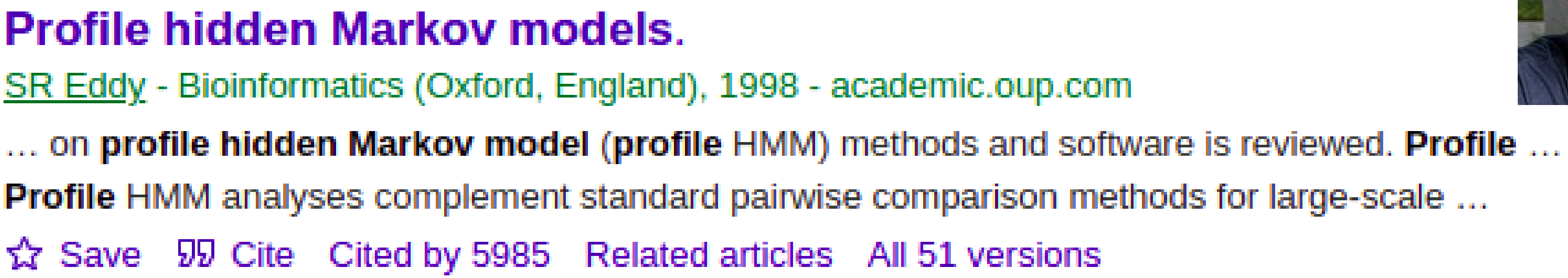
Email:

Visit Phyre at Google Groups

Follow @Phyre2server

Sean R. Eddy





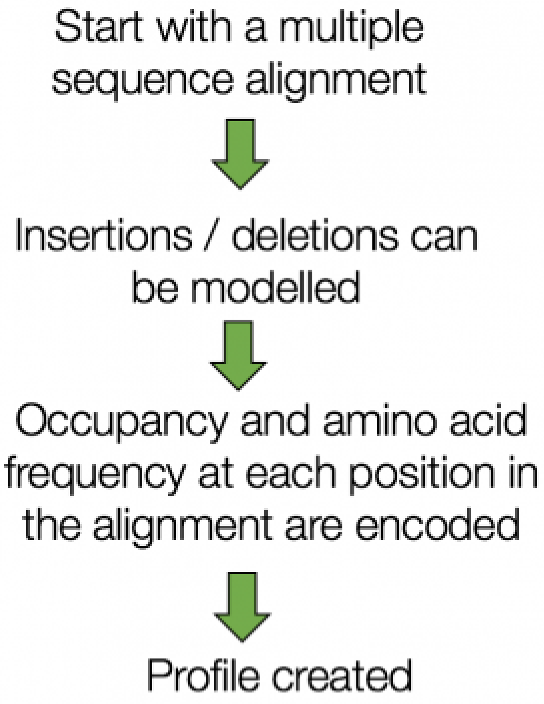
Profile hidden Markov models.

SR Eddy - Bioinformatics (Oxford, England), 1998 - academic.oup.com

… on **profile hidden Markov model** (**profile** HMM) methods and software is reviewed. **Profile** …
Profile HMM analyses complement standard pairwise comparison methods for large-scale …

☆ Save 99 Cite Cited by 5985 Related articles All 51 versions

HMMER biosequence analysis using profile hidden Markov models

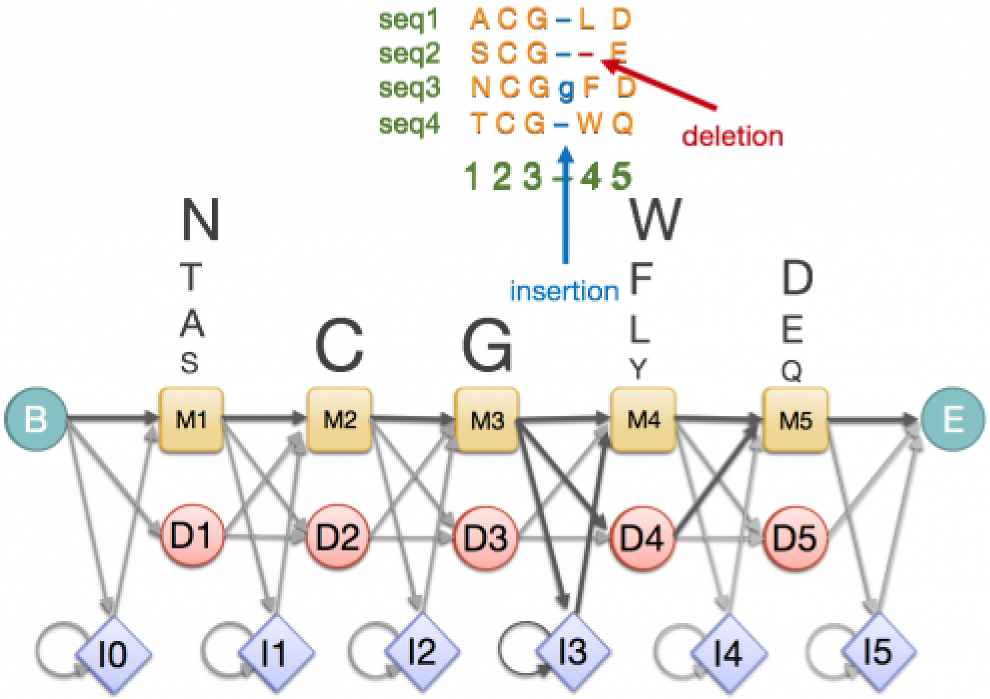


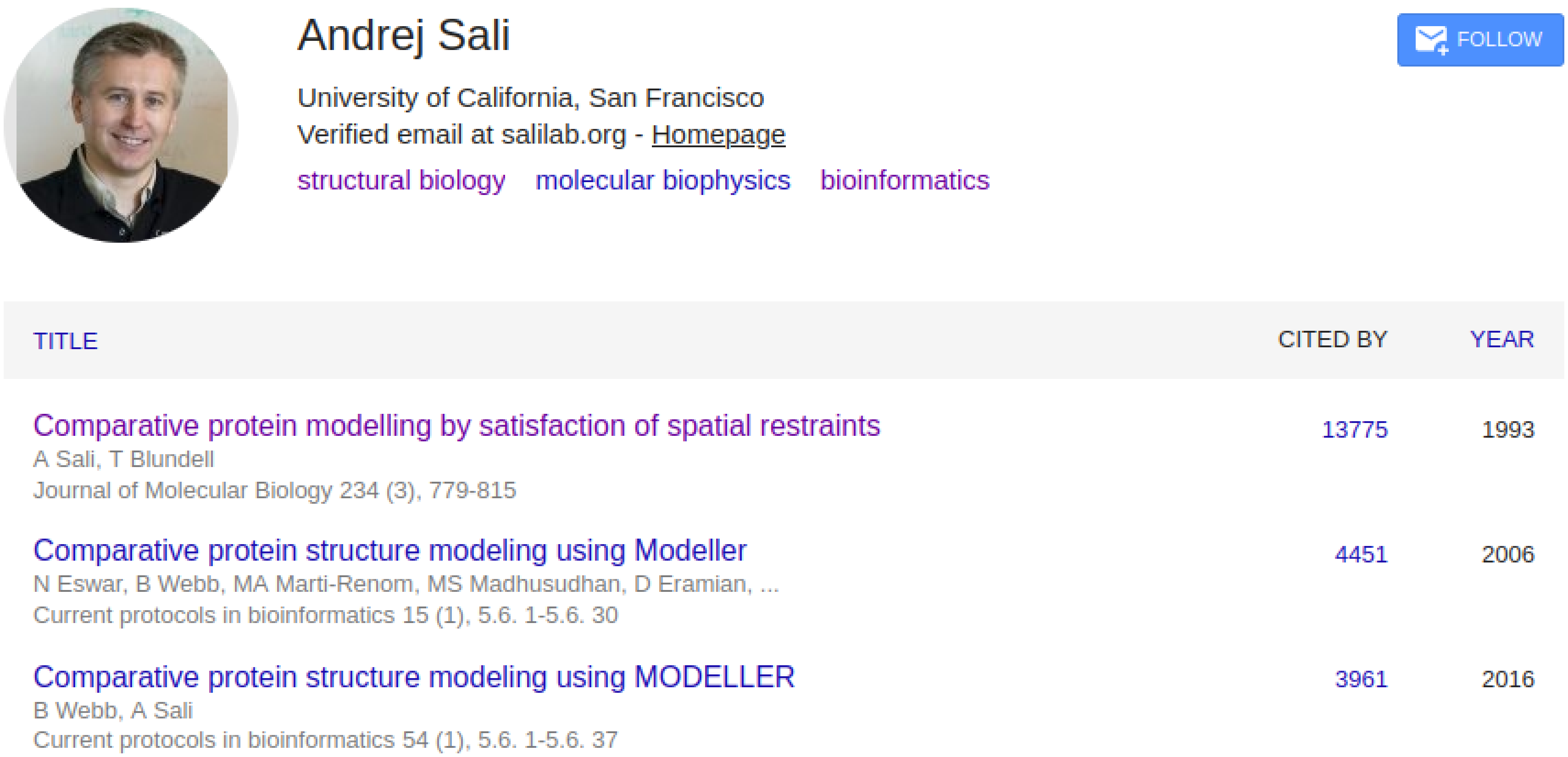
Start with a multiple sequence alignment

Insertions / deletions can be modelled

Occupancy and amino acid frequency at each position in the alignment are encoded

Profile created



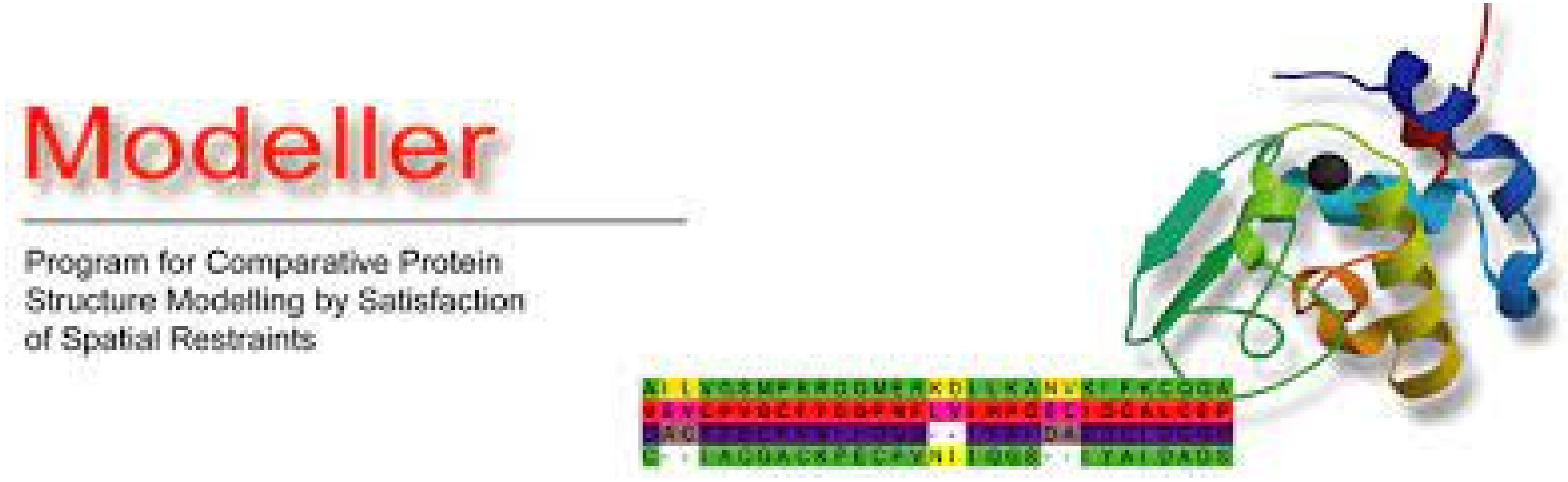


Andrej Sali

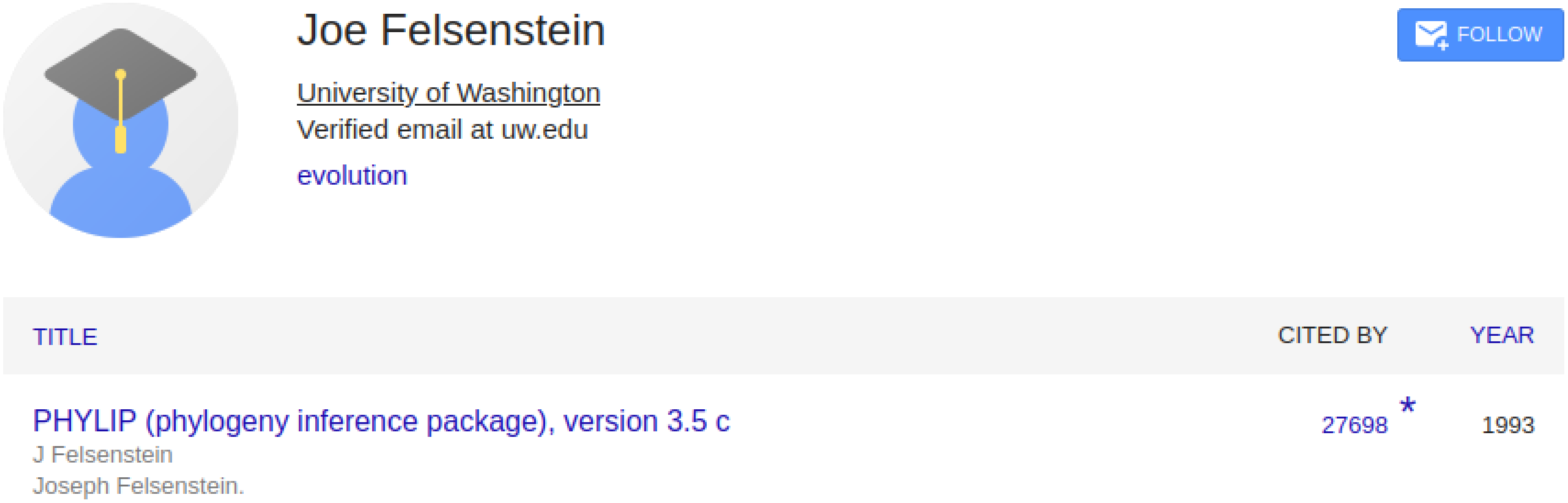
University of California, San Francisco
Verified email at salilab.org - Homepage

structural biology   molecular biophysics   bioinformatics

TITLE	CITED BY	YEAR
Comparative protein modelling by satisfaction of spatial restraints A Sali, T Blundell Journal of Molecular Biology 234 (3), 779-815	13775	1993
Comparative protein structure modeling using Modeller N Eswar, B Webb, MA Marti-Renom, MS Madhusudhan, D Eramian, ... Current protocols in bioinformatics 15 (1), 5.6. 1-5.6. 30	4451	2006
Comparative protein structure modeling using MODELLER B Webb, A Sali Current protocols in bioinformatics 54 (1), 5.6. 1-5.6. 37	3961	2016


Modeller
Program for Comparative Protein
Structure Modelling by Satisfaction
of Spatial Restraints

https://www.youtube.com/watch?v=Zb98mmfnsvg

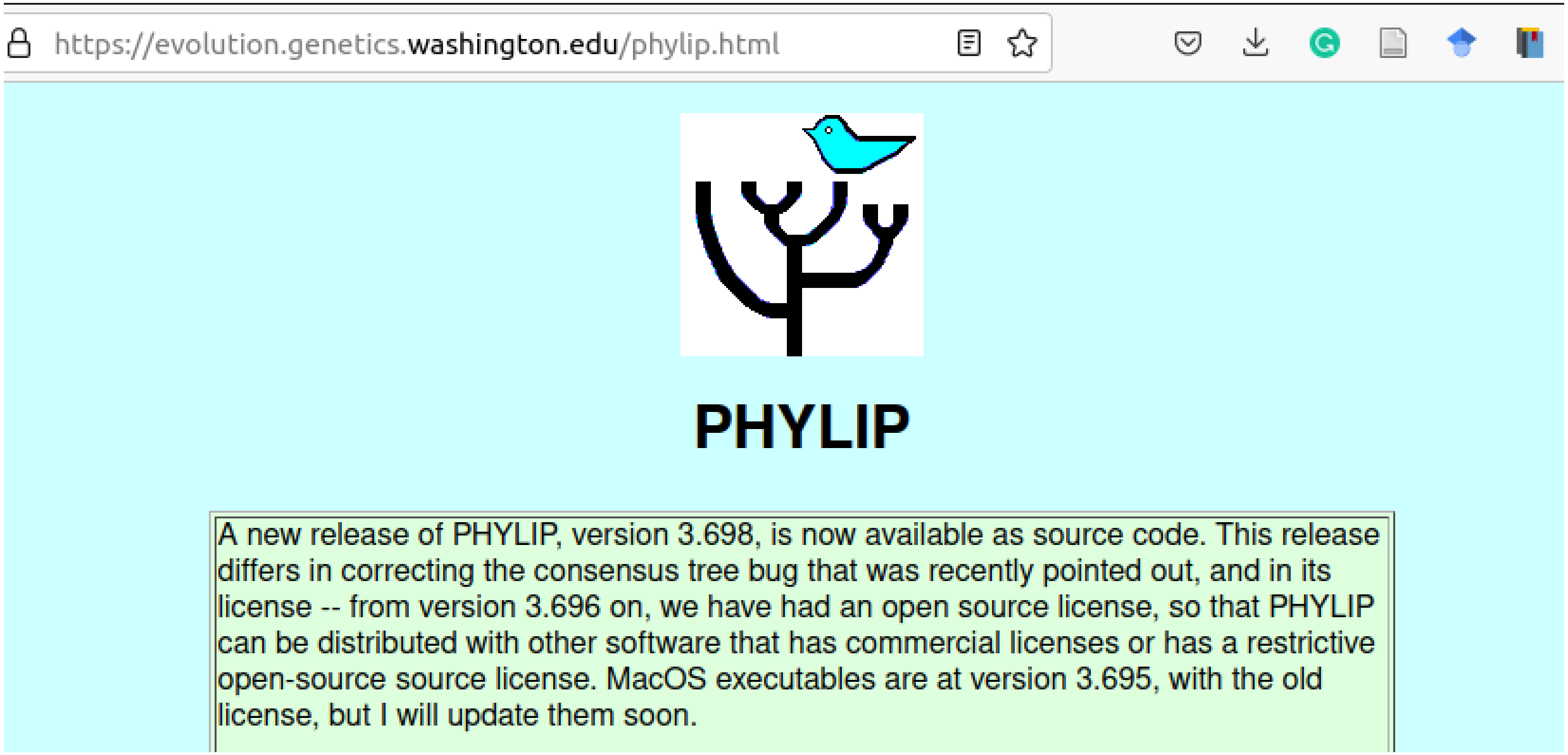


Joe Felsenstein

University of Washington
Verified email at uw.edu

evolution

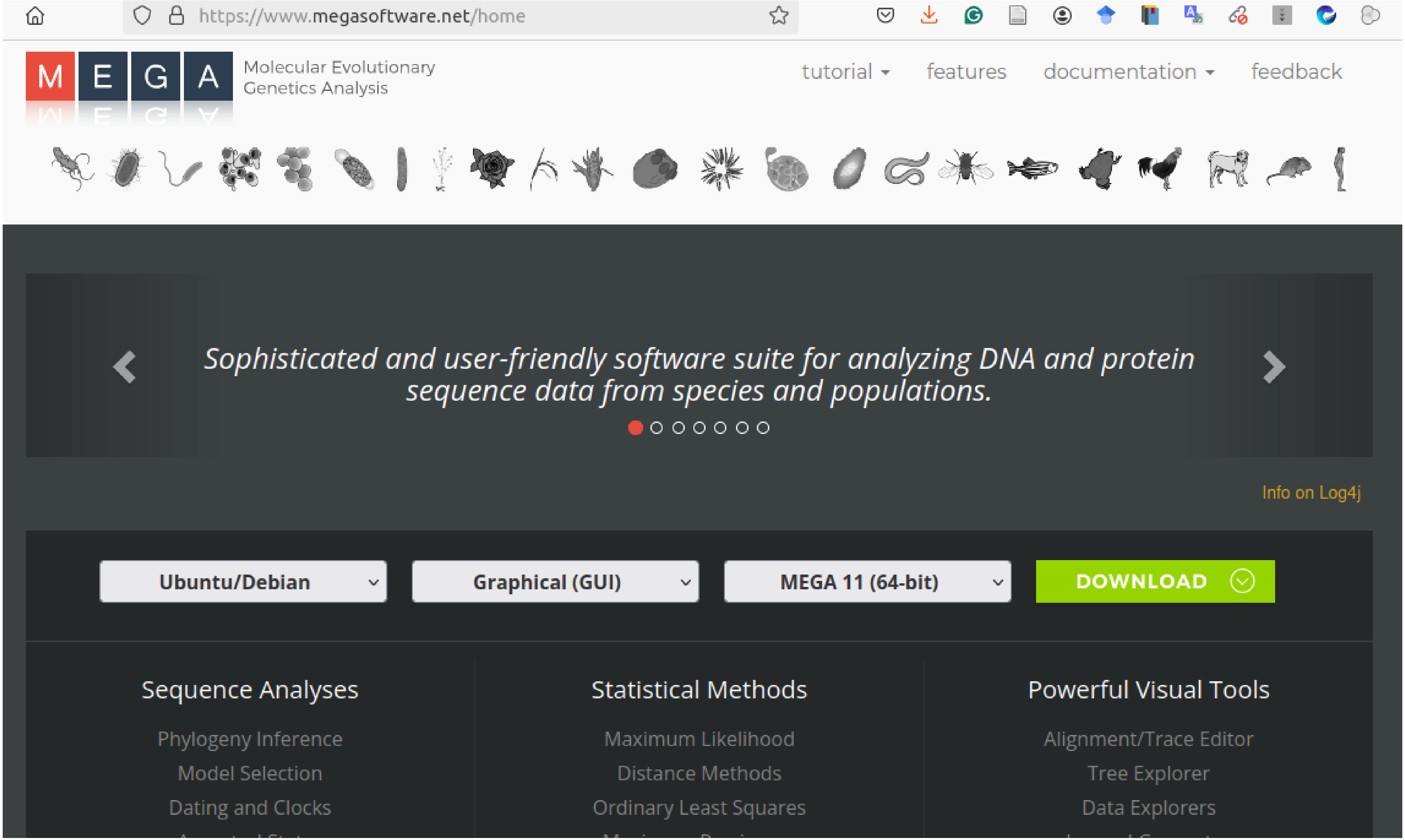
TITLE	CITED BY	YEAR
PHYLIP (phylogeny inference package), version 3.5 c J Felsenstein Joseph Felsenstein.	27698 *	1993



https://evolution.genetics.washington.edu/phylip.html

PHYLIP

A new release of PHYLIP, version 3.698, is now available as source code. This release differs in correcting the consensus tree bug that was recently pointed out, and in its license -- from version 3.696 on, we have had an open source license, so that PHYLIP can be distributed with other software that has commercial licenses or has a restrictive open-source source license. MacOS executables are at version 3.695, with the old license, but I will update them soon.



https://www.megasoftware.net/home
MEGA
Molecular Evolutionary Genetics Analysis
tutorial
features
documentation
feedback

Sophisticated and user-friendly software suite for analyzing DNA and protein sequence data from species and populations.

Ubuntu/Debian

Graphical (GUI)

MEGA 11 (64-bit)

Sequence Analyses

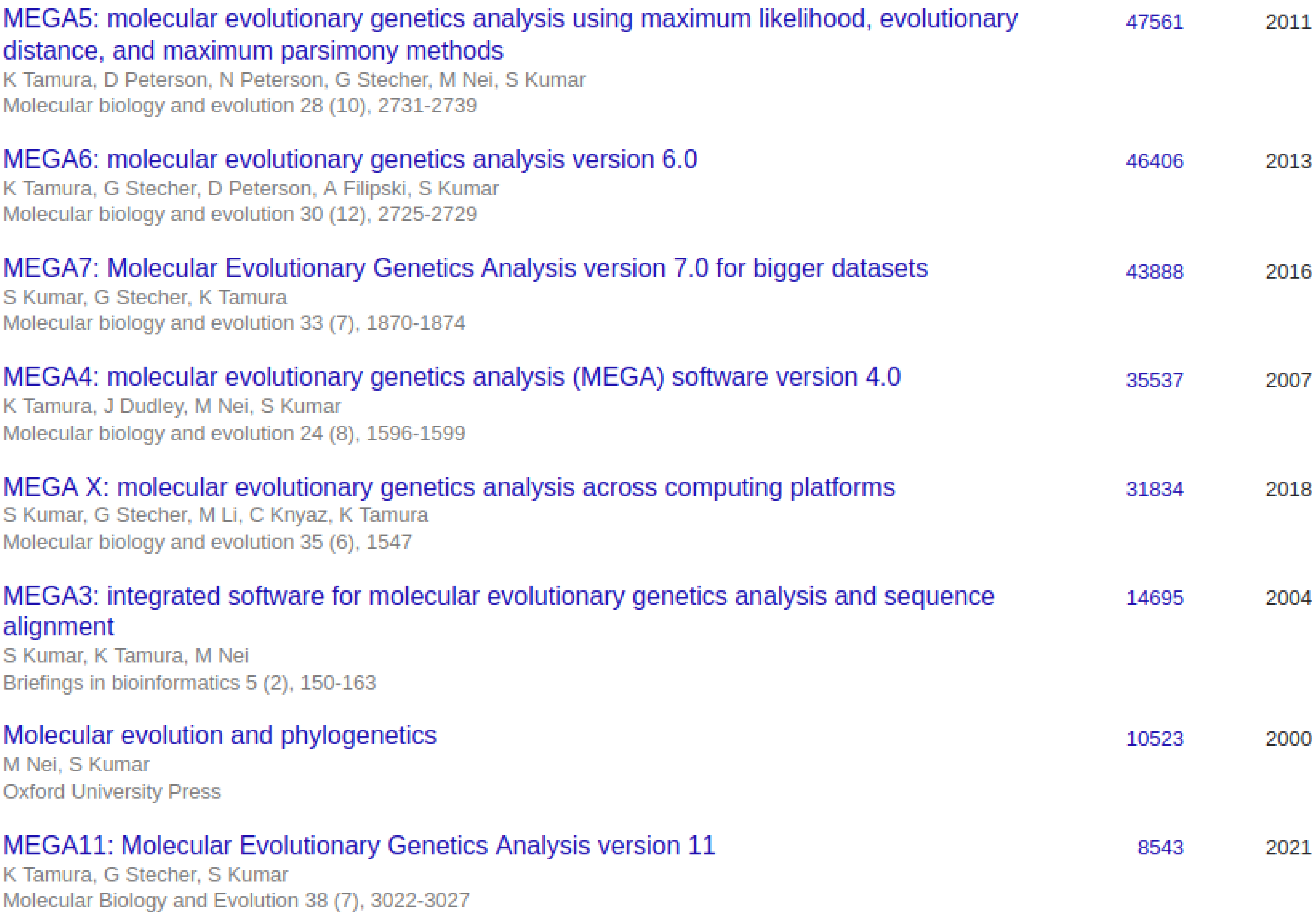
- Phylogeny Inference
- Model Selection
- Dating and Clocks

Statistical Methods

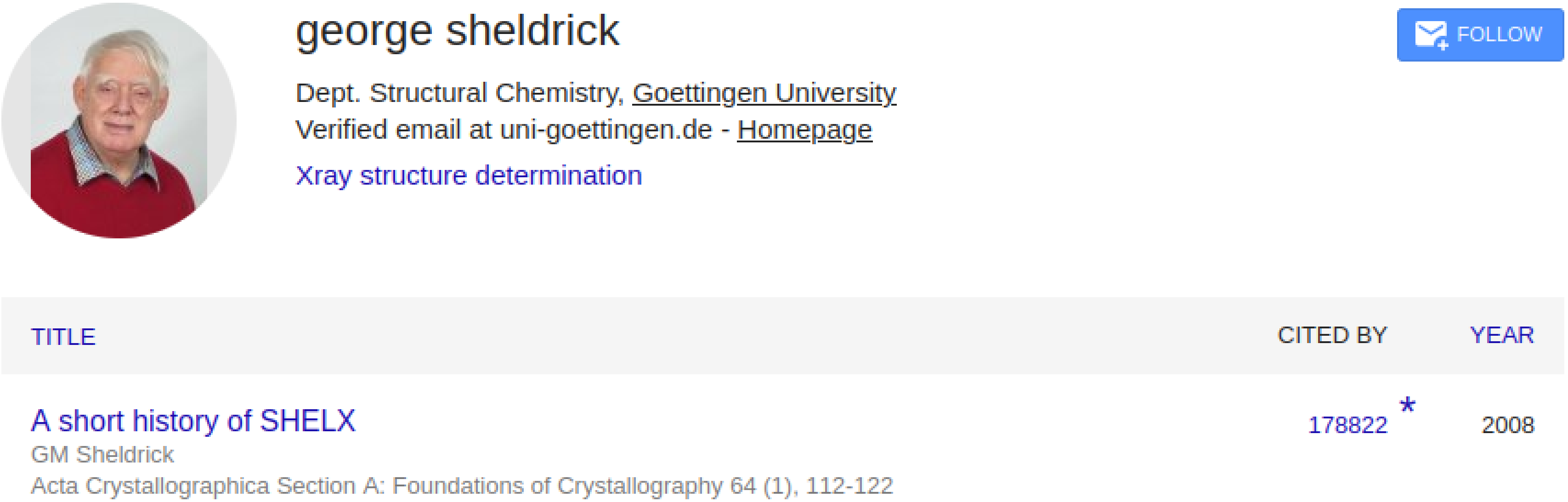
- Maximum Likelihood
- Distance Methods
- Ordinary Least Squares

Powerful Visual Tools

- Alignment/Trace Editor
- Tree Explorer
- Data Explorers



MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods K Tamura, D Peterson, N Peterson, G Stecher, M Nei, S Kumar Molecular biology and evolution 28 (10), 2731-2739	47561	2011
MEGA6: molecular evolutionary genetics analysis version 6.0 K Tamura, G Stecher, D Peterson, A Filipski, S Kumar Molecular biology and evolution 30 (12), 2725-2729	46406	2013
MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets S Kumar, G Stecher, K Tamura Molecular biology and evolution 33 (7), 1870-1874	43888	2016
MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0 K Tamura, J Dudley, M Nei, S Kumar Molecular biology and evolution 24 (8), 1596-1599	35537	2007
MEGA X: molecular evolutionary genetics analysis across computing platforms S Kumar, G Stecher, M Li, C Knyaz, K Tamura Molecular biology and evolution 35 (6), 1547	31834	2018
MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment S Kumar, K Tamura, M Nei Briefings in bioinformatics 5 (2), 150-163	14695	2004
Molecular evolution and phylogenetics M Nei, S Kumar Oxford University Press	10523	2000
MEGA11: Molecular Evolutionary Genetics Analysis version 11 K Tamura, G Stecher, S Kumar Molecular Biology and Evolution 38 (7), 3022-3027	8543	2021

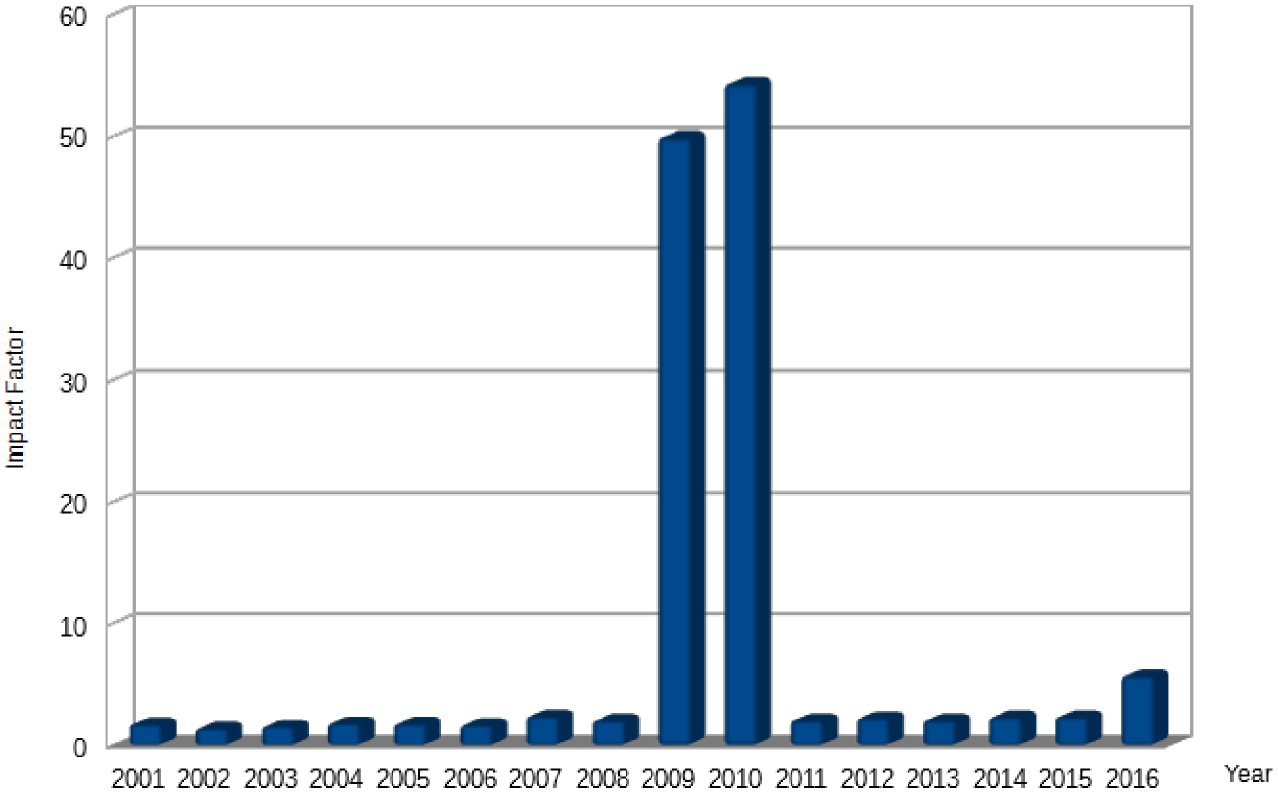


george sheldrick

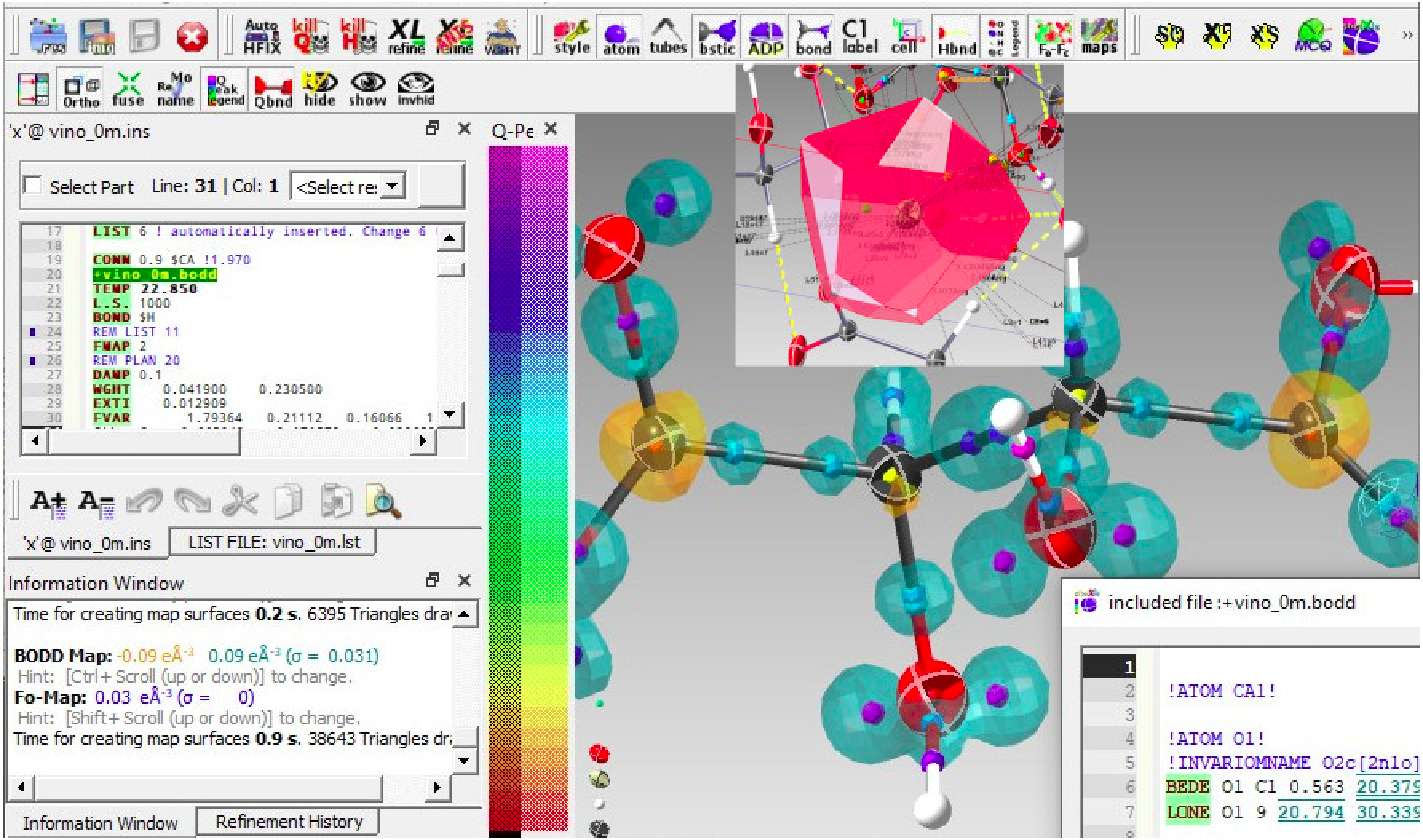
Dept. Structural Chemistry, Goettingen University
Verified email at uni-goettingen.de - Homepage

Xray structure determination

TITLE	CITED BY	YEAR
A short history of SHELX GM Sheldrick Acta Crystallographica Section A: Foundations of Crystallography 64 (1), 112-122	178822 *	2008



SHELX



'x'@ vino_0m.ins

☐ Select Part Line: **31** | Col: **1**

```
LIST 6 ! automatically inserted. Change 6 !

CONN 0.9 $CA !1.970
+vino_0m.bodd
TEMP 22.850
L.S. 1000
BOND $H
REM LIST 11
FMAP 2
REM PLAN 20
DAMP 0.1
WGHT      0.041900     0.230500
EXTI      0.012909
FVAR         1.79364    0.21112    0.16066   1.
```

Information Window

Time for creating map surfaces **0.2 s.** 6395 Triangles dra

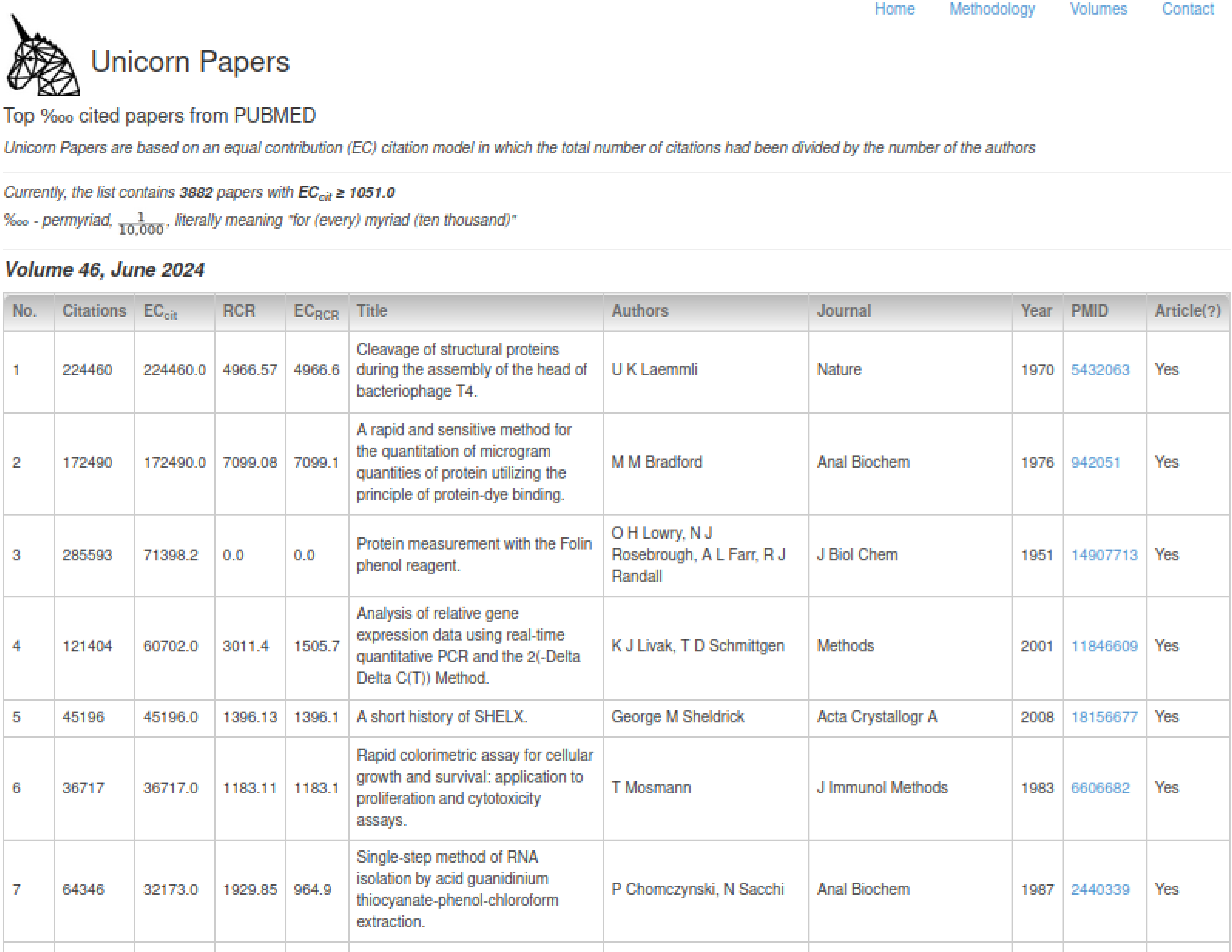
BODD Map: -0.09 eÅ^{-3} 0.09 eÅ^{-3} (σ = 0.031)
Hint: [Ctrl+Scroll (up or down)] to change.
Fo-Map: 0.03 eÅ^{-3} (σ = 0)
Hint: [Shift+Scroll (up or down)] to change.
Time for creating map surfaces **0.9 s.** 38643 Triangles dr

included file :+vino_0m.bodd

```

!ATOM CA1!

!ATOM O1!
!INVARIOMNAME O2c[2n1o]
BEDE O1 C1 0.563 20.379
LONE O1 9 20.794 30.339
```



Unicorn Papers

Top ‱ cited papers from PUBMED

Unicorn Papers are based on an equal contribution (EC) citation model in which the total number of citations had been divided by the number of the authors

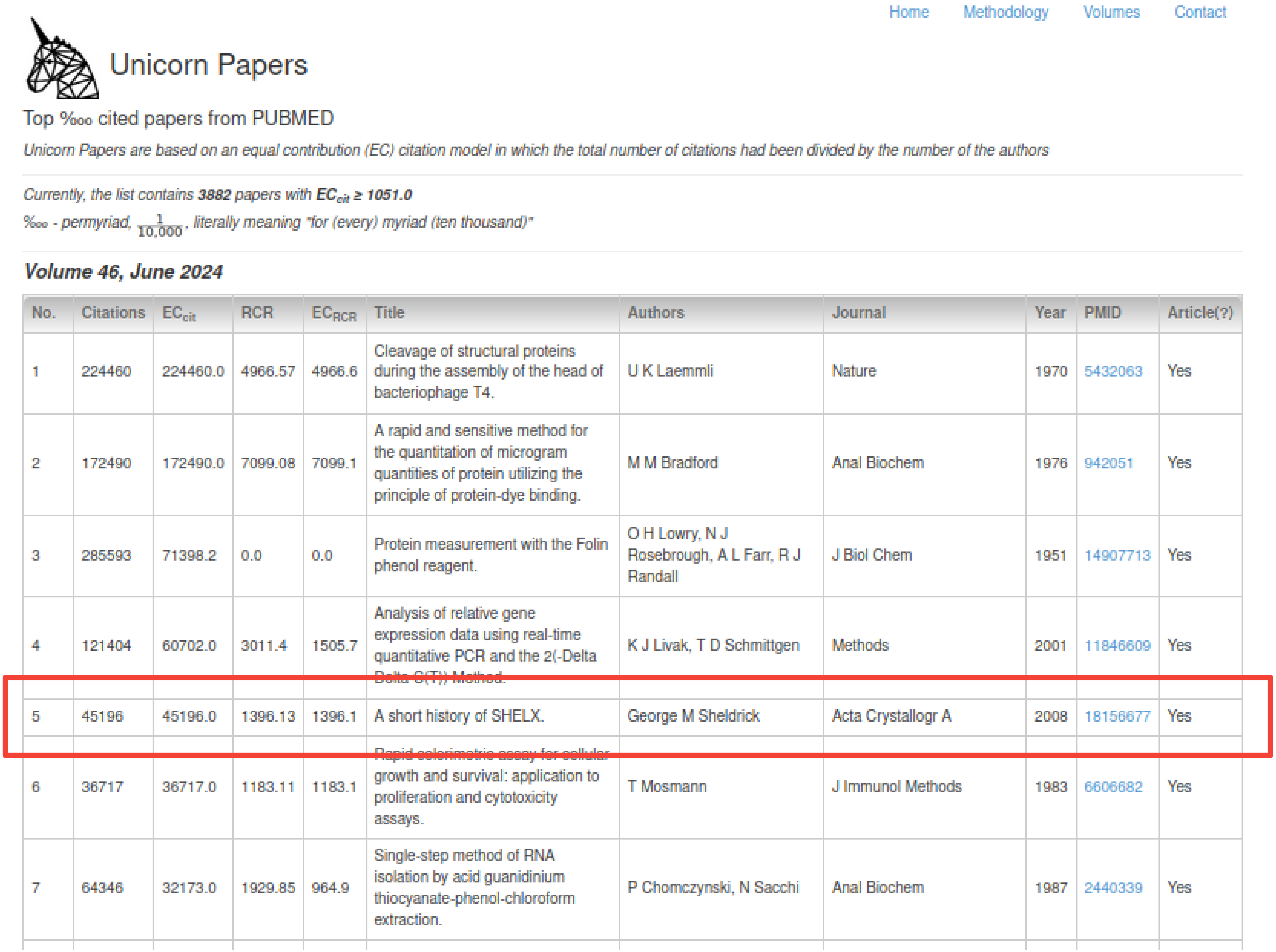
Currently, the list contains ***3882*** *papers with* $\boldsymbol{EC_{cit} \geq 1051.0}$

‱ - permyriad, $\frac{1}{10{,}000}$*, literally meaning "for (every) myriad (ten thousand)"*

Volume 46, June 2024

No.	Citations	EC_{cit}	RCR	EC_{RCR}	Title	Authors	Journal	Year	PMID	Article(?)
1	224460	224460.0	4966.57	4966.6	Cleavage of structural proteins during the assembly of the head of bacteriophage T4.	U K Laemmli	Nature	1970	5432063	Yes
2	172490	172490.0	7099.08	7099.1	A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.	M M Bradford	Anal Biochem	1976	942051	Yes
3	285593	71398.2	0.0	0.0	Protein measurement with the Folin phenol reagent.	O H Lowry, N J Rosebrough, A L Farr, R J Randall	J Biol Chem	1951	14907713	Yes
4	121404	60702.0	3011.4	1505.7	Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.	K J Livak, T D Schmittgen	Methods	2001	11846609	Yes
5	45196	45196.0	1396.13	1396.1	A short history of SHELX.	George M Sheldrick	Acta Crystallogr A	2008	18156677	Yes
6	36717	36717.0	1183.11	1183.1	Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays.	T Mosmann	J Immunol Methods	1983	6606682	Yes
7	64346	32173.0	1929.85	964.9	Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.	P Chomczynski, N Sacchi	Anal Biochem	1987	2440339	Yes

http://unicornpapers.org



Unicorn Papers

Top ‱ cited papers from PUBMED

Unicorn Papers are based on an equal contribution (EC) citation model in which the total number of citations had been divided by the number of the authors

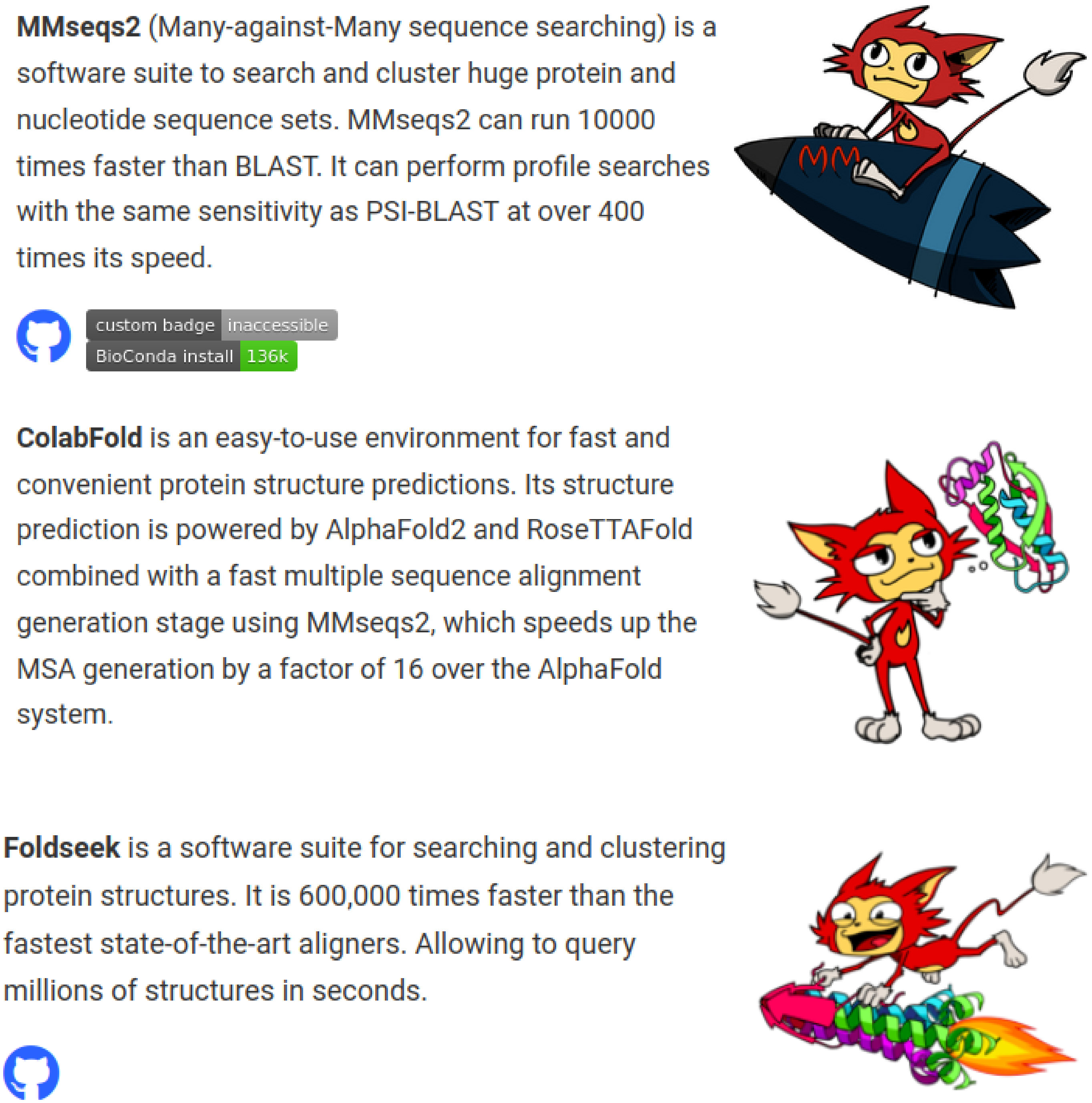
Currently, the list contains ***3882*** *papers with* ***$EC_{cit} \geq 1051.0$***

‱ - permyriad, $\frac{1}{10,000}$, literally meaning "for (every) myriad (ten thousand)"

Volume 46, June 2024

No.	Citations	EC_{cit}	RCR	EC_{RCR}	Title	Authors	Journal	Year	PMID	Article(?)
1	224460	224460.0	4966.57	4966.6	Cleavage of structural proteins during the assembly of the head of bacteriophage T4.	U K Laemmli	Nature	1970	5432063	Yes
2	172490	172490.0	7099.08	7099.1	A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.	M M Bradford	Anal Biochem	1976	942051	Yes
3	285593	71398.2	0.0	0.0	Protein measurement with the Folin phenol reagent.	O H Lowry, N J Rosebrough, A L Farr, R J Randall	J Biol Chem	1951	14907713	Yes
4	121404	60702.0	3011.4	1505.7	Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.	K J Livak, T D Schmittgen	Methods	2001	11846609	Yes
5	45196	45196.0	1396.13	1396.1	A short history of SHELX.	George M Sheldrick	Acta Crystallogr A	2008	18156677	Yes
6	36717	36717.0	1183.11	1183.1	Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays.	T Mosmann	J Immunol Methods	1983	6606682	Yes
7	64346	32173.0	1929.85	964.9	Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.	P Chomczynski, N Sacchi	Anal Biochem	1987	2440339	Yes

http://unicornpapers.org



MMseqs2 (Many-against-Many sequence searching) is a software suite to search and cluster huge protein and nucleotide sequence sets. MMseqs2 can run 10000 times faster than BLAST. It can perform profile searches with the same sensitivity as PSI-BLAST at over 400 times its speed.

custom badge | inaccessible

BioConda install | 136k

ColabFold is an easy-to-use environment for fast and convenient protein structure predictions. Its structure prediction is powered by AlphaFold2 and RoseTTAFold combined with a fast multiple sequence alignment generation stage using MMseqs2, which speeds up the MSA generation by a factor of 16 over the AlphaFold system.

Foldseek is a software suite for searching and clustering protein structures. It is 600,000 times faster than the fastest state-of-the-art aligners. Allowing to query millions of structures in seconds.

Martin Steinegger

Extra lecture on YouTube

Martin Steinegger

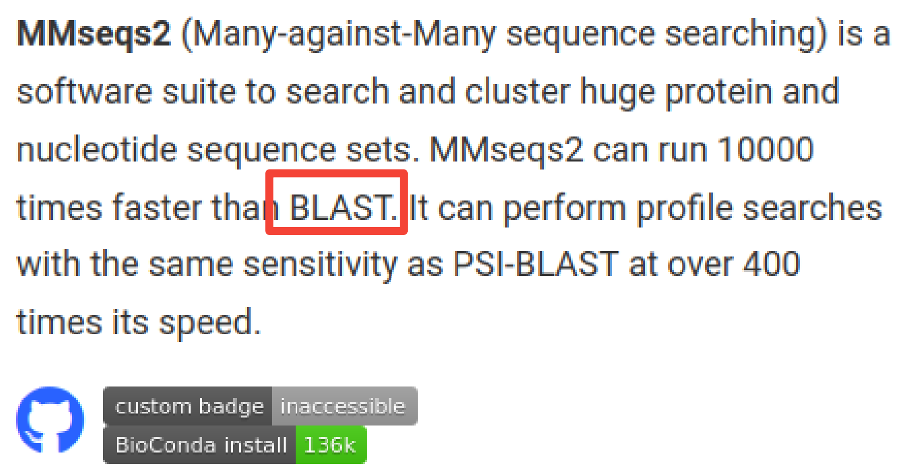
MMseqs2 (Many-against-Many sequence searching) is a software suite to search and cluster huge protein and nucleotide sequence sets. MMseqs2 can run 10000 times faster than BLAST. It can perform profile searches with the same sensitivity as PSI-BLAST at over 400 times its speed.

custom badge | inaccessible

BioConda install | 136k

ColabFold is an easy-to-use environment for fast and convenient protein structure predictions. Its structure prediction is powered by AlphaFold2 and RoseTTAFold combined with a fast multiple sequence alignment generation stage using MMseqs2, which speeds up the MSA generation by a factor of 16 over the AlphaFold system.

Foldseek is a software suite for searching and clustering protein structures. It is 600,000 times faster than the fastest state-of-the-art aligners. Allowing to query millions of structures in seconds.

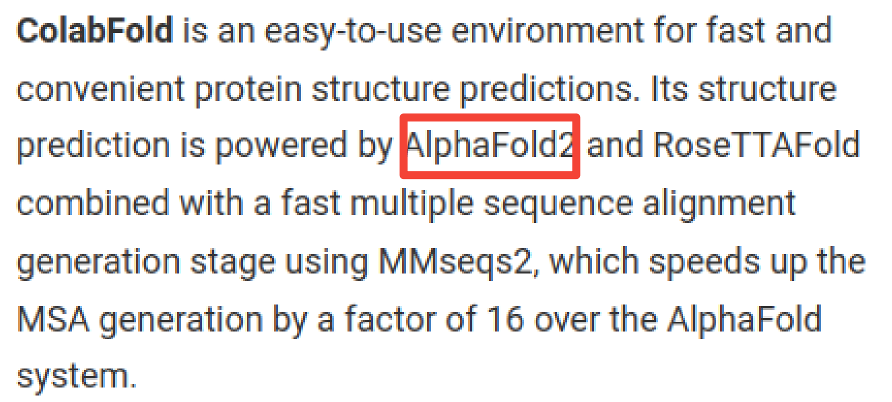


MMseqs2 (Many-against-Many sequence searching) is a software suite to search and cluster huge protein and nucleotide sequence sets. MMseqs2 can run 10000 times faster than BLAST. It can perform profile searches with the same sensitivity as PSI-BLAST at over 400 times its speed.

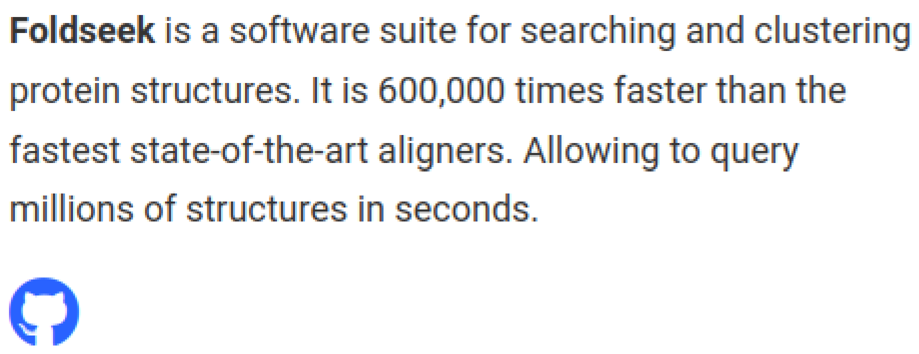
custom badge inaccessible

BioConda install 136k

Martin Steinegger



ColabFold is an easy-to-use environment for fast and convenient protein structure predictions. Its structure prediction is powered by AlphaFold2 and RoseTTAFold combined with a fast multiple sequence alignment generation stage using MMseqs2, which speeds up the MSA generation by a factor of 16 over the AlphaFold system.



Foldseek is a software suite for searching and clustering protein structures. It is 600,000 times faster than the fastest state-of-the-art aligners. Allowing to query millions of structures in seconds.

Extra lecture

YouTube

https://www.youtube.com/watch?v=k5Rbi22TtOA

Bioinformatics (especially large scale projects usually require serious computer resources)

AlphaFold installed locally ~3 TB

AF2DB >50TB

PDB ~1TB

UniProt - just TrEMBL 104 GB

...

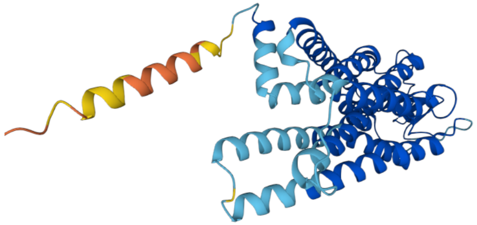
Bioinformatics (especially large scale projects usually require serious computer resources)

ALPHAFOLD
PROTEIN STRUCTURE
DATABASE

25TB (tar.gz)

3 x 214M files

AlphaFold installed locally ~3 TB
AF2DB >50TB
PDB ~1TB
UniProt - just TrEMBL 104 GB
...



ALPHAFOLD
PROTEIN STRUCTURE
DATABASE

Google Cloud

25TB (tar.gz) **~ 3 weeks to download**

1,015,797 sharded proteome tar files
containing
from 1 to 10,000*
protein structure models

3 x 214M files

>90% cases just 1, but some proteomes divided into multiple shards

https://github.com/deepmind/alphafold/blob/main/afdb/README.md

UniProt

MGnify

> 500M

> 2.4B

sequences

ALPHAFOLD

PROTEIN STRUCTURE

DATABASE

DeepMind

ESM Atlas

Metagenomics

Protein Dataset

Meta AI

214M

617M

structures

PART 1

PART 2

UniProt

> 500M

MGnify

> 2.4B

sequences

ALPHAFOLD

PROTEIN STRUCTURE DATABASE

DeepMind

214M (189M)

25TB

ESM Atlas

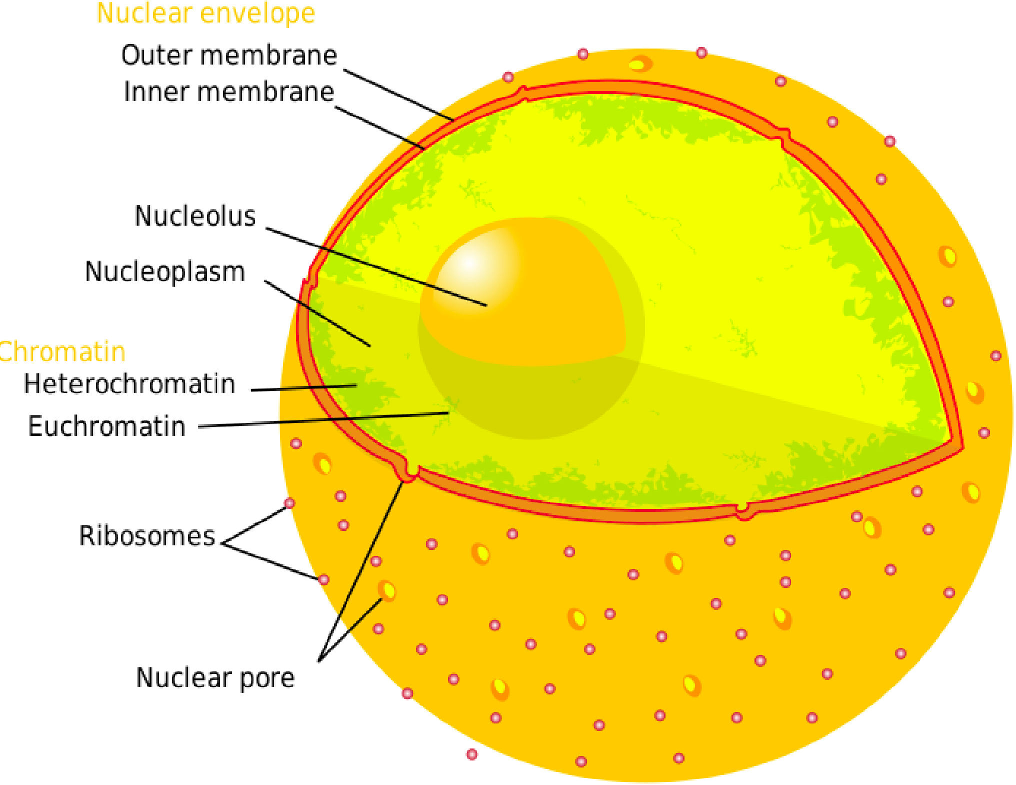
Metagenomics Protein Dataset

Meta AI

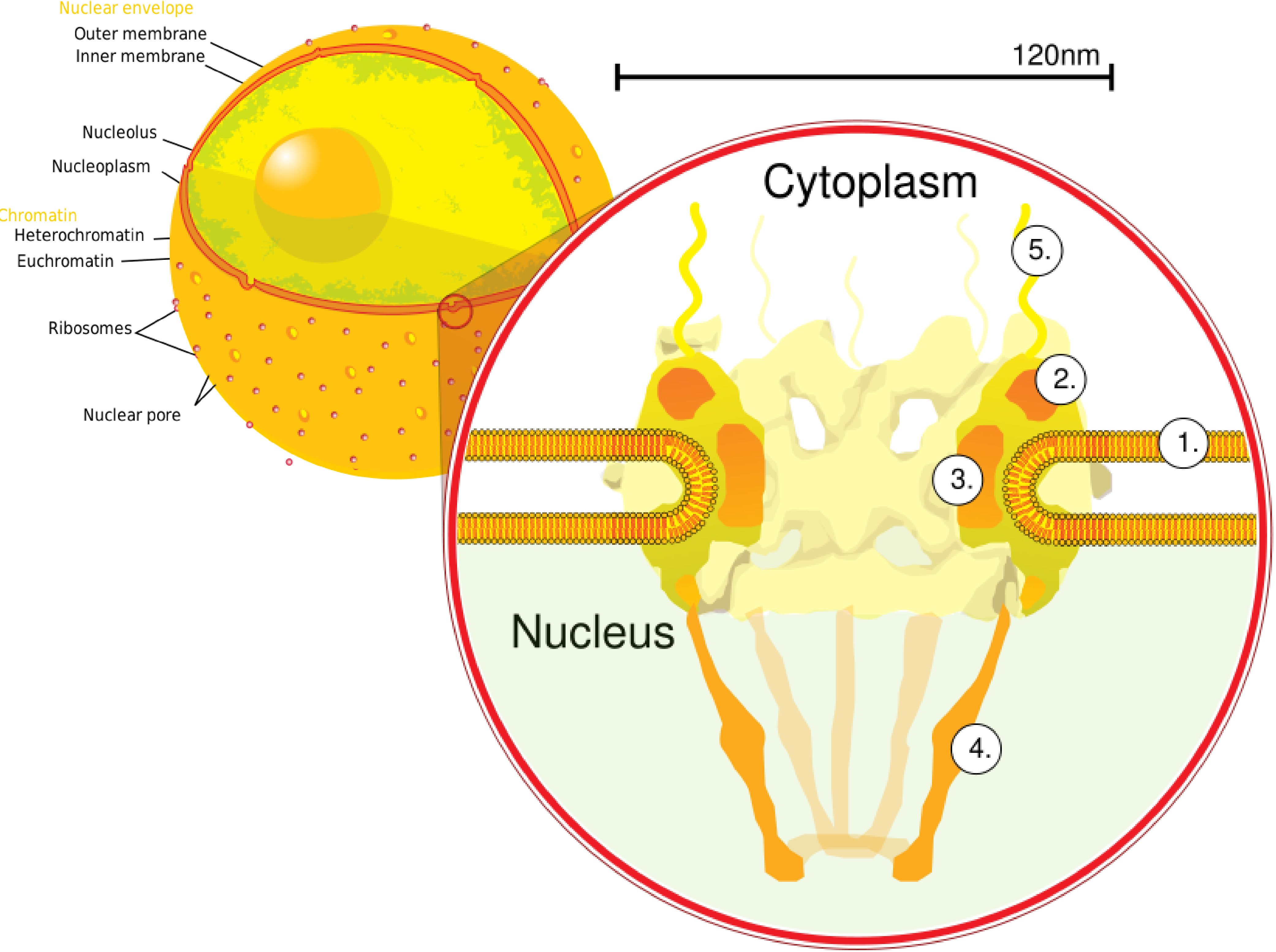
617M

15TB

structures



https://en.wikipedia.org/wiki/Nuclear_pore


Nuclear envelope
Outer membrane
Inner membrane
Nucleolus
Nucleoplasm
Chromatin
Heterochromatin
Euchromatin
Ribosomes
Nuclear pore
120nm
Cytoplasm
5.
2.
1.
3.
Nucleus
4.

Jan Kosinski

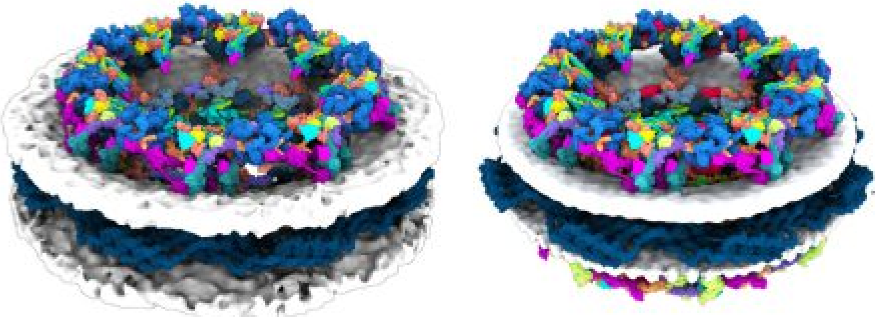
Group Leader

EMBL

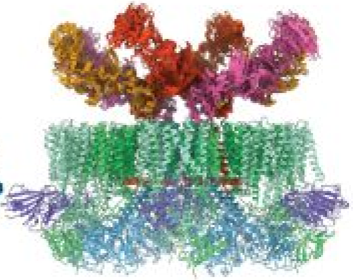
Complexes modeled using Assembline

Human pore complex
(Science, 2016)

Human pore complex (Science, 2022)

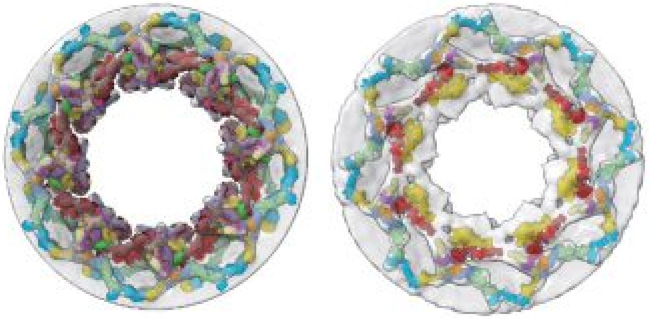


Type VII secretion system
(Science Advances, 2021)

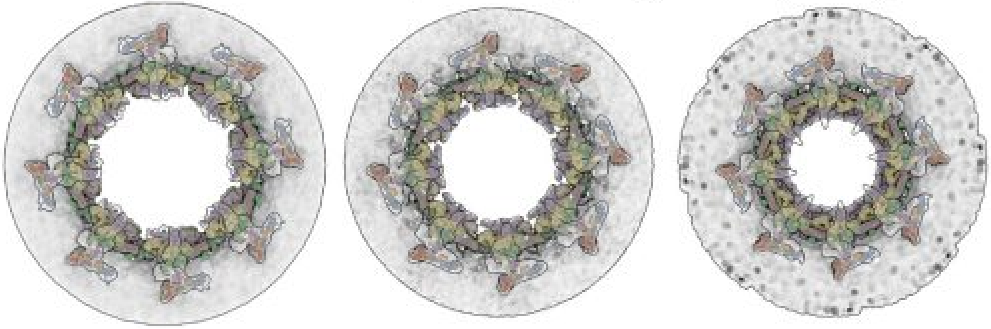


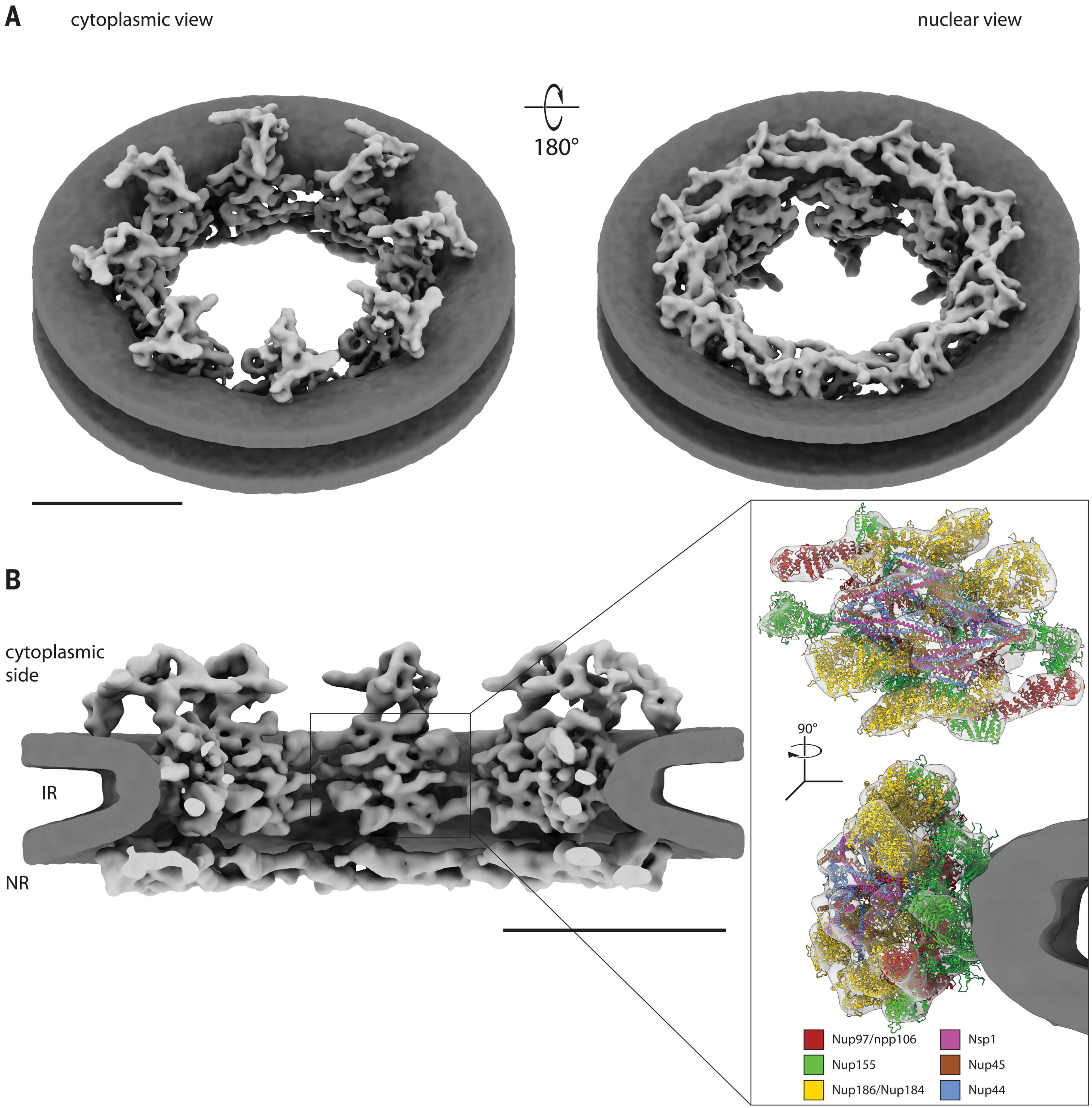
Elongator complex
(EMBO Reports, 2017)

Budding yeast nuclear pore complex (Nature, 2020)



Fission yeast nuclear pore complex (Science, 2021)




A
cytoplasmic view
nuclear view
180°
B
cytoplasmic side
IR
NR
90°
🟥 Nup97/npp106
🟩 Nup155
🟨 Nup186/Nup184
🟪 Nsp1
🟫 Nup45
🟦 Nup44

Thank you for your time
and
See you at the next lecture

Presentation of the projects

Any other
questions & comments

lukaskoz@mimuw.edu.pl