

Rešitve naloge

- najprej poiščemo ORF z npr. ORF finder (<https://www.ncbi.nlm.nih.gov/orffinder/>):

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF1	+	1	421	4344	3924 1307
ORF8	+	3	4191	>4829	639 212
ORF22	-	2	1496	1197	300 99
ORF13	-	1	2820	2539	282 93
ORF19	-	1	252	>1	252 83
ORF25	-	2	239	>3	237 78
ORF10	-	1	3921	3712	210 69
ORF9	-	1	4227	4030	198 65
ORF12	-	1	3117	2920	198 65
ORF15	-	1	1063	892	192 63
ORF4	+	2	881	1054	174 57

Dobimo kar lepo število možnih ORF. Odločimo se za vidno najdaljšega, saj je zelo malo verjetno, da bi naključna sekvenca nukleotidov generirala tako dolg ORF. Ugotovimo, da vključuje nukleotide od 421 do 4344, torej je ORF dolg 3924 nukleotidov in se prevede v 1307 aminokislinskih ostankov. Vsega skupaj najdemo 27 ORF.

- v levem oknu lahko takoj dobimo FASTA format tega ORF, skopiramo ga in uporabimo BLAST:

Bolj specifično uporabimo blastn, saj želimo najti ostale nukleotidne sekvence na podlagi naše. Za iskanje uporabimo bazo nr in iščemo le zelo podobne sekvence.

BLAST® » blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

AAAGCTCGCAATGACCAATTA
CAKAGCAGTAAAGCAATGACCGATCTATAAGACTATGCTGATCA
TGTCTCCAACCTGGTACAG
CTTGACTGGGAGATCTATCTGCGACATAGACTCTATCGTAAAGGAAAA

Or, upload file Choose File No file chosen

Job Title samo_Cas12a_sekvenca_nt

Choose Search Set

Database Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus Experimental databases

Core nucleotide database (core_nt)

Organism Optional Enter organism name or id-completions will be suggested exclude Add Organism

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez® Query Optional Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn) Choose a BLAST algorithm

BLAST Search database core_nt using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastn suite » results for RID-ZVE6H9JH016

Job Title: samo_Cas12a_sekvenca_nt
 RID: ZVE6H9JH016
 Program: BLASTN
 Database: core_nt
 Query ID: IcljQuery_2697085
 Description: samo_Cas12a_sekvenca_nt
 Molecule type: dna
 Query Length: 3924

Filter Results
 Organism: only top 20 will appear
 Percent Identity: [] to []
 E value: [] to []
 Query Coverage: [] to []

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Synthetic construct Cpfl1-GENEWRITE gene, complete cds	synthetic construct	7241	7241	100%	0.0	100.00%	7235	MZ493918.1
Acidimicrococcus intestini DS-0019-208 DNA, complete genome	Acidimicrococcus intestini	7203	7203	100%	0.0	99.80%	2204130	AP031434.1
Segatella copri strain HDCC1 chromosome 1	Segatella copri	62.1	62.1	1%	8e-04	94.87%	3117565	CP156891.1

Dobili smo tri zadetke, od katerih sta le dva relevantna, saj tretji ni tako identičen kot sta prva dva, hkrati pa je tudi E-vrednost pri tretjem kar zelo visoka v primerjavi z ostalima, kjer je praktično 0.

- oba zadetka si ogledamo v GenBank:

Nucleotide Search

GenBank - Synthetic construct Cpfl1-GENEWRITE gene, complete cds

GenBank: MZ493918.1

FASTA Graphics

Go to: []

LOCUS: HD493918 7235 bp DNA linear SYN 10-AUG-2021
 DEFINITION: Synthetic construct Cpfl1-GENEWRITE gene, complete cds.
 ACCESSION: HD493918
 VERSION: HD493918.1
 KEYWORDS: other sequences; artificial sequences.
 SOURCE: synthetic construct
 ORGANISM: synthetic construct
 REFERENCE: 1 (bases 1 to 7235)
 AUTHORS: Kuhlman, T.E.
 TITLE: Direct Submission
 JOURNAL: Submitted (01-JUL-2021) Physics and Astronomy, University of California Riverside, 900 University Avenue, Riverside, CA 92521, USA

FEATURES

source Location/Qualifiers
 1..7235
 /organism="synthetic construct"
 /mol_type="other DNA"
 /db_xref="taxon:32830"
 /note="TTCpfl1-ORF2pINT1MS16whis"
 1..19
 /regulatory_class="promoter"
 51..7235
 /transl_table=1
 /product="Cpfl1-GENEWRITE"
 /protein_id="QVX72835.1"
 /translation="MGRRRKAMPFKLSENKYLAKVEVMTQEGFTNLVQSKTLRF
 ELIPQKTLINHQDQFIEIGKGRNPKVLEKPIIDREYKTHQCLVQLVNLIS
 AAISYRKEKTEETRNALIEEQATYRNALINDVYIGRDTDLDAINRKHAEIYKGLFKA
 ELFNGKVLQLQDTVTTTEHNLKRSFKFTTYVSGPYEMRNVPVSAEDISTAPHR
 VQNRPKFENKCFITLITAVPLSLRDFRNNKALGEPVYTSIEVSPFPINQLT

Recent activity

- Synthetic construct Cpfl1-GENEWRITE gene, complete cds
- Acidimicrococcus intestini DS-0019-208 DNA, complete genome
- CRISPR-associated protein Cpfl1, subtype PREFRAN [Acidimicrococcus sp. BV3L Protein
- Acidimicrococcus sp. BV3L6 contig00028, whole genome shotgun sequence
- cas12a (14892)

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information Log in

Nucleotide Help

Advanced

GenBank Send to:

Acidaminococcus intestini i35-0019-2B8 DNA, complete genome

GenBank: AP031434.1 Selected region

[FASTA](#) [Graphics](#) from: 59942 to: 103865

Go to:

LOCUS AP031434 3924 bp DNA linear BCT 24-JUL-2024

DEFINITION Acidaminococcus intestini i35-0019-2B8 DNA, complete genome.

ACCESSION AP031434 REGION: 99942..103865

VERSION AP031434.1

DBLINK BioProject: PRJ20817661

BIOSAMPLE SRR200759819

Sequence Read Archive: DRR538051, DRR538169

KEYWORDS .

SOURCE Acidaminococcus intestini

ORGANISM Acidaminococcus intestini

Bacteria; Bacillati; Bacillota; Negativicutes; Acidaminococcales; Acidaminococaceae; Acidaminococcus.

REFERENCE 1

AUTHORS Furuichi,M., Kawaguchi,T., Pust,M., Yasuma,K., Plichta,D., Haragawa,M., Ohta,T., Shattaraei,S., Sasajima,S., Aoto,Y., Tuganbaev,T., Yaginuma,M., Ueda,H., Okahashi,R., Anefuji,K., Kiridoshii,Y., Sugita,K., Strazar,M., Skelly,A., Suda,W., Hattori,M., Nakamoto,M., Caballero,S., Norman,J., Olla,B., Tanoue,T., Arita,M., Buccì,V., Atarashi,K., Xavier,R. and Honda,K.

TITLE Defined microbial consortia suppress multidrug-resistant proinflammatory Enterobacteriaceae via ecological control

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 3924)

AUTHORS Suda,W., Shindo,C., Furuichi,M., Aoto,Y. and Kiridoshii,Y.

TITLE Direct Submission

JOURNAL Submitted (16-APR-2024) Contact@Hunehiro Furuichi Keio University school of medicine, Department of Microbiology and Immunology; Shinanomachi 35, Shinjuku-ku, Tokyo 166-8582, Japan

COMMENT Annotated by DFAST <https://dfast.dobj.nig.ac.jp/>

#Genome-Assembly-Data-START##

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Related information

Assembly

BioProject

BioSample

Protein

Taxonomy

Gene

Identical RefSeq

Recent activity Turn Off Clear

Acidaminococcus intestini i35-0019-2B8 DNA, complete genome Nucleotide

V prvem zadetku razberemo, da gre za protein, ki ga kodira gen Cpf1, ime proteina pa je tudi Cpf1. V drugem zadetku ugotovimo, da se ta genski zapis sicer pojavi v organizmu Acidaminococcus intestine. To je gram negativna bakterija.

[Download](#) [GenBank Graphics](#)

Acidaminococcus intestini i35-0019-2B8 DNA, complete genome

Sequence ID: AP031434.1 Length: 2204130 Number of Matches: 1

Range 1: 99942 to 103865 [GenBank](#) [Graphics](#)

Score	Expect	Identities	Gaps	Strand
7203 bits(3900)	0.0	3916/3924(99%)	0/3924(0%)	Plus/Minus
Query 1		ATGACC AATTG AAGGTTT TACCAATTT ATACAAGTT TCGAAGCCCTCGTTT GAA		60
Sbjct 103865		ATGACC AATTG AAGGTTT TACCAATTT ATACAAGTT TCGAAGCCCTCGTTT GAA		103806
Query 61		CTGATTC CCAAGGAAAACACT CAACATATCCAGGAC AAGGGTTCATTGAGGAGGAT		120
Sbjct 103805		CTGATTC CCAAGGAAAACACT CAACATATCCAGGAC AAGGGTTCATTGAGGAGGAT		103746
Query 121		AAAGCTCGCAATGAC CATTACAAGAGTTAAACCAATCATTGACCGCATATAAGACT		180
Sbjct 103745		AAAGCTCGCAATGAC CATTACAAGAGTTAAACCAATCATTGACCGCATATAAGACT		103686
Query 181		TATGCTGATCAATGTC CCAAGCTGACAGCTTGC TGGGAAATCATATCGACGCATA		240
Sbjct 103685		TATGCTGATCAATGTC CCAAGCTGACAGCTTGC TGGGAAATCATATCGACGCATA		103626
Query 241		GACTCCTATGTAAGGAAAACCAAGAAACACGAATCGCTGATTGAGGACCAAGCA		300
Sbjct 103625		GACTCCTATGTAAGGAAAACCAAGAAACACGAATCGCTGATTGAGGACCAAGCA		103566
Query 301		ACATATAGAAATGCGGATTCATGACTACTTTATAGTTCGGACGATAATCGACAGATGCC		360
Sbjct 103565		ACATATAGAAATGCGGATTCATGACTACTTTATAGTTCGGACGATAATCGACAGATGCC		103506
Query 361		ATAAATAGCCCATGCTGAAATCTATAAGGACTTTTAAAGCTGAACTTTTCAATGGA		420
Sbjct 103505		ATAAATAGCCCATGCTGAAATCTATAAGGACTTTTAAAGCTGAACTTTTCAATGGA		103446
Query 421		AAAGTTTTAAAGCAATTAGGACCGTAACCCAGCAGAAATGAAATGCTCTACTCCGT		480
Sbjct 103445		AAAGTTTTAAAGCAATTAGGACCGTAACCCAGCAGAAATGAAATGCTCTACTCCGT		103386
Query 481		TCGTTTGACAAATTTAGCACTATTTTTCCGGCTTTTATGAAACCGAAAATGCTTT		540
Sbjct 103385		TCGTTTGACAAATTTAGCACTATTTTTCCGGCTTTTATGAAACCGAAAATGCTTT		103326
Query 541		AGCGCTGAAGATATCAGCAGGCAATTC CCAATCGAATCGTCCAGGCAATTTCCCTAAA		600
Sbjct 103325		AGCGCTGAAGATATCAGCAGGCAATTC CCAATCGAATCGTCCAGGCAATTTCCCTAAA		103266
Query 601		TTTAGGAAAACCTGCCATATTTTACAAGATGATAACCGAGTTCCTCTTGGGGAG		660
Sbjct 103265		TTTAGGAAAACCTGCCATATTTTACAAGATGATAACCGAGTTCCTCTTGGGGAG		103206
Query 661		CATTTTGAATATCAAAAAGCCATTGGAACTTTTGTAGTACGCTATTGAAAGACTC		720
Sbjct 103205		CATTTTGAATATCAAAAAGCCATTGGAACTTTTGTAGTACGCTATTGAAAGACTC		103146

Če se vrnemo nazaj na BLAST vidimo kateri nukleotidi so se poravnali (Range na zgornji zaslonski sliki).

```

AAHSAIDQRRYSTHKNILQSLFVESSTYGLIGKIDLFDSSSTKTLIERKKQIKTZYDGY
IFQIYQGVFCHIENGFQVDHLQVLSLDHHRKYDISLPKDDQPMHFQKFEQLLKEHRQFS
LSSFQKQKQKCLLCHIEPACDRSL"
gene complement(8994..103865)
/locus_tag="I350019288_00960"
CDS complement(8994..103865)
/locus_tag="I350019288_00960"
/inference="COORDINATES: ab initio
prediction:MetaGeneAnnotator"
/codon_start=1
/transl_table=11
/product="hypothetical protein"
/protein_id="BPK75607.1"
/translacion="MTQFEGFTNLYQVSKTLRFELIPQGKTLKHIQEQGFIEEDKARN
DNHYKELKPIIDRIYKTYADQCLQVLDWENLSAAIDSYRKEKTEETRNALIEEQATY
RNAIHDYFIGRDTNLDIAINKRHAETKGLFKAELFNGKVLKQLGTVTTTEHENALLR
SFDKFTTYFSGFYENRKNVSAEDISTAIPHRIQDNFQPKFKENCHIFRILITAVPSL
REHFENVKKAIGIFVSTSIIEEVFSPFYQLLTQTQIDLYNQLGGISREAGTEKIKG
LNEVLNLAIQKNDETAHIIASLPHRFIPLFKQILSDRNTLSFIEEFKSDEEVIQSFCK
KYKTLNRNENLEAEALFNLNSIDLTHFISHKLETISSALCOHMDLRLNLYER
RISELTGKIKSAEKVQSLKHEDINLQETISAAKGLSEAFKQKTEILSHAAH
DQPLPTTLKQEKELKQSLDGLVHLLDFAVDESNEVDPEFSARLTGKLEHE
PSLSFYNKARNYATKKPVSVEKFLNQPTLASGDVNEKKNAGILFVKNGLYLG
IMPQKGRYKALSFEPTKTESEGFQKHYDYFDDAAKIPKCKSTQLKAVTAHFQTHTT
PILLSMFIPELITKEIYDLWPEKPKKFTAYAKKIGQKGYREALCKMIDFTRD
FLSKYKTTSIDLSSLRPSSQKDLGEVYAEWPLLYHESQRIAEKEIDMVAETGKL
YLFQVYKDFAKGHGKPHLHTLYTGLFSPENLAKTISIKLNGQAELEFVPRKSRMKNV
AHLRGEKMLNKKLDQKTPPTDQLVQVYVNRHLSHDLSEARALPMVITKEVSH
EIIKDRRFTSKDFFHVPITLNYQANSKFNQRVNYLKEHPETPIIGIDRGRNL
IYITVIDSTGKLEQRSLTIQQDYQKLDNREKVAARQAHSVGTIKDLKQGYL
SQVHEIVDLHIHYQAVLVLENLNFQKSKRTGIAEKAVYQFQEKMLIDKLNKLVKD
YPAEKVGGVLPYQLTDQTSFAKMGTSQGFVYVPAPYTSKIDPLTGFVDPPVWIKTI
KHEKGRVLEHGFVHYVKTGDFILHFQWRNLSFQRGLPGFPAHDIIVFKNETQ
FDAKGFPIAGKRIVPVIEHRRFTGRADLYPANELIALLEEKGFVFRDGSNIIPLKLL
ENDSSHAIDTIVALIRSVLQIRNSNAATGEDYINSVRDLNMGVCFDSRFQPEHPIDA
DMSGYHIALKGLLLNPLKESKDLKLGINSNQDMLAYIQELRN"
gene complement(104104..104742)
/locus_tag="I350019288_00970"
CDS complement(104104..104742)
/locus_tag="I350019288_00970"
/inference="COORDINATES: ab initio
prediction:MetaGeneAnnotator"
/inference="similar to AA sequence:RefSeq:WP_011228142.1"
/codon_start=1
/transl_table=11
/product="uracil-DNA glycosylase"
/protein_id="BPK75608.1"

```

Z orodjem za iskanje poiščemo mesto zaporedja v genomu in dobimo aminokislinsko zaporedje, v katerega se prevede.

- za poravnavo zaporedij uporabimo algoritem Needle, saj pričakujemo da sta si zaporedji zelo podobni in ju poravnamo globalno. Če ne bi naredili globalne poravnave nam dodatnega dela sintetičnega konstrukta ne bi prikazalo.

Input sequence ⓘ

Sequence type
 Protein DNA

Paste your first sequence here - or use the example sequence

```
>complete_genome
MTQFEGFTNLYQVSKTLRFELIPQGKTLKHIQEQGFIEEDKARNDNHYKELKPIIDRIYK
YADQCLQVLDWENLSAAIDSYRKEKTEETRNALIEEQATYRNAIHDYFIGRDTNLDIA
INKRHAETKGLFKAELFNGKVLKQLGTVTTTEHENALLRSFDKFTTYFSGFYENRKNV
SAEDISTAIPHRIQDNFQPKFKENCHIFRILITAVPSLREHFENVKKAIGIFVSTSIIEV
FSPFYQLLTQTQIDLYNQLGGISREAGTEKIKGLNEVLNLAIQKNDETAHIIASLPH
RFIPLFKQILSDRNTLSFIEEFKSDEEVIQSFCKYKTLNRNENLEAEALFNLNSID
.....
```

Choose File No file chosen

Paste your second sequence here - or use the example sequence

```
>synthetic_construct
MSRRRKNPTKLSENAKKLAKEVENMTQFEGFTNLYQVSKTLRFELIPQGKTLKHIQEQG
FIEEDKARNDNHYKELKPIIDRIYKTYADQCLQVLDWENLSAAIDSYRKEKTEETRNAL
IEEQATYRNAIHDYFIGRDTNLDIAINKRHAETKGLFKAELFNGKVLKQLGTVTTTEH
ENALLRSFDKFTTYFSGFYENRKNVSAEDISTAIPHRIQDNFQPKFKENCHIFRILITAV
PSLREHFENVKKAIGIFVSTSIIEEVFSPFYQLLTQTQIDLYNQLGGISREAGTEKIK
GLNEVLNLAIQKNDETAHIIASLPHRFIPLFKQILSDRNTLSFIEEFKSDEEVIQSFCK
.....
```

Choose File No file chosen

Use the example Clear sequence More example inputs

complete_geno	1	-----MTQFEGFTNLYQVSKTLRFELIPQG	25	synthetic_con	1201	LYPANELIALLEEKGIVFRDGSNILPKLLENDSDHAIDTMVALIRSVLQM	1250
synthetic_con	1	MSRRRKNPTKLSNAKKLAKEVENMTQFEGFTNLYQVSKTLRFELIPQG	50	complete_geno	1226	RNSNAATGEDYINSPVRDLNGVCFDSRFQNPPEWMDADANGAYHIALKGQ	1275
complete_geno	26	KTLLKHIQEQGFIEEDKARNHDHYKELKPIIDRIYKYADQCLQLVLDWEN	75	synthetic_con	1251	RNSNAATGEDYINSPVRDLNGVCFDSRFQNPPEWMDADANGAYHIALKGQ	1300
synthetic_con	51	KTLLKHIQEQGFIEEDKARNHDHYKELKPIIDRIYKYADQCLQLVLDWEN	100	complete_geno	1276	LLLNLKESKDLKLONGISNQDWAYIQELRN-----	1307
complete_geno	76	LSAAIDSYRKEKTEETRNALIEEQATYRNAIHDFYIGRTDNLDAINKRH	125	synthetic_con	1301	LLLNLKESKDLKLONGISNQDWAYIQELRNNGGGGGGGGGKILQTSRS	1350
synthetic_con	101	LSAAIDSYRKEKTEETRNALIEEQATYRNAIHDFYIGRTDNLDAINKRH	150	complete_geno	1308	-----	1307
complete_geno	126	AEIYKGLFKAELFNGKVLKQLGTVTTTEHENALLRSFDKFTTYFSGFYEN	175	synthetic_con	1351	TTWKLNLLNDYVWHNEMKAEIKMFFETENKDTTYQNLWDAFKAIVCRG	1400
synthetic_con	151	AEIYKGLFKAELFNGKVLKQLGTVTTTEHENALLRSFDKFTTYFSGFYEN	200	complete_geno	1308	-----	1307
complete_geno	176	RKNVFAEDISTAIPHRIVQDNFQPKFKENCHIFTRLITAVPSLREHFENV	225	synthetic_con	1401	KFIALNAYKRKQERSKIDTLTSQLEKEQEQTHSKASRRQEIITKIRAEI	1450
synthetic_con	201	RKNVFAEDISTAIPHRIVQDNFQPKFKENCHIFTRLITAVPSLREHFENV	250	complete_geno	1308	-----	1307
complete_geno	226	KKAIIGIFVSTSIIEEVFFPFYQLLTQTQIDLQYLLGGISREAGTEKIK	275	synthetic_con	1451	KEIETQKTLQKINESRSMFFERINKIDRPLARLIKKKREKNQIDITKNDK	1500
				complete_geno	1308	-----	1307

Že takoj vidimo, da sta se zaporedji lepo poravnali, vendar je sintetični konstrukt bistveno daljši na C- in N- koncu. Ostali ak-ostanki, pa so identični ali pa zelo podobni.

- Z analizo C- in N- konca daljšega zaporedja ugotovimo, da je prisoten nekakšen ekspresijski tag ali pa zgolj artefakt na N-koncu, saj ga pri naravni varianti ni. Na C-koncu pa opazimo heksahistidinsko oznako, ki služi pri čiščenju proteina (afinitetna kromatografija).
- Da je protein fuzijski lahko sklepamo, ker se poravna le prvi del zaporedja, sledi glicinsko-serinski linker (glicin omogoča fleksibilnost, serin pa topnost). Ko z BLAST poiščemo, kateri protein je drugi dobimo UniProt AC: O00370 (to je prvi zadetek na BLAST) Protein ima aktivnost reverzne transkriptaze.
- S poravnavo ak zaporedja, ki se prevede iz genoma bakterije iščemo po BLAST in pridemo do zadetkov:

Job Title **ak_sekvenca**

RID [01CFB8GK014](#) Search expires on 05-12 00:27 am [Download All](#)

Program BLASTP [Citation](#)

Database swissprot [See details](#)

Query ID |cl|Query_2293748

Description unnamed protein product

Molecule type amino acid

Query Length 1307

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to E value to Query Coverage to

[Filter](#) [Reset](#)

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments [Download](#) [Select columns](#) Show

select all 2 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> RecName: Full=CRISPR-associated endonuclease Cas12a; AltName: Full=AsCpf1; AltName: Full=CRISPR-associat...	<i>Acidaminococcus</i> ...	2701	2701	100%	0.0	99.77%	1307	U2UMQ6
<input checked="" type="checkbox"/> RecName: Full=CRISPR-associated endonuclease Cas12a; AltName: Full=CRISPR-associated endonuclease Cpf1...	<i>Francisella tulare</i> ...	673	673	100%	0.0	34.67%	1300	A0Q7O2.1

Očitno gre za CRISPR-associated endonukleazo Cas12a. Vzeli bomo zgornji zadetek, saj smo ugotovili, da se naša sekvenca pojavi v tem organizmu. Koda dostopa na UniProt: U2UMQ6.

- Na UniProt piše da je to endonukleaza Cas12a in ne Cpf1, vendar dilemo hitro rešimo, saj slednje ime najdemo pod alternativnimi. Zdaj lahko sklepamo da je na epici pisalo CRISPR ("CR") in spodaj Cas12a ("2a")
- Skupen izvor lahko preverimo na več načinov. Že v UniProt lahko razberemo informacijo o homologih in ortologih, ampak v primeru našega proteina tega ne opazimo. Potem s pomočjo naše sekvence iščemo s pomočjo BLAST. Sam pri uporabi organizma, ki je naveden in uporabe različnih izvedenk (blastp, tblastn, PSI-blast ...) nisem uspel ugotoviti, da bi bila proteina evolucijsko kakorkoli povezana.
- V Uniprot izberemo več organizmov, ki vsebujejo protein Cas12a in naredimo dokument z vsemi zaporedji v format fasta (Download).

Tools • Download (18) Add View: Cards Table Customize columns Share • 18 rows selected out of 75

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input checked="" type="checkbox"/> U2UMQ6	CS12A_ACISB	CRISPR-associated endonuclease Cas12a[...]	cas12a_cpf1, HMPREF1246_0236	Acidaminococcus sp. (strain BV3L6)	1,307 AA
<input checked="" type="checkbox"/> A0Q7Q2	CS12A_FRATN	CRISPR-associated endonuclease Cas12a[...]	cas12a_cpf1, FTN_1397	Francisella tularensis subsp. novicida (strain ATCC 15482 / CCUG 33449 / U112)	1,300 AA
<input checked="" type="checkbox"/> A0ABF7PPZ3	A0ABF7PPZ3_9FIRM	CRISPR-associated endonuclease Cas12a		Anaeroglobus	1,365 AA
<input checked="" type="checkbox"/> A0ABY8MDZ0	A0ABY8MDZ0_9SPIO	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_P0082_06825	Candidatus Halotiospira prima	1,330 AA
<input checked="" type="checkbox"/> A0ABT1WNM7	A0ABT1WNM7_9LACT	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_NPA36_05700	Granulicatella seriolae	1,278 AA
<input checked="" type="checkbox"/> A0ABU3ZBP9	A0ABU3ZBP9_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_RVY80_09915	Veillonella absiana	1,162 AA
<input checked="" type="checkbox"/> A0ABV1F9R1	A0ABV1F9R1_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_WMO39_07215	Ruminococcoides intestinalis	1,245 AA
<input checked="" type="checkbox"/> A0A7T3RDH8	A0A7T3RDH8_9SPIR	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_IWA51_12190	Treponema peruense	1,334 AA
<input checked="" type="checkbox"/> A0ABT2M0G8	A0ABT2M0G8_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_N5B56_01285	Eubacterium album	1,175 AA
<input checked="" type="checkbox"/> A0ABY6F4M0	A0ABY6F4M0_9GAMM	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_LU297_00890	Moraxella nasicae	1,264 AA
<input checked="" type="checkbox"/> A0ABY7JNS4	A0ABY7JNS4_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	cas12a_OOR46_07965	Peptostreptococcus equinus	1,345 AA

Zaporedja poravnamo s programom Clustal Omega

Tool output

Download

CLUSTAL O(1.2.4) multiple sequence alignment

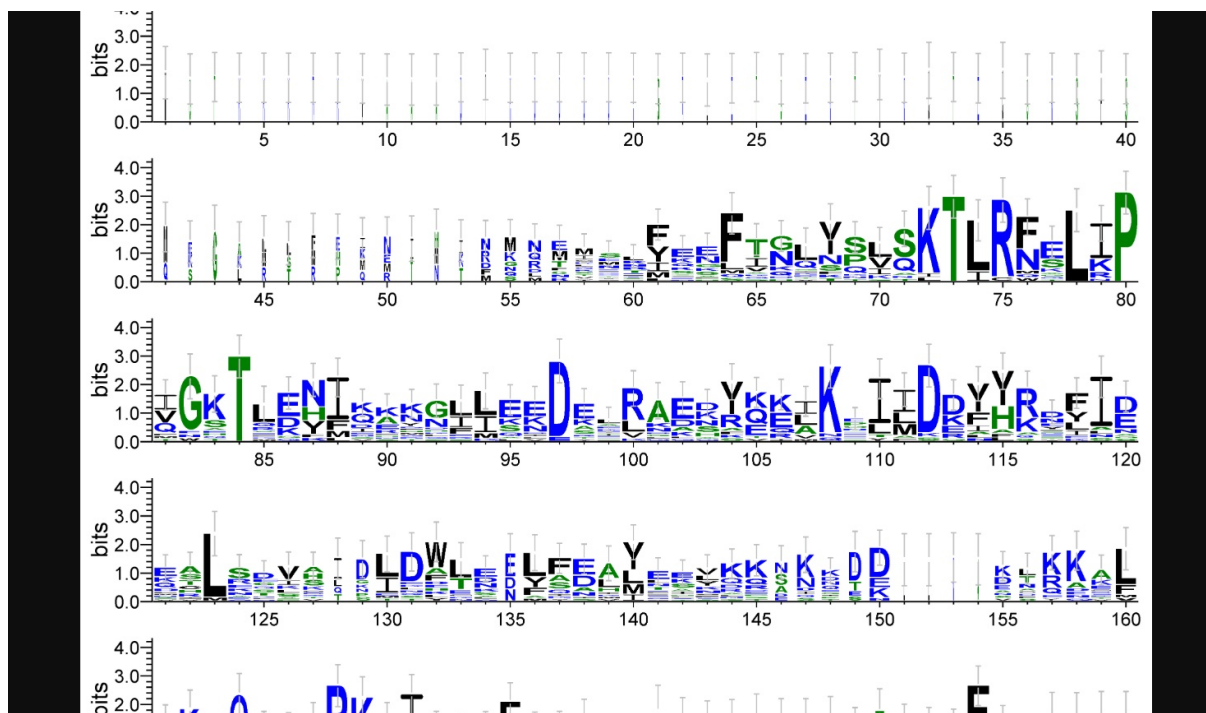
```

tz|A0ABT2M0G8|A0ABT2M0G8_9FIRM      -----MNQET      5
tz|A0ABT1WNM7|A0ABT1WNM7_9LACT      -----MVYN      4
sp|U2UMQ6|CS12A_ACISB                -----MTQ      3
tz|A0ABF7PPZ3|A0ABF7PPZ3_9FIRM      MGPKKRKRVAADYKDDDDKSRLEPGEKPKYKCEGKFSQSGALTRHQRTHTRMTMVT 60
tz|A0ABU3ZBP9|A0ABU3ZBP9_9FIRM      -----      0
tz|A0ABY7JNS4|A0ABY7JNS4_9FIRM      -----MIRKLEKSS  9
tz|A0A7T3RDH8|A0A7T3RDH8_9SPIR      -----MRTIDE      6
tz|A0A7C9H0Z9|A0A7C9H0Z9_9FIRM      -----M--NGNRIIV  8
tz|A0AAP3Q149|A0AAP3Q149_9FIRM      -----MNNGTNN      7
tz|A0ABR7NBZ6|A0ABR7NBZ6_9FIRM      -----MEGKRSFEKN--NQNMYNR 17
tz|A0ABT2SLB2|A0ABT2SLB2_9FIRM      -----ME--DKQFLER  9
tz|A0ABV1F9R1|A0ABV1F9R1_9FIRM      -----MQERKK      6
tz|A0ABV1HU47|A0ABV1HU47_9FIRM      -----MKGLMLPINKFSDCECRK 20
tz|A0ABV5GTK2|A0ABV5GTK2_9FLA0      -----      0
tz|A0ABW8U4L5|A0ABW8U4L5_9GAMM      -----ML      2
tz|A0ABY6F4M0|A0ABY6F4M0_9GAMM      -----ML      2
sp|A0Q7Q2|CS12A_FRATN                -----MSI      3
tz|A0ABY8MDZ0|A0ABY8MDZ0_9SPIO      -----MSSL      4

tz|A0ABT2M0G8|A0ABT2M0G8_9FIRM      IEKMAGLNKKTLTICQKLPVGGKTRNIDKFNMAEDEFVKANKEKINTLIKKAASEKID 65
tz|A0ABT1WNM7|A0ABT1WNM7_9LACT      LKESIGTKLSKTLRFTLIPQNTREYLSKNLLKEE--EVDLQFEQAKLLDGIYRKIE 62
sp|U2UMQ6|CS12A_ACISB                FEGFTNLVQVSKTLRFELIPQKTLKHIQEQGFIEEDKARNHDKLPIDRIYKYTYAD 63
tz|A0ABF7PPZ3|A0ABF7PPZ3_9FIRM      FENFTKQVQVSKTLRFELIPQKTLLENMKRDGIIISVDRQRNDYQKAKGILDKLYKILD 120
tz|A0ABU3ZBP9|A0ABU3ZBP9_9FIRM      MSKFQNLTYINKTLRFGLKPFKGLLENFNKTNLLQLDEYKAKHRKEVQRLFDENFKQIE 60
tz|A0ABY7JNS4|A0ABY7JNS4_9FIRM      YESFTQLFPKQITLRNELIPWETKDNMISHKEIEIDKNRAEDYKRIKSIIDDFYRVLIN 69
tz|A0A7T3RDH8|A0A7T3RDH8_9SPIR      FCGQKNGVYLSKTLRNKLPVGGKTEDKIKEYHLMENDYGRAAAVVEVKNLIDDFHRSFIA 66
tz|A0A7C9H0Z9|A0A7C9H0Z9_9FIRM      YREFVGVTPVAKTLRNELRPIGHTQEIIHNGLIQEDELROEKSTELKNIMDDYRYEID 68
tz|A0AAP3Q149|A0AAP3Q149_9FIRM      FQNFIGISSLKQTLRNALIPETTQQFIVKNGIIEKDELRENQILKDIMDDYRYGFIS 67
tz|A0ABR7NBZ6|A0ABR7NBZ6_9FIRM      YRELIGLSSSKTLRNSLIPVGSCLKYIKKHGILEKDTLRAEKREELKALIMDDYRYNID 77
+||A0ABT2M0G8|A0ABT2M0G8_9FIRM      VVEETFLIEKVTIENELTRFVETIYVTEDEVTIEEELIDAKVDEELVETMDDVYRYNIE 60
    
```

in prenesemo rezultate, da jih uporabimo za izdelavo WebLogo. Prav tako si shranimo rezultat pod podoknom Phylogenetic Tree.

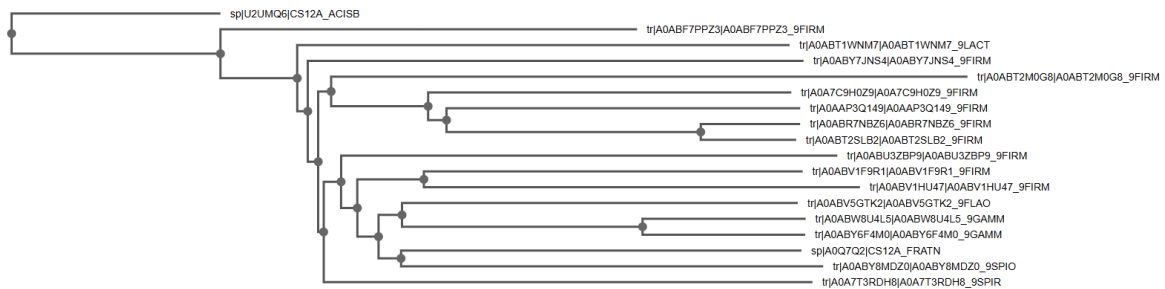
Del WebLogo, ki sem ga sam dobil je prikazan spodaj (glej tudi dodatni dokument z rešitvami):



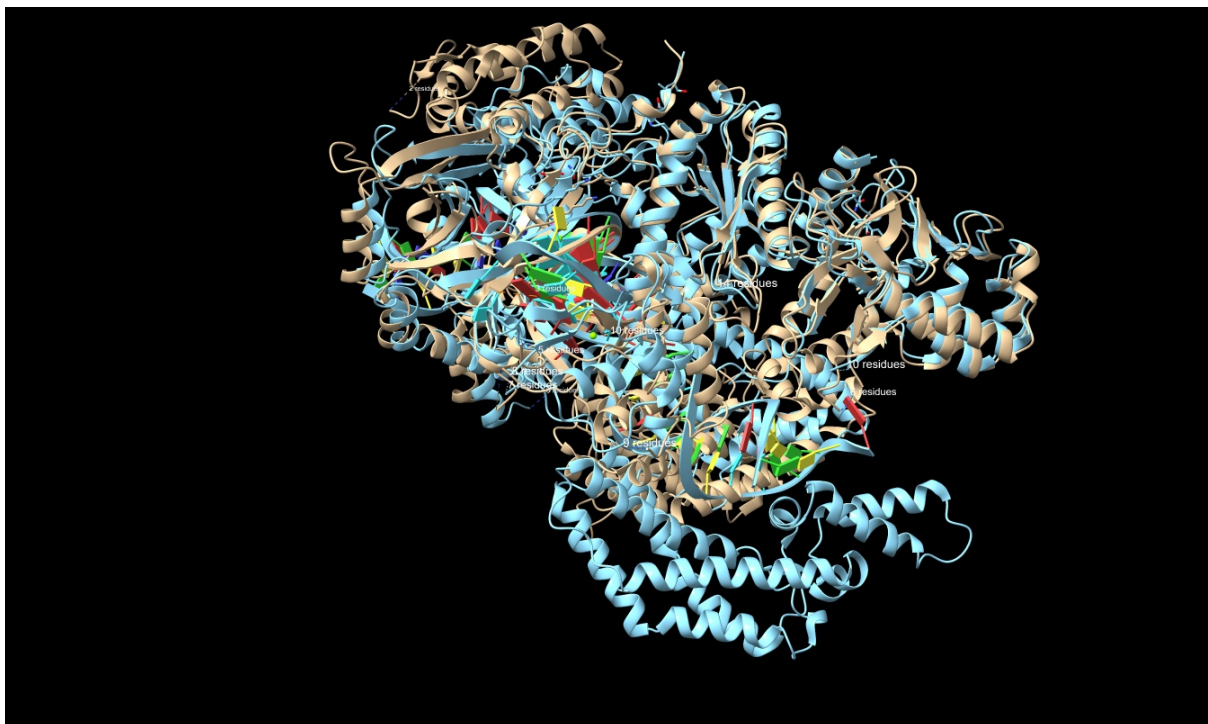
Ko pogledamo in preračunamo kje v poravnavi se začne protein iz našega organizma ugotovimo na katerih mestih so zelo ohranjeni aminokislinski ostanki v primerjavi z drugimi istoimenskimi proteini iz drugih organizmov.

Vidimo na primer Arg18, ki je potreben za vezavo crRNA (na sliki 75), potem lahko opazimo, da vezavno mesto od 47-51 ak ostanka ni tako ohranjeno, ampak je ohranjen bolj samo K51 (na sliki K108, pri nadaljni analizi pa tudi sam nisem več našel še kakšnega izrazito ohranjenega aminokislinskega ostanka s pomembno funkcijo, tako da očitno motivi niso tako specifični.

Ko narišemo filogenetsko drevo (npr. v programu phylo.io) in koreninjenjem glede na naš organizem dobimo:



iz česar razberemo, da je najbolj podoben Cas12a iz organizma z AC: A0ABF7PPZ3. Za potrditev podobnosti proteinov je zelo učinkovito narediti tudi strukturno poravnavo (npr. v ChimeraX). Koda PDB sta že podani. Z orodjem Matchmaker pa naredimo poravnavo danih proteinov. Dobimo sledeče:



Iz poravnave opazimo, da sta si strukturi res kar podobni, vsaj v določenih delih. V drugih delih pa so zvitja prav tako dokaj podobna, le da so konformacije sekundarnih struktur glede na ostale dele proteina malo drugačne. Iz strukture prav tako vidimo

kako se protein veže na DNA in da vezava v teh dveh proteinih ni identična (glej sejo, ki je med dodatnimi rešitvami).