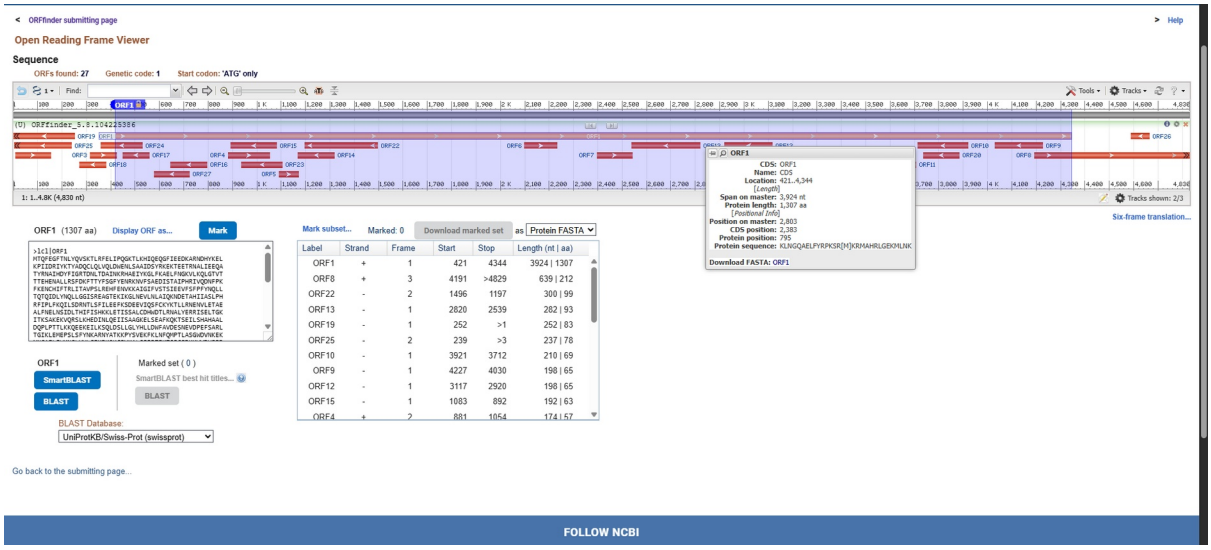
Rešitve naloge

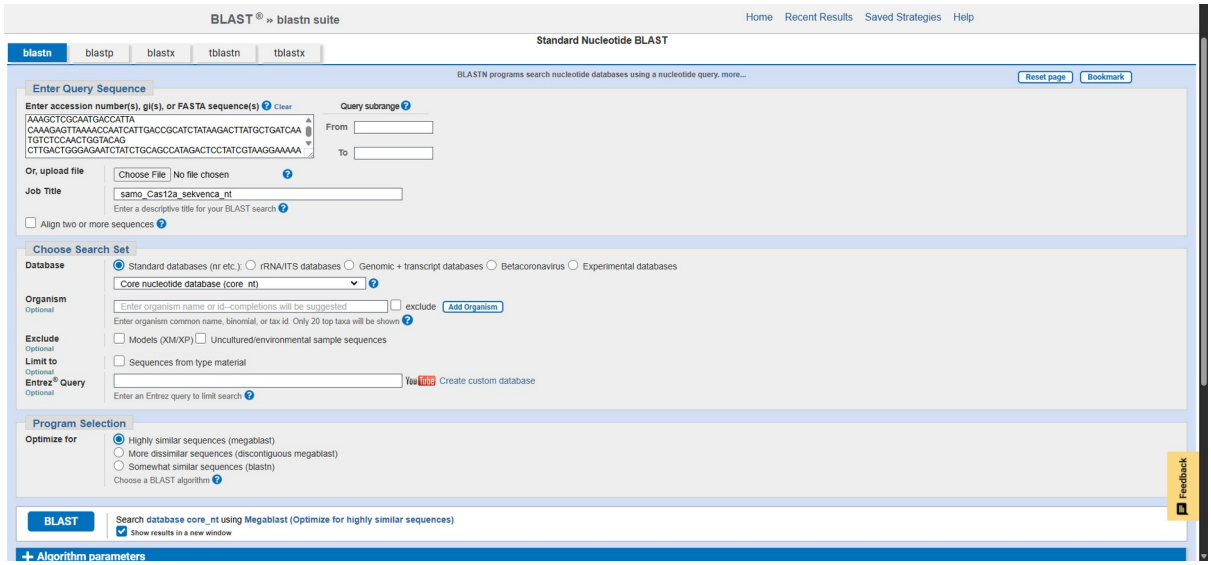
- najprej poiščemo ORF z npr. ORF finder (https://www.ncbi.nlm.nih.gov/orffinder/):

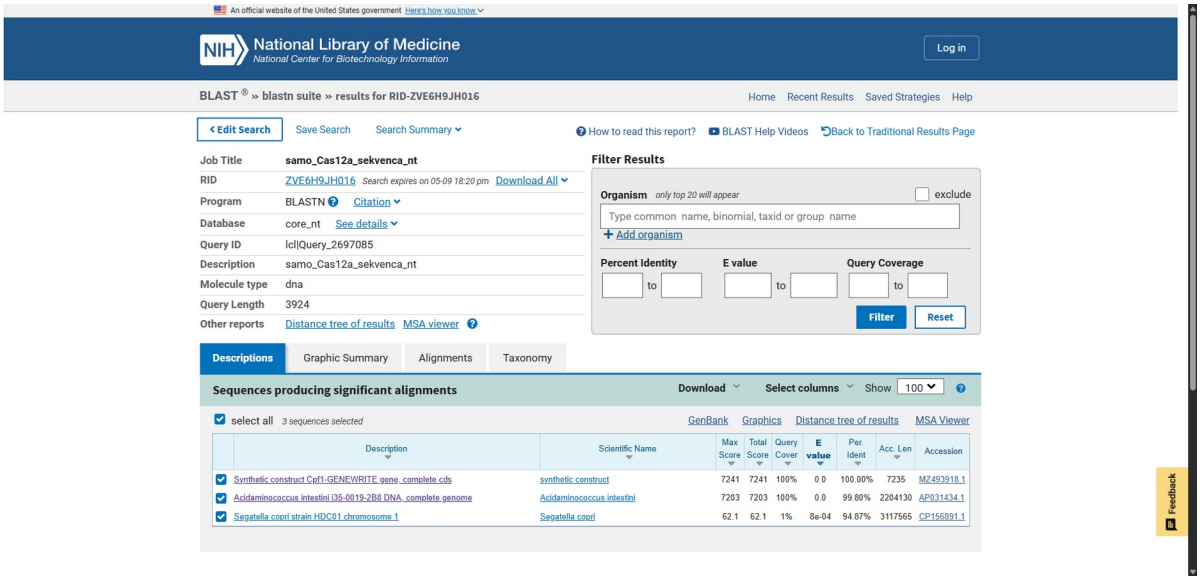


Dobimo kar lepo število možnih ORF. Odločimo se za vidno najdaljšega, saj je zelo malo verjetno, da bi naključna sekvenca nukleotidov generirala tako dolg ORF. Ugotovimo, da vključuje nukleotide od 421 do 4344, torej je ORF dolg 3924 nukleotidov in se prevede v 1307 aminokislinskih ostankov. Vsega skupaj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

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

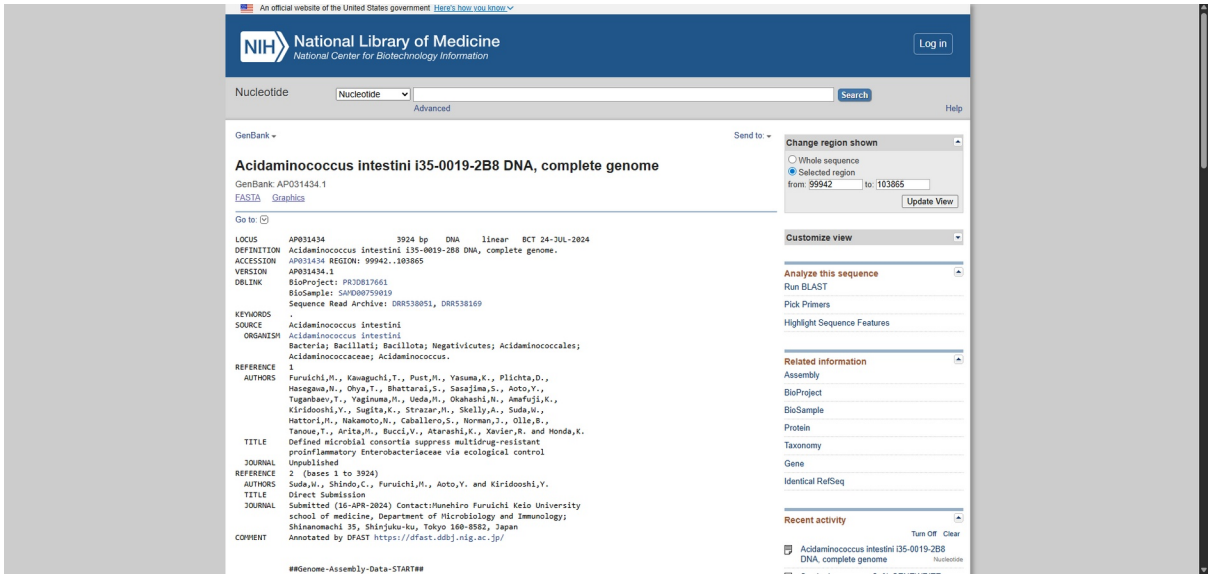
Bolj specifično uporabimo blastn, saj želimo najti ostale nukleotidne sekvence na podlagi naše. Za iskanje uporabimo bazo nr in iščemo le zelo podobne sekvence.



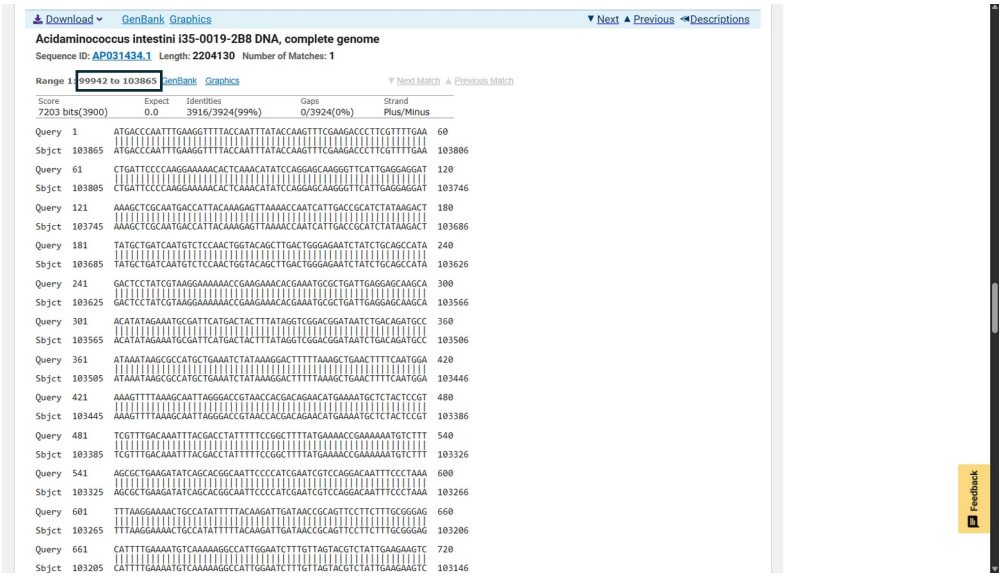


Dobili smo tri zadetke, od katerih sta le dva relevantna, saj tretji ni tako identičen kot sta prva dva, hkrati pa je tudi E-vrednost pri tretjem kar zelo visoka v primerjavi z ostalima, kjer je praktično 0.

- oba zadetka si ogledamo v GenBank:



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.

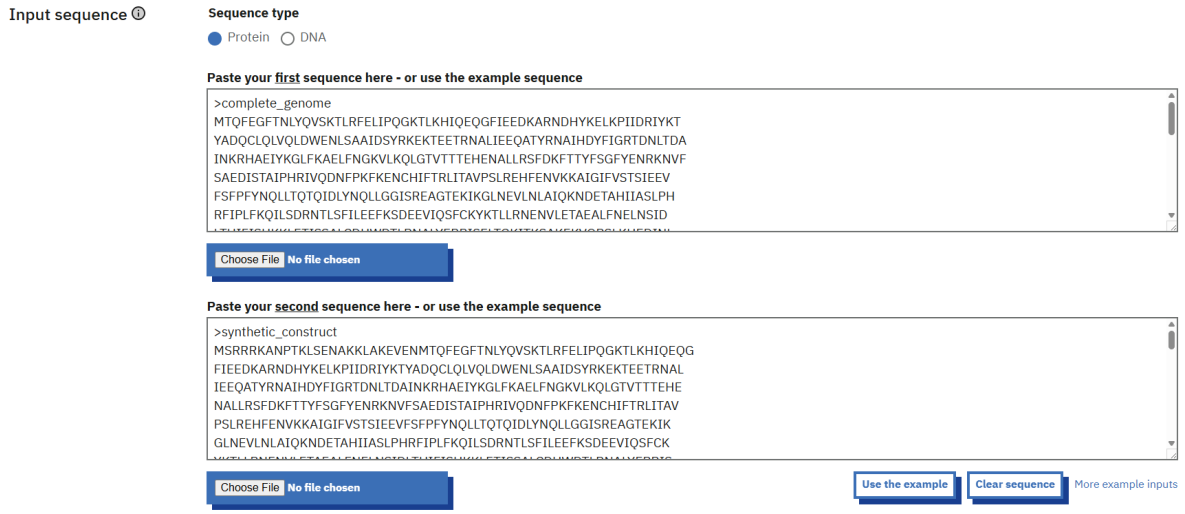


Če se vrnemo nazaj na BLAST vidimo kateri nukleotidi so se poravnali (Range na zgornji zaslonski sliki).

```
                AAHSAIDQRRYSTHKNILQSLFVESSTYGLIGKIDLFDSSTKTLIERKKQIKTIYDGY
                IFQIYGQYFCMIEMGFQVDHLQLYSLDDHRKYDISLPKDDQPMFQKFEQLLKEMRQFS
                LSSFKQQNQKKCLHCIYEPACDRSLL"
gene            complement(99942..103865)
                /locus_tag="I3500192B8_00960"
CDS             complement(99942..103865)
                /locus_tag="I3500192B8_00960"
                /inference="COORDINATES: ab initio
                prediction:MetaGeneAnnotator"
                /codon_start=1
                /transl_table=11
                /product="hypothetical protein"
                /protein_id="BFK75607.1"
                /translation="MTQFEGFTNLYQVSKTLRFELIPQGKTLKHIQEQGFIEEDKARN
                DHYKELKPIIDRIYKTYADQCLQLVQLDWENLSAAIDSYRKEKTEETRNALIEEQATY
                RNAIHDYFIGRTDNLTDAINKRHAEIYKGLFKAELFNGKVLKQLGTVTTTEHENALLR
                SFDKFTTYFSGFYENRKNVFSAEDISTAIPHRIVQDNFPKFKENCHIFTRLITAVPSL
                REHFENVKKAIGIFVSTSIEEVFSFPFYNQLLTQTQIDLYNQLLGGISREAGTEKIKG
                LNEVLNLAIQKNDETAHIIASLPHRFIPLFKQILSDRNTLSFILEEFKSDEEVIQSFC
                KYKTLLRNENVLETAEALFNELNSIDLTHIFISHKKLETISSALCDHWDTLRNALYER
                RISELTGKITKSAKEKVQRSLKHEDINLQEIISAAGKELSEAFKQKTSEILSHAHAAL
                DQPLPTTLKKQEEKEILKSQLDSLLGLYHLLDWFAVDESNEVDPEFSARLTGIKLEME
                PSLSFYNKARNYATKKPYSVEKFKLNFQMPTLASGWDVNKEKNNGAILFVKNGLYYLG
                IMPKQKGRYKALSFEPTEKTSEGFDKMYYDYFPDAAKMIPKCSTQLKAVTAHFQTHTT
                PILLSNNFIEPLEITKEIYDLNNPEKEPKKFQTAYAKKTGDQKGYREALCKWIDFTRD
                FLSKYTKTTSIDLSSLRPSSQYKDLGEYYAELNPLLYHISFQRIAEKEIMVAVETGKL
                YLFQIYNKDFAKGHHGKPNLHTLYWTGLFSPENLAKTSIKLNGQAELFYRPKSRMKRV
                AHRLGEKMLNKKLKDQKTPIPDTLYQELYDYVNHRLSHDLSDEARALLPNVITKEVSH
                EIIKDRRFTSDKFFFHVPITLNYQAANSPSKFNQRVNAYLKEHPETPIIGIDRGERNL
                IYITVIDSTGKILEQRSLNTIQQFDYQKKLDNREKERVAARQAWSVVGTIKDLKQGYL
                SQVIHEIVDLMIHYQAIVVLENLNFGFKSKRTGIAEKAVYQQFEKMLIDKLNCLVLKD
                YPAEKVGGVLNPYQLTDQFTSFAKMGTQSGFLFYVPAPYTSKIDPLTGFVDPFVWKTI
                KNHESRKHFLEGFDFLHYDVKTGDFILHFKMNRNLSFQRGLPGFMPAWDIVFEKNETQ
                FDAKGTPFIAGKRIVPVIENHRFTGRYRDLYPANELIALLEEKGIVFRDGSNILPKLL
                ENDDSHAIDTMVALIRSVLQMRNSNAATGEDYINSPVRDLNGVCFDSRFQNPEWPMDA
                DANGAYHIALKGQLLLNHLKESKDLKLQNGISNQDWLAYIQELRN"
gene            complement(104104..104742)
                /locus_tag="I3500192B8_00970"
CDS             complement(104104..104742)
                /locus_tag="I3500192B8_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="BFK75608.1"
```

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.

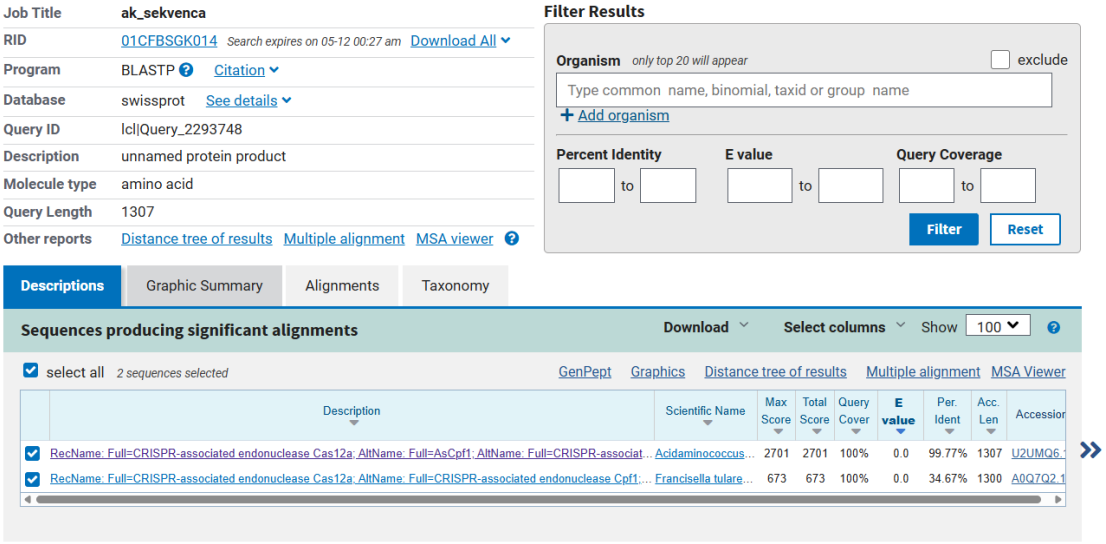


```
complete_geno      1 -------------------------MTQFEGFTNLYQVSKTLRFELIPQG     25
                                               ||||||||||||||||||||||||
synthetic_con      1 MSRRRKANPTKLSENAKKLAKEVENMTQFEGFTNLYQVSKTLRFELIPQG     50

complete_geno     26 KTLKHIQEQGFIEEDKARNDHYKELKPIIDRIYKTYADQCLQLVQLDWEN     75
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
synthetic_con     51 KTLKHIQEQGFIEEDKARNDHYKELKPIIDRIYKTYADQCLQLVQLDWEN    100

complete_geno     76 LSAAIDSYRKEKTEETRNALIEEQATYRNAIHDYFIGRTDNLTDAINKRH    125
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
synthetic_con    101 LSAAIDSYRKEKTEETRNALIEEQATYRNAIHDYFIGRTDNLTDAINKRH    150

complete_geno    126 AEIYKGLFKAELFNGKVLKQLGTVTTTEHENALLRSFDKFTTYFSGFYEN    175
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
synthetic_con    151 AEIYKGLFKAELFNGKVLKQLGTVTTTEHENALLRSFDKFTTYFSGFYEN    200

complete_geno    176 RKNVFSAEDISTAIPHRIVQDNFPKFKENCHIFTRLITAVPSLREHFENV    225
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
synthetic_con    201 RKNVFSAEDISTAIPHRIVQDNFPKFKENCHIFTRLITAVPSLREHFENV    250

complete_geno    226 KKAIGIFVSTSIEEVFSFPFYNQLLTQTQIDLYNQLLGGISREAGTEKIK    275
```

```
synthetic_con   1201 LYPANELIALLEEKGIVFRDGSNILPKLLENDDSHAIDTMVALIRSVLQM   1250

complete_geno   1226 RNSNAATGEDYINSPVRDLNGVCFDSRFQNPEWPMDADANGAYHIALKGQ   1275
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
synthetic_con   1251 RNSNAATGEDYINSPVRDLNGVCFDSRFQNPEWPMDADANGAYHIALKGQ   1300

complete_geno   1276 LLLNHLKESKDLKLQNGISNQDWLAYIQELRN----------------   1307
                     ||||||||||||||||||||||||||||||||
synthetic_con   1301 LLLNHLKESKDLKLQNGISNQDWLAYIQELRNGGGGGGGGGGKNLTQSRS   1350

complete_geno   1308 -------------------------------------------------   1307

synthetic_con   1351 TTWKLNNLLLNDYWVHNEMKAEIKMFFETNENKDTTYQNLWDAFKAVCRG   1400

complete_geno   1308 -------------------------------------------------   1307

synthetic_con   1401 KFIALNAYKRKQERSKIDTLTSQLKELEKQEQTHSKASRRQEITKIRAEL   1450

complete_geno   1308 -------------------------------------------------   1307

synthetic_con   1451 KEIETQKTLQKINESRSWFFERINKIDRPLARLIKKKREKNQIDTIKNDK   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.

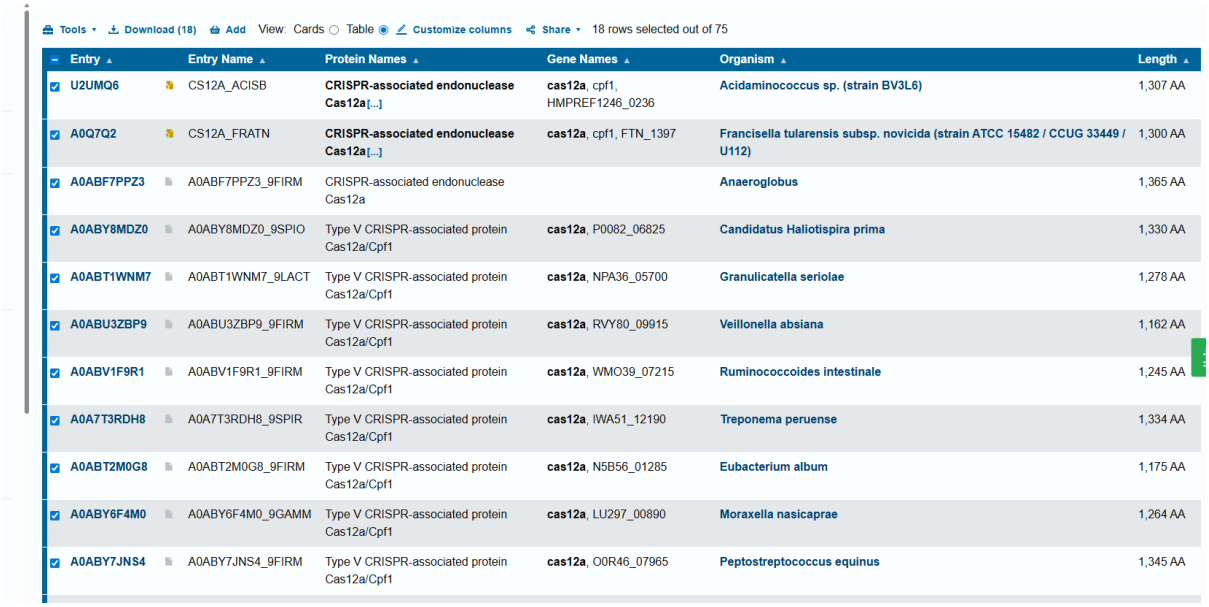
- Z analizo C- in N- konca daljšega zaporedja ugotovimo, da je prisoten nekakšen ekspresijski tag ali pa zgolj artefakt na N-koncu, saj ga pri naravni varianti ni. Na C-koncu pa opazimo heksahistidinsko oznako, ki služi pri čiščenju proteina (afinitetna kromatografija).
- Da je protein fuzijski lahko 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:



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
U2UMQ6	CS12A_ACISB	**CRISPR-associated endonuclease Cas12a**[...]	**cas12a**, cpf1, HMPREF1246_0236	Acidaminococcus sp. (strain BV3L6)	1,307 AA
A0Q7Q2	CS12A_FRATN	**CRISPR-associated endonuclease Cas12a**[...]	**cas12a**, cpf1, FTN_1397	Francisella tularensis subsp. novicida (strain ATCC 15482 / CCUG 33449 / U112)	1,300 AA
A0ABF7PPZ3	A0ABF7PPZ3_9FIRM	CRISPR-associated endonuclease Cas12a		Anaeroglobus	1,365 AA
A0ABY8MDZ0	A0ABY8MDZ0_9SPIO	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, P0082_06825	Candidatus Haliotispira prima	1,330 AA
A0ABT1WNM7	A0ABT1WNM7_9LACT	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, NPA36_05700	Granulicatella seriolae	1,278 AA
A0ABU3ZBP9	A0ABU3ZBP9_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, RVY80_09915	Veillonella absiana	1,162 AA
A0ABV1F9R1	A0ABV1F9R1_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, WMO39_07215	Ruminococcoides intestinale	1,245 AA
A0A7T3RDH8	A0A7T3RDH8_9SPIR	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, IWA51_12190	Treponema peruense	1,334 AA
A0ABT2M0G8	A0ABT2M0G8_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, N5B56_01285	Eubacterium album	1,175 AA
A0ABY6F4M0	A0ABY6F4M0_9GAMM	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, LU297_00890	Moraxella nasicaprae	1,264 AA
A0ABY7JNS4	A0ABY7JNS4_9FIRM	Type V CRISPR-associated protein Cas12a/Cpf1	**cas12a**, O0R46_07965	Peptostreptococcus equinus	1,345 AA

Zaporedja poravnamo s programom Clustal Omega

Tool Output | Alignments | Guide Tree | Phylogenetic Tree | Results Viewers | Result File

Tool output

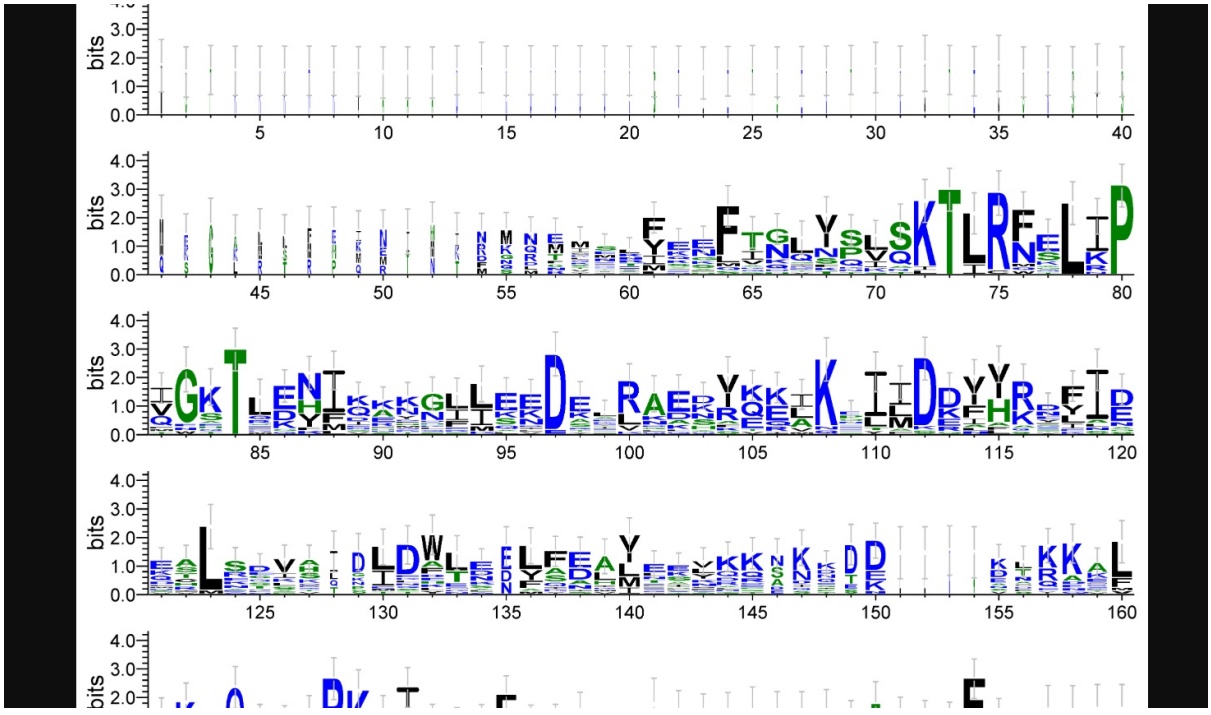
Download

CLUSTAL O(1.2.4) multiple sequence alignment

```
tr|A0ABT2M0G8|A0ABT2M0G8_9FIRM     ----------------------------------------------------MNQET       5
tr|A0ABT1WNM7|A0ABT1WNM7_9LACT     ----------------------------------------------------MVYN        4
sp|U2UMQ6|CS12A_ACISB              ----------------------------------------------------MTQ         3
tr|A0ABF7PPZ3|A0ABF7PPZ3_9FIRM     MGPKKKRKVAAADYKDDDDKSRLEPGEKPYKCPECGKSFSQSGALTRHQRTHTRMRTMVT       60
tr|A0ABU3ZBP9|A0ABU3ZBP9_9FIRM     ------------------------------------------------------------       0
tr|A0ABY7JNS4|A0ABY7JNS4_9FIRM     ----------------------------------------------------MIRKLEKSS       9
tr|A0A7T3RDH8|A0A7T3RDH8_9SPIR     ----------------------------------------------------MRTIDE       6
tr|A0A7C9H0Z9|A0A7C9H0Z9_9FIRM     ----------------------------------------------M---NGNRIIV       8
tr|A0AAP3Q149|A0AAP3Q149_9FIRM     ----------------------------------------------------MNNGTNN       7
tr|A0ABR7NBZ6|A0ABR7NBZ6_9FIRM     ---------------------------------------MEGKRSFEKN---NQNMYNR       17
tr|A0ABT2SLB2|A0ABT2SLB2_9FIRM     ----------------------------------------------ME---DKQFLER       9
tr|A0ABV1F9R1|A0ABV1F9R1_9FIRM     ----------------------------------------------------MQERKK       6
tr|A0ABV1HU47|A0ABV1HU47_9FIRM     ---------------------------------------MKGLMLMPININKFSDECRK       20
tr|A0ABV5GTK2|A0ABV5GTK2_9FLAO     ------------------------------------------------------------       0
tr|A0ABW8U4L5|A0ABW8U4L5_9GAMM     ------------------------------------------------------------ML       2
tr|A0ABY6F4M0|A0ABY6F4M0_9GAMM     ------------------------------------------------------------ML       2
sp|A0Q7Q2|CS12A_FRATN              ------------------------------------------------------------MSI       3
tr|A0ABY8MDZ0|A0ABY8MDZ0_9SPIO     ------------------------------------------------------------MSSL       4

tr|A0ABT2M0G8|A0ABT2M0G8_9FIRM     IEKMAGLNKKTLTICQKLIPVGKTRENIDKFNSMAEDEFVKANKEKINTLIKKAASEKID       65
tr|A0ABT1WNM7|A0ABT1WNM7_9LACT     LKESIGTKSLSKTLRFTLIPQNQTREYLSKNLLKEE--EVDLQFEQAKLILDGIYRKIIE       62
sp|U2UMQ6|CS12A_ACISB              FEGFTNLYQVSKTLRFELIPQGKTLKHIQEQGFIEEDKARNDHYKELKPIIDRIYKTYAD       63
tr|A0ABF7PPZ3|A0ABF7PPZ3_9FIRM     FENFTKQYQVSKTLRFELIPQGKTLENMKRDGIISVDRQRNDDYQKAKGILDKLYKYILD       120
tr|A0ABU3ZBP9|A0ABU3ZBP9_9FIRM     MSKFQNLYTINKTLRFGLKPFGKTLENFNKTNLLQLDEYKAKHRKEVQRLFDENFKQLIE       60
tr|A0ABY7JNS4|A0ABY7JNS4_9FIRM     YESFTQLFPKQITLRNELIPMWETKDNMISMKEIEIDKNRAEDYKRIKSIIDDFYRVLIN       69
tr|A0A7T3RDH8|A0A7T3RDH8_9SPIR     FCGQKNGYSLSKTLRNKLIPVGKTEDKIKEYHLMENDYGRAAAYVEVKNLIDDFHRSFIA       66
tr|A0A7C9H0Z9|A0A7C9H0Z9_9FIRM     YREFVGVTPVAKTLRNELRPIGHTQEHIIHNGLIQEDELRQEKSTELKNIMDDYYREYID       68
tr|A0AAP3Q149|A0AAP3Q149_9FIRM     FQNFIGISSLQKTLRNALIPTETTQQFIVKNGIIKEDELRGENRQILKDIMDDYYRGFIS       67
tr|A0ABR7NBZ6|A0ABR7NBZ6_9FIRM     YRELIGLSSSSKTLRNSLIPVGSTLKYIQKHGILEKDTLRAEKREELKAIMDDYYRNYID       77
```

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

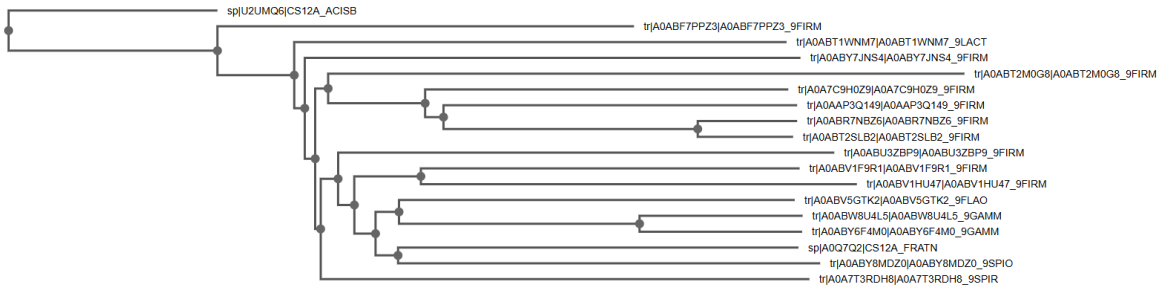
Del WebLogo, ki sem ga sam dobil je prikazan spodaj (glej tudi dodatni dokument z rešitvami):



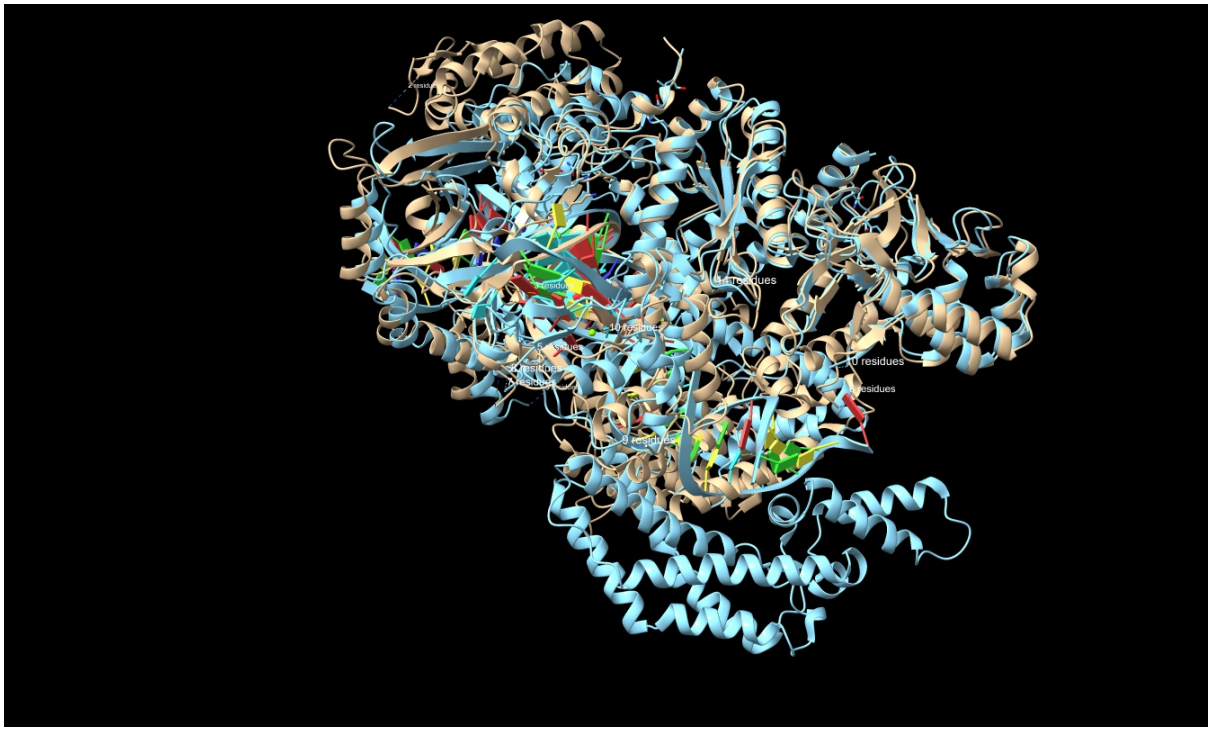
Ko pogledamo in preračunamo kje v poravnavi se začne protein iz našega organizma ugotovimo na katerih mestih so zelo ohranjeni aminokislinski ostanki v primerjavi z drugimi istoimenskimi proteini iz drugih organizmov.

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). Kodi PDB sta že podani. Z orodnjem 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).