Genome Annotation Project

A.M. Ismailov

Federal State Autonomous Educational Institution for Higher Education «National Research University Higher School of Economics», Russian Federation, 101000, Moscow

*** Contact Information: Ismailov Aly Mekhtiyevich, Master's degree, HSE FCS Data Analysis in Biology and Medicine**

Email: neuro.promotion@gmail.com

Introduction:
The main goal of this project is to annotate the specified genomic region of an unknown microorganism, which is 21,001 base pairs long.

Annotation of this genome fragment includes:

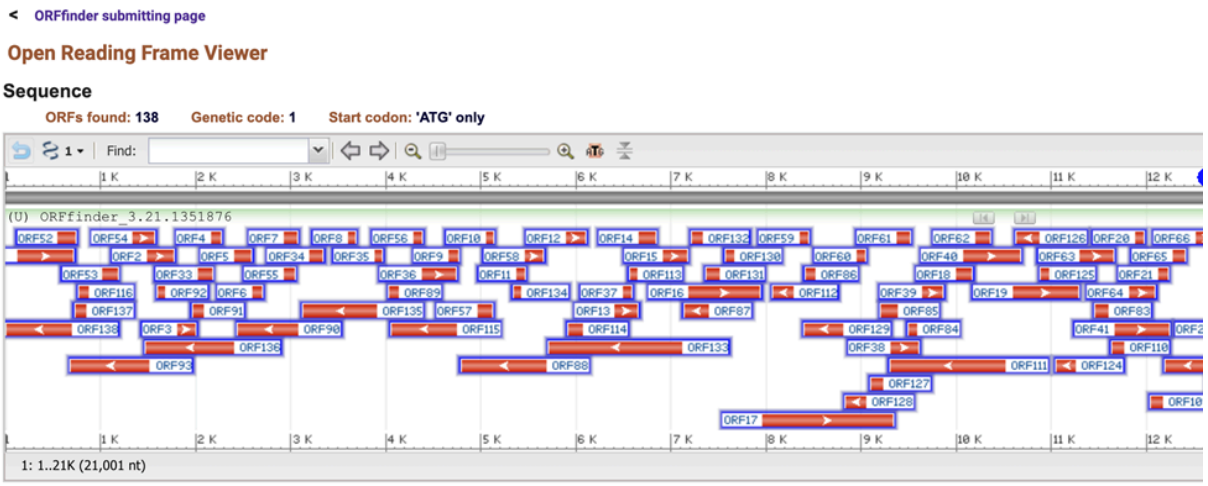
- Primary annotation of the sequence
- Analysis of functional gene products
- Analysis of operons and their structure
- Analysis of non-coding RNAs (if present)
- Identification of regulatory regions
- Detection of genes acquired through horizontal gene transfer

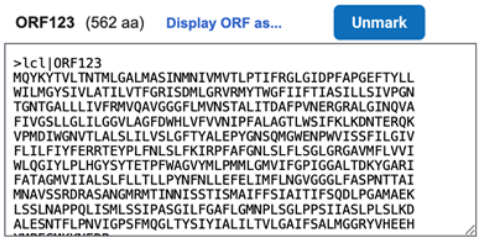
Materials and Methods:
Open reading frames (ORFs) were identified using the ORFfinder tool, followed by BLASTP analysis. tRNA detection was performed using tRNAscan. Automatic annotation of the region was conducted using the Prokka and RAST programs with standard settings. InterPro was utilized for domain structure identification and family assignment. Operon prediction was carried out using the Operon Mapper tool. Regulatory element search was conducted using the RSAT tool. Data analysis was performed using the Python programming language, including libraries such as pandas and Biopython.

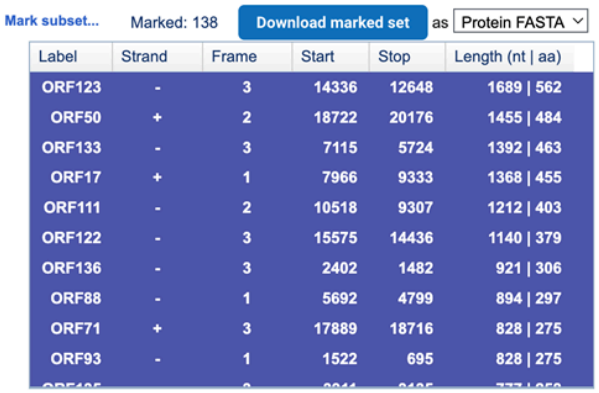
Results:

● Primary sequence annotation

A total of 138 open reading frames (ORFs) were found, with a minimum length of 75 base pairs and a start codon ATG.







Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF123	-	3	14336	12648	1689 | 562
ORF50	+	2	18722	20176	1455 | 484
ORF133	-	3	7115	5724	1392 | 463
ORF17	+	1	7966	9333	1368 | 455
ORF111	-	2	10518	9307	1212 | 403
ORF122	-	3	15575	14436	1140 | 379
ORF136	-	3	2402	1482	921 | 306
ORF88	-	1	5692	4799	894 | 297
ORF71	+	3	17889	18716	828 | 275
ORF93	-	1	1522	695	828 | 275

The search for similar alignments was conducted using BLASTP. The assessment of the findings was primarily based on the percentage of identity and E-value. A finding was considered good if it had more than 70% identity and an E-value < 0.001. The process proceeded in three stages: first, the SwissProt database was used, resulting in the discovery of 9 good findings. The remaining ORFs not found in the SwissProt database were then queried against the RefSeq database, resulting in 0 good findings. Finally, the remaining ORFs were queried against the non-redundant protein sequences (nr) database, resulting in 0 good findings.

Brief summary table of the best findings for each ORF:

Swissprot:

ORF\feature	Description	Organism	E-val	Per. Ident.
16	Uncharacterized metal-dependent hydrolase	Methanocaldococcus jannaschii	3e-43	33.86%
16	D-aminoacyl-tRNA deacylase	Synechocystis sp.	2e-09	29.58%
17	Signal recognition particle 54 kDa protein	Thermoplasma volcanium	0	56.89%
19	Polyamine aminopropyltransferase	Putrescine aminopropyltransferase	2e-09	24.50%
19	Methyltransferase-like protein 13	Homo sapiens	1e-07	32.77%
31	30S ribosomal protein S4	Picrophilus oshimae	1e-66	58.14%
41	Putative zinc metalloprotease MJ0611	Methanocaldococcus jannaschii	3e-26	40.91%
48	30S ribosomal protein S13	Pyrococcus horikoshii	7e-50	53.42%

50	Lysyl-tRNA synthetase 2	Methanosarcina mazei	2e-159	50.21%
51	Uncharacterized ABC transporter ATP-binding protein MJ0412	Methanocaldococcus jannaschii	5e-78	45.93%
51	Aliphatic sulfonates import ATP-binding protein SsuB	Bacillus cereus	1e-66	44.17%
63	Uncharacterized protein MJ1618	Methanocaldococcus jannaschii	2e-05	29.41%
70	30S ribosomal protein S11	Picrophilus oshimae	2e-58	71.55%
71	DNA-directed RNA polymerase subunit D	Thermoplasma acidophilum	6e-107	56.36%
80	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	Thermoplasma acidophilum	3e-42	38.97%
88	Phospholipid-transporting ATPase ABCA7	Homo sapiens	2e-35	38.53%
90	Fibrillarin-like rRNA/tRNA 2'-O-methyltransferase	Nanoarchaeum equitans Kin4-M	9e-60	49.26%
93	UPF0273 protein PF1931	Pyrococcus furiosus	0.007	42.86%
111	Low silicon protein 3	Oryza sativa Japonica Group	1e-26	39.55%

122	Probable L-galactonate transporter	Escherichia coli	4e-11	26.34%
123	Riboflavin transporter RibZ	Clostridioides difficile	2e-35	30.36%
133	Riboflavin transporter RibZ	Clostridioides difficile	3e-34	27.32%
135	Undecaprenyl diphosphate synthase	Thermoplasma acidophilum	3e-103	55.25%
136	Galactowaldenase	Methanocaldococcus jannaschii	4e-41	34.25%
138	UPF0273 protein MK0039	Methanopyrus kandleri	3e-39	35.39%

Refseq:

ORF\feature	Description	Organism	E-val	Per. Ident.
12	hypothetical protein	Caldivirga sp.	2e-04	53.19%
82	MarR family transcriptional regulator	Pseudodonghicola xiamenensis	4e-09	33.33%
115	ABC transporter permease	Thermogymnomonas acidicola	3e-95	61.95%

NR:

ORF\feature	Description	Organism	E-val	Per. Ident.
24	hypothetical protein FDG2_6054	Candidatus Protofrankia californiensis	7e-07	40.82%
40	Uncharacterised protein	Clostridioides difficile	4e-05	31.91%
94	Protein of uncharacterised function	Bordetella pertussis	3e-07	48.39%
129	MAG: hypothetical protein A4E43_00718	Methanosaeta sp. PtaB.Bin005	2e-10	58.33%

As you can see, only ORF 70 yielded the desired result. Let's take a closer look at the findings regarding this genomic region:

Swissprot (ORF 70):

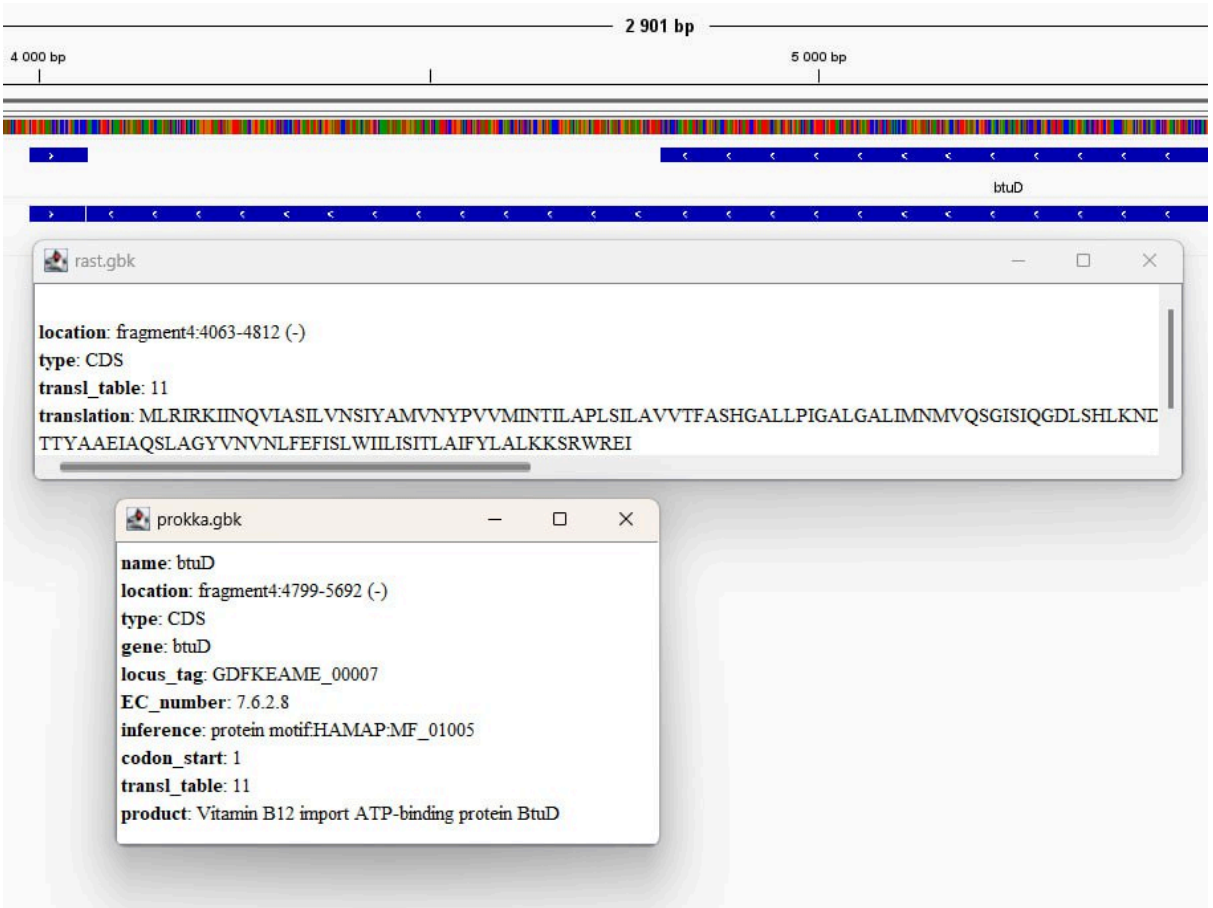
Description	Scientific name	E-value	Perc. indent
30S ribosomal protein S11	Thermococcus sibiricus MM 739	2e-50	71.79%
30S ribosomal protein S11	Picrophilus oshimae DSM 9789	2e-58	71.55%
30S ribosomal protein S11	Methanosarcina mazei Go1	2e-56	71.07%
30S ribosomal protein S11	Thermococcus kodakarensis KOD1	2e-49	70.94%
30S ribosomal protein S11	Thermococcus gammatolerans EJ3	3e-48	70.94%
30S ribosomal protein S11	Methanosarcina barkeri str. Fusaro	3e-55	70.25%
30S ribosomal protein S11	Methanosarcina acetivorans C2A	8e-56	70.25%
30S ribosomal protein S11	Pyrococcus furiosus DSM 3638	1e-46	70.09%
30S ribosomal protein S11	Pyrococcus abyssi GE5	4e-47	70.09%
30S ribosomal protein S11	Methanocella arvoryzae MRE50	2e-58	69.6%

All organisms in the table are extremophilic archaea. Notably, Picrophilus oshimae has one of the lowest known optimal pH values for growth.

Automatic annotation using RAST and Prokka:

Using Prokka, 23 coding sequences were identified, while RAST identified 24. Let's compare them using the Integrative Genomics Viewer (IGV):

Most of the coding regions are identical, except for one finding present in RAST but not in Prokka. Let's take a closer look:



Here is the amino acid sequence of the potential protein found by RAST but not by Prokka:
MLRIRKIINQVIASILVNSIYAMVNYPVVMINTILAPLSILAVVTFASHGALLPIGALGALIMNMVQS
GISIQGDLSHLKNDMKLQDMVVSSPTSAGIYIFGMAISEIVYSLPTITLLLILNFLFVKASIIAWILIFL
DMALIFTFSIALGFLLSTFSSDIVQSWAFAGILTPLLSTIPPVYYPITYIPLPFRYISYLSPTTYAAEIAQS
LAGYVNVNLFEFISLWIILISITLAIFYLALKKSRWREI

According to RAST data, this is an "Efflux ABC transporter, permease protein." The gene is unidentified. Let's try using BLASTP to determine its origin:

The protein was found in the Thermoplasmatales archaeon, and its name is "ABC transporter permease." The percentage of identity is 100%, with an E-value of 5e-156, and the accession number is MBD6955528.1.

Let's look at the table of identified coding sequences and the BLASTP results:

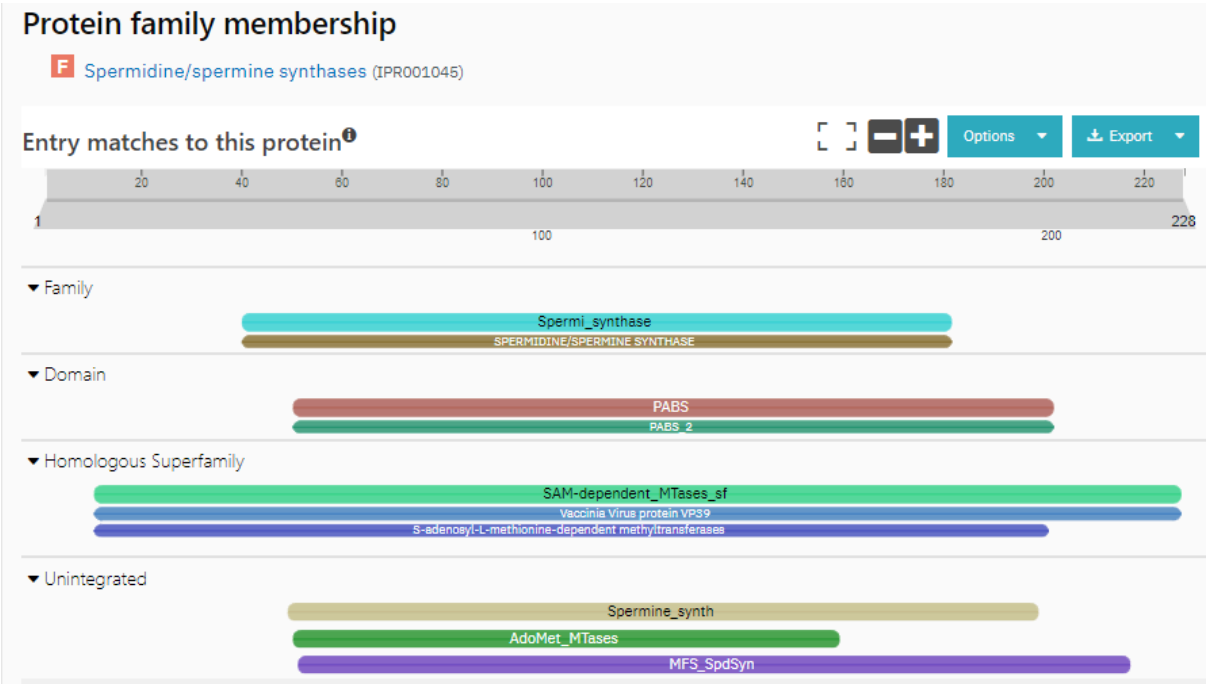
N	Location	Gene[RAST]	Product[RAST]	Gene[Prokka]	Product[Prokka]	BLASTP best findings: - protein name - organizm name - percent identity
1	8:695	Unknown	hypothetical protein	kaiC	Circadian clock protein kinase KaiC	-KaiC domain-containing protein -**Thermoplasmatales archaeon** -100%
2	694:1498 [RAST] 694:1522 [Prokka]	Unknown	hypothetical protein	Unknown	hypothetical protein	-recombinase RecA -**Thermoplasmatales archaeon** -100%
3	1481:2456	Unknown	UDP-glucose 4-epimerase	galE1	UDP-glucose 4-epimerase	-NAD-dependent epimerase/dehydratase family protein -**Thermoplasmatales archaeon** -100%
4	2452:3088	Unknown	Box C/D RNA-guided RNA methyltransferase subunit fibrillarin	Unknown	hypothetical protein	-fibrillarin-like rRNA/tRNA 2'-O-methyltransferase -**Thermoplasmatales archaeon** -99.53%
5	3134:3911	Unknown	Undecaprenyl diphosphate synthase	Unknown	(2Z,6E)-farnesyl diphosphate synthase	-Di-trans,poly-cis-decaprenylcistransferase -**Thermoplasmatales archaeon** -100%
-	4062:4812	Unknown	Efflux ABC transporter, permease protein	-	-	-ABC transporter permease -**Thermoplasmatales archaeon** -100%
6	4798:5692	Unknown	hypothetical protein	btuD	Vitamin B12 import ATP-binding protein BtuD	-ABC transporter ATP-binding protein -**Thermoplasmatales archaeon** -99.66%
7	5723:7115	Unknown	hypothetical protein	ribZ	Riboflavin transporter RibZ	-MFS transporter -**Thermoplasmatales archaeon** -100%
8	7188:7941	Unknown	hypothetical protein	dtd3	D-aminoacyl-tRNA deacylase	- hydrolase TatD - **Aciduliprofundum sp.** - 100%

9	7965:9333	Unknown	Signal recognition particle protein Ffh	ffh	Signal recognition particle protein	- signal recognition particle protein -**Thermoplasmatales archaeon** -100%
10	9306:10518	Unknown	putative anion permease	Unknown	putative transporter	- anion transporter -**Thermoplasmatales archaeon** -100%
11	10593:11280	Unknown	hypothetical protein	speE	Polyamine aminopropyltransferase	-hypothetical protein -**Thermoplasmatales archaeon** -100%
12	11303:11654	Unknown	putative carbohydrate binding protein	Unknown	hypothetical protein	-cupin domain-containing protein -**Thermoplasmatales archaeon** -100%
13	11662:12241	Unknown	hypothetical protein	Unknown	hypothetical protein	-site-2 protease family protein -**Thermoplasmatales archaeon** -100%
14	12202:12661	Unknown	hypothetical protein	Unknown	hypothetical protein	-MarR family transcriptional regulator -**Thermoplasmatales archaeon** -98.68%
15	12647:14336	Unknown	Permease, multidrug efflux	mdtD	Putative multidrug resistance protein MdtD	-MFS transporter -**Thermoplasmatales archaeon** -100%
16	14435:15575	Unknown	hypothetical protein	ycaD	putative MFS-type transporter YcaD	-MFS transporter -**Thermoplasmatales archaeon** -100%
17	15589:16156	Unknown	hypothetical protein	pspB	Putative phosphoserine phosphatase 2	-histidine phosphatase family protein -**Thermoplasmatales archaeon** -100%
18	16483:16945	Unknown	SSU ribosomal protein S18e (S13p)	rpsM	30S ribosomal protein S13	-30S ribosomal protein S13 -**Thermoplasmatales archaeon** -100%
19	16944:17487	Unknown	SSU ribosomal protein S9e (S4p)	Unknown	hypothetical protein	-30S ribosomal protein S4 -**Thermoplasmatales archaeon** -100%
20	17483:17867	Unknown	SSU ribosomal protein S14e (S11p)	rpsK	30S ribosomal protein S11	-30S ribosomal protein S11 -**Thermoplasmatales archaeon** -100%
21	17888:18716	Unknown	DNA-directed RNA polymerase subunit D	Unknown	hypothetical protein	-DNA-directed RNA polymerase subunit D -**Thermoplasmatales archaeon** -100%
22	18721:20176	Unknown	Lysyl-tRNA synthetase (class II)	lysS	Lysine--tRNA ligase	-lysine--tRNA ligase -**Thermoplasmatales archaeon** -100%
23	20203:20956	Unknown	ABC transporter,	cmpD	Bicarbonate	-ABC transporter ATP-binding protein

			ATP-binding protein		transport ATP-binding protein CmpD	-**Thermoplasmatales archaeon** -100%

Thus, practically all findings correspond to the **Thermoplasmatales archaeon**, except for the eighth coding sequence, which corresponds to the hydrolase TatD protein of the microorganism **Aciduliprofundum sp.**

In the case of the 11th sequence, the function of the protein was not determined using BLASTP and RAST. However, Prokka identified this protein as D-aminoacyl-tRNA deacylase, which is an enzyme playing a crucial role in the protein translation mechanism. This enzyme specifically recognizes and removes D-amino acids that may be erroneously added to transfer RNA (tRNA) during protein synthesis. Let's use the InterPro service to determine the domain characteristics of the protein:



The family of enzymes Spermidine/spermine synthase belongs to the group of enzymes involved in the biosynthesis of polyamines such as spermidine and spermine. Polyamines play a crucial role in cellular growth, division, and differentiation, as well as in the regulation of gene expression. PABS (polyamine biosynthesis domain) is a structural element present in enzymes responsible for polyamine synthesis.

● Analysis of gene functional products

1. *KaiC domain-containing protein* - the protein "KaiC domain-containing protein" is found in organisms with internal clocks, such as cyanobacteria, bacteria, and some archaea. It is a key component of bacterial circadian systems, such as the KaiABC system in cyanobacteria, where it regulates the phosphorylation and dephosphorylation of other proteins in this system.
2. *recombinase RecA* - plays a key role in genetic recombination, DNA repair, and regulation of stress responses. It is capable of catalyzing the exchange of genetic material between two identical or nearly identical DNA segments, leading to the formation of crossover structures and allowing the resolution of problems associated with DNA damage or defects. RecA also participates in processes of repairing double-strand breaks and restoring damaged DNA.
3. *UDP-glucose 4-epimerase* - It is an enzyme that catalyzes the epimerization reaction of UDP-glucose to UDP-galactose. This is an important step in the biosynthesis of galactose, which is necessary for the formation of galactose from glucose. Galactose plays a key role in various biological processes, such as the formation of glycoproteins, glycolipids, and glycans, and can be used as an energy source.
4. *fibrillarin-like rRNA/tRNA 2'-O-methyltransferase* - This enzyme catalyzes the methylation of the 2'-O-ribose group in ribosomal RNA (rRNA) and transfer RNA (tRNA). This process of modification of ribosomal and transfer RNA plays an important role in their structural and functional stability, as well as in the accuracy of genetic information translation.
5. *Di-trans,poly-cis-decaprenylcistransferase* - This enzyme participates in the biosynthesis of terpenes
6. *ABC transporter permease, ABC transporter ATP-binding protein* - This is a superfamily of transport systems, which is one of the largest and possibly one of the oldest gene families. The ATPase subunits utilize the energy from binding and hydrolyzing adenosine triphosphate (ATP) to provide the necessary energy for transporting substrates across membranes, both for substrate uptake and export.
7. *MFS transporter* - The main superfamily of facilitators is a superfamily of membrane transport proteins that facilitate the movement of small dissolved substances across cell membranes in response to chemiosmotic gradients.
8. *hydrolase TatD* - It often participates in various biological processes, such as RNA and DNA processing and degradation, as well as in the repair of damaged nucleic acids.
9. *signal recognition particle protein* - This is a common cytosolic universally conserved ribonucleoprotein that recognizes and directs specific proteins to the endoplasmic reticulum in eukaryotes and to the plasma membrane in prokaryotes.
10. *anion transporter* - Anion transporter protein

11. *cupin domain-containing protein* - It belongs to the family of proteins containing a characteristic structural domain called a cupin. The cupin domain is one of the most widely distributed structural motifs in nature and is found in various classes of enzymes and proteins. It typically consists of two β-sheets forming a beta-barrel, which can vary in size and shape. Proteins with the cupin domain perform diverse functions, including catalyzing chemical reactions, binding to metals and other ligands, and participating in various biological processes.
12. *site-2 protease family protein* - It belongs to the group of proteases capable of cleaving protein chains at specific sites. The serine protease family includes enzymes that play an important role in proteolytic cascades, regulating the processing and activation of various proteins.
13. *MarR family transcriptional regulator*- It is a member of the family of proteins involved in gene expression regulation. It typically acts as a transcriptional repressor, controlling gene activity by binding to specific DNA sequences.
14. *histidine phosphatase family protein* - It is a member of the family of enzymes capable of hydrolyzing phosphate from histidine residues in proteins. Histidine phosphatases participate in the regulation of various cellular signaling pathways, as histidine residues play an important role in signal transduction inside cells.
15. *30S ribosomal protein S13, S11, S4* - ribosomal RNA proteins
16. *DNA-directed RNA polymerase subunit D* - protein subunit D of DNA-dependent RNA polymerase.
17. *lysine--tRNA ligase* - It's the enzyme that catalyzes the attachment of the amino acid lysine to its corresponding transfer RNA (tRNA).

● The analysis of operons and their structure

To analyze operons, the initial genome fragment in FASTA format and the GFF file obtained from the automatic annotation by the RAST service were used. This choice was made because RAST predicted more coding sequences than Prokka.

Operon	Start	Stop	Strand	Function
1	9	695	-	-RecA-superfamily ATPases implicated in signal transduction
	695	1498	-	-RecA-superfamily ATPases implicated in signal transduction
	1482	2456	-	-Nucleoside-diphosphate-sugar epimerases -Fibrillarin-like rRNA methylase
	2453	3088	-	-Undecaprenyl pyrophosphate synthase
	3135	3911	-	

2	4063 4799 5724	4812 5692 7115	- - -	-Uncharacterized membrane-associated protein -ABC-type Na+ transport system, ATPase component -Na+/melibiose symporter and related transporters
3	7189 7966	7941 9333	+ +	-Predicted metal-dependent hydrolases with the TIM-barrel -Signal recognition particle GTPase
4	9307	10518	-	-Na+/H+ antiporter NhaD and related arsenite
5	10594 11304 11663	11280 11654 12241	+ + +	-Spermidine synthase -Mannose-6-phosphate isomerase -Zn-dependent proteases
6	12203 12648 14436 15590	12661 14336 15575 16156	- - - -	-Transcriptional regulators -Nitrate/nitrite transporter -Sugar phosphate permease -Fructose-2,6-bisphosphatase
7	16484 16945 17484 17889 18722 20204	16945 17487 17867 18716 20176 20956	+ + + + + +	-Ribosomal protein S13 -Ribosomal protein S4 and related proteins -Ribosomal protein S11 -DNA-directed RNA polymerase, alpha subunit/40 kD -Lysyl-tRNA synthetase (class II) -ABC-type nitrate/sulfonate/bicarbonate transport system, ATPase component
8	3989	4061	+	tRNA

● Non-coding RNA analysis

One transfer RNA (tRNA) was found during the analysis.

Results

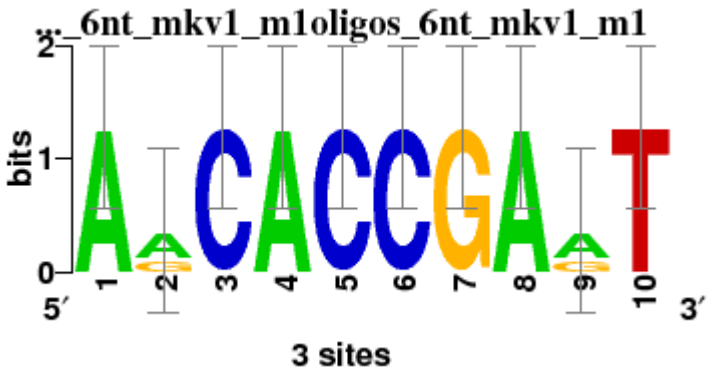
Download as text

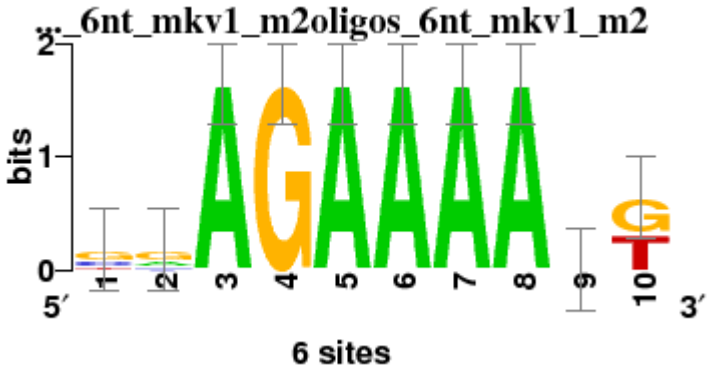
Sequence Name	tRNA #	Predicted tRNA Structure	Similar tRNAs in GtRNAdb	tRNA Begin	tRNA End	tRNA Type	Anticodon	Intron Begin	Intron End	Infernal Score	Isotype Model	Isotype Score	Note
Muhammad_Daha_Garba	1	View	View	3989	4061	Gly	GCC	0	0	64.0	Arg	57.7	IPD:-0.40

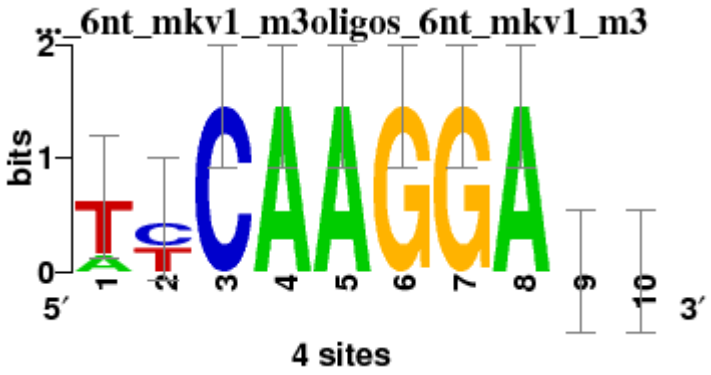
Isotype-Specific Model Scores:

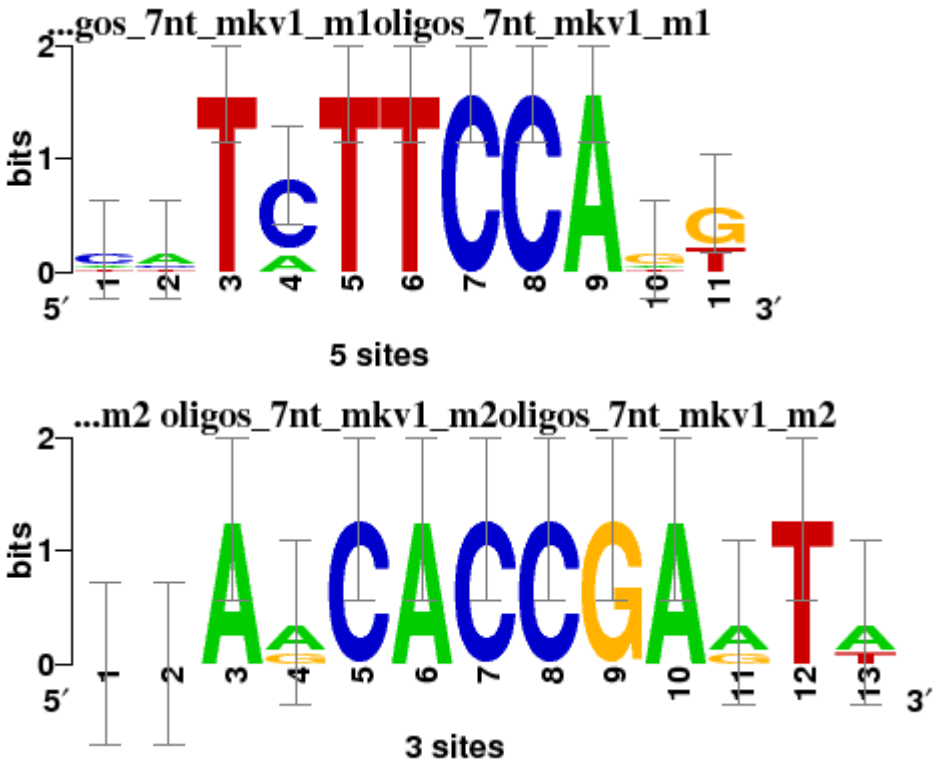
- Search for regulatory regions

Five regulatory regions have been identified in the genome:









● Detection of genes acquired through horizontal gene transfer.

Most of the protein-coding genes correspond to proteins from Thermoplasmatales archaeon. However, one coding sequence corresponds to another microorganism species. Aciduliprofundum sp. is an archaeal species that inhabits extreme environments such as deep-sea geysers and volcanic formations on the ocean floor. This microorganism is adapted to high temperatures, high pressure, acidity, and other extreme conditions typical of such environments. The "Hydrolase TatD" protein is an enzyme belonging to the hydrolase class, capable of catalyzing the hydrolysis of chemical bonds using water. The TatD enzyme belongs to the TatD/TatD-related hydrolase family and plays a role in various biological processes such as metabolism, DNA replication, and gene expression regulation. I'll speculate and suggest that this gene was likely acquired through horizontal gene transfer.

Conclusion

The genome fragment corresponds to the **Thermoplasmatales archaeon**. Thermoplasmatales archaeon is a general term for archaea from the order Thermoplasmatales. These are archaeal organisms that inhabit extreme environments with

high temperatures, pH levels, and osmotic stress. They are known to inhabit various locations, including hot springs, volcanic formations, hydrothermal vents, and other extreme environments.