Skoltech
Skolkovo Institute of Science and Technology

Bioinformatics course

Archaean mystery: the final project

Nikita Vaulin

Introduction

Thermoplasmata is a species of archaea in the phylum Euryarchaeota [1, 2, 3]. The study of archaea has become especially relevant recently, as more information is appearing about the involvement of colonies of extremophilic organisms in the global cycle of substances[1, 2, 3]. However, their metabolic pathways remain incompletely understood. A more detailed understanding of these processes could be useful from both ecological and biotechnological perspectives. A special factor in this case may be the horizontal gene transfer (HGT) occurring in microbial communities.

Recently, a genome of an archaea from microbial communities from various locations in Russia was sequenced. In this work, a fragment of this sequencing is used for a detailed analysis.

Materials and methods

Genomic data

For the project the part of unclassified Thermoplasmata genome 22 521 bp long was used. The data is available in the GitHub repository (see Supplemental materials).

Gene annotation

To perform gene prediction and annotation Prokka v. 1.14.6 was used with parameters `--kingdom Archaea --rfam` [4]. For the homologs search command line version blast tools were used v. 2.13.0 [5] with the SwissProt database [6]. A database for blast search was created with makeblastdb tool. For the evolutionary analysis web version BLAST was used with the reporting results as a distant tree built by Neighbor Joining method.

RNA-coded genes prediction was conducted with the barrnap v. 0.9 [7] and tRNAscan-SE v. 2.0.9 [8].

For the protein domains prediction HMMER v. 3.3.2 was used with the Pfam database [9] and the 10^{-4} E-value cut-off.

Secondary metabolites search was conducted with antiSMASH web server [10]. Prokaryotic defence systems elements were searched with PADLOC web tool [11].

Operon identification

For the operon identification Operon-mapper web server and *merge* function from the bedtools v. 2.30.0 package with the parameters `-d 149 -s -c 7,4 -o distinct,count` were used [12, 13]. For the gene synteny analysis SynTax web tool were used against *Candidatus Thermoplasmatota* chromosomes [14].

№	Start	Stop	Strand	Genes
1	2	6917	-	*iolD, 02, tdh, 04, mdtD*
2	10834	14887	+	*08, IMPDH, tldD, tldD/pmbA*
3	14091	15885	-	*rlmE, gcvH*
4	16475	19174	-	*MJ0916, tnpB, 17, cntAB*
5	19268	22472	+	*19, aglaA, exuT*

Table 1. Found operons (with more than 1 gene in the operon)

Results

Protein annotation

Gene prediction with Prokka yielded 21 coding sequences (see Table 2). The structure of the locus under study is shown in the Fig. 1. Proteins functions was determined for 10 of the coding sequences (Fig.1, solid outline). For 5 genes without defined function, homologs were found in the SwissProt database (Fig. 1, dashed outline with annotation). The numerical annotation corresponds to the ID of the entry in the Prokka results. Protein domains were identified for 4 of their 5 genes without defined functions or homologs. For 1 protein, no definite function or any protein domains could be predicted.

Operon structure

For the locus under this study 5 operons were identified, which are described in the Table 1 and marked in different colors in the Fig. 1. No upstream regulatory motifs were identified for these operons (MEME search resulted only in hits with E-value > 0).

Functional loci and RNA-coding genes

No loci associated with prokaryotic defense systems or secondary metabolites were found for this nucleotide sequence. Also, no genes encoding any RNA were found.

Identification of closest relatives

Since there were no rRNA genes or at least simply housekeeping genes there are in the sample, the closest relatives were identified with the blastp search coupled with the distant tree building. According to this, the *Thermoplasmatales archaeon*, *[Aciduliprofundum sp.* and *[Euryarchaeota archaeon* are the closes relatives (Fig. 2).

Also a synteny-based approach was used. Several predicted genes were used to search for taxones with similar locus structure. Among them, the most frequent close relatives were *Thermoplasma volcanium*, *Thermoplasma acidophilum* and *Candidatus Methanoplasma termitum*.

Compiled on: December 25, 2022.

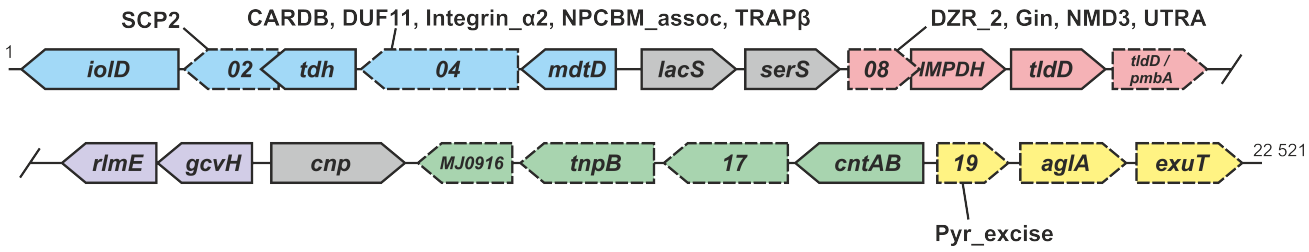

Figure 1. Genetic structure of the locus under study. Different colors indicate different groups of operons, the gray color indicates genes that are not included in any operon. Annotation is given for genes predicted with Prokka (solid outline) or for the nearest homolog defined with BLAST (dotted line). Numerically signed genes for which no function was found with Prokka and BLAST; for such genes, where possible, protein domains predicted with HMMER are signed.

№	Gene	Prokka feature	BLAST Hit	BLAST Acc.	HHMER domains
01	*iolD*	3D-(3,5/4)-trihydroxy-cyclohexane-1,2-dione hydrolase	Probable acetolactate synthase large subunit	Q57725	
02	-	-	-	-	SCP2
03	*tdh*	L-threonine 3-dehydrogenase	Uncharacterized zinc-type alcohol dehydrogenase-like protein YdjJ	P77280	
04	-	-	-	-	CARDB, DUF11, Integrin_α2, NPCBM, TRAPβ
05	*mdtD*	Putative multidrug resistance protein MdtD	Tetracenomycin C resistance and export protein	P39886	
06	*lacS*	Beta-galactosidase	Beta-galactosidase	P50388	
07	*serS*	Serine–tRNA ligase	Serine–tRNA ligase	O58441	
08	-	-	-	-	DZR_2, Gin, NMD3, UTRA
09	*IMPDH*	Inosine-5'-monophosphate dehydrogenase	Uncharacterized protein MJ1404	Q58799	
10	*tldD*	Zinc metalloprotease TldD	Metalloprotease MTH_856	O26944	
11	*pmpbA*	-	Metalloprotease MJ0996	Q58403	
12	*rlmE*	Ribosomal RNA large subunit methyltransferase E	Ribosomal RNA large subunit methyltransferase E	O28228	
13	*gcvH*	Glycine cleavage system H protein	Glycine cleavage system H protein	B0K24	
14	*cnp*	RNA 2',3'-cyclic phosphodiesterase	RNA 2',3'-cyclic phosphodiesterase	Q58963	
15	*MJ0916*	-	Uncharacterized protein MJ0916	Q58326	
16	*tnpB*	-	TnpB-like protein MJ0751	Q58161	
17	-	-	-	-	-
18	*cntAB*	Carnitine monooxygenase oxygenase subunit	holesterol 7-desaturase nvd	F7J188	
19	-	-	-	-	Pyr_excise
20	*aglA*	-	Alpha-glucosidase	O33830	
21	*exuT*	-	Hexuronate transporter	O34456	

Table 2. Genes found in the loci of interest, a dash means the lack of data

Search for horizontal gene transfer events

According to the blast protein search, every protein except for the Prokka ID 17 one have some more or les similiar homologs in the highlighted closest relatives. Also, a synteny-based resulted no similar operons for the Prokka ID 17 protein.

Dicussion

Operon structure

In this work, the operon structure of the studied locus of the Archaea genome was determined. The predicted operons and protein functions are well correlated with the use of different methods. Nevertheless, it cannot be called a complete *bona fide* operon, since no regulatory sequences were found. Also, I am unable to clearly identify monofunctional operons. All of the ones presented here have genes with quite different functions.

For example, the operon highlighted in purple in Fig. 1 (the *rlmE* and *gcvH* genes) encodes two functional group transfer proteins. The rlmE protein transfers a methyl group to the 23S rRNA [15]. The gcvH protein is a member of GCS system and transfers the methylamine group from pyridoxal phosphate-dependent glycine decarboxylase to tetrahydrofolate-requiring aminomethyltransferase [16]. Both proteins have a localization associated with mitochondria in eukaryotes [15, 16]. Nevertheless, it cannot be clearly stated that they belong to any single metabolic pathway.

At the same time, the operon highlighted in yellow in Fig. 1 has three proteins - homologs of α-glucosidase and hexuronat transporter and a protein containing the pyrimidine dimer DNA glycosylases domain. Most likely, either the protein annotation is not accurate enough, or it is not some functionally single operon.

Cellular functions genes

It can be said that the locus under study does not represent a separate special region in the Archaean genome. It does not contain a protective cassette or a set of genes for a particular metabolic pathway. Rather, it has genes associated with various cellular activities.

Horizontal gene transfer evidences

With some uncertainty, we can assume that horizontal gene transfer is observed in this example. In particular, it is associated with protein with Prokka ID 17. Firstly, no similar operons have been found for it. Secondly, a BLAST search yielded only one result with the closest relative. Moreover, no functionally known homologs or at least protein domains could be identified for this protein. Moreover, the presence of the tnpB transposase protein in the immediate

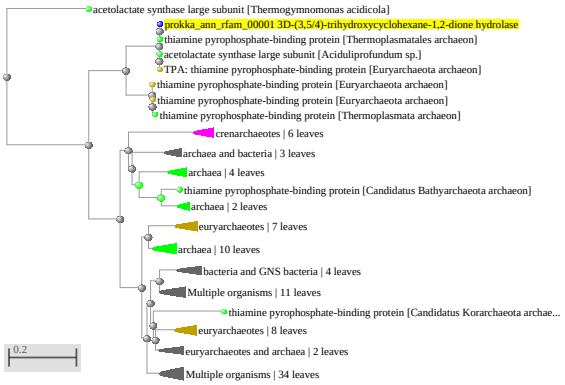

Figure 2. Distant tree built for BLAST search with the iolD sequence query with the Neighboor-Joining method

vicinity may serve to confirm this hypothesis. We can assume that during the transposon transfer, this gene was cut and inserted in some unfortunate way, which deprived the resulting protein of any important domains and similarity to anyone else. Nevertheless, this is only a hypothesis and is based only on circumstantial facts.

Conslusion

During this work, a small region of the Archaea genome was annotated and analyzed. We can say that it does not encode any special sets of genes. At the same time, it may be a potential example of horizontal gene transfer.

Supplemental materials

GitHub repository of the project: Sk_Bioinf_Project
Working notes: Notion

References

1. Yuan Y, Liu J, Yang TT, Gao SM, Liao B, Huang LN. Genomic Insights into the Ecological Role and Evolution of a Novel Thermoplasmata Order,"Candidatus Sysuiplasmatales". Applied and environmental microbiology 2021;87(22):e01065–21.
2. Hu W, Pan J, Wang B, Guo J, Li M, Xu M. Metagenomic insights into the metabolism and evolution of a new Thermoplasmata order (Candidatus Gimiplasmatales). Environmental Microbiology 2021;23(7):3695–3709.
3. Zinke LA, Evans PN, Santos-Medellín C, Schroeder AL, Parks DH, Varner RK, et al. Evidence for non-methanogenic metabolisms in globally distributed archaeal clades basal to the Methanomassiliicoccales. Environmental Microbiology 2021;23(1):340–357.
4. T S. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–9.
5. et al ASF. Basic local alignment search tool. Journal of Molecular Biology 1990;215(3):403–410.
6. Duvaud S, Gabella C, Lisacek F, Stockinger H, Ioannidis V, Durinx C. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. Nucleic Acids Research 2021;49(W1):W216–W227.
7. T S, Barrnap v0.9; 2018. https://github.com/tseemann/barrnap/blob/master/README.md.
8. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic acids research 2016;44(W1):W54–W57.
9. UniProt: the universal protein knowledgebase in 2021. Nucleic acids research 2021;49(D1):D480–D489.
10. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema M, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research 2021 05;49(W1):W29–W35. https://doi.org/10.1093/nar/gkab335.
11. Payne LJ, Todeschini TC, Wu Y, Perry BJ, Ronson CW, Fineran PC, et al. Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. Nucleic acids research 2021;49(19):10868–10878.
12. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 01;26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.
13. Taboada B, Estrada K, Ciria R, Merino E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics 2018;34(23):4118–4120.
14. Oberto J. SyntTax: a web server linking synteny to prokaryotic taxonomy. BMC bioinformatics 2013;14(1):1–10.
15. Lopez Sanchez MIG, Cipullo M, Gopalakrishna S, Khawaja A, Rorbach J. Methylation of ribosomal RNA: a mitochondrial perspective. Frontiers in Genetics 2020;11:761.
16. Leung KY, De Castro SC, Galea GL, Copp AJ, Greene ND. Glycine cleavage system H protein is essential for embryonic viability, implying additional function beyond the glycine cleavage system. Frontiers in Genetics 2021;12:625120.