Characterization of 30 megabase-long piece of *Tepidisphaera mucosa* genome.

Dmitry Biba

Introduction

Tepidisphaera mucosa is a newly sequenced species of bacteria from terrestrial hot springs of Baikal Lake. The similarity of its genome to its closest relatives is around 80%, which is very low, so it was proposed to be a member of novel order *Tepidisphaerales.* Characterization of such extraordinary genome is challenging, but is crutial for complete description of bacterial diversity (Kovaleva et al., 2015).

Matherials and methods

I downloaded the piece of *Tepidisphaera mucosa* genome from Skoltech server (mg.uncb.iitp.ru). As a first step, I annotated it with prokka (prokka --outdir annotation_prokka/ --force data/3.fasta --genus Tepidisphaera --species mucosa ; 3.fasta – the file with the piece of genome). Then I extracted CDSs from the piece of genome using gff-file (output of prokka) and a custom python scrupt (hereafter CPS, avaliable at https://github.com/Captain-Blackstone/Sk_bioinf_final_project, as well as the rest of the matherials) . To obtain some insight into the function of genes annotated as hypothetical proteins by prokka, I performed a search for conserved domains. To do that I sumbitted the file all_proteins_translated.fasta to Batch-CD search (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi). This file contains amino acid sequences, also obtained with the help of CPS. The results of Batch-CD search were used to complete prokka annotation: hypothetical proteins were substituted with lists of domains found by Batch-CD where it was possible. To get further insite in the functions of annotated proteins I also attempted to find transmembrane domains in the set of the annotated proteins with the help of TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), but no such domains were found (you may see the results in the github repository).

Using gff-file outputted by prokka, I extracted operon structure of the given piece of genome. The set of genes was considered to be an operon if it met 3 conditions: (1) the distance between neighbouring genes was not greater than 150 b.p., (2) all the genes in the set were located on the same strand, (3) the number of genes in the set was greater than two. To assign functions to the found operons we manually inspected literature (mainly wikipedia.org, uniprot.org and domain descriptions in Batch-CD results).

To infer horizontal gene transfer (HGT) we used two approaches. First one was to blast obtained CDSs versus NCBI nt database (we used blastn and tblastx, merging their results with CPS) and find the phylum of the closest hit (using https://lpsn.dsmz.de, in case of *Candidatus Solibacter* https://microbewiki.kenyon.edu/ was used). If the phylum was not the same as that of *Tepidisphaera mucosa* (namely, *Planctomycetes*) and e-value was lower that threshold (1e-10) HGT was assumed. The second approach was to measure the deviation of gc-content of a gene from the mean gc-content of a given genome piece. For each gene we measured its gc-content, compared it with the distribution of gc-contents of all the possible regions of the same size in the given piece of genome. If the gene in question resided among 5% of the most deviating regions, HGT was assumed. One may notice that the problem of multiple testing arises. If Bonferroni correction was applied, no significant genes were found, so this test wasn't used to infer HGT, rather it was used to support the HGT assumption from the previous test. The same procedure was done with whole operons rathere than genes, because it is assumed that operons may be horizontally transferred as a distinc entity. All the procedures concerning gc-content were done with the help of CPS.

I used barrnap to find rRNAs in the piece of genome (barrnap data/3.fasta --outseq barrnap/rRNAs), aragorn to find tRNAs (aragorn -t data/3.fasta -o aragorn_out.txt) and CRISPRCasFinder to find CRISPRCas genes (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index). I used antiSMASH to find genes associated with secondary methabolytes synthesis ((https://antismash.secondarymetabolites.org/#!/start).

Results

Prokka annotation and conserved domains

36 CDSs were found by prokka, among them 16 hypothetical proteins. Batch-CD search found domains for 11 of 16 hypothetical proteins (the comprehensive list of these domains and their description may be found in the github repository: Hypothetical_proteins_annotation_batch_CD.txt and Prokka_annotated_proteins_annotation_batch_CD.txt). The results of these annotations are presented on Table 1.

Prokka ID	coordinates	strand	Prokka annotation	Batch CD annotation
GFKBJCAA_00001	43-771	+	hypothetical protein	TauE
GFKBJCAA_00002	843-1520	+	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	Glyco_tranf_GTA_type superfamily
GFKBJCAA_00003	1514-2143	+	SPBc2 prophage-derived endonuclease YokF	SNc superfamily
GFKBJCAA_00004	2221-2622	+	hypothetical protein	DUF4870
GFKBJCAA_00005	2716-4011	+	hypothetical protein	PEP_TPR_lipo superfamily; PRK11447 superfamily
GFKBJCAA_00006	4101-5120	+	General stress protein 69	AKR_AKR12A1_B1_C 1
GFKBJCAA_00007	5162-6310	-	Aldose sugar dehydrogenase YliI	GSDH
GFKBJCAA_00008	6500-7870	+	Poly(A) polymerase I	PcnB
GFKBJCAA_00009	7892-8200	-	Flagellar motor switch protein FliN	FliMN_C
GFKBJCAA_00010	8310-9551	-	3-oxoacyl-[acyl-carrier-protein] synthase 2	fabF
GFKBJCAA_00011	9651-9890	-	Acyl carrier protein	acpP
GFKBJCAA_00012	10315-11820	+	hypothetical protein	-
GFKBJCAA_00013	11915-12445	-	Orotate phosphoribosyltransferase	PyrE
GFKBJCAA_00014	12453-13469	+	Epoxyqueuosine reductase	TIGR00276 superfamily
GFKBJCAA_00015	13466-13945	+	hypothetical protein	Dabb
GFKBJCAA_00016	14040-14477	-	18 kDa heat shock protein	ACD_sHsps-like
GFKBJCAA_00017	14416-14589	-	hypothetical protein	-
GFKBJCAA_00018	14680-15582	+	tRNA dimethylallyltransferase	miaA
GFKBJCAA_00019	15597-16658	+	3-isopropylmalate dehydrogenase	Iso_dh superfamily
GFKBJCAA_00020	16663-17310	-	hypothetical protein	S2P-M50_like_1
GFKBJCAA_00021	17419-18015	-	Histidine biosynthesis bifunctional protein HisB	hisB

GFKBJCAA_00022	18044-18421	-	Undecaprenol kinase	UDPK_IM_like
GFKBJCAA_00023	18414-19181	-	hypothetical protein	EEP superfamily
GFKBJCAA_00024	19331-19786	+	hypothetical protein	nfrB superfamily
GFKBJCAA_00025	19704-20465	-	hypothetical protein	-
GFKBJCAA_00026	20517-21383	-	Branched-chain-amino-acid aminotransferase	PLPDE_IV superfamily
GFKBJCAA_00027	21578-21961	+	Aspartate 1-decarboxylase	Asp_decarbox
GFKBJCAA_00028	22007-22489	-	putative peptidyl-prolyl cis-trans isomerase	PpiB
GFKBJCAA_00029	22563-23513	-	All-trans-phytoene synthase	SQS_PSY
GFKBJCAA_00030	23566-24402	+	hypothetical protein	DUF1444 superfamily
GFKBJCAA_00031	24465-25412	+	hypothetical protein	Abhydrolase superfamily
GFKBJCAA_00032	25418-26356	-	hypothetical protein	metallo-dependent_hydrolases superfamily
GFKBJCAA_00033	26360-27823	-	L-Rhamnulokinase	FGGY_RhuK
GFKBJCAA_00034	27920-28186	+	hypothetical protein	-
GFKBJCAA_00035	28393-28773	+	hypothetical protein	-
GFKBJCAA_00036	28871-29617	-	hypothetical protein	TS_Pyrimidine_HMase superfamily

Table 1. The results of Prokka and Batch-CD annotations.

Operon structure

We have detected 3 operons in the given piece of genome (Table 2, Figure 1). The scan through literature revealed that most of genes of the first operon are involved in synthesis of cell wall components: GFKBJCAA_00001 in taurine syntheiss, GFKBJCAA_00002 in the synthesis of isoprenoid precursors, GFKBJCAA_00005 in cellulose syntheiss. Two of three genes in the second operon (GFKBJCAA_00010 and GFKBJCAA_00011) are involved in fatty acids biosnythesis. The genes of the third operon lacked any explicit uniting feature - GFKBJCAA_00020 is involved in cleavage of TM-domains, GFKBJCAA_00021 in histidine biosnthesis, GFKBJCAA_00022 in peptidoglycan biosynthesis and GFKBJCAA_00023 turned out to be some kind of an exonuclease.

operon	coordinates	genes	possible function
1	43-5120	00001; 00002; 00003; 00004; 00005; 00006	cell wall building
2	7892-9890	00009; 00010; 00011	fatty acid synthesis
3	16663-19181	00020; 00021; 00022; 00023	undefined

Table 2. The operon structure of the given piece of genome. In the “genes” column the “GFKBJCAA” prefix of each gene is omited for the sake of clarity of representation.

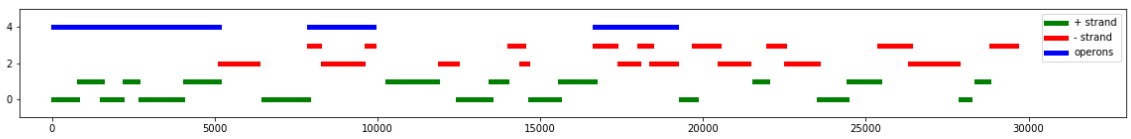

Figure 1. A visualization of genes along the given piece of genome. Green lines represent genes on the “+”-strand, red ones – on “-”-strand. Blue lines represent operons.

Non-coding RNAs and secondary methabolytes synthesis genes

Aragorn and barrnap have not found any tRNAs or rRNAs in the given piece of genome. AntiSMASH found one gene associated with secondary methabolytes syntheiss – GFKBJCAA_00029, synthase of all-trans-phytoene – a carotene compound.

HGT inference

The results of blast search for honologous genes are presented in Table 3. Overall, 20 genes produced significant hits. We considered a gene to be a HGT candidate only if log10 e-value of its hit was less than -10 and the phylum of this hit was not Planctomycetes. There are 6 such genes: GFKBJCAA_00004, GFKBJCAA_00010, GFKBJCAA_00015, GFKBJCAA_00021, GFKBJCAA_00022 and GFKBJCAA_00028.

Prokka ID	method	log10_e_val	Prokka annotation	organism	phylum
GFKBJCAA_00002	tblastx	-14	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	Planctomycetes bacterium	Planctomycetes
GFKBJCAA_00003	tblastx	-5	SPBc2 prophage-derived endonuclease YokF	Shewanella sp.	Proteobacteria
GFKBJCAA_00004	tblastx	-27	hypothetical protein	Pseudoalteromonas issachenkonii	Proteobacteria
GFKBJCAA_00006	blastn	-3	General stress protein 69	Hymenobacter sp.	Bacteroidetes
GFKBJCAA_00009	tblastx	-17	Flagellar motor switch protein FliN	Phycisphaerae bacterium	Planctomycetes
GFKBJCAA_00010	blastn	-20	3-oxoacyl-[acyl-carrier-protein] synthase 2	Aeromonas salmonicida	Proteobacteria
GFKBJCAA_00011	tblastx	-35	Acyl carrier protein	Planctomycetales bacterium	Planctomycetes
GFKBJCAA_00013	tblastx	-54	Orotate phosphoribosyltransferase	Phycisphaerae bacterium	Planctomycetes
GFKBJCAA_00014	blastn	-5	Epoxyqueuosine reductase	Caldilinea aerophila	Chloroflexi
GFKBJCAA_00015	tblastx	-13	hypothetical protein	Methylococcus capsulatus	Proteobacteria
GFKBJCAA_00016	tblastx	-27	18 kDa heat shock protein	Roseimaritima ulvae	Planctomycetes
GFKBJCAA_00017	tblastx	-1	hypothetical protein	Pseudomonas rhizosphaerae	Proteobacteria
GFKBJCAA_00021	tblastx	-67	Histidine biosynthesis bifunctional protein HisB	Alpha proteobacteria	Proteobacteria
GFKBJCAA_00022	tblastx	-13	Undecaprenol kinase	Bacillus circulans	Firmicutes
GFKBJCAA_00024	tblastx	-12	hypothetical protein	Planctomycetales bacterium	Planctomycetes
GFKBJCAA_00026	tblastx	-126	Branched-chain-amino-acid	Thermogutta	Planctomycetes

			aminotransferase	terrifontis	
				Planctomycetales	
GFKBJCAA_00027	tblastx	-36	Aspartate 1-decarboxylase	bacterium	Planctomycetes
			putative peptidyl-prolyl cis-	Stenotrophomona	
GFKBJCAA_00028	blastn	-19	trans isomerase	s rhizophila	Proteobacteria
				Candidatus	
GFKBJCAA_00033	blastn	-7	L-Rhamnulokinase	Solibacter	Acidobacteria
				Lateolabrax	
GFKBJCAA_00034	tblastx	0	hypothetical protein	maculatus	fish

Table 3. The results of blast search in NCBI nt database for genes annotated with prokka.

The results of tests on gc-content deviation for genes are presented in Figure 2 and for operons in Figure 3. Only 2 of genes had notably extraordinary gc-conent – GFKBJCAA_00034 and GFKBJCAA_00036. However, the effect disappears when applying Bonferroni correction. For the operons no deviations were obeserved.

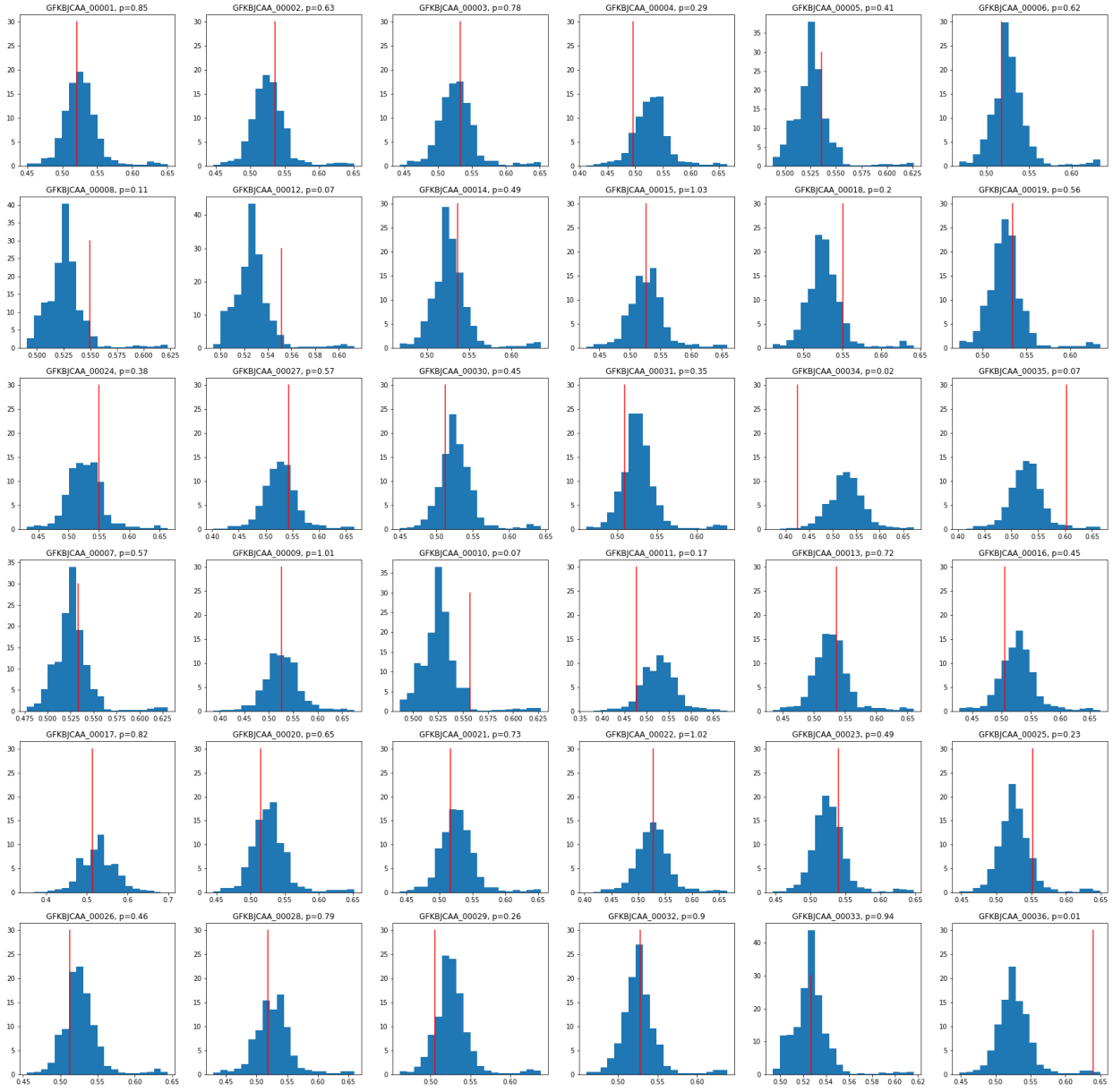

Figure 2. GC-content deviation in prokka annotated genes. Red lines represent gc-content of the gene of interest. Blue histograms represent the distribution of gc-content in all the regions of a given piece of a genome.

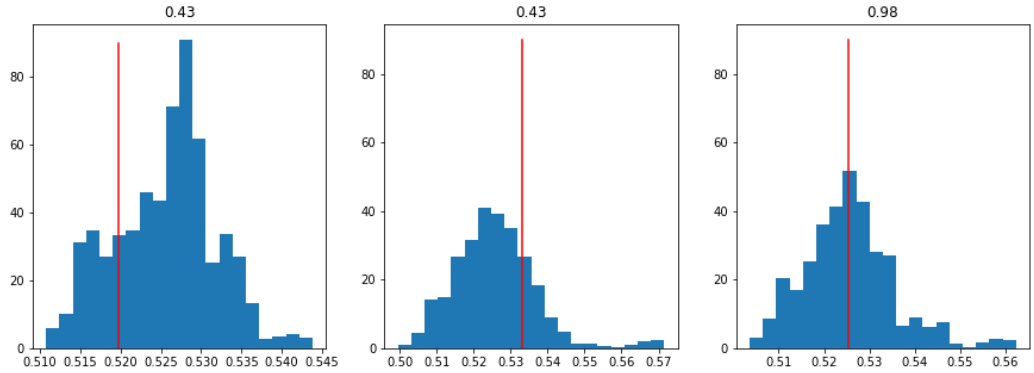

Figure 3. GC-content deviation in operons. The notation is the same as in Figure 2.

Discussion

I have found 3 operons in the given piece of genome, and managed to assign functions to two of them. The first one conains the following genes: GFKBJCAA_00001 is involved in taurine metabolism (constituent of cell surface polymers, Smiley & Brian, 1983), GFKBJCAA_00002 is involved in isoprenoid precursor synthesis (participating in cell wall synthesis, Rodríguez-Concepción, 2012) and GFKBJCAA_00005 contains cellulose synthase subunit and domain associated with exopolysaccharide production genes (which take part in cell wall synthesis, Haft et al., 2006). The second operon contains GFKBJCAA_00010, 3-oxoacyl-[acyl-carrier-protein] synthase 2 and GFKBJCAA_00011, acyl-carrier-protein, both parts of fatty acid synthesis pathways (InterPro and wikipedia). It is interesting, that the third one, with unassigned function contains two candidates on HGT out of four genes assigned to this operon, so, it probably is not an operon at all.

The two tests on HGT that I peroformed yield inconsistent results. One reason I can think of is that in hot springs it is crutial for bacteria to have some precise (and elevated) gc-content in order to not get its DNA melted, so even if gene is horizontally transferred, its gc-content rapidly changes to suite the genomic level. The problem with this interpretation is that genomic average gc-content does not appear to be elevated (it is equal to 0.52), as one would expect from hot spring bacteria. Another explanation might be that gc-content of transferred gene donors was the same as that of *Tepidisphaera mucosa*.

More specifically, GFKBJCAA_00034 gene, which has very deviant gc-content, produced the best hit in Lateolabrax maculatus – a fish, which, probably, hints its HGT origin. The preoblem here is that this hit is extremely unreliable with e-value ~ 2.5. The rest of the hits produced by this gene have also very high e-value, but all are from various vertebrates. GFKBJCAA_00036 – the second gene with deviant gc-content – is annotated as hypothetical protein and no conserved domains were found there, so I can't say anything about it.

The genes which produced hits ourside the *Planctomycetes* phylum and had the lowest e-value are the most likely candidates to be horizontally transferred. However, if they are parts of the operons, their horizontal origin is questionable, since entire pathways are unlikely to be formed independenty and should probably be transferred together. However, only GFKBJCAA_00004, GFKBJCAA_00021 and GFKBJCAA_00022 genes among best HGT candidates are parts of various operons. Among them, GFKBJCAA_00004 was annotated as a hypothetical protein and the conserved domain found there doesn't have known function, so this gene may even not have a

function in this operon. Two other genes, as already has been noted, belong to an operon without an obvious function, which hints that this operon may not be real at all.

Conclusions

I have described the region of *Tepidisphaera mucosa* genome. I found 36 genes there, at the end only 5 of them lacked annotated function and any domains. I also found 3 operons, 2 of which had more or less distinc function. Lastly, I have found 6 candidate genes which were probably obtained by this bacterium via HGT.

Supplementary materials

One may find the data I used, the script, all the intermediate files and my journal in the github repository: https://github.com/Captain-Blackstone/Sk_bioinf_final_project