CHSHEGLOVA ALINA VASILIEVNA
Taya23226@gmail.com

RESEARCH Open Access

Thiolapillus brandeum review

Chsheglova Alina

Abstract

Background: The bacterium was isolated from Japan, Okinawa Trough, Minami-Ensei Knoll hydrothermal active field, no. 7. Currently , 10 types of organism metabolism are known.

Introduction data:

Optimal growth conditions: 37 C

Complete genome size: 3129.662 kbp

Genome structure: 3.19Mb chromosome and two plasmids (pTBH10 - 10.9Kb, and pTBH13 - 13.4Kb) [2].

Statistics: median total length (Mb): 2.0208

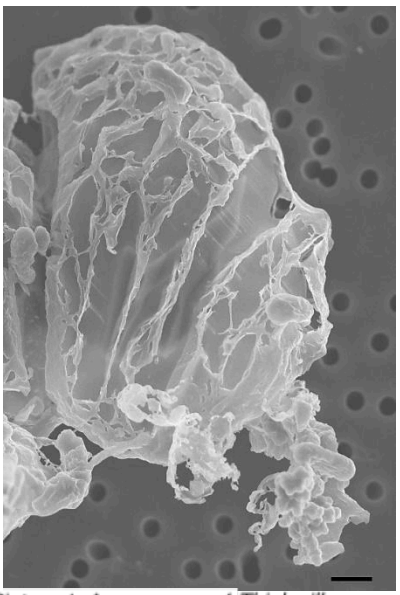
median protein count: 2102

median GC%: 56.7826

total count of proteins: 2929

The mesophilic anaerobe was studied, and composition of the genome- 1 chromosome and 2 plasmids, their size, quantitative and qualitative composition was determined. Chargaff's second rule on the quantitative equality of nitrogenous bases was confirmed. The predominance of GC-nucleotides allows the bacteria to be in conditions with high temperature. A graph was constructed on the distribution of protein lengths, where the medial length was 322.5. It was revealed that only 1 restriction site is located in the chromosome, while in the plasmid pTBH10 - 28, and in pTBH13-33. The largest protein was transporter substrate-binding domain-containing protein, which is mostly represented by leucine. While using GC-skew, the replication start point is determined, which corresponds to the first codon, and the end of replication correlated to 1.547.001 position.

Introduction



Picture 1. Appearance of *Thiolapillus brandeum [3]*

Bacteria's name etymology can be associated with Christian brandeum, which covered elemental sulfur. The layer consists of polysaccharide-like substance.

Bacterium acts as a symbiont of a gastropod and is supposed to use sulfur, O2 oxidation, nitrate reduction metabolism pathways [3]. Basic taxonomy: cellular organism; Bacteria; Pseudomonatoda; Gammaproteobacteria; Gammaproteobacteria incertae sedis; Thiolapillus; *Thiolapillus brandeum* [4].

The main condition for choosing this bacterium was its thermal habitat, which is of personal interest in studying the metabolism of an organism capable of functioning at high temperatures. The study will bring scientific data and will allow them to be adapted to improve human life and, possibly, open up the opportunity to visit hot planets. Also, the etymology of the bacteria name is associated with the Christian concept of brandeum, which means the cloth in which the relics of the saint are wrapped. Biological meaning in the formation of a film covering elemental sulfur.

Methods

Genome data from NCBI [5] were analyzed, a histogram of protein lengths was constructed using Google Sheets [6], and data about the number of different types of RNA were obtained [7].The Python language was used to write a program for counting GC-nucleotide and for getting dates about cumulative GC-skew [8]. Work was also carried out in the Gene Viewer program on predicting sites for restrictions and obtaining data on the amino acid composition of proteins.

Results

Standard dates about genome:

The genome of Thiolapillus brandeum consist of two plasmids (pTBH10- linear, 10.9Kb, and pTBH13-circular - 13.4Kb) and circular chromosome [2].

Table 1 presents main characteristic components of the genome. Length and GC-percent was obtained during the analysis file assembly_stats.txt. Thiolapillus brandeum is a mesophilic bacterium which is characterized by a 50% content of GC-nucleotides in the genome, as we may notice in *table1*.

Table 1. Standard dates about genome

\	Length (bp.)	GC percent
Chromosome	3129962	56.6 %
Plasmid pTBH10	10497	54 %
Plasmid pTBH13	13478	50 %

Table2 shows that the quantitative composition of nucleotides corresponds to the second Chargaff's rule, where the amount of adenine is approximately equal to the amount of thymine, and the amount of guanine is approximately equal to the amount of cyanine.

Table 2. Nucleotide composition of genomic DNA

DNA	A	T	G	C
Chromosome	979	852	889	880
Plasmid pTBH10	2433	2404	2928	2732
Plasmid pTBH13	3238	3496	3233	3511
Total count	377257	377822	823955	822295

By analyzing the content of such groups of proteins as: ribosomal, transport, hypothetical– from the data of the CDS table of genome features, was found that the percentage of ribosomal proteins from the total number of proteins is the smallest and does not exceed 2.5% [table 3].

Table 3. Data about proteins selected groups

Protein	Number	Percentage of all proteins
Ribosomal proteins	60	2,04%
Hypothetical proteins	448	15.29%
Transport proteins	209	7.13%

There are 45 RNA genes in the genome [9,list "RNA"], which is 0.75% of all genes. tRNA – 38, rRNA - 3 genes [Table 4],[2]. The genome contains the 6S, 16S, 23S rRNA genes in Svedberg units, which is typical for prokaryotic organisms. 16S rRNA is conservative and least susceptible to changes during evolution that can be used for phylogenetic analysis.

Table 4. Data about RNA selected groups

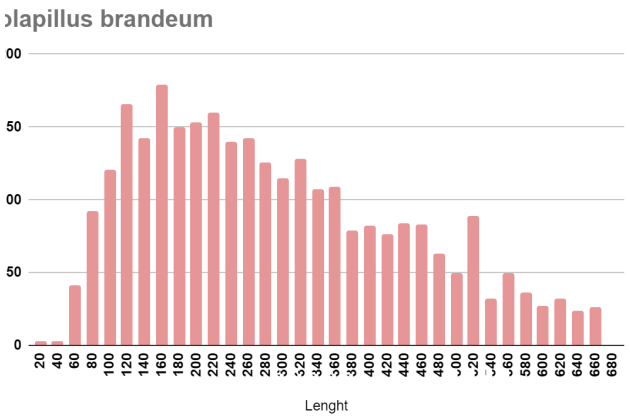
RNA	Number	Percentage of all RNAs
tRNA	38	84.5%
rRNA	3	6.7%

Statistical data about proteome proteins.

During constructing the length diagram in Excel spreadsheets[6], statistical data on the average, maximum and minimum length of proteins were obtained. As we can notice, most proteins are 100-140 bp. For a more detailed study, *Table 5* was compiled with statistical data on the distribution of protein length.

Table 5. Statistical parameters of protein length distribution

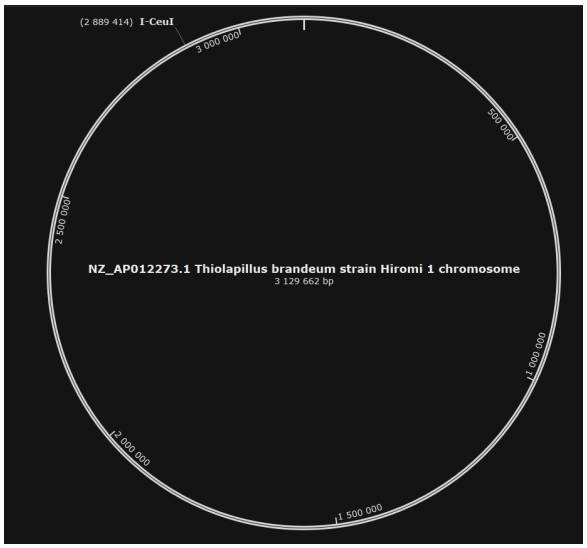
Medium length	322.5
Standard deviation	
Median	268
Maximum value	2692
Minimum value	34



Picture 2. Protein lengths histogram.

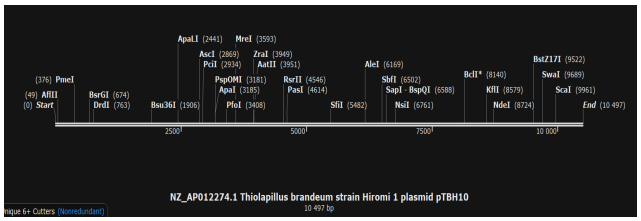
Genome analysis for quantitative and qualitative composition of enzymes

The genome of the bacterial chromosome contains only one site for the enzymes, and this one is I-CeuI.[pic.3]

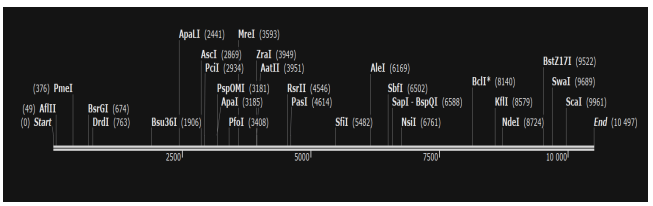


Picture 3. Chromosomal genome map

The I-CeuI endonuclease is a member of the family of homing endonucleases that catalyse mobility of introns group I by making a double-strand break at the homing site of these introns in cognate intronless alleles during genetic crosses [10].The number of enzymes sites in pTBH10 - 28 [pic.4], in pTBH13-33 [pic.5]



Picture 4. Plasmid pTBH10 genome map



Picture 5. Plasmid pTBH13 genome map

Restrictases marked with an asterisk mean that more than one restrictase is specific to one binding site.

Analysis of the largest protein

The longest protein is transporter substrate-binding domain-containing protein.Protein belongs to the family, the vast majority of which transports a variety of substances.

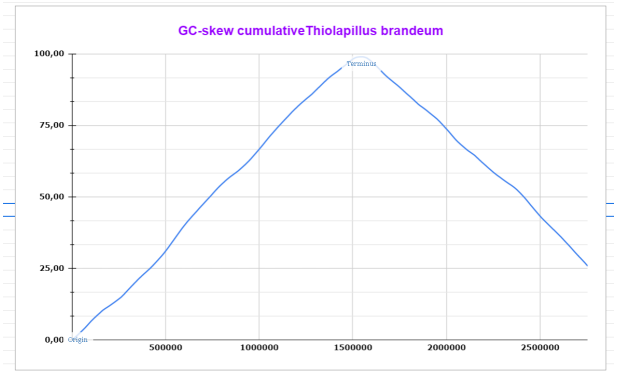
Amino Acid			Number	Percent
A	Ala	Alanine	128	7,95
C	Cys	Cysteine	5	0,31
D	Asp	Aspartic Acid	96	5,96
E	Glu	Glutamic Acid	129	8,01
F	Phe	Phenylalanine	56	3,48
G	Gly	Glycine	112	6,95
H	His	Histidine	31	1,92
I	Ile	Isoleucine	84	5,21
K	Lys	Lysine	68	4,22
L	Leu	Leucine	200	12,41
M	Met	Methionine	44	2,73
N	Asn	Asparagine	55	3,41
P	Pro	Proline	62	3,85
Q	Gln	Glutamine	84	5,21
R	Arg	Arginine	114	7,08
S	Ser	Serine	98	6,08
T	Thr	Threonine	60	3,72
V	Val	Valine	118	7,32
W	Trp	Tryptophan	23	1,43
Y	Tyr	Tyrosine	44	2,73

Picture 6. Amino acid composition of transporter substrate-binding domain-containing protein.

Protein is mostly represented by the amino acid leucine, and least of all in the composition of cysteine [pic.6].

Analyzing genomes with cumulative skew graph

The skew of GC-nucleotides is defined as (G-C) /(G+C) and allows you to predict the position of the terminal, which corresponds to the maximum value and the place of the beginning of replication - the origin corresponding to the minimum value.
After receiving data on the cumulative skew of GC-nucleotides, a *Cumulative GC-skew graph* was constructed where the Ox axis is responsible for the position on the chromosome from 3' - 5', and the Oy axis determines the deviation value. The starting point of replication is defined in the first codon, and the end of replication corresponds to the position 1.547.001. These two points are located from each other at a distance approximately equal to half the length of the genome.
The main reason for the asymmetry can be called the location of most genes on the leading chain and mutational pressure. From the data obtained follows that a positive bias occurs first, and after the termination section it changes to a negative one, which indicates that it is a leading chain.



Picture 7. Cumulative GC-skew graph

Conclusion

During the study, the possibility of living at high temperatures was substantiated.Restriction sites in the chromosome and plasmids were analyzed, and the qualitative composition of RNA was revealed, which can be used for phylogenetic determination of bacteria.

References

1.Japan Collection of Microorganism (JCM) ; Curators JCM;doi:10.13145/bacdive130815.20220

2. NCBI genome sequences https://www.ncbi.nlm.nih.gov/genome/?term=Thiolapil ndeum

3.Nunoura T, Takaki Y, Kazama H, Kakuta J, Shimamu Makita H, et al. (2014) Physiological and Genomic Fea a Novel Sulfur-Oxidizing Gammaproteobacterium Belo to a Previously Uncultivated Symbiotic Lineage Isolate a Hydrothermal Vent. PLoS ONE 9(8): e104959. https://doi.org/10.1371/journal.pone.0104959

4. NCBI taxonomy https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwv ?id=1076588

5. NCBI https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/828/(F_000828615.1_ASM82861v1/GCF_000828615.1_AS 1v1_assembly_stats.txt

6.Practice 8

7.Parte, A.C., Sardà Carbasse, J., Meier-Kolthoff, J.P., I L.C. and Göker, M.: List of Prokaryotic names with St in Nomenclature (LPSN) moves to the DSMZ. IJSEM (DOI 10.1099/ijsem.0.004332)

8. cum GC-skew

9. CDS pr7

10.https://www.researchgate.net/publication /242824959_The_I-_Ceu_I_endonuclease_ recognizes_a_sequence_of_19_base _pairs_and_preferentially_cleaves_ the_coding_strand_of_the_Chlamydomonas _moewusii_chloroplast_large_subunit_rRNA _gene