Genome annotation

Final Project

For the project I received a file with a fragment (30000 nt long) of genomic DNA of *Tepidiforma bonchosmolovskayae* gen. nov., sp. nov.. It is a moderately thermophilic bacterium isolated from a Chukotka hot spring. However, the genome of it was not well characterized because of its recent discovery. In addition, *T. bonchosmolovskayae* is an organism which belongs to a new class *Tepidiforma* within *Chloroflexi*, and no other organisms from this class have been sequenced so far. As a result, it is relevant to annotate genes of this bacterium and represent enough biological information about it using all the tools which were studied during the course.

The project is divided into the several steps:

1. Annotation of all coding and non-coding genes
2. Identification of functions for the found hypothetical proteins
3. Description of the operon structure and for the long operons (longer than 4 genes) identification of the known regulatory regions if it is possible
4. Finding of genes obtained by the bacteria through horizontal gene transfer (HGT)
5. Finding of genes of the bacteria associated with secondary metabolites

1. Annotation of all coding and non-coding genes

For annotation I took the file with my variant (27.fasta) from the server and then used Prokka which is appropriate for the annotation of prokaryotic genomes.

1) I identified the number of coding sequences (CDS) for coding genes with the command

```
prokka --outdir prokka_result --locustag
prokka --kingdom Bacteria 27.fasta
```

and found the result

```
organism: Genus species strain
contigs: 1
bases: 30000
CDS: 32
```

2) I identified the number of coding sequences (CDS) for non-coding genes with the command

```
prokka --outdir prokka_non-coding --rfam --
locustag prokkar --kingdom Bacteria 27.fasta
```

and found the result

organism: Genus species strain
contigs: 1
bases: 30000
CDS: 32

The results are the same, it means that there are no any non-coding genes in the fragment.

With Prokka I identified that some of the proteins are hypothetical ones and one of the aims is to recognise which functions hypothetical proteins have.

2. Identification of functions for the found hypothetical proteins

Prokka found 18 hypothetical proteins and their functions were found with the tools (1 – BLAST, 2 – Pfam, 3 – TMHMM).

For each protein the steps of identification are described:

1) With BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi for hypothetical protein #1 the result is

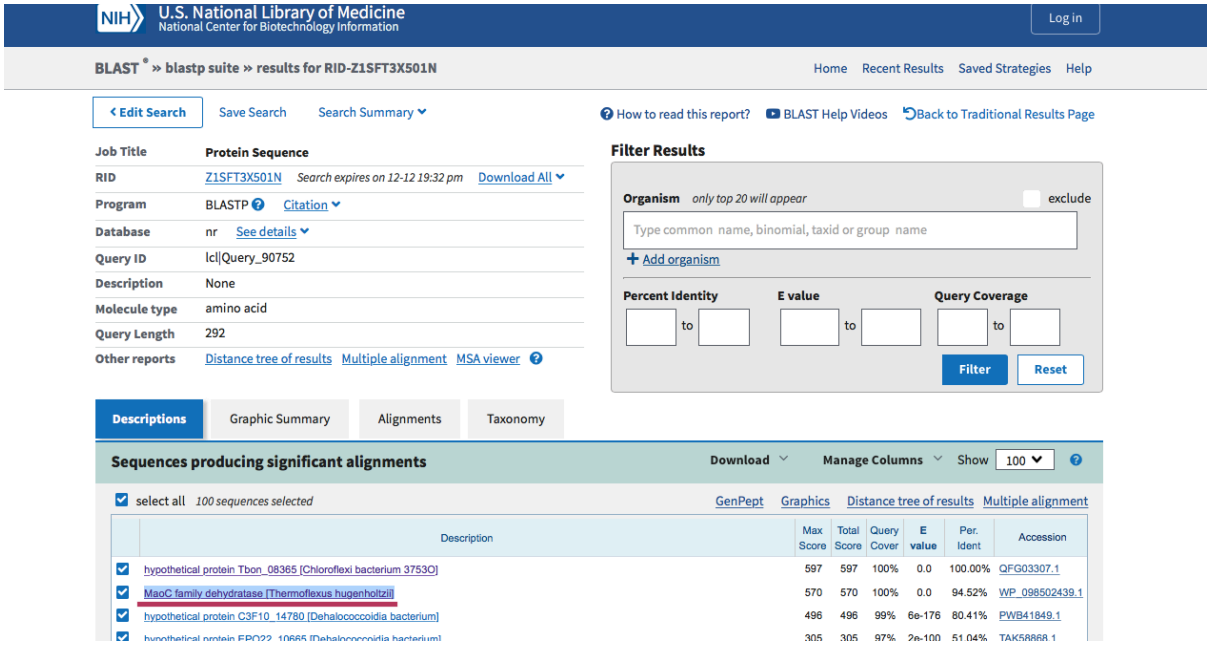


Figure 1 – Results of BLAST for hypothetical protein #1

The result is with good identity (94.52%), E-value (0.0), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as dehydratase [Thermoflexus hugenholtzii].

2) The result for hypothetical protein #2

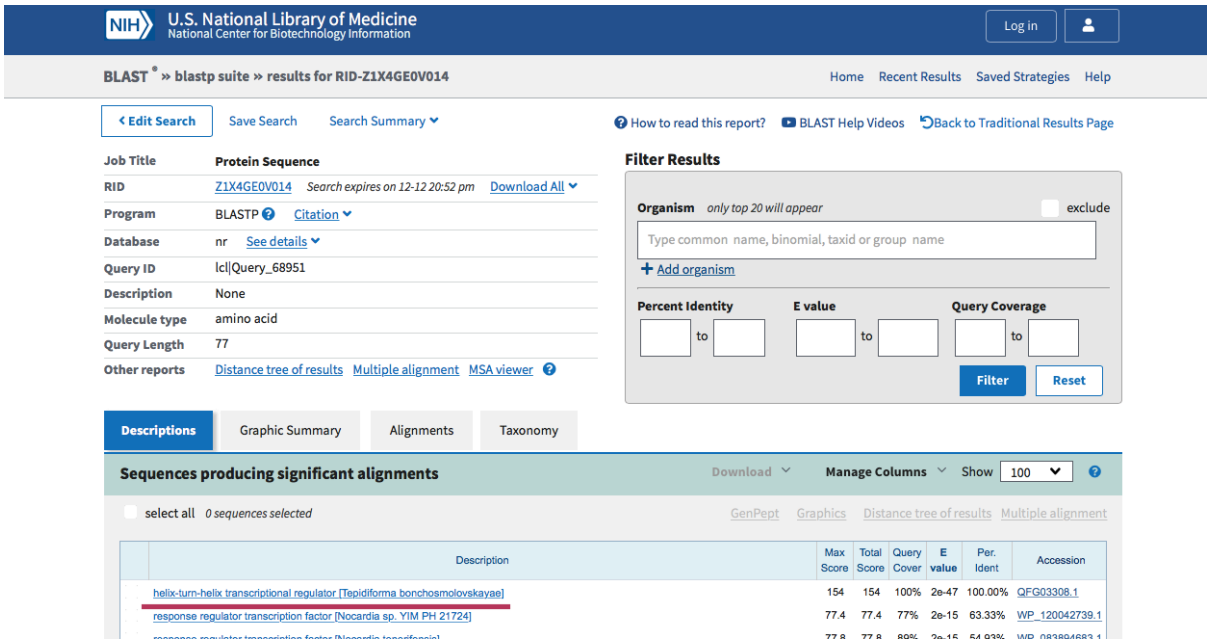


Figure 2 – Results of BLAST for hypothetical protein #2

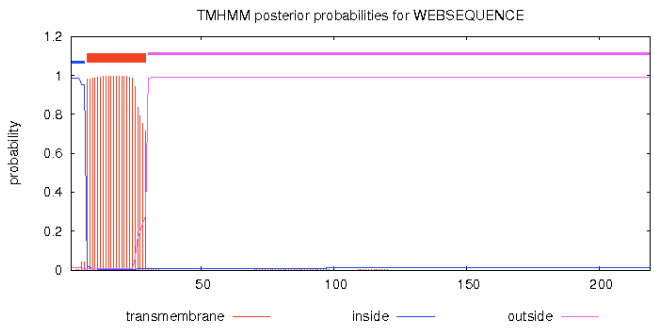
The result is with good identity (100%), E-value (2e-47), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as helix-turn-helix transcriptional regulator [Tepidiforma bonchosmolovskayae].

3) For hypothetical protein #3 there are no any relevant results with BLAST (only hypothetical proteins were found) and Pfam (no any Pfam-A matches), but with TMHMM the result is

TMHMM result

HELP with output formats

```
# WEBSEQUENCE Length: 219
# WEBSEQUENCE Number of predicted TMHs:  1
# WEBSEQUENCE Exp number of AAs in TMHs: 21.96946
# WEBSEQUENCE Exp number, first 60 AAs:  21.95223
# WEBSEQUENCE Total prob of N-in:        0.98589
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        inside       1     6
WEBSEQUENCE     TMHMM2.0        TMhelix      7    29
WEBSEQUENCE     TMHMM2.0        outside     30   219
```



plot in postscript, script for making the plot in gnuplot, data for plot

Figure 3 – Results of TMHMM for hypothetical protein #3

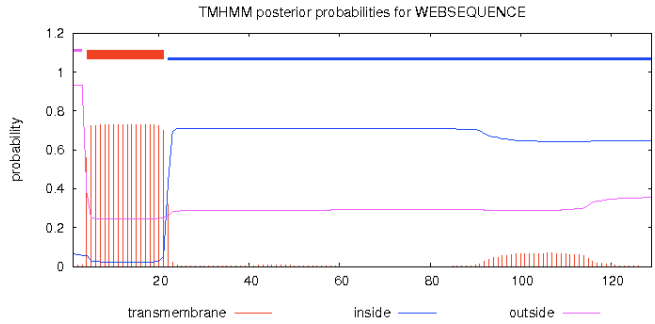
With TMHMM a transmembrane segment was found, it means that hypothetical protein #3 has the functions of transmembrane proteins.

4) For hypothetical protein #4 the relevant result was found by TMHMM

TMHMM result

HELP with output formats

```
# WEBSEQUENCE Length: 129
# WEBSEQUENCE Number of predicted TMHs:  1
# WEBSEQUENCE Exp number of AAs in TMHs: 14.9171
# WEBSEQUENCE Exp number, first 60 AAs:  13.39293
# WEBSEQUENCE Total prob of N-in:        0.06615
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        outside      1     3
WEBSEQUENCE     TMHMM2.0        TMhelix      4    21
WEBSEQUENCE     TMHMM2.0        inside      22   129
```



plot in postscript, script for making the plot in gnuplot, data for plot

Figure 4 – Results of TMHMM for hypothetical protein #4

With TMHMM a transmembrane segment was found, it means that hypothetical protein #4 has the functions of transmembrane proteins.

5) For hypothetical protein #5 the relevant result was found by Pfam

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant)

IMS

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
IMS	impB/mucB/samB family	Family	n/a	10	216	13	143	**8**	**139**	150	46.5	3.8e-12	n/a	Show

Pfam is part of the ELIXIR infrastructure

Pfam is an Elixir service Read more

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.

European Molecular Biology Laboratory

Figure 5 – Results of Pfam for hypothetical protein #5

The protein belongs to impB/mucB/samB family (the proteins of it are involved in UV protection).

6) For hypothetical protein #6 no any appropriate functions with all of the described tools were found.

7) For hypothetical protein #7 the relevant result was found by Pfam

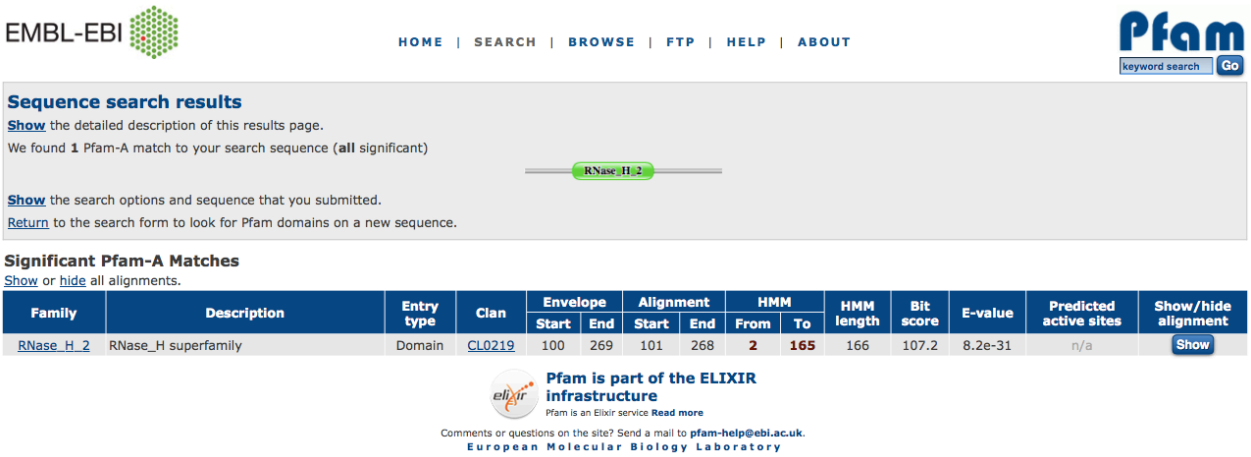


Figure 6 – Results of Pfam for hypothetical protein #7

The protein belongs to RNase_H superfamily and has the functions like the other proteins of this family – numerous enzymes which are involved in nucleic acid metabolism and implicated in many biological processes, including replication, homologous recombination, DNA repair, transposition and RNA interference.

8) For hypothetical protein #8 the result is

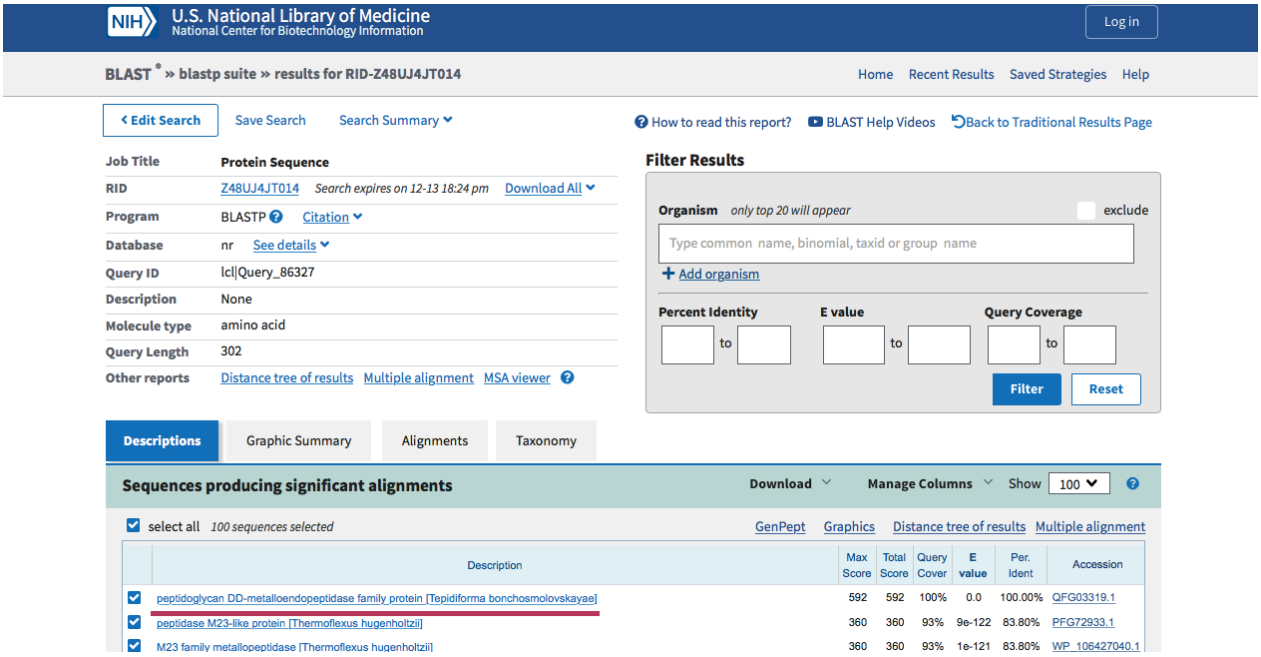


Figure 7 – Results of BLAST for hypothetical protein #8

The result is with good identity (100%), E-value (0.0), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as peptidoglycan DD-metalloendopeptidase family protein [Tepidiforma bonchosmolovskayae].

9) For hypothetical protein #9 no any appropriate functions with all of the described tools were found.

10) For hypothetical protein #10 no any appropriate functions with all of the described tools were found.

11) For hypothetical protein #11 the result is

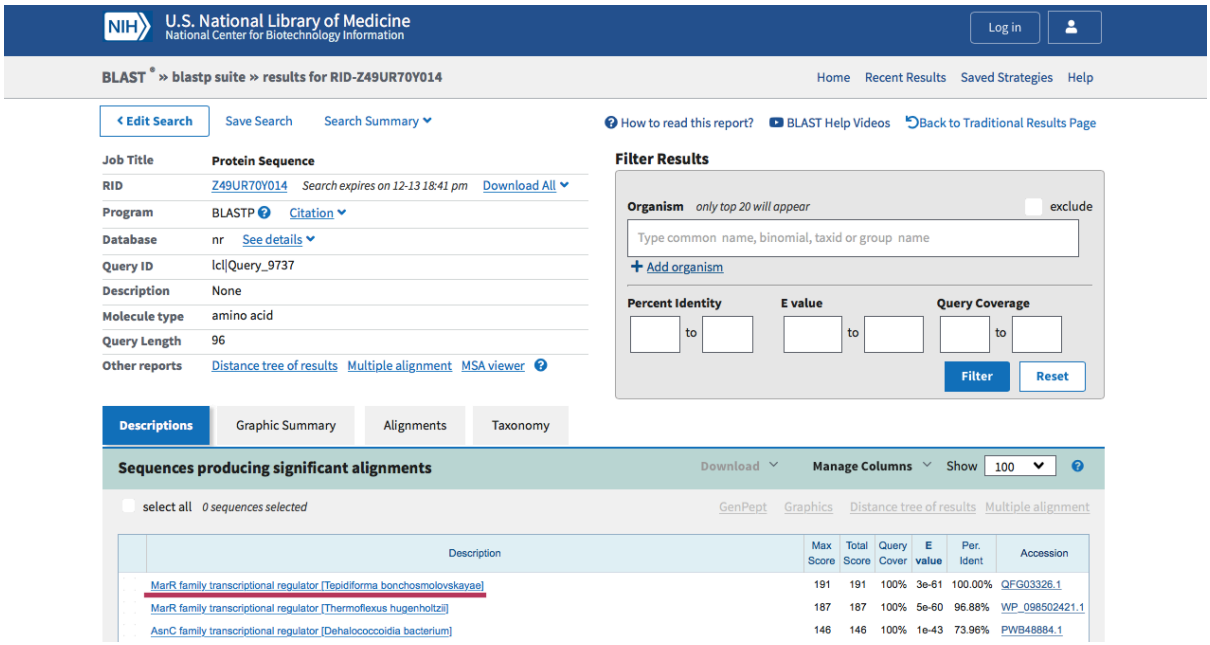


Figure 8 – Results of BLAST for hypothetical protein #11

The result is with good identity (100%), E-value (3e-61), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as MarR family transcriptional regulator [Tepidiforma bonchosmolovskayae].

12) For hypothetical protein #12 the result is

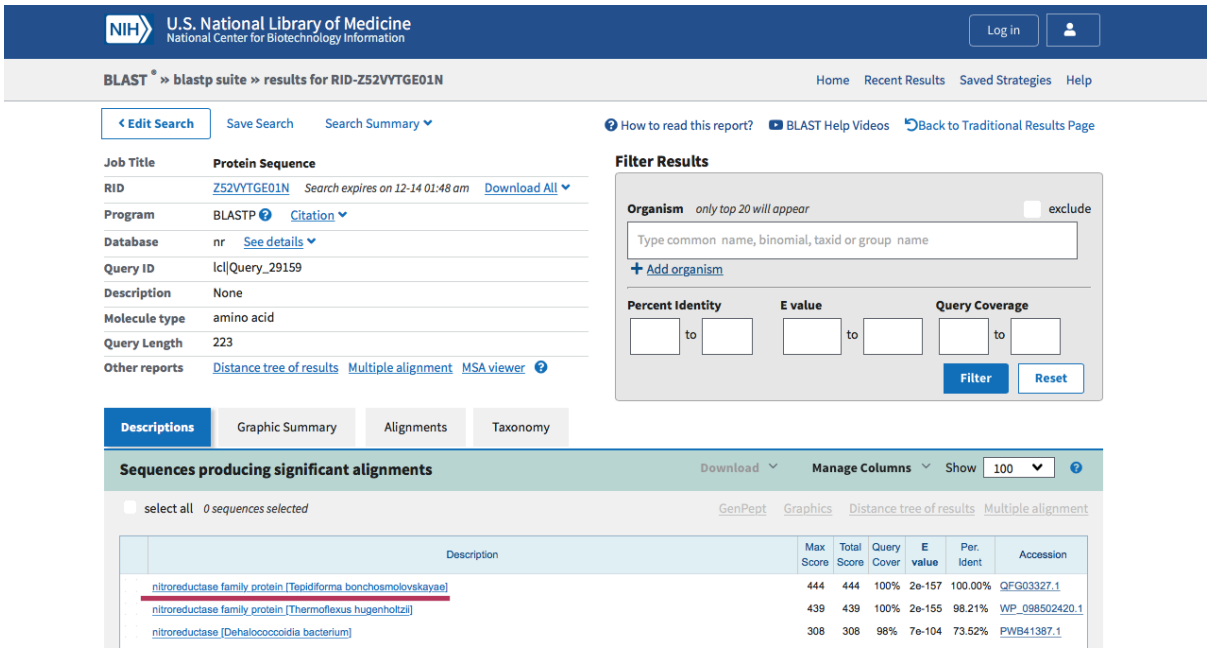


Figure 9 – Results of BLAST for hypothetical protein #12

The result is with good identity (100%), E-value (2e-157), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as nitroreductase family protein [Tepidiforma bonchosmolovskayae].

13) For hypothetical protein #13 the result is

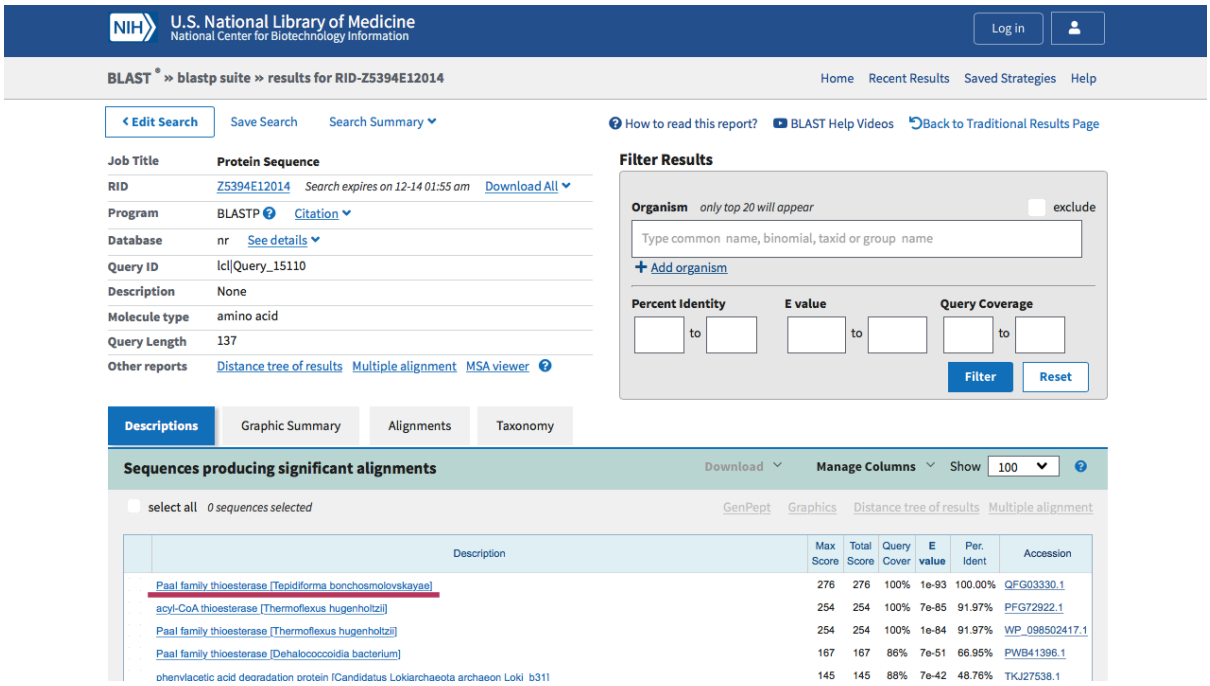


Figure 10 – Results of BLAST for hypothetical protein #13

The result is with good identity (100%), E-value (1e-93), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as PaaI family thioesterase [Tepidiforma bonchosmolovskayae].

14) For hypothetical protein #14 no any appropriate functions with all of the described tools were found.

15) For hypothetical protein #15 the relevant result was found by Pfam

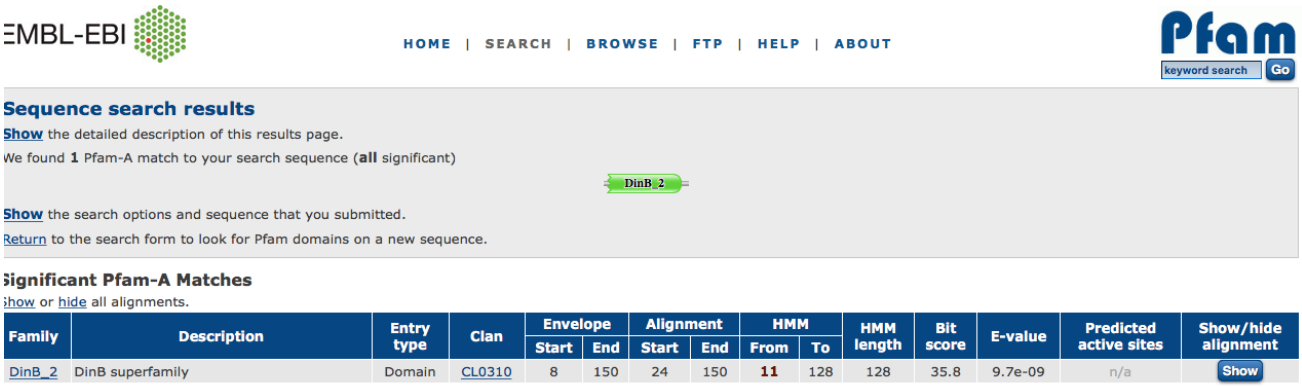


Figure 11 – Results of Pfam for hypothetical protein #15

The protein belongs to DinB superfamily and has the functions like the other proteins of this family – metalloenzymes to perform functions such as redox reactions.

16) For hypothetical protein #16 no any appropriate functions with all of the described tools were found.

17) For hypothetical protein #17 the result is

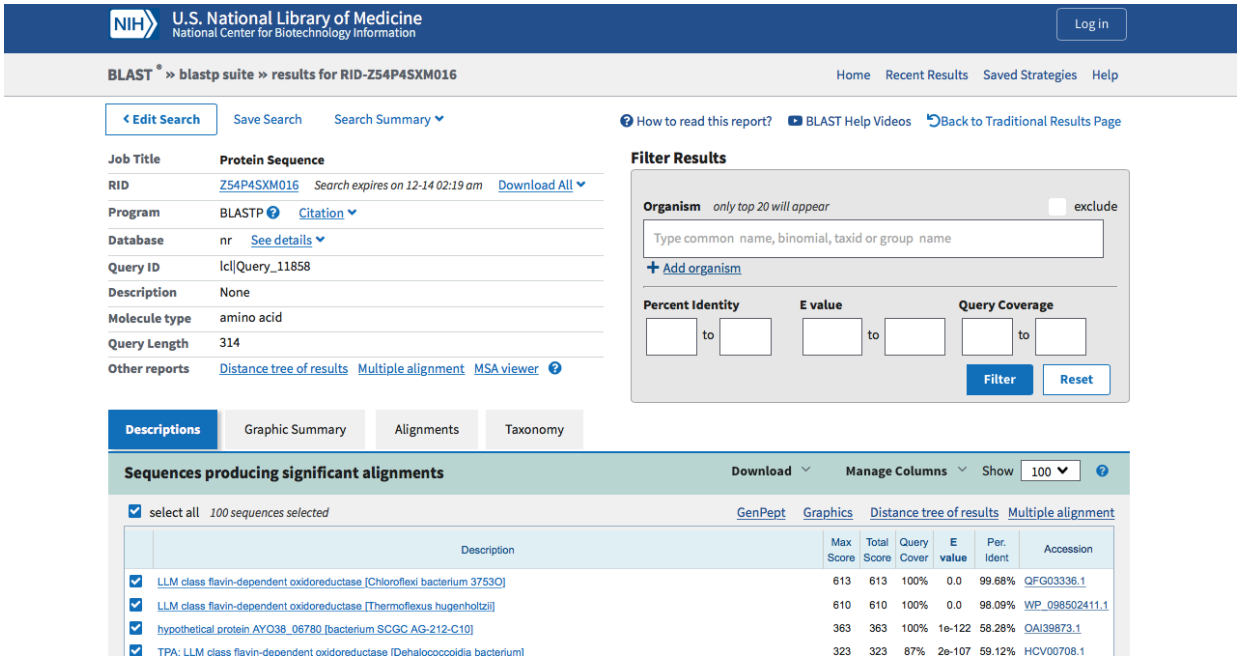


Figure 12 – Results of BLAST for hypothetical protein #17

The result is with good identity (99.68%), E-value (0.0), query cover (100%) and it can be suggested that our hypothetical protein has the same functions as LLM class flavin-dependent oxidoreductase [Chloroflexi bacterium 3753O].

18) For hypothetical protein #18 no any appropriate functions with all of the described tools were found.

3. Description of the operon structure

From Prokka results it is shown that the such operons exist

1-7 (long operon), 8, 9, 10, 11, 12-14, 15-16, 17, 18-23 (long operon), 24-25, 26, 27-29, 30-31, 32.

The distance between the genes in the same operon <150bp.

```
 GNU nano 2.2.6                    File: PROKKA_12112019.gff

##gff-version 3
##sequence-region Tepidiforma 1 30000
Tepidiforma     Prodigal:2.6    CDS     423     1892    .       -       0       ID=result_non-coding_00001;eC_num
Tepidiforma     Prodigal:2.6    CDS     1971    2849    .       -       0       ID=result_non-coding_00002;infere
Tepidiforma     Prodigal:2.6    CDS     2898    3131    .       -       0       ID=result_non-coding_00003;infere
Tepidiforma     Prodigal:2.6    CDS     3219    3878    .       -       0       ID=result_non-coding_00004;infere
Tepidiforma     Prodigal:2.6    CDS     3940    4329    .       -       0       ID=result_non-coding_00005;infere
Tepidiforma     Prodigal:2.6    CDS     4442    7150    .       -       0       ID=result_non-coding_00006;eC_num
Tepidiforma     Prodigal:2.6    CDS     7249    8142    .       -       0       ID=result_non-coding_00007;infere
Tepidiforma     Prodigal:2.6    CDS     8262    10760   .       +       0       ID=result_non-coding_00008;eC_num
Tepidiforma     Prodigal:2.6    CDS     10777   10989   .       -       0       ID=result_non-coding_00009;infere
Tepidiforma     Prodigal:2.6    CDS     11049   12125   .       -       0       ID=result_non-coding_00010;eC_num
Tepidiforma     Prodigal:2.6    CDS     12839   13699   .       +       0       ID=result_non-coding_00011;eC_num
Tepidiforma     Prodigal:2.6    CDS     13868   15103   .       -       0       ID=result_non-coding_00012;infere
Tepidiforma     Prodigal:2.6    CDS     15141   16484   .       -       0       ID=result_non-coding_00013;eC_num
Tepidiforma     Prodigal:2.6    CDS     16588   17496   .       -       0       ID=result_non-coding_00014;infere
Tepidiforma     Prodigal:2.6    CDS     17602   17814   .       +       0       ID=result_non-coding_00015;infere
Tepidiforma     Prodigal:2.6    CDS     17804   18304   .       +       0       ID=result_non-coding_00016;eC_num
Tepidiforma     Prodigal:2.6    CDS     18286   18900   .       -       0       ID=result_non-coding_00017;eC_num
Tepidiforma     Prodigal:2.6    CDS     19007   20614   .       +       0       ID=result_non-coding_00018;Name=g
Tepidiforma     Prodigal:2.6    CDS     20717   21157   .       +       0       ID=result_non-coding_00019;infere
Tepidiforma     Prodigal:2.6    CDS     21177   21479   .       +       0       ID=result_non-coding_00020;eC_num
Tepidiforma     Prodigal:2.6    CDS     21568   21858   .       +       0       ID=result_non-coding_00021;infere
Tepidiforma     Prodigal:2.6    CDS     21903   22574   .       +       0       ID=result_non-coding_00022;infere
Tepidiforma     Prodigal:2.6    CDS     22682   23476   .       +       0       ID=result_non-coding_00023;eC_num
Tepidiforma     Prodigal:2.6    CDS     23473   23955   .       -       0       ID=result_non-coding_00024;Name=y
Tepidiforma     Prodigal:2.6    CDS     23945   24358   .       -       0       ID=result_non-coding_00025;infere
Tepidiforma     Prodigal:2.6    CDS     24418   24756   .       +       0       ID=result_non-coding_00026;infere
Tepidiforma     Prodigal:2.6    CDS     24753   25244   .       -       0       ID=result_non-coding_00027;infere
Tepidiforma     Prodigal:2.6    CDS     25292   25480   .       -       0       ID=result_non-coding_00028;infere
Tepidiforma     Prodigal:2.6    CDS     25487   26923   .       -       0       ID=result_non-coding_00029;eC_num
Tepidiforma     Prodigal:2.6    CDS     27101   28414   .       +       0       ID=result_non-coding_00030;Name=t
Tepidiforma     Prodigal:2.6    CDS     28491   29435   .       +       0       ID=result_non-coding_00031;db_xre
Tepidiforma     Prodigal:2.6    CDS     29461   29703   .       -       0       ID=result_non-coding_00032;infere
##FASTA
```

Figure 13 – Results of Prokka for operon structure task

For long operons (1-7 and 18-23) the regulatory regions should be identified if possible. With software Softberry it can be done (http://www.softberry.com).

For 18-23 operon (+)

```
Length of sequence-     30000
Threshold for promoters -  0.20
Number of predicted promoters -     16
Promoter Pos:  17602 LDF-  3.74
-10 box at pos.  17587 ccgtacact Score    50
-35 box at pos.  17570 ttgact    Score    61
Promoter Pos:  21533 LDF-  2.14
-10 box at pos.  21518 cgtcatatt Score    43
-35 box at pos.  21496 gagata    Score   -12
Promoter Pos:  26953 LDF-  2.08
-10 box at pos.  26938 cggtgtact Score    42
-35 box at pos.  26917 tcgcca    Score    24
Promoter Pos:   8261 LDF-  2.06
-10 box at pos.   8246 ggctacgat Score    53
-35 box at pos.   8223 ttgctg    Score    47
Promoter Pos:  18975 LDF-  1.43
-10 box at pos.  18960 tgctaatat Score    67
-35 box at pos.  18939 ttagca    Score    15
Promoter Pos:  23658 LDF-  1.33
-10 box at pos.  23643 gggtgattt Score    12
-35 box at pos.  23624 ttgcat    Score    50
Promoter Pos:  12838 LDF-  1.15
-10 box at pos.  12823 ccccaaaat Score    37
-35 box at pos.  12803 tcgaca    Score    29
Promoter Pos:  29474 LDF-  0.84
-10 box at pos.  29459 ccctacaat Score    54
-35 box at pos.  29437 ctcccg    Score     0
Promoter Pos:   4444 LDF-  0.79
-10 box at pos.   4429 tggaaagat Score    35
-35 box at pos.   4408 tggacg    Score    23
Promoter Pos:    220 LDF-  0.48
-10 box at pos.    205 gggcaccat Score    31
-35 box at pos.    180 tgccca    Score     4
Promoter Pos:  27604 LDF-  0.46
-10 box at pos.  27589 cggtatcgt Score    48
-35 box at pos.  27569 atgctc    Score     1
Promoter Pos:  15513 LDF-  0.40
-10 box at pos.  15498 cggcacaat Score    48
```

Figure 14 – Results of Softberry for operon 18-23

If we look at Figure 13 and coordinates of operons we see that right promoter has

pos: 18975

LDF - 1.43

However, no TF were found for this promoter

```
Oligonucleotides from known TF binding sites:

For promoter at  17602:
     rpoH3:  CTCCCCCT at position   17574 Score -   9
    rpoS17:  CCCCCTCC at position   17576 Score -  17
For promoter at  21533:
      metR:  TTTTTTCA at position   21510 Score -   8
    rpoD15:  TTTTCACG at position   21512 Score -   6
     ompR:  TCATATTT at position   21520 Score -  11
     argR2:  CATATTTT at position   21521 Score -   8
For promoter at  26953:
       ihf:  ATCATACA at position   26953 Score -  13
No such sites for promoter at   8261
No such sites for promoter at  18975
For promoter at  23658:
      cysB:  TCTTGCAT at position   23622 Score -  10
    rpoD16:  TTCAATCT at position   23650 Score -   7
No such sites for promoter at  12838
For promoter at  29474:
    rpoD16:  CCTACAAT at position   29460 Score -  12
No such sites for promoter at   4444
For promoter at    220:
    rpoS17:  CCCCCTCC at position     195 Score -  17
No such sites for promoter at  27604
No such sites for promoter at  15513
For promoter at   3905:
    rpoD17:  TTCCTCCT at position    3883 Score -   8
    rpoD19:  TCCTGCTA at position    3887 Score -   7
No such sites for promoter at  22932
No such sites for promoter at   7019
No such sites for promoter at  26402
```

Figure 15 – Results of Softberry for operon 18-23

For 1-7 operon we should reverse our sequence (-) and then use Softberry

```
> test sequence
 Length of sequence-     30000
 Threshold for promoters -  0.20
 Number of predicted promoters -     11
 Promoter Pos:  21832 LDF-  2.90
 -10 box at pos.  21817 ttctattct Score    57
 -35 box at pos.  21797 ttgcag    Score    49
 Promoter Pos:   6566 LDF-  2.37
 -10 box at pos.   6551 gggtaaact Score    74
 -35 box at pos.   6526 tagccg    Score    18
 Promoter Pos:   3055 LDF-  2.03
 -10 box at pos.   3040 gtgtatgat Score    68
 -35 box at pos.   3023 ctcacg    Score     5
 Promoter Pos:  17225 LDF-  1.89
 -10 box at pos.  17210 tgctacact Score    62
 -35 box at pos.  17192 atgtcg    Score    17
 Promoter Pos:   9471 LDF-  1.02
 -10 box at pos.   9456 gaggaccat Score     7
 -35 box at pos.   9436 gtgacg    Score    27
 Promoter Pos:  16315 LDF-  0.89
 -10 box at pos.  16300 ggttattcg Score    28
 -35 box at pos.  16282 ttgccg    Score    55
 Promoter Pos:  23476 LDF-  0.71
 -10 box at pos.  23461 agctgtact Score    35
 -35 box at pos.  23441 ttacct    Score    33
 Promoter Pos:   8484 LDF-  0.66
 -10 box at pos.   8469 gtgaaaaat Score    41
 -35 box at pos.   8451 gtgtca    Score    20
 Promoter Pos:  16890 LDF-  0.58
 -10 box at pos.  16875 ggtgaccat Score    20
 -35 box at pos.  16855 ttcagt    Score    18
 Promoter Pos:  13503 LDF-  0.37
 -10 box at pos.  13488 cggcagaat Score    47
 -35 box at pos.  13468 cggccg    Score   -17
 Promoter Pos:  10287 LDF-  0.37
 -10 box at pos.  10272 ctggaggat Score    14
 -35 box at pos.  10252 atgacg    Score    30

 Oligonucleotides from known TF binding sites:
```

Figure 16 – Results of Softberry for operon 1-7

It is shown that no promoters for operon was found.

4. Finding of genes obtained by the bacteria through horizontal gene transfer (HGT)

All the found proteins with high identity are from very close related species. I suppose horizontal gene transfer (HGT) in my fragment sequence has not been concerned.

5. Finding of genes of the bacteria associated with secondary metabolites

KEGG tool was used for finding of metabolic pathways https://www.kegg.jp.

1. Trehalose-6-phosphate synthase is in metabolic pathway – trehalose synthesis
2. Valine - - tRNA ligase - synthesis of tRNA
3. Prodigiosin synthesizing transferase PigC – biosynthesis of the red antibiotic prodigiosin

…

Visualisation part

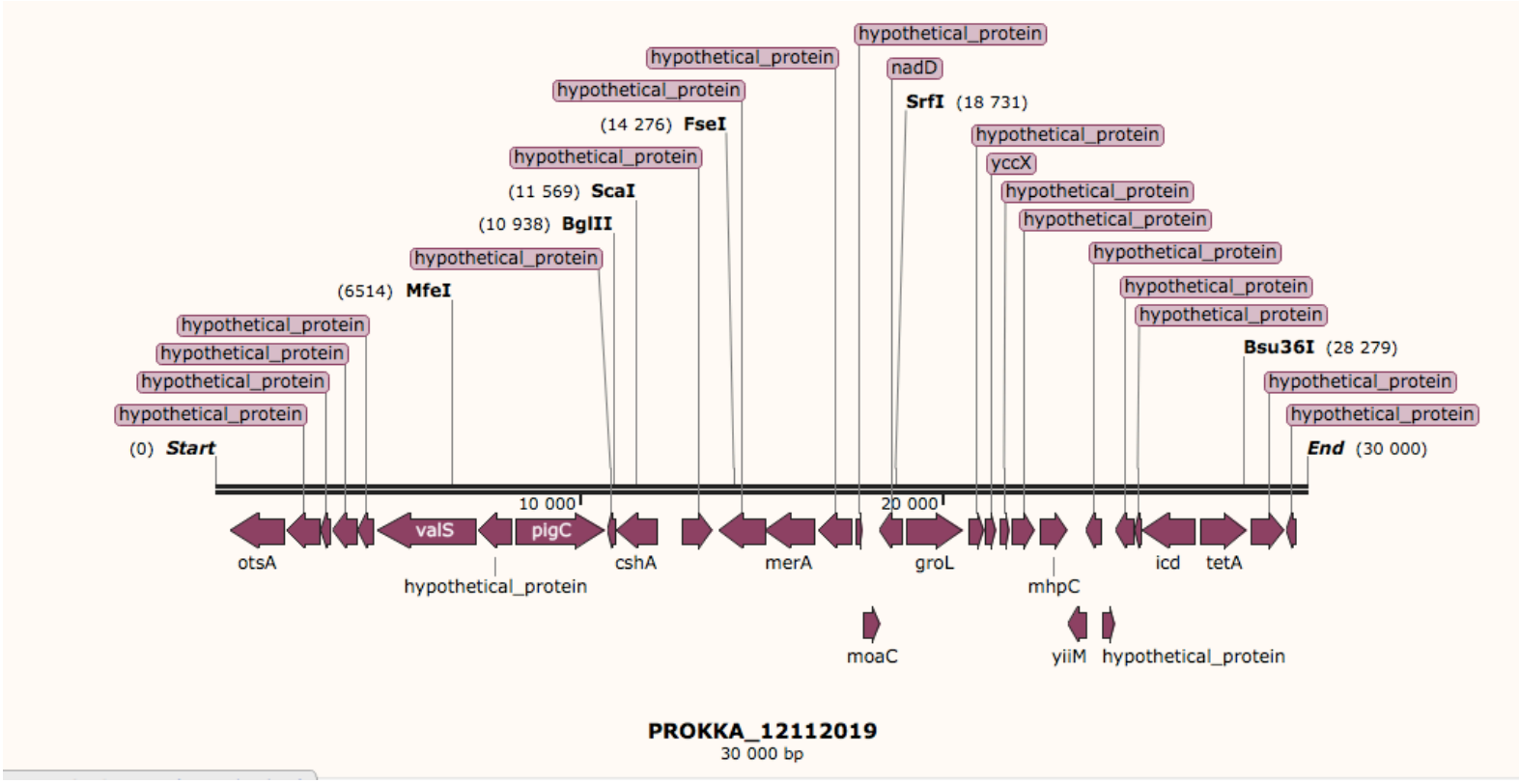


Figure 17 – SnapGene visualisation