HSE / Comparative Genomics course

# Final project: Genome annotation

Student: Katerina Oleynikova
Teacher: Mikhail Gelfand

March, 2022

**Abstract**

This report contains information about the project work of the course "Comparative genomics". The main aim of the project is to annotate a genome of a bacteria of a new species from a new phylum that recently was sequenced. The bacteria is not characterized yet. A fasta file with a fragment (30000 nt long) of genomic DNA of this bacteria is assigned. The aim is to annotate the given region and write a structured report based on the results and findings.

## 1 Introduction

Genome annotation is the process of obtaining the functional and structural information of a gene or protein from raw dataset by use of various analysis, estimation, comparison, precision, and other techniques.

Genome annotation consists of three basic steps. The first one is a nucleotide-level annotation, where the main aim is to find the physical location of DNA sequences to identify where components such as genes, RNAs, and repetitive elements are located. At this stage, sequencing and/or assembly errors can result in false pseudogenes through indels (insertions and deletions). The second category is a protein-level annotation, in which it is needed to to determine the possible gene functions (identifying which one a given organism does or does not have). The third step is a process-level annotation, which seeks to identify the processes and pathways in which different genes can interact (assembling an efficient functional annotation). In the last two levels, sequencing and/or assembly errors may compromise the conclusion about the true function of gene because of reduced similarity (Stein, 2001; Reeves et al., 2009; Miller et al., 2010).

Using the approaches of genome annotation, genes or proteins that can belong to a particular genome can be predicted. Functional annotation of such new genes or proteins can then be performed by identifying their similarity with known experimentally verified sequences that are available in the databases.

Genome annotation is necessary because the the genome or DNA sequencing produces information about the sequence without its functional role. After the genome is sequenced, it is needed to be annotated to bring more information about its functional roles and structural features (Salzberg, 2019).

## 2 Methods

Genome annotation of the given bacterial sequence (to be precise, its fragment) includes several aspects and can be divided into the main sections:

1. Annotation of all coding and non-coding genes;
2. Identification of the functions for coding genes (especially hypothetical ones);
3. Description of the structure of found operons (if any exists);
4. For the long operons (longer than 4 genes): try to identify whether they have some known regulatory regions;
5. Finding genes that were obtained by the bacteria through horizontal gene transfer (HGT);
6. Finding genes associated with secondary metabolites.

To annotate a prokaryotic genome, it is relevant to use Prokka; it is a software tool used for annotation of archaeal, bacterial, and viral genomes [1]. It can be downloaded from GitHub [2] and used then to describe all coding and non-coding genes of a sequence fragment. Prokka can be launched from terminal both on Mac and Linux operational systems.

To identify the number of coding sequences (CDS) for coding genes by Prokka it is possible to do this with the following commands:

*conda install -c bioconda prokka* (1)
*prokka –outdir prokka-result –locustag prokka –kingdom Bacteria genslice13.fasta* (2).

To identify the number of coding sequences (CDS) for non-coding genes it is needed to run this command:

*prokka –outdir prokka-res-n –rfam –locustag prokkar –kingdom Bacteria genslice13.fasta* (3).

After that, we can use Artemis tool which is used for .gff outfiles of Prokka visualisation [3]. Artemis is a genome browser and annotation tool which allows visualisation of next generation data, sequence features, and the results of analyses within the context of a sequence. This tool can be downloaded from the Sanger-pathogens GitHub [3] on any OS that you want.

To found the functions for coding genes (in particular, to investigate the hypothetical proteins), it is possible to look at the Prokka .faa outfile, in which the list of all the possible hypothetical and well-known proteins is represented. Then, it is needed to use BLAST to identify the possible functions for all the hypothetical proteins. BLAST is the Basic Local Alignment Search Tool that allows to find similarity regions between biological sequences (it compares nucleotide or protein sequences to sequence databases and calculates the statistical significance). BLAST can be launched via website page and it does not need to be installed on a laptop or a local machine [4].

Whether operons exist, it is possible to check by Prokka (.gff outfile contains all the needed information). If the long operons exist, their regulatory regions may be checked by use of Softberry tool BPROM. It is best used in regions immediately upstream from ORF start for improved gene and operon prediction in bacteria. Fasta file with bacterial sequence (or its fragment) can be pasted into the special window on the website to run the process of operons identification [5].

To find genes associated with secondary metabolites, we use SnapGene for pathway visualisation (it can be downloaded via the SnapGene website on a local machine) [6].

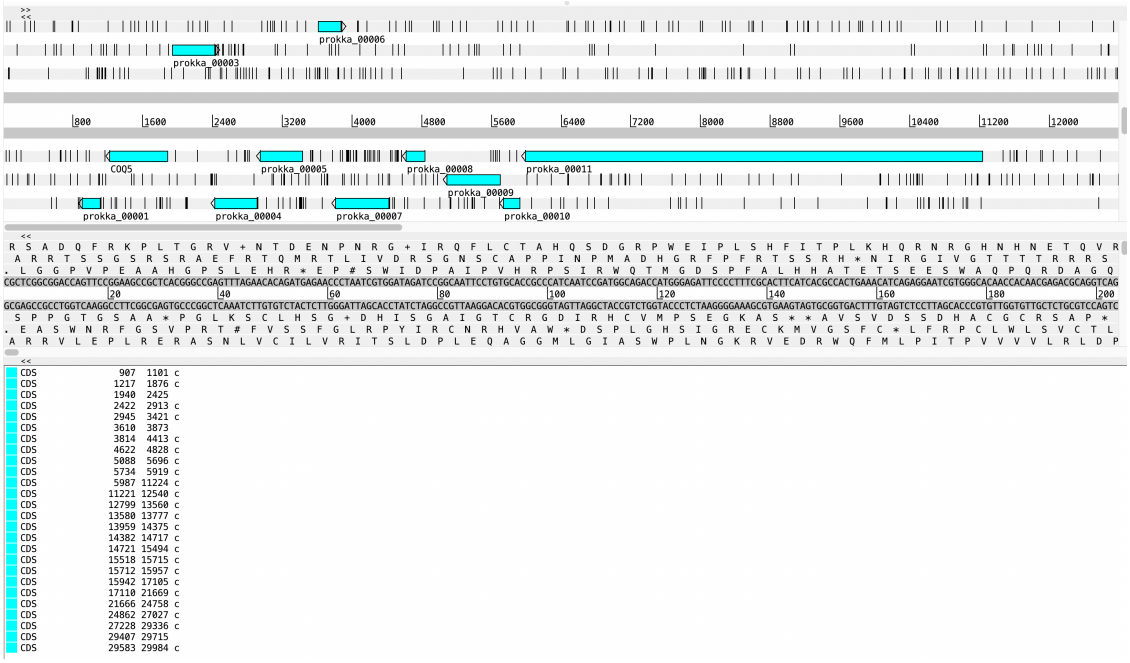


Figure 1: Visualisation of CDS (highlighted in blue) for protein coding genes by Artemis tool.

# 3 Results and Discussion

## 3.1 Annotation of all coding and non-coding genes

As it was previously mentioned, this is possible to annotate both all coding and non-coding genes by running commands 1-2 and 3, respectively; thus, we get the outfiles with the required information about the number of CDS of the bacterial sequence fragment:

CDS: 26 (for coding genes),

CDS: 26 (for non-coding genes).

The results (the number of CDS) are the same; it means there are no any non-coding genes in the fragment.

Using Artemis tool, we get the visualisation of CDS for coding genes (Figure 1). The found CDS are highlighted in blue, the figures that are represented below show the coordinates (start, end) for each coding region of a gene.

## 3.2 Identification of the functions for coding genes

Prokka identifies the 23 hypothetical proteins plus the 3 well-known proteins: 1) 2-methoxy-6-polyprenyl-1,4-benzoquinol methylase, mitochondrial; 2) RNA polymerase-associated protein RapA; 3) Metal-pseudopaline receptor CntO (the list of all found proteins is represented in Supplements).

Then, it is possible to investigate the functions of the found hypothetical proteins by BLAST (blastp) [4]. The Table 1 below represents BLAST (blastp) results (with identity percentage, E-value, and query cover) for all the hypothetical proteins that we have. The step-by-step results by BLAST for the 1st hypothetical protein (as the example) are provided in Supplements); we may suggest that the first hypothetical protein has the same functions as dihydrofolate reductase family protein [Pyrinomonas methylaliphatogenes].

*Def.* Percent Identity: is a number which describes how similar the query sequence is to the target sequence (i.e., how many characters in each sequence are identical). The higher the percentage of identity is, the more significant the match.

*Def.* The Expect value (E-value): is a parameter which describes the number of hits we can expect to see by chance when investigation a database of a particular size (it decreases exponentially as the score (S) of the match increases; essentially, the E-value describes the random background noise).

*Def.* Query Cover: is a number which describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100% (this tells us how long the sequences are, relative to each other).

## 3.3 Identification of operons and description of their structures

The three operons were identified by Prokka (the list (retrieved from .gff Prokka outfile) is shown in Supplements): 1-16, 17, 18.

*Def.* An operon is a cluster of genes that are transcribed together to get a single messenger RNA (mRNA) molecule, which encodes multiple proteins.

### 3.3.1 Identification of regulatory regions (for long operons).

To investigate the existence of promoters, BPROM was used. In our case, we reverse the long operon sequence (-) (the code was implemented in Python using Biopython module) and paste it in the special window. As a result, checking coordinates of long operon we identify that the promoter has position - 5393. Then, we find out about TF (transcription factor) existence and obtain information about it: for promoter at 5393 position TF - rpoH2: CCCTTTAA: at position 5376.

*Def.* In molecular biology, a transcription factor (TF) is a protein that controls the rate of genetic information transcription from DNA to mRNA by binding to a specific DNA sequence.

## 3.4 Finding genes obtained by the bacteria through horizontal gene transfer (HGT)

All the found proteins (from Section 3.2) with high identity are retrieved from very close related species. We may suppose horizontal gene transfer (HGT) in the bacterial fragment sequence has not been concerned.

## 3.5 Finding genes of the bacteria associated with secondary metabolites

According to SnapGene results (it is shown on Figure 2), we identify that the following enzymes take part in metabolic pathways: HpaI, ClaI, BspDI, PsiI, AvrII, DraI, BsrGI, AhdI, SspI, MauBI, KflI (type II restriction enzymes). As we know based on theory the biosynthesis of secondary metabolites is catalysed by enzymes organized into complexes.

*Def.* Secondary metabolism is a term for pathways and small molecule products of metabolism

Table 1: Possible functions of identified hypothetical proteins (according to BLAST results)

| Hypothetical protein | BLAST result | Per. id., % | E-val. | Query cov., % |
|---|---|---|---|---|
| Hypothetical protein 1 | dihydrofolate reductase family protein [Pyrinomonas methylaliphatogenes] | 96.61 | 2e-31 | 92 |
| Hypothetical protein 2 | not found | - | - | - |
| Hypothetical protein 3 | DUF1579 domain-containing protein [Actinomadura bangladeshensis] | 68.18 | 3e-59 | 80 |
| Hypothetical protein 4 | TPA: DUF1697 domain-containing protein [Chlorobi bacterium] | 100.0 | 4e-113 | 100 |
| Hypothetical protein 5 | not found | - | - | - |
| Hypothetical protein 6 | DUF4145 domain containing protein [Azonexus hydrophilus] | 49.22 | 1e-59 | 96 |
| Hypothetical protein 7 | Magnesium-transporting ATPase, P-type 1 [bacterium HR08] | 82.46 | 2e-18 | 83 |
| Hypothetical protein 8 | not found | - | - | - |
| Hypothetical protein 9 | TPA: DUF3387 domain-containing protein [Betaproteobacteria bacterium] | 77.05 | 2e-13 | 100 |
| Hypothetical protein 10 | DEAD/DEAH box helicase [Chlorobi bacterium NICIL-2] | 99.94 | 0.0 | 100 |
| Hypothetical protein 11 | ATP-binding protein [Methylacidiphilum sp. Yel] | 92.03 | 0.0 | 100 |
| Hypothetical protein 12 | not found | - | - | - |
| Hypothetical protein 13 | N-6 DNA methylase [Chloroflexi bacterium] | 78.79 | 6e-11 | 50 |
| Hypothetical protein 14 | TPA: PIN domain-containing protein [Chlorobi bacterium] | 100.0 | 2e-95 | 100 |
| Hypothetical protein 15 | not found | - | - | - |
| Hypothetical protein 16 | TPA: TIGR04255 family protein [Chlorobi bacterium] | 100.0 | 0.0 | 100 |
| Hypothetical protein 17 | N-6 DNA methylase [Chloroflexi bacterium] | 82.86 | 7e-13 | 53 |
| Hypothetical protein 18 | Type IIS restriction enzyme Eco57I [bacterium HR18] | 76.0 | 6e-17 | 61 |

that are involved in ecological interactions but are not absolutely required for the survival of the organism.

| Hypothetical protein | BLAST result | Per. id., % | E-val. | Query cov., % |
|---|---|---|---|---|
| Hypothetical protein 19 | TPA: MBL fold metallo-hydrolase [Chlorobi bacterium] | 100.0 | 0.0 | 100 |
| Hypothetical protein 20 | N-6 DNA methylase [Limisphaera ngatamarikiensis] | 56.56 | 0.0 | 76 |
| Hypothetical protein 21 | not found | - | - | - |
| Hypothetical protein 22 | not found | - | - | - |
| Hypothetical protein 23 | not found | - | - | - |

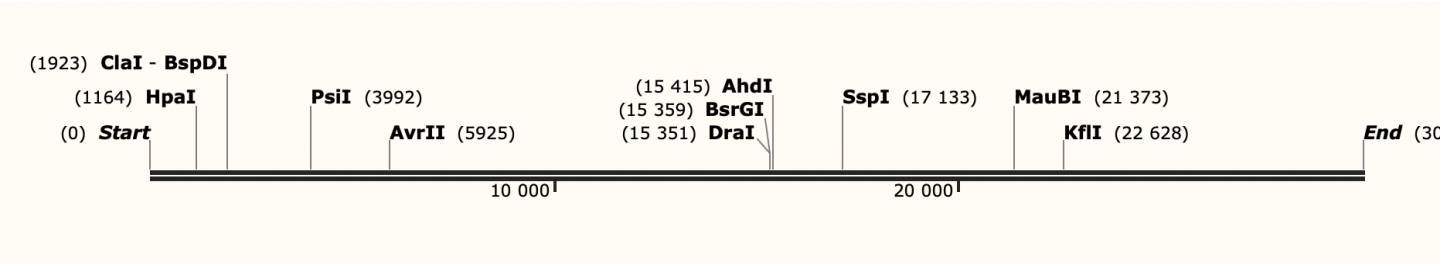


Figure 2: Visualisation of CDS (highlighted in blue) for protein coding genes by Artemis tool.

# 4 Conclusion

In the present project work, several of the genome analysis options allow us to consider the dataset (the bacterial fragment sequence) in terms of its annotation.

The goal of this project was to illustrate how the ideas of genome analysis can be applied to investigate the real novel data.

We only provided a few analysis options, and many more options could have been included. In addition, our choice of analysis options was highly personal and subjective.

# References

Prokka: rapid prokaryotic genome annotation. Retrieved from: https://academic.oup.com/bioinformatics/article/30/14/2068/2390517

Prokka: rapid prokaryotic genome annotation GitHub page. Retrieved from: https://github.com/tseemann/prokka

Artemis Software GitHub page. Retrieved from: https://github.com/sanger-pathogens/Artemis

Basic Local Alignment Search Tool. Retrieved from: https://blast.ncbi.nlm.nih.gov/Blast.cgi

BPROM Softberry tool. Retrieved from: http://www.softberry.com

# 5 Supplements

```
>prokka_00001 hypothetical protein
>prokka_00002 2-methoxy-6-polyprenyl-1,4-benzoquinol methylase, mitochondrial
>prokka_00003 hypothetical protein
>prokka_00004 hypothetical protein
>prokka_00005 hypothetical protein
>prokka_00006 hypothetical protein
>prokka_00007 hypothetical protein
>prokka_00008 hypothetical protein
>prokka_00009 hypothetical protein
>prokka_00010 hypothetical protein
>prokka_00011 hypothetical protein
>prokka_00012 hypothetical protein
>prokka_00013 hypothetical protein
>prokka_00014 hypothetical protein
>prokka_00015 hypothetical protein
>prokka_00016 hypothetical protein
>prokka_00017 hypothetical protein
>prokka_00018 hypothetical protein
>prokka_00019 hypothetical protein
>prokka_00020 hypothetical protein
>prokka_00021 hypothetical protein
>prokka_00022 RNA polymerase-associated protein RapA
>prokka_00023 hypothetical protein
>prokka_00024 Metal-pseudopaline receptor CntO
>prokka_00025 hypothetical protein
>prokka_00026 hypothetical protein
```

Figure 3: The list of all the found proteins by Prokka.

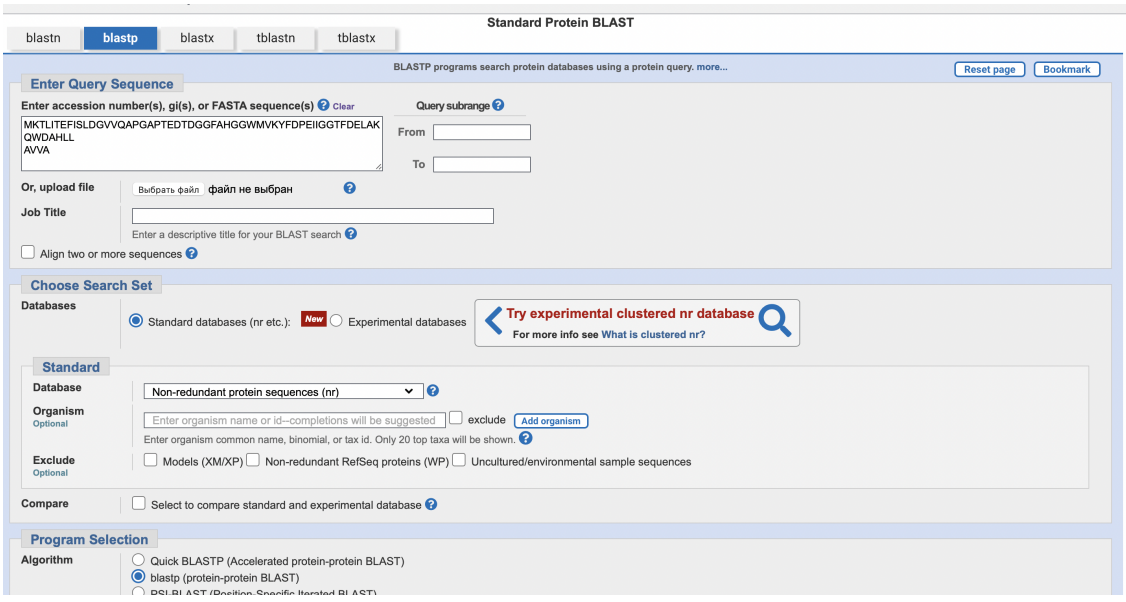


Figure 4: The BLAST searching parameters example.

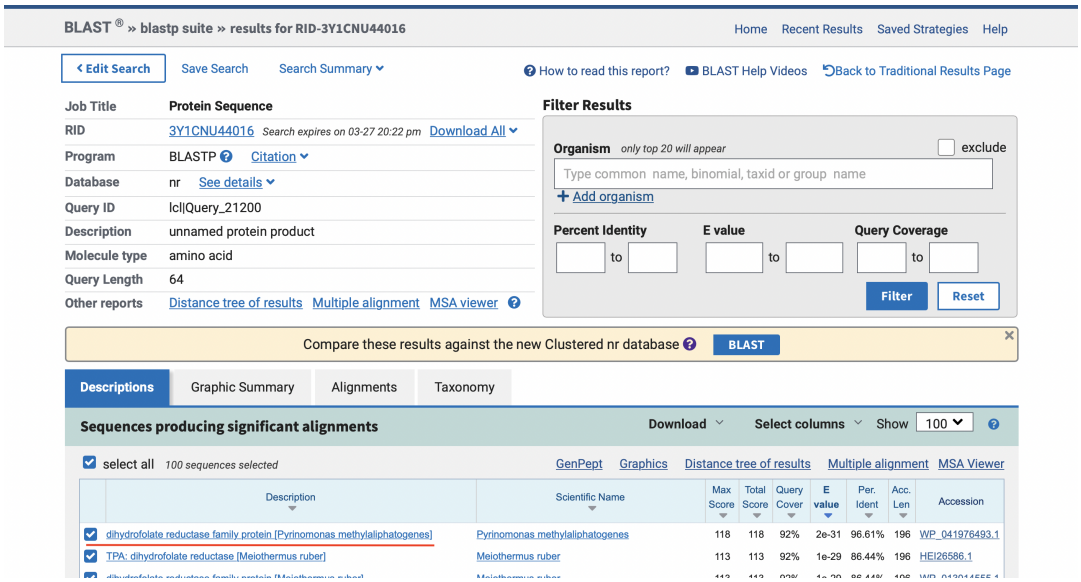


Figure 5: The BLAST result example (for hypothetical protein 1).

```
part    Prodigal:002006 CDS     5088    5696    .       -       0       ID=prokka_00009;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00009;product=hypothetical protein
part    Prodigal:002006 CDS     5734    5919    .       -       0       ID=prokka_00010;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00010;product=hypothetical protein
part    Prodigal:002006 CDS     5987    11224   .       -       0       ID=prokka_00011;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00011;product=hypothetical protein
part    Prodigal:002006 CDS     11221   12540   .       -       0       ID=prokka_00012;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00012;product=hypothetical protein
part    Prodigal:002006 CDS     12799   13560   .       -       0       ID=prokka_00013;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00013;product=hypothetical protein
part    Prodigal:002006 CDS     13580   13777   .       -       0       ID=prokka_00014;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00014;product=hypothetical protein
part    Prodigal:002006 CDS     13959   14375   .       -       0       ID=prokka_00015;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00015;product=hypothetical protein
part    Prodigal:002006 CDS     14382   14717   .       -       0       ID=prokka_00016;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00016;product=hypothetical protein
part    Prodigal:002006 CDS     14721   15494   .       -       0       ID=prokka_00017;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00017;product=hypothetical protein
part    Prodigal:002006 CDS     15518   15715   .       -       0       ID=prokka_00018;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00018;product=hypothetical protein
part    Prodigal:002006 CDS     15712   15957   .       -       0       ID=prokka_00019;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00019;product=hypothetical protein
part    Prodigal:002006 CDS     15942   17105   .       -       0       ID=prokka_00020;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00020;product=hypothetical protein
part    Prodigal:002006 CDS     17110   21669   .       -       0       ID=prokka_00021;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00021;product=hypothetical protein
part    Prodigal:002006 CDS     21666   24758   .       -       0       ID=prokka_00022;eC_number=3.6.4.-;Name=rapA;gene=rapA;inference=ab initio prediction:Prodigal:002006,protein motif:HAMAP:MF_01821;l
part    Prodigal:002006 CDS     24862   27027   .       -       0       ID=prokka_00023;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00023;product=hypothetical protein
part    Prodigal:002006 CDS     27228   29336   .       -       0       ID=prokka_00024;Name=cntO;gene=cntO;inference=ab initio prediction:Prodigal:002006,similar to AA sequence:UniProtKB:A0A0H2ZI93;locu
part    Prodigal:002006 CDS     29407   29715   .       +       0       ID=prokka_00025;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00025;product=hypothetical protein
part    Prodigal:002006 CDS     29583   29984   .       -       0       ID=prokka_00026;inference=ab initio prediction:Prodigal:002006;locus_tag=prokka_00026;product=hypothetical protein
```

Figure 6: Operons identification by Prokka.

```
Promoter Pos:   5393 LDF-  0.61
-10 box at pos.   5378 CTTTAACGT Score     33
-35 box at pos.   5361 TCGATG    Score     15
```

Figure 7: Promoters identification by Softberry.

```
For promoter at   5393:
     rpoH2:  CCCTTTAA at position    5376 Score -  10
```

Figure 8: TF identification by Softberry.