Mini-Review

Genome and proteome analysis of the facultative anaerobic bacterium *Erysipelothrix larvae*

Yuliya A. Mironenko[1*]

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

* To whom correspondence should be addressed.

Contact: juliamir3112@gmail.com

Abstract: *Erysipelothrix larvae* is a novel, Gram-stain positive, facultative anaerobic bacterium, isolated from the larval gut of the rhinoceros beetle, *Trypoxylus dichotomus*. Unlike other *Erysipelothrix* strains, this bacterium is arsenic-resistant. In this mini-review, some genome and proteome parameters have been analyzed. Using descriptive statistics methods, protein length distribution has been analyzed. Codon usage analysis was made. Specific arsenic-resistant genes were calculated. Replication origin and terminus position have been predicted using a cumulative GC skew method.

Keywords: *Erysipelothrix larvae, Trypoxylus dichotomus*

1 INTRODUCTION

Erysipelothrix species are known to be pathogenic, causing erysipelas in a wide range of hosts, including mammals and birds (Chirico et al., 2003). As Bang BH et al. (2015) described, *Erysipelothrix larvae* (strain LV19^T) is a modern, facultative anaerobic, non-motile and straight to curve rod shaped bacterium, isolated from the larval gut of the rhinoceros beetle, *Trypoxylus dichotomus*, which was collected from Yeong-dong, Chuncheongbuk-do, South Korea. According to the article (Bang BH et al., 2015), the colonies of the new isolate were convex, circular, cream white in color and 1–2 mm in diameter after 3 days incubation on Tryptic Soy Agar at 37 °C. Based on the 16S rRNA gene sequence similarity, the new isolate was most closely related to *E. inopinata, E. rhusiopathiae* and *E. tonsillarum* (94.8, 93.8 and 93.7 % similarity, respectively).

In concordance with S. Lim et al. article (2016), virulence- and chemical tolerance-related genes were compared between *Erysipelothrix larvae* LV19^T and other *Erysipelothrix rhusiopathiae* strains.

Interestingly, none of the virulence-related genes was found in LV19^T. However, the following 6 genes were identified: arsenical pump-driving ATPase (EC 3.6.3.16), arsenical resistance operon repressor, arsenical resistance operon trans-acting repressor (ArsD), arsenical-resistance protein (ACR3), internalin putative and an internalin-like protein (LPXTG motif) Lmo0409 homolog. Most of these genes are associated with arsenic resistance and have not been detected in other pathogenic *Erysipelothrix* strains (S. Lim et al., 2016).

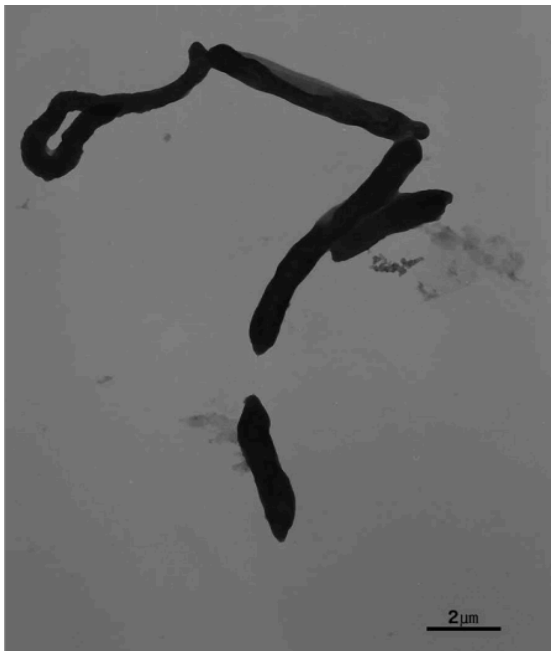

Fig. 1 Transmission electron micrograph of *Erysipelothrix larvae* (Bang BH et al., 2015)

2 METHODS

2.1 Genome analysis

In order to make this research BASH, Google Sheets, Microsoft Visual Studio (Python 3.9) programs and NCBI database were used.

2.1.1 Genome's structure

The information about chromosome and plasmid's length was taken from the NCBI database (Supplementary Materials 1). The percentage of GC-content was counted by My Python program (Supplementary Materials 3).

2.1.2 Codon usage analysis

My Python program (Supplementary Materials 3) was used to analyze the frequency of codons, which encode different amino acids. Stop-codons were not counted by the program. Using

Google Sheets (Supplementary Materials 2), a table with the results was made. All codons, except stop-codons were sorted in the alphabet order. To find the acid which is encoded by the specific codon amino acid codon mapping table was used (Supplementary Materials 5).

2.1.3 Analysis of the nucleotide composition of genomic DNA
In order to find the number of nucleotides in the chromosome and the plasmid My Python program (Supplementary Materials 3) was used. Sequences of *Erysipelothrix larvae's* chromosome and plasmid from fasta files, which were obtained from the NCBI database (Supplementary Materials 1), were used. The pie-charts were made with the help of Google Sheets (Supplementary Materials 2).

2.1.4 Replication origin and terminus prediction
The cumulative GC skew method was used to predict the position of replication origin and terminus. Minimum and maximum cumulative GC skew were analyzed by using the online version of the Genskew program (Supplementary Materials 4). GC skew was calculated using the formula below:
GC skew = (G - C) / (G + C)

2.1.5 Protein distribution in + and - DNA strands
Sorting CDs list in Google Sheets (Supplementary Materials 2) and using the function "СЧЁТЕСЛИМН", the numbers of genes in the chromosome and the plasmid were counted. The statistical significance for chromosome was found by using the following formula:
"=2*BINOM.DIST(МИН(B2;B3);B2+B3;0,5;ИСТИНА)"
Similarly, for the plasmid.

2.2 Proteome analysis
2.2.1 Protein length distribution analysis
Google Sheets with sorting CDs list and the following function (for the cell A4):
"=СЧЁТЕСЛИМН(CDS!H:H;">="&A3;CDS!H:H;"<"&A4)"
were used in order to make a histogram. And after that, using opportunities of Google Sheets, the diagram of protein length distribution and protein length statistics were made (Supplementary Materials 2).

2.2.2 Protein distribution by their functions analysis
For protein distribution analysis BASH was used.
In order to find the total number of genes:

```
tail -n +2 ./genome/*_feature_table.txt |  wc –l
```

Where " ./genome/*_feature_table.txt" is *Erysipelothrix larvae's* feature table txt file with a BASH mask. "./genome/" means that this file is stored in "genome" folder, and "genome" folder is stored in the current directory.

The number of genes on straight and complementary chains:

```
tail -n+2 ./genome/*_feature_table.txt | cut -f10 | sort | uniq -c
```

In order to find the total number of proteins, the total number of all RNA, the number of tRNA (transport RNA) and rRNA (ribosomal RNA):

```
tail -n+2 ./genome/*_feature_table.txt | cut -f1 | sort | uniq -c
```

The number of ribosomal proteins:

```
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'ribosomal protein'
```

The number of transport proteins:

```
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'transport'
```

The number of hypothetical proteins:

```
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'hypothetical'
```

The numbers of genes, which make *Erysipelothrix larvae* arsenic resistant:

```
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'ATPase'
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'ArsD'
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'ACR3'
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'LPXTG'
tail -n+2 ./genome/*_feature_table.txt | cut -f14 | grep -c 'ArsR'
```

3 RESULTS

3.1 Genome analysis
3.1.1 Genome's structure
The *Erysipelothrix larvae's* genome consists of a single circular DNA chromosome and one unnamed plasmid. For each of these molecules GC-content was calculated. The results of calculations and molecules' lengths can be found in Table 1.

Table 1 Genome's structure

DNA	Length (bp)	GC-content (%)
Chromosome	2,495,108	37.427
Plasmid	16,378	36.781

3.1.2 Codon usage analysis
The result of the codon usage analysis is represented in Table 2. The numbers in table cells show how many times every codon was found. It was revealed that codons in *Erysipelothrix larvae* are not used with the same probability. It is consistent with the synonymous codon usage bias conception (Ermolaeva MD et al., 2001).
As it can be seen from the table, the most popular codons are TTT, AAA, ATT and AAT, which encode Phe, Lys, Ile and Asn, respectively.

Table 2 Codon usage analysis results sorted in the alphabet order

AAA 28578	CAA 21195	GAA 17646	TAC 11314
AAC 15601	CAC 11140	GAC 7198	TAT 20060
AAG 17831	CAG 8331	GAG 6834	TCA 20743
AAT 26091	CAT 18805	GAT 18834	TCC 9960
ACA 15853	CCA 12364	GCA 11763	TCG 7854
ACC 10131	CCC 5796	GCC 4036	TCT 14144
ACG 7476	CCG 3787	GCG 5465	TGC 12002
ACT 12058	CCT 8222	GCT 8001	TGG 11623
AGA 12674	CGA 7706	GGA 9621	TGT 16770
AGC 7716	CGC 5450	GGC 4043	TTA 19951
AGG 7962	CGG 3818	GGG 5854	TTC 19263
AGT 12317	CGT 7686	GGT 10174	TTG 22476
ATA 20580	CTA 7805	GTA 11706	TTT 30694
ATC 19125	CTC 7108	GTC 7530	TAA 0
ATG 18501	CTG 8569	GTG 11412	TAG 0
ATT 26327	CTT 18729	GTT 16667	TGA 0

3.1.3 Analysis of the nucleotide composition of genomic DNA

The results of counting nucleotides in *Erysipelothrix larvae's* chromosome and plasmid are represented in Table 3 and Table 4, respectively. As it can be seen from Table 3 and Fig. 2, the number of A nucleotides in the chromosome is approximately equal to the number of T nucleotides (774296 versus 797068). And the number of C is approximately equal to the number of G (471577 versus 468445). This means the confirmation of the Chargaff's second rule.

However, Table 4 and Fig. 3 illustrate significant differences between numbers of A (5067, 30.94%) and T (5287, 32.28%), and numbers of C (2800, 17.10%) and G (3224, 19.68%). These differences can be explained by the small length of the plasmid for making this test reliable.

Table 3 Nucleotide distribution of the chromosome

The chromosome

	count	percentage (%)
A	**774396**	**30.83**
T	**797068**	**31.74**
C	**471577**	**18.78**
G	**468445**	**18.65**

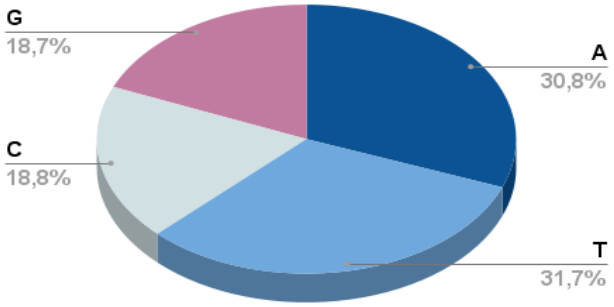

Fig. 2 Pie-chart based on the results from Table 3

Table 4 Nucleotide distribution of the plasmid

The plasmid

	count	percentage (%)
A	**5067**	**30.94**
T	**5287**	**32.28**
C	**2800**	**17.10**
G	**3224**	**19.68**

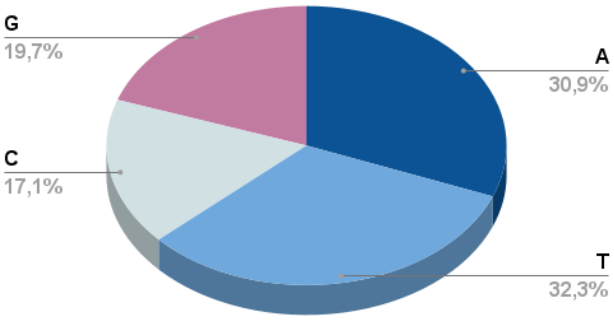

Fig. 3 Pie-chart based on the results from Table 4

3.1.4 Replication origin and terminus prediction

The oriC (start of DNA replication) and ter (termination of DNA replication) were found in the chromosome of *Erysipelothrix larvae* (Fig. 4). They were estimated by analyzing the Gen-skew plot for *Erysipelothrix larvae's* sequence. The position where GC-skew was the smallest was determined as oriC (approximately 2495934) and the position where GC-skew was the biggest was determined as ter (approximately 1067175).

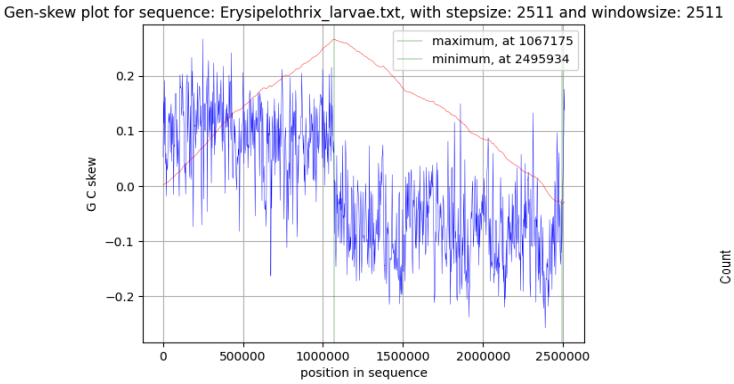

Fig. 4 Gen-skew plot for the chromosome

3.1.5 Protein distribution in + and - DNA strands

As it can be seen from Table 5, there is quite a big difference between the numbers of protein genes on "+" and "-" strands, both in the chromosome and unnamed plasmid. The statistical significance is also really small. In the chromosome it is only 0,00000003003, and in unnamed plasmid approximately 0,00154, which is much bigger than in the chromosome, but still too small. The statistical significance in the chromosome and plasmid is less than 0,05 which means that genes between the chains are distributed randomly.

Table 5 The protein distribution in "+" and "-" strands in the chromosome and unnamed plasmid

	chromosome	unnamed
+	1013	20
-	1279	4
statistical significance	0,00000003003	0,001543879509

3.2 Proteome analysis

3.2.1 Protein length distribution analysis

The protein length distribution diagram is represented in Fig. 5. Some statistical parameters of protein length distribution are represented in Table 6. The mean value of protein length is 316, which is really close to the value of 320 amino acids, calculated for bacterial proteins (Tiessen A et al., 2012). The median value of protein length is 277, which is also close to the value of 273 amino acids, calculated for bacterial proteins (Tiessen A et al., 2012). Other results of protein length distribution in *Erysipelothrix larvae* are also predictable and common for Bacteria.

Table 6 Protein length statistics

Mean value	316
Standard deviation	230
Median	277
Minimum	36
Maximum	4 515

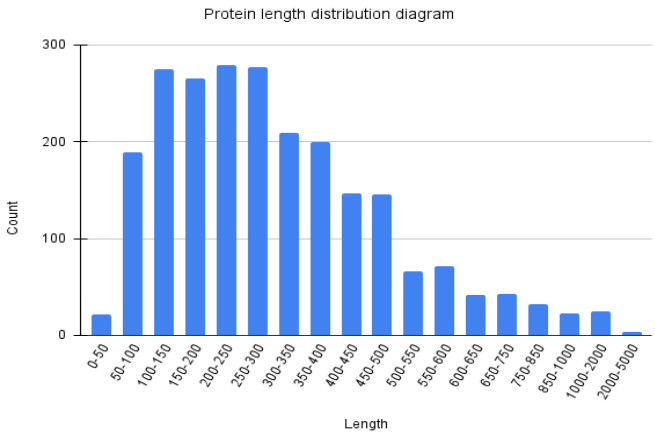

Fig. 5 Protein length distribution diagram

3.2.2 Protein distribution by their functions analysis

The results of counting proteins, genes and specific arsenic-resistant genes are represented in Table 7. As it can be seen from the table, there is a significant difference between the numbers of genes on straight and complementary chains (2128 versus 2640). 3 RNA, which are not included in rRNA and tRNA were found. 39 special genes, which make *Erysipelothrix larvae* arsenic resistant (ATPase - 32, ArsD - 2, ACR3 - 1, LPXTG - 2 and ArsR - 2) were found. These genes are unique for *Erysipelothrix larvae* LV19^T and can not be found in other pathogenic *Erysipelothrix* strains (S. Lim et al., 2016).

Table 7 The number of proteins and genes divided by their functions

Total number of genes	**4768**
Straight chain genes	**2128**
Genes on complementary chain	**2640**
Total number of protein genes	**2384**
Ribosomal proteins	**50**
Hypothetical proteins	**409**
Transport proteins	**247**
ATPase	**32**
ArsD	**2**
ACR3	**1**
LPXTG	**2**
ArsR	**2**

Total RNA	**52**
Ribosomal RNA	**12**
Transport RNA	**37**
Other RNA (indefinite)	**3**

4 CONCLUSION

Erysipelothrix larvae LV19^T is a bacterium with a single circular DNA chromosome and one unnamed plasmid. In this research the approximate oriC (2495934 position) and ter (1067175 position) were found in the chromosome. The second Chargaff's rule was verified. The codon usage analysis was made. Different types of protein distribution analysis were made. Special genes, which make the bacterium arsenic resistant and separate it from other *Erysipelothrix* strains, were calculated.

ACKNOWLEDGEMENTS

I thank my parents, sister and friends for their support and encouragement. I express deep gratitude to MSU teachers of bioinformatics for their help and interesting lessons.

SUPPLEMENTARY MATERIALS

1. All NCBI database for *Erysipelothrix larvae* LV19^T Index of /genomes/all/GCF/001/545/095/GCF_001545095.1_ASM154509v1 (nih.gov)
2. All results made with a help of Google Sheets https://docs.google.com/spreadsheets/d/1wz6_tgPhJqJBqMCldyIdmAJXtQlXvULh8KMmAH2_SEY/edit?usp=sharing
3. All My Python scripts in py format https://drive.google.com/drive/folders/1oxZbhNCPFv_TR8S8iLB6SfLPuaViZinu?usp=share_link
4. The online version of Gen-skew program https://genskew.csb.univie.ac.at/webskew
5. Amino acid codon mapping table Codon Table (genscript.com)

REFERENCES

Bang, B.-H., Rhee, M.-S., Chang, D.-H., Park, D.-S., Kim, B.-C., 2015. Erysipelothrix larvae sp. nov., isolated from the larval gut of the rhinoceros beetle, Trypoxylus dichotomus (Coleoptera: scarabaeidae). Antonie Van Leeuwenhoek 107, 443–451.

Chirico, J., Eriksson, H., Fossum, O., Jansson, D., 2003. The poultry red mite, Dermanyssus gallinae, a potential vector of Erysipelothrix rhusiopathiae causing erysipelas in hens. Med. Vet. Entomol. 17, 232–234.

S. Lim et al. / Journal of Biotechnology 223 (2016) 40–41.

Ermolaeva MD. Synonymous codon usage in bacteria. Curr Issues Mol Biol. 2001 Oct;3(4):91-7. PMID: 11719972.

Tiessen A et al. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Res Notes. 2012 Feb 1;5:85. doi: 10.1186/1756-0500-5-85. PMID: 22296664; PMCID: PMC3296660.