Mini-review

Mini-review of the bacterium *Leuconostoc citreum,* strain KM20 proteome

Emir Radkevich

Moscow State University, School of Bioengineering and Bioinformatics, 1-st course

ABSTRACT

This review is devoted to research of *Leuconostoc citreum KM20* proteome. The analysis of the gene's distribution on straight and complementary strings is shown. Cross-genes patterns, quasioperons principles are studied. The hypothesis of the relative distribution of genes on straight and complementary strings is checked.

1 INTRODUCTION

Leuconostoc citreum KM20 belongs to the Leuconostoc genus. There are 14 species in this genus (except "sp." species), 9 plasmids in "citreum" species and 25 plasmids in "Leuconostoc" genus. As a result the percentage of "KM20" plasmids is 55.56% among all species and 20% among all genus. KM20 strain is reviewed in this paper; the genome of this bacterium was completely sequenced in the 2004. [1]

Classification of the bacterium: [2]

>Superkingdom: Bacteria
>Phylum: Firmicutes
>Class: Bacilli
>Order: Lactobacillales
>Family: Leuconostocaceae
>Genus: Leuconostoc
>Species: Leuconostoc citreum
>Strain: *Leuconostoc citreum KM20*

The genome includes one circular chromosome and four circular plasmids (pLCK1, pLCK2, pLCK3, pLCK4). There are 1.820 CDS genes, which encode proteins and 82 RNA genes, which encode rRNAs and tRNAs. During the work there were constructed:

- Bar chart of the protein distribution (percentage);
- Bar chart of the protein distribution (length);
- Line graph of possible p-value of RNA-genes, CDS-genes and all genes;
- Line graph of the length of cross-genes between strings;
- Line graph of the length of cross-genes inside strings.

The main purpose of my work is to analyze bacterium proteome, show some unique features that belong to the bacterium.

2 MATERIALS AND METHODS

Microsoft Excel 2013 was used as a main program for data processing. There are some basic computing functions using which it was possible to calculate binomial distribution (BINOM.DIST), the length of cross-genes(COUNTIF, COUNTIFS), construct figures.

Data was loaded using NCBI's FTP server (open access database). [3*]

3 RESULTS

There are three subparagraphs in this subsection. The first subparagraph is a short description of the protein distribution in bacterium proteome (two figures: percentage, lengths). The second subparagraph is short review of a protein distribution between straight and complementary strings. The third - is an additional information about bacterium's quasioperons and cross-genes.

3.1 Protein distribution in *Leuconostoc citreum KM20* proteome

Protein lengths distribution is illustrated as a two histograms.

I. On the *Fig.1*, you can see the percentage distribution of all the proteome (chromosome and four plasmids). As you can see, the biggest amount of proteins lays in range between 0-300 aa's while the smallest – 900-1000 aa's. Thus, it can be said without prejudice, that amount of short proteins (peptides) is greater than long-length proteins. Also this figure shows the relative protein distribution among all genetic material (chromosome and four plasmids).

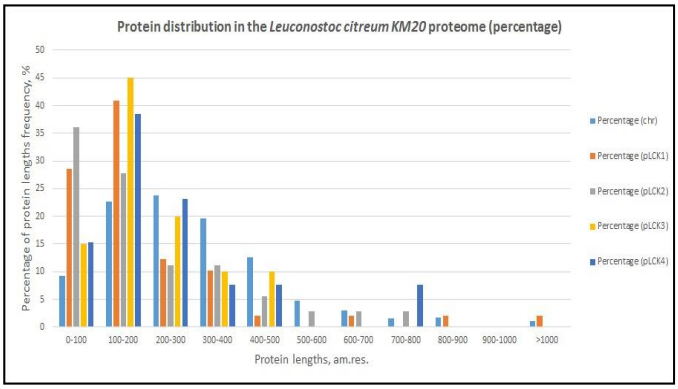

Figure 1. Percentage distribution.

II. On the *Fig.2 and Fig.3*, you can see protein lengths distribution in chromosome and plasmids respectively. These histograms are extremely similar to the previous one. Description of the protein distribution is the same as in previous figure. The main reason to construct more than one histogram is to show protein distribution from different perspectives.

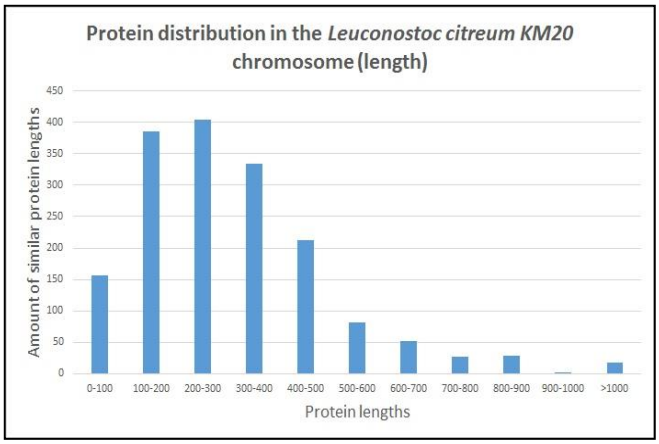

Figure 2. Protein lengths distribution in chromosome.

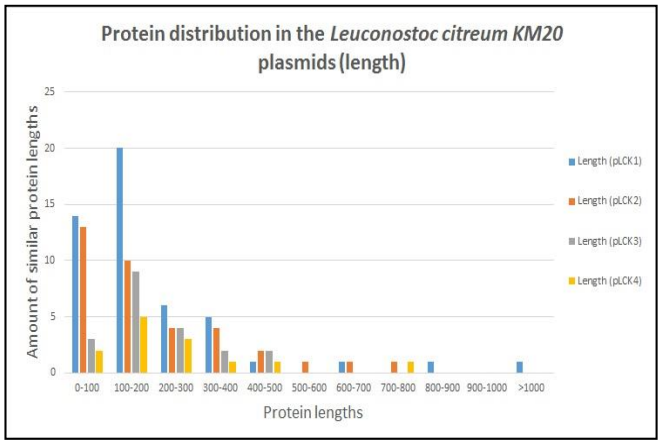

Figure 3. Protein lengths distribution in plasmids.

3.2 Gene distribution between straight and complementary strings.

Data of the number of genes located on different strings are shown in the *Table 1.*

Table 1. Gene's distribution, that encode proteins, rRNAs and tRNAs on straight and complementary strings.

DNA type	*String direction*	*Proteins*	*tRNAs, rRNAs*
Chromosome	Straight	838	30
	Complementary	864	52
pLCK1	Straight	40	-
	Complementary	9	-
pLCK2	Straight	17	-
	Complementary	19	-
pLCK3	Straight	14	-
	Complementary	6	-
pLCK4	Straight	11	-
	Complementary	2	-

As you can see, there are no RNA-coding genes in plasmids, protein-coding genes only. Next step of processing data was to calculate binomial distribution for protein- and RNA-coding genes, genes in total. As a result of using BINOM.DIST function, the following results were obtained:

- P-value of protein-coding genes equals to 0.54454
- P-value of RNA-coding genes equals to 0.10364
- P-value of genes in total equals to 0.98171

According to the results of this test hypothesis is true for genes in total, protein-coding genes, RNA-coding genes, because of results, that more than 0.05. Line graphs (possible p-value) were constructed for deeper understanding of the problem. Boolean expression that determines the type of function was changed from cumulative to mass function. Due to the changing, you can see clearer picture of the binomial distribution (*Fig.4 and Fig. 5*). Black dots on figures represent p-values of obtained data. As a result of constructing additional graphs it can be said without prejudice, that hypothesis is fully confirmed.

3.3 Quasioperons and cross-genes.

Data of the number of quasioperons and cross-genes are shown in the *Table 2.*

Table 2. Amount of quasioperons and cross-genes in *Leuconostoc citreum KM20* genome.

DNA type	*Genes amount*	*Quasioperons*	*Cross-genes(between strings/in strings)*
Chromosome	1784	353	6/290
pLCK1	49	5	2/10
pLCK2	36	19	-/2
pLCK3	20	14	-/1
pLCK4	13	9	-/2

As you can see the biggest amount of quasioperons is in chromosome, the least – pLCK1. It is not obviously because of plasmids size. pLCK1 (49 genes) is much bigger than pLCK3 (20 genes). If you will take a look on percentage of quasioperons, You can see a big difference between plasmids. pLCK1 has much smaller percentage of quasioperons than other plasmids.

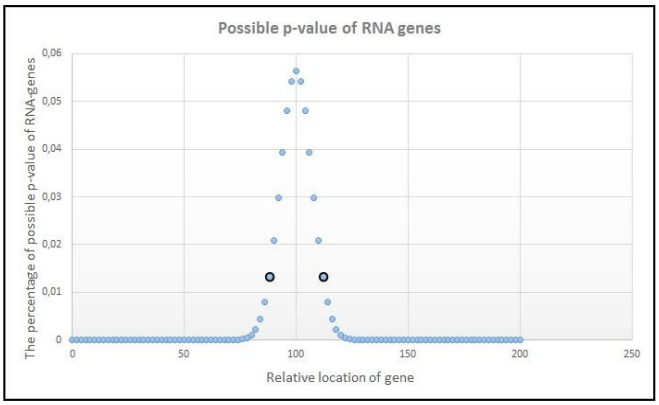

Figure 4. Possible p-value of RNA-coding genes.

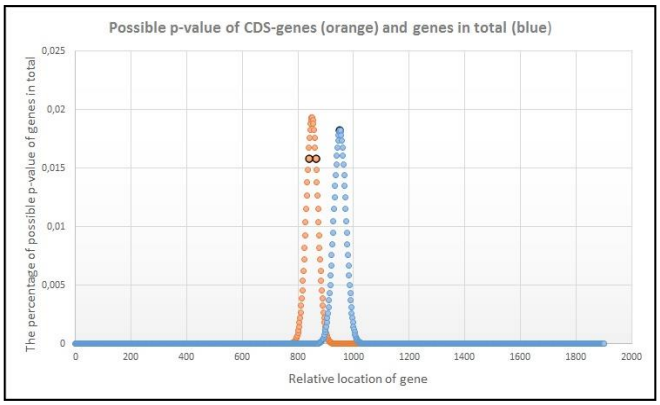

Figure 5. Possible p-value of protein-coding genes and genes in total

After looking at all data in the *Table 2* it becomes obvious, that amount of cross-genes between strings is much less than their amount in strings. No other features are not observed.
(*Additional graphs you can see in chromosome.xlsx)

4 DISCUSSIONS

More than 75% of all proteins have length in range of 1 to 500 aa's. It is a normal distribution among bacteria. There are only one protein, which length more than 2000 aa's and four proteins, which lengths less than 50 aa's. The largest protein called Alternansucrase. [4] It is an enzyme that transfers an alpha-D-glucosyl residue from sucrose alternately to the non-reducing terminal residue alpha-D-glucan. Protein of 40 aa's is a hypothetical protein that is not completely defined. Protein of 43 aa's. is a 50s ribosomal protein L34, binding the 23S rRNA is its main function. [5]

5 CONCLUSION

The hypothesis that the distribution of genes on strands is accidental is confirmed.
The file chromosome.xlsx shows the number of quasioperons, cross-genes and their proportion of total number of genes. Proportion of quasioperons for plasmids in common more than 50%. Such big percent can be explained by the relatively small size of plasmids that have to perform certain function. Therefore, the genes must be located compactly in plasmids.

6 SUPPLEMENTARY MATERIALS

chromosome.xlsx – file with all calculations, tables, figures, such as protein distribution, gene's distribution, quasioperons, cross-genes and other necessary information.
data.xlsx – file with all data that were given from the NCBI's FTP server (5 .ptt files and 1 .rnt file)

7 ACKNOWLEDGEMENTS

I thank teachers of bioinformatics and English language for excellent teaching of subjects in this term.

8 REFERENCES

[1] - http://www.ncbi.nlm.nih.gov/nuccore/DQ489736.1 (*Leuconostoc citreum KM20*, complete genome)
[2] - http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=349519 (Bacterium classification)
[3*]ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Leuconostoc_citreum_KM20_uid58481/ (NCBI's data)
[4] - https://en.wikipedia.org/wiki/Alternansucrase (Alternansucrase description)
[5] - http://www.rcsb.org/pdb/protein/Q9RSH2 (50S ribosomal protein L34 description)