Journal of Bioinformatics and Computational Biology
♥ Imperial College Press

ROSEATELES DEPOLYMERANS KCTC 42856 GENOME AND PROTEOME OVERVIEW

Anna Rozina

Faculty of bioengineering and bioinformatics, Lomonosov Moscow State University, Leninskiye gory 1-73
Moscow, Russia
anarou@inbox.ru

In this study I analyzed genome of bacteria Roseateles depolymerans KCTC 42856. The goal of the study was to structuralize information from the full genome sequence to provide useful data for future research and for practical use. It was found that protein length distribution deviates from normal in extreme short and extreme long values. Distribution of genes by strands appeared to be close to the expected, and quantities of genes coding different products corresponded naturally.

Keywords: genome; sequence; roseateles

1. Introduction

Roseateles depolymerans is obligate-aerobic b-proteobacteria originally isolated from river water in Japan [1]. This species is of much interest because one of its stains is capable of degrading poly(hexamethylene carbonate) and some other plastics [2]. However, *Roseateles depolymerans* has been under research only in context of photosynthesis regulation [3].

Length of its genome is 5681722 bp, its genome contains approximately 855,37 genes per mln bp (g.v. supplementary materials). The goal of this study was to partly structuralize full genome sequence information to obtain some traits of *Roseateles depolymerans* genome and proteome that might be useful for future laboratory study and in practical use, in particular to:

(1) Estimate distribution of different gene groups by DNA strands
(2) Compare quantities of genes coding different products
(3) Analyze protein length distribution

2. Methods

Strain KCTC 42856 full genome sequence is open at NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_001483865.1/). All the data was taken from full genome refseq file. Excel was used as a basal software. Specific methods that were used to analyze genome and proteome will be observed in the following sub-paragraphs.

2.1. *Genome*

To calculate number of protein coding, RNA coding genes and pseudogenes on both strands I juxtaposed information from "strand" and "class" columns of original file. Only those strings were under consideration that contained the word "gene" in "#feature" column.

I counted gene as:

(1) protein coding if it had "protein_coding" in "class" column;
(2) pseudogene if it contained "pseudogene";
(3) RNA coding if there were "ncRNA", "tRNA", "tmRNA", "SRP_RNA", "rRNA", "RNase_P_RNA" in "class" column.

The results are presented in table 1.

The same method was used to identify number of DNA sequences coding different types of products (table 2). If string contained "cds" in column "#feature" CDS was counted as coding:

(1) transport protein if it contained "transporter" or "transport protein" and did not contain "binding" in column "name"
(2) ribosomal protein if it contained "ribosomal" and did not contain "transferase" in column "name"
(3) hypothetical protein if it contained "hypothetical" in column "name" and "with_protein" in column "class"
(4) other protein if it did not match any listed requirements (quantity = all cds – (1) – (2) – (3))

For RNA coding genes the same combination of indexes in columns as for table 1 was used: "gene" in column "#feature" and a type of RNA accordingly in "class" column

2.2. *Proteome*

To build protein length histogram I used values from column "product_length" in original refseq file. The histogram was built in Excel. Protein length statistical marks (minimal, maximal length etc.), that are presented in table 3, were identified using Excel either. Q-Q chart was made using in-built facilities of IBM SPSS Statistics.

3. Results and discussion

3.1. *Genome overview*

3.1.1. *Distribution of different gene groups by DNA strands*

One of the goals of the study was to analyze distribution of genes by strands. Table below (table 1) shows quantities of genes on DNA strands.

Table 1. Gene distribution

	gene class		
	protein coding	RNA coding	pseudogenes
+ strand	2276 (48.3%)	45 (60.8%)	30 (42.25%)
- strand	2439(51.7%)	29 (39.2%)	41 (57.75%)

The percentage of the total number of genes in the gene class is indicated in parentheses. It is seen that there is no extreme deviation from accidental (50%) distribution in protein coding genes group. In other grops deviation is significant, but they contain less than a hundred samples, so the estimation of distribution randomness is not justified.
It is also clear from this table that protein coding genes group is much bigger than RNA coding. It is natural as proteins are more varied in cell. Quantity of genes coding different RNA types is indicated in table 2.

3.1.2. *Comparison of quantities of genes coding different products*

Table 2 shows quantities of genes coding coding different classes of products.

Table 2. Number sequences coding different products

	product	count
protein coding DNA sequences	ribosomal protein	55
	transport protein	232
	hypothetical protein	1342
	other	3157
RNA coding genes	tRNA	58
	rRNA	12
	others	4

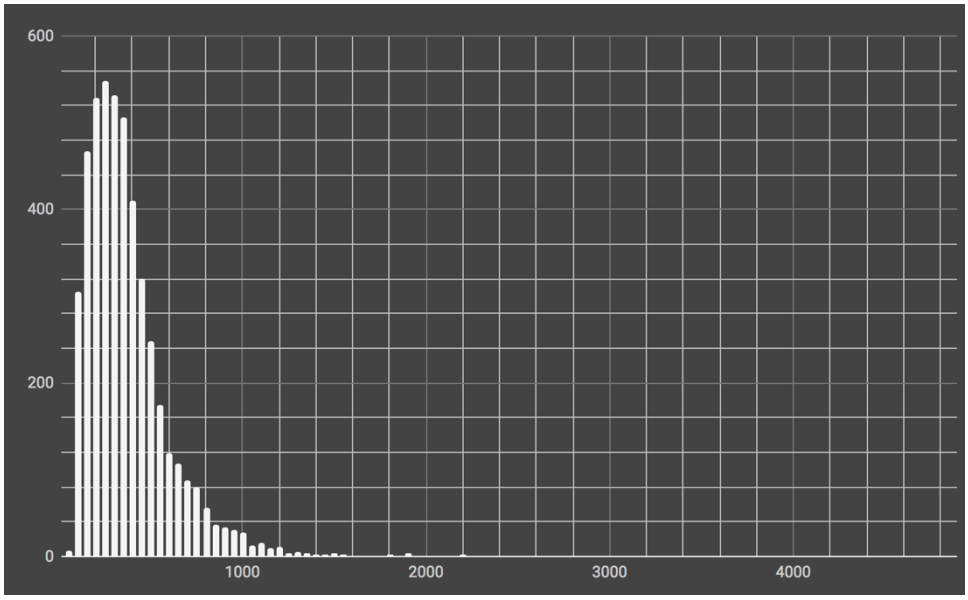
Here small number of not transport or ribosomal RNA coding genes is not surprising as they are normally not varied in cell. tRNA database (lowelab.ucsc.edu/GtRNAdb/) indicates that such a quantity of tRNA coding genes is natural for betaproteobacteria.

3.2. *Proteome overview*

3.2.1. *Protein length distribution analyze*

The chart below (fig.1) illustrates protein length distribution. Each column corresponds to the number of proteins of relevant length (each gap is of 50 a.a). It is clear from the histogram that there is one peak of the most common protein length (200-250 a.a), and the frequency decreases gradually. XLSX file with histogram is available in supplementary materials.

Fig.1 Protein length histogram



It is worth to estimate the normality of the distribution more thoroughly. Q-Q chart below (fig.2) visualizes deviation of protein length distribution from the normal distribution. Here straight line shows expected (normal) allocation while complex of points indicates observed one. It is seen that protein length distribution is close to normal with an exception of unexpected number of extremely short and extremely long proteins (minimal and maximal lengths are listed in table 3). It can be logically explained: very short and very long proteins are needed in cell only for some specific purposes, so they don't match normal distribution of proteins with more optimal length. In other words, optimality of protein length doesn't decrease normally in extreme (very little or very big) values

Fig.2 Q-Q chart

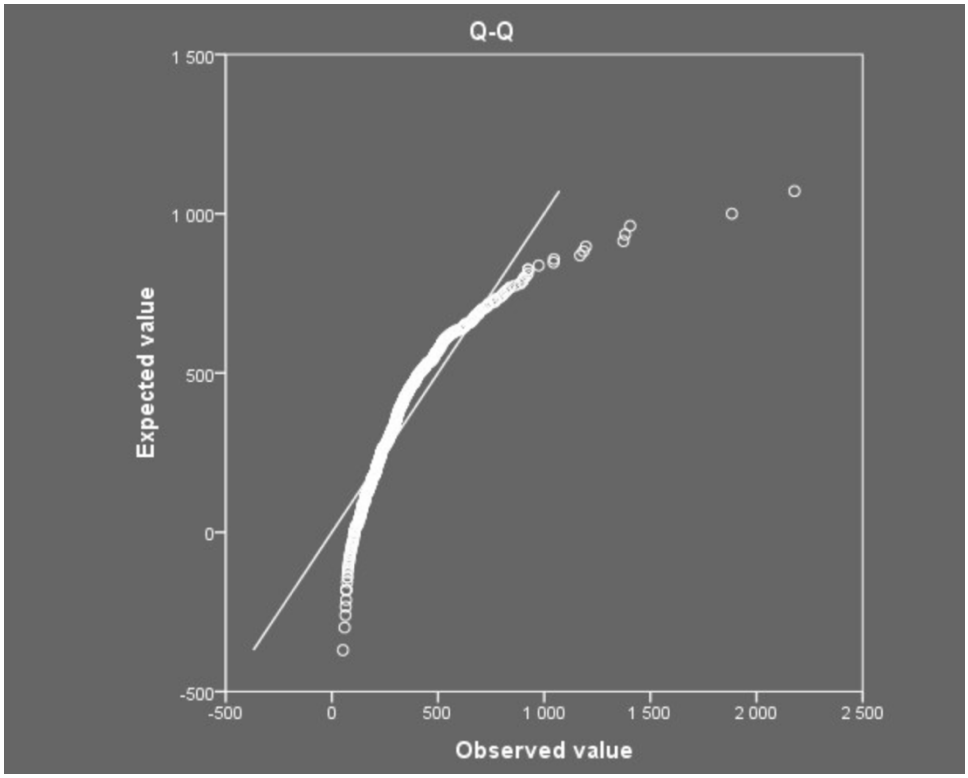


Table 3 provides some additional characteristics of proteome.

Table 3. Protein length characteristics

minimal length	29 a.a
maximal length	4828 a.a
average length	348,82 a.a
median	297
standard deviation	258,59

4. Conclusion

As a result of the research it was found that protein length distribution is close to normal for medium values and deviates from normal the stronger the further from medium values. Distribution of genes by strands appeared to be close to expected for big group of protein coding genes (48-52% with expected 50). Deviation was bigger in smaller groups of

pseudogenes and RNA coding genes (39-61%). Quantities of genes coding different types of products corresponded naturally.

5. Acknowledgments

I would like to thank Grigory Novikov, HSE student for teaching me how to use IBM SPSS Statistics

6. Supplementary materials

All the excel data including figures is available here:
http://kodomo.fbb.msu.ru/~anrozina/term1/supplementary.xlsx

References

1. Suyama, T., T. Shigematsu, S. Takaichi, Y. Nodasaka, S. Fujikawa, H. Hosoya, Y. Tokiwa, T. Kanagawa, and S. Hanada, *Roseateles depolymerans* gen. nov., sp. nov., a new bacteriochlorophyll *a*-containing obligate aerobe belonging to the β-subclass of the *Proteobacteria*, *Int. J. Syst. Bacteriol*. **49**:449-457, 1999
2. Yutaka Tokiwa, Buenaventurada P. Calabia, Charles U. Ugwu, Seiichi Aiba, Biodegradability of Plastics *Int J Mol Sci*., pp. 3722–3742, 2009 Sep; 10(9):
3. Suyama, T., T. Shigematsu, S. Takaichi, Y. Nodasaka, S. Fujikawa, H. Hosoya, Y. Tokiwa, T. Kanagawa, and S. Hanada, *Roseateles depolymerans* gen. nov., sp. nov., a new bacteriochlorophyll *a*-containing obligate aerobe belonging to the beta-subclass of the Proteobacteria, *Int. J. Syst. Bacteriol*. 49**:**449-457, 1999