

## Analysis of genome size, CG content and characteristics of microorganisms' environment habitats

V.V. Suslov<sup>1</sup>, A.V. Tsukanov<sup>1,2</sup>, Y.L. Orlov<sup>2,3\*</sup>

<sup>1</sup>Novosibirsk State University, Novosibirsk, Russia

<sup>2</sup>Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

<sup>3</sup>A.O. Kovalevsky Institute of Marine Biology Research RAS, Sevastopol, Russia

\* e-mail: orlov@bionet.nsc.ru

**Key words:** genomics, bacteria, microbiology, sequencing, environments, GC content, bioinformatics

*Motivation and Aim:* We aimed to analyze microorganisms habitats and environment dependence by genome size and context characteristics [1]. Research on evolution of early unicellular organisms relies on complete genome sequencing data. Rapid growth of data banks allows us reexamine sequence features necessary for minimal genome size and minimal gene set as unit of evolution. Several theoretical and experimental studies have endeavored to derive the minimal set of genes that are necessary and sufficient to sustain a functioning cell (unicellular organism). Minimal genome size rise questions about relation of environment and genome nucleotide characteristics.

*Methods and Algorithms:* We have downloaded from NCBI ftp site thousands of complete and assembled whole genome sequences of bacteria and achaea. We used original software for GC content estimation [2] comparing genome size with gene number and GC content separately by groups of organisms.

*Results:* For whole genome assemblies we found high linear correlation between genome size and gene number (0.8), that been expected, and, more interestingly, between genome size and GC content (0.46). Thus, larger genome size is strongly related to higher fraction of G and C nucleotides. We have considered grouping of species by habitat and environment and found same common trend of correlation between genome size and GC content. But we found bias for some groups. Fusobacteria and planctomycetes have lower GC content that might be expected. *Deinococcus-thermus* have higher GC content.

*Conclusion:* Overall correlation of GC content to genome size follows the same trend in archaeal and bacterial groups. Comparing oxygen requirement, genome size and GC content we found that aerobic organism have larger average genome size and corresponding GC content. It is interesting to note, that smaller genome size is associated with specialized habitat, hyperthermophilic temperature and microaerophilic oxygen requirement. In all cases it is related to lower GC content. Thus, genome content has restriction for organism adaptation.

*Acknowledgements:* The research has been supported by the Ministry of Education and Science of the Russian Federation grant No. 14.W03.31.0015. Computing was done at Siberian Supercomputer center SB RAS.

### References

1. Suslov V.V., Afonnikov D.A., Podkolodny N.L., Orlov Y.L. (2013) Genome features and GC content in prokaryotic genomes in connection with environmental evolution. *Paleontological Journal*. 47(9):1056-1060.
2. Babenko V.N., Bogomolov A.G., Babenko R.O., Galieva E.R., Orlov Y.L. (2018) CpG islands' clustering uncovers early development genes in the human genome. *Computer Science and Information Systems*. 15(2):473-485.