DOI 10.18699/BioGenEvo-2018-17

Evolutionary study of low complexity glycine-arginine rich domains

A. Kotyurgin[1]*, A. Alexeevski[1, 2]
[1] *Lomonosov Moscow State University, A.N. Belozersky Institute of Physico-Chemical Biology and Faculty of Bioengineering and Bioinformatics, Moscow, Russia*
[2] *Scientific Research Institute for System Analysis RAS, Moscow, Russia*
* *e-mail: alekoksan@gmail.com*

Key words: low-complexity regions, glycine-arginine rich domains, neutral evolution

Motivation and Aim: Low complexity glycine-arginine rich (GAR, also known as RGG) domains are abundantly presented in many RNA-binding proteins [1]. The study of evolution of GAR domains is complicated by the impossibility to use standard methods of phylogeny due to low complexity corruption of alignment. Previously, it was discovered that GAR domains of the mammalian fibrillarin have some specific features that make it possible to classify GAR domains by taxa at the order rank. The aim of this work is to study of the diversity, occurrence and evolution of GAR domains.

Methods and Algorithms: We created python script to find all GAR-containing proteins in the Uniref100 database (126152540 sequences) and describe their functions using such databases as Gene Ontology, Pfam and eggNOG. While standard methods of phylogeny do not work, we used taxonomic cladograms with an indicated proportion of GAR-containing proteins in different taxa to study the evolution of GAR domains. Such characteristics as frequencies of amino acids and specific number of glycine repeats per length were counted.

Results: 141608 GAR-containing proteins were found in Uniref100 database. Annotation on GO, Pfam and eggNOG databases confirmed that the most common activities of GAR-containing proteins are RNA binding, helicase activity and other functions associated with nucleic acid. Cladograms analysis of the most represented COGs showed that GAR domains are not ubiquitous in COGs and could emerge independently during evolution. Frequency distribution of amino acids seems to be result of the interaction of two modes of evolution, such as neutral evolution determined by the most probable path of non-synonymous mutations and directional selection according to physicochemical properties of residues (hydrophobicity, affinity to nucleic acids). The first type of evolution is assumed in the cases of alanine and serine, the second in the case of phenylalanine, proline, tyrosine and non-alanine hydrophobic residues.

Conclusion: A functional association of GAR-containing proteins with the metabolism of nucleic acid was shown on a much larger sample than in any previous study. Multiple independent emergency of GAR domains during evolution seems to be the most typical scenario. Amino acid composition of GAR is presumably determined by both neutral evolution and directional selection. We intend to complete sequence based GAR classification compatible with their evolution.

Acknowledgements: the work was partially supported by RSF grant 16-14-10319.

References

1. Thandapani P. et al. (2013) Defining the RGG/RG motif. Molecular Cell. 50(5):613-23.