Prediction and analysis of separated restriction-modification systems in prokaryotic genomes.

A.S. Ershova[1,2,3], A.N. Lyaschuk[2], V.G. Lunin[2,3], A.S. Karyagina[1,2,3], S.A. Spirin [1,4,5], A.V. Alexeevski [1,4,5]

[1]Moscow State University, Belozersky Institute of Physical and Chemical Biology, Moscow, 119992, Russia; [2]Gamaleya Institute of Epidemiology and Microbiology, Gamaleya st. 18, Moscow 123098, Russia, [3]Institute of Agricultural Biotechnology, Academy of Agricultural Sciences, Moscow, 127550, [4]Scientific Research Institute for System Studies (NIISI RAS), Nakhimovsky Prosp., 36, 1, Moscow, 117281, [5]Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow, 119992, Russia

asershova@gmail.com

Prokaryotic restriction-modification (R-M) systems defend host cell from invasion of a foreign DNA. They comprise two enzymatic activities, specific DNA cleavage activity and DNA methylation activity preventing the cleavage. Typically, these activities are provided by two separated enzymes, a DNA methyltransferase and a restriction endonuclease. Usually, genes of the R-M system are linked in the genome [1]. There are only a few reported cases for Type I R-M systems, in which genes of DNA methyltransferase and restriction endonuclease from one R-M system are not linked [2]. However, in Restriction Enzyme database (REBASE, http://rebase.neb.com) we found 272 solitary restriction endonuclease genes, i.e. genes of putative restriction endonucleases without nearly located genes of DNA methyltransferase. For each of them, all MTase-like ORFs in the vicinity of the gene were found using TBLASTN [12], with E-value threshold 0.01 and program *getorf* from the EMBOSS package. No new ORFs, except for those annotated in REBASE, were identified. We tried to exclude from consideration the most obvious pseudogenes or gene fragments. RE was considered presumably functional if its gene was not annotated as a "pseudogene" in genome entry, and any of the following conditions were true: (i) activity of the RE was shown experimentally or (ii) there is a homologous RE (with identity >20%) with an experimentally shown activity, and the length of a homologue's sequence is approximately the same as the length of the solitary RE (difference in lengths is less than 20%) or (iii) there is a homologous RE within an annotated in REBASE complete R-M system, and the length of a homologue's sequence is approximately the same as the length of the solitary RE (the same thresholds as in the previous condition).

We found that 99 of 272 solitary RE genes are likely to be fragments of RE genes and we excluded them from analysis.

For each solitary RE we found all orthologs of the RE in all bacterial genomes and selected genomes containing the found orthologs. Two REs from different genomes were considered

orthologous if their amino acid (aa) sequences had >40% identity over 80% of length of the longer sequence. For MTases, the thresholds of >50% identity over 80% of length were used due to typically higher sequence similarity of MTases. All MTases from the selected genomes were subdivided into orthologous groups. Each orthologous group with more than one member determines a group of RE/MTase orthologous pairs in corresponding genomes, each pair including the ortholog of the initial solitary RE or the initial solitary RE itself. If more than one group of RE/MTase orthologous pairs containing the initial solitary RE were detected, then we selected the group with the greatest number of pairs as a group of putative orthologous R-M systems.

For 57 from rest 173 solitary RE genes (group 1) we predicted corresponding MTase genes located distantly in a genome. For the remaining 116 (group 2) putative solitary restriction endonuclease genes we did not find any well-defined explanations [3].

We started work on cloning and studying of the activity of putative restriction endonucleases and DNA methyltransferases genes from predicted separated R-M systems.

The separated R-M systems, in case of experimental evidence of their functionality, represent the new form of R-M systems existence in prokaryotic genomes.

The work was partially supported by Russian Foundation of Basic Research grants Nos.11-04-91340 and 12-04-31594.

1. Wilson, G.G. (1991). *Nucleic Acids Res*. **19**, P. 2539–2566.

2. Waldron, D.E., Lindsay, J.A. (2006). *J. Bacteriol*. **188** (15), P. 5578-85.

3. Ershova A.S., Karyagina A.S., Vasiliev M.O., Lyashchuk A.M., Lunin V.G., Spirin S.A., Alexeevski A.V. (2012). *Nucleic Acids Res*. **40** (20). P. 10107-10115.