Occurrence of Restriction-Modification systems' recognition sites in genomes of bacteria, archaea and their viruses

I.S. Rusinov[1], A.S. Ershova[2, 3, 4], A.S. Karyagina[2, 3, 4], S.A. Spirin[1, 3, 5], A.V. Alexeevski[1, 3, 5]

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

aba@belozersky.msu.ru

Restriction-Modification systems (R-M systems) consist of two components as a rule. One component (restriction endonuclease) cleaves DNA near or within unmethylated recognition site, another component (DNA methyltransferase) methylates certain nucleotide base within the same site to prevent DNA cleavage. These systems in prokaryotes are thought to defend their hosts from invasion of foreign DNA. Sometimes R-M systems are also considered as selfish genetic elements. Prokaryotes and their viruses tend to avoid R-M recognition sites in their genomes [1].

Goal of this work is to study R-M system sites avoidance in all available genomes of prokaryotes, bacteriophages and archaeal viruses.

We analyzed 1556 prokaryotic genomes, 1349 bacteriophages genomes and 60 archaeal viruses genomes available from EBI on 18.09.2012. List of 4128 R-M systems encoded in these genomes was obtained from REBASE version 209 (www.rebase.neb.com). We used 3617 genomes of eukaryotic viruses as a control set because these viruses do not meet R-M systems in their life.

To estimate expected number of R-M recognition sites in genomes we used the method suggested in the work of Karlin et. al [2]. For each pair (recognition site, genome) we calculated Kr as ratio of observed sites number to expected sites number. Following [2], we supposed that recognition site is underrepresented if Kr is less than 0.78.

Examining all 4128 sites, we confirmed R-M site avoidance in genomes of prokaryotes by comparison with the occurrences in eukaryotic virus genomes. Next, for each genome we selected 'actual' recognition sites, i.e. proven sites of R-M systems encoded in this genome. We found that about 50 % of actual sites are underrepresented in prokaryotic genome, which seems to be the first statistical estimation of actual R-M site avoidance. However, the number of underrepresented sites in genomes is higher than the number of all R-M systems encoded in them: on average there are about 9 underrepresented recognition sites and only 3 R-M systems.

We showed that at least partially, this difference can be explained by the footprint of lost R-M systems. Indeed, recognition sites of R-M systems encoded in close relatives, but not in the analyzed genome, are underrepresented more often than recognition sites of R-M systems from more distant species.

The same trend was observed for bacteriophages and archaeal viruses, which avoid about 50% of recognition sites of host R-M systems. At the same time, the number of underrepresented sites in bacteriophages and archaeal viruses genomes on average is greater (23), than in prokaryotic genomes (9). It could be explained by wide spectrum of bacteriophage hosts. Additionally, we found that about 25 % of all underrepresented sites are completely absent in bacteriophage genomes (Kr=0, whereas the expected number of sites is greater than 5). Seems, in these cases exclusion of specific sites is effective bacteriophage strategy against host R-M systems.

To explain the results, we speculate that in process of phage – bacteria – R-M-system coevolution bacterial and phage genomes adapt to R-M system. As a result, this system becomes less effective in preventing common phage invasions and thus, can be eliminated from the genome. Therefore, underrepresented recognition site can be a footprint of such lost R-M systems. On the other hand, recognition sites of recently acquired R-M systems are not necessarily underrepresented in genomes.

The work was partially supported by Russian Foundation of Basic Research grants no. 11-04-91340.

1. Tock M.R., Dryden D.T.F (2005) The biology of restriction and anti-restriction. *Current Opinion in Microbiology*, 8:466–472.

2. Karlin S., Cardon L.R. (1994) Computational DNA sequence analysis. *Annu. Rev. Microbiol.* 48:619–654.