Estimation of selection pressure on degenerate sequences in genomes: choice of method

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

[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, Moscow 123098, Russia; [4]Institute of Agricultural Biotechnology, the Russian Academy of Sciences, Moscow 127550, Russia; [5]Scientific Research Institute for System Studies, the Russian Academy of Science (NIISI RAS), Moscow 117281, Russia.

isrusinov@gmail.com

Several methods are widely used for detection of short sequences (words) that are under evolutionary pressure (exceptional words) in genomes. Two most common methods are S. Karlin's method [1] and a method, based on maximal order Markov model (Mmax) (see S. Schbath [2]). The former takes into account observed frequencies of all subwords of the word, including discontinuous, for expected frequency estimation while the latter considers only subwords obtained by deleting one letter from either 5', 3' or both word ends.

We compared these two methods in terms of detection of restriction sites avoided in a prokaryotic genome. Recognition sites of restriction-modification systems were chosen for the methods comparison as target short words because of high specificity of restriction-modification systems. Also, avoidance of many restriction sites in prokaryotic genomes was previously shown [3]. We used 2141 complete prokaryotic genomes for the comparison.

We estimated the bias in sites representation with the observed to expected number ratio, called "contrast". If the ratio is less 1 then there are less sites in a genome sequence than expected, and vice versa. Mmax based method and Karlin's one were used for computation of expected site number. A site was considered to be underrepresented if Karlin's contrast value was less 0.78 (was taken from original Karlin's work [1]). The analogous cutoff for Mmax based method was selected 0.6, because of the same percent of underrepresented sites between two methods on the same data set. We found that two lists of underrepresented sites differ by 40%. Thus, the method used has significant impact on the obtained results.

Karlin's method takes into account observed frequencies of discontinuous subwords of the word unlike Mmax based one. Therefore, we suggest the greatest difference between the methods to be in case of discontinuous word under selection. Thus, among all restriction sites only degenerate ones, like CCNNGG where N is any nucleotide, were selected for a comparison of methods precision.

If a degenerate restriction site (like CCNNGG) is under pressure to reduce number of its occurrences, then we expect to observe this site more avoided then its non-degenerate variants (CCATGG, CCTGGG, and so on), especially non-palindromic ones. Hence, the method seems to be better if it matches this rule more often. We found that Karlin's contrast showed better results than Mmax based one: 87% vs. 38% of cases matching the assumption (only non-palindromic variants were taken into account).

We known at least one experimentally confirmed case approving this result. Gelfand and Koonin [3] predicted specificity of MjaIV restriction-modification system as GTYRAC based on site underrepresentation in the genome of *Methanocaldococcus jannaschii*. Mmax based method was used through Z-score calculation. MjaIV system was experimentally characterized [4] and its specificity in REBASE is GTNNAC. Notably, Karlin's contrast reveals underrepresentation of GTNNAC but not GTYRAC and non-degenerate variants of GTNNAC.

In addition, we evaluated the methods dependence on subword frequency biases with simulated selection against certain word in 100000 bp random Bernoulli sequence and tracing effects on representation of other short words in the sequence with both methods. In case of degenerative word under selection results are just the same that were observed for the degenerative restriction sites in the real genome sequences. We also found, that the Mmax based ratio almost always diverges from 1 worse than the Karlin's one. The most part of short words in random Bernoulli sequence should have contrast values close to 1. Thus, the greater dispersion in case of Mmax based method seems to be a result of greater dependence of the method on subword frequency biases. This effect could be explained with much more information about sequence use in case of Karlin's method.

In summary Karlin's method is more reliable for detection of exceptional words in genome sequences, probably due to use of all site subwords frequencies for the representation evaluation.

The work is partially supported by RFBR grant 14-04-91350.

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

2. S.Schbath, B.Prum, E.de Turckheim. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences, *J Comput Biol,* **2(3):**417–437.

3. M.Gelfand, E.Koonin. (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes, *Nucleic Acids Res,* **25(12):**2430–2439.

4. Y.Zheng et al. (2009) Using shotgun sequence data to find active restriction enzyme genes, *Nucleic Acids Res,* **37(1):**e1.