DOI 10.18699/MM-HPC-BBB-2018-16

The performance improvement of the permutation test algorithm for GSEA

M. Grishchenko[1]*, A. Yakimenko[1, 2], M. Khairetdinov[1, 2], A. Lazareva[2]
[1] *Institute Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State Technical University, Novosibirsk, Russia*
* *e-mail: mikhail.grishch@gmail.com*

Key words: resampling, randomization, permutation test, GSEA

Motivation and Aim: Processing of genetic data for the analysis genetic determination of traits is very important problem for modern biology. Resampling method are widely used to solve this problem. Resampling methods combine three different approaches: permutation test, “jack-knife” method and bootstrap [1]. In this work, permutation test method is considered. The basic idea of this method is to randomly permute rows or columns of observed values table [2]. It is important that the size of the table and the number of samples do not change during permutations. It allows analyzing multiple hypotheses simultaneously without correction of the statistical significance level. However, permutation test method requires much computational resources. The aim of this paper is to determine a minimal number of iterations of the permutation test algorithm to calculate steady p-value depending on the input data.

Methods and Algorithms: Permutation test algorithm allows us to calculate the p-value simultaneously for all characteristics of the gene sequence. The process of computing p-value is an iterative, in which the values of the computed statistics gradually converge to the stable value of the neighborhood of a certain value p*. The average number of iterations was estimated to achieve a stable p-value, with a given confidence interval. It was shown that the average number of iterations is 27500–28500 iterations and in most cases, it does not depend on the amount of input data. It could be used this number of iterations. However, this approach has two drawbacks: 1) not all p-values achieved their stable values; 2) are cases when this number of iterations is not enough. Another approach is to use the maximum number of iterations, when all p-values reach their stable values.

Results: We investigated the permutation test algorithm aimed at finding statistically significant overrepresented gene characteristics under different external and/or internal conditions. It was obtained that the necessary number of iterations does not depend on the number of genes in the input data, but depends on the number of properties of the genes. In addition, we replace algorithm of random permutations to Fisher-Yates shuffle algorithm [3].

Acknowledgements: This work was supported by the Russian foundation for basic research (Grant No. 16-37-00240)

References

1. Efron B. (1988) Nontraditional methods of statistical analysis. Moscow: Finansy i statistika. 263 p.
2. Yakimenko A.A., Gunbin K.V., Khairetdinov M.S. (2014) Search for the Overrepresented Gene Characteristics: The Experience of Implementation of Permutation Tests Using GPU. Optoelectronics, Instrumentation and Data Processing. 50(1):123-129.
3. Knuth D.E. (1969) Seminumerical algorithms. The Art of Computer Programming. 2. Reading, MA: Addison–Wesley. pp. 139-140. OCLC 85975465.