doi 10.18699/bgrs2024-1.4-22

DEAr – tool for differential expression gene analysis

Karpenko D.*
National Medical Research Center for Hematology, Moscow, Russia
** d_@list.ru*

Key words: DEAr; differential expression; housekeeping genes; statweight; recursive calculations; ranged values; missing values; normalization

Motivation and Aim: Differential gene expression is a standard way to study and to describe cells and tissues. Directed studies or clinical tests often use real-time PCR with limited number of measured genes. For individual samples, we can't be sure about quantity and quality of analyzed RNA. An important stage in the proceeding is a normalization. Normalization methods can be generalized to the idea of establishing a baseline from a subset of stable genes, and then representing gene expressions as differences from this baseline. If we compare different cells under different circumstances, some genes are considered to establish a baseline could be unstable in particular samples or conditions. In case of collecting clinical data for years or from multiple centers, datasets could include different subsets for normalization, forming that way a mosaic data. This leads to natural difficulties in data processing. There are no automated solutions or the strict protocol, especially for inconsistent real-time PCR data. Therefore, I decided to create an automated differential expression calculation tool that could handle mosaic data and be resistant to individual outliers.
Methods and Algorithms: Instead of selecting preferable for normalization genes, the original algorithm identifies common expression pattern for genes in analysis. Pattern identification is based on expanded data of pairwise comparisons between genes of a sample. It negates system shift in data from RNA quantity, but increases data complexity. Through several cycles of adjusting and convolution, cohesive data for the whole dataset is formed. In each cycle, the algorithm identifies how to adjust each expression value in samples to reach a common pattern. As the algorithm changes all values simultaneously, recursive calculations are required to reach a stable result. Due to multiple outliers in data, high robustness is required from the algorithm. I used statweight to make the necessary distinction between noise and reproducible patterns. A statweight is a measure of how reproducible a particular element of the data, and is interpreted as an effective number of measurements. Applied statweights are formed as an exponential penalty for deviations from the common pattern. An additional penalty is applied according to the inaccuracy of a measurement, implemented in such a way that each value is considered as a range. This allows both accurate and inaccurate values to be used in the same calculation, while distinguishing their influence. While the statistical weights are based on the reproducibility of the data, I have also applied predictive weights based on external knowledge of the stability and instability of genes. This allows the algorithm to operate on a combination of information from external and internal data. This means that the algorithm can work on unknown gene profiles, but additional information can be used to improve performance. Alike outliers and uncertain values, missing values don't contribute to a common pattern and that way provide no obstacles for analysis. This follow general theme of making tools for problematic data with ranged

and missing values [1]. In summary, the DEAr (Differential Expression Analyzer) tool uses a highly recursive algorithm developed in Python v.3.7, packaged in an executable file. I presented algorithm on previous BGRS2022 [2]. The extended algorithm presentation is available in a preprint [3].

Results: I present ready to use software available within the preprint version of the study [3]. Analysis of the real data from another study [4] showed that genes considered as stable were unstable for certain tissues. On the data from the experiment with mosaic deletion, DEAr performed calculations as for initial data. Standard protocols don't provide a ready to use solution and require manual data preparation to compensate for missing data by inputting or discarding of whole gene for all samples. In the testing data, the single value for the gene of interest was manually shifted as it was contaminated with PCR product. DEAr recognized it as an outlier from three technical replicates due to its highly robust core algorithms using statweight, while standard geometric mean demonstrated shift in the resulting evaluation. DEAr allows aggregation of data or knowledge not only from same experiments but also from external sources, evaluated as how we are certain for a value in sample or for a gene expression stability in general. The program can be used as a single executable file that gets and returns data in Excel format. No additional programming skills are required.

Conclusion: I present the original tool DEAr for differential expression analysis and the original implementation of the robust computations. The tool utilizes information of different origin to provide a conclusion. The applied approach uses information of each measurement accuracy. DEAr works in the Microsoft Windows OS using data in Excel format. As no additional programming skills are required, DEAr will be a handy tool for classical biologists or medical scientists. The program is available to any independent researcher who wishes to use it for non-commercial purposes, provided that the name of the program and the article [3] are correctly cited. Use by non-academics requires a license.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Karpenko D.V., Bigildeev A.E. Small groups in multidimensional feature space: two examples of supervised two-group classification from biomedicine. *J Bioinform Comput Biol*. 2023;21(6):2350025. doi 10.1142/S0219720023500257
2. Karpenko D.V. Recursive matrix algorithm for calculating differential expressions. In: BGRS/SB-2022. Novosibirsk, 2022;1101-1102. doi 10.18699/SBB-2022-660
3. Karpenko D.V. DEAr – Differential expression analyzer. 2023. doi 10.21203/RS.3.RS-2957165/V3
4. Karpenko D., Dorofeeva A., Petinati N., Shipounova I., Drize N., Bigildeev A. Functional characteristics of the mouse Il1b promoter in various tissues before and after irradiation. *DNA Cell Biol*. 2020;39:790-800. doi 10.1089/DNA.2019.5310