

Recursive matrix algorithm for calculating differential expressions

Karpenko D.V.

National Research Center for Hematology, Moscow, Russia

d_@list.ru

Key words: differential expression, statweight

Motivation: Differential gene expression calculations involve comparing the expression of a gene of interest and one or more housekeeping genes whose expression level is considered stable. Stability is defined as the maintenance of the pattern of gene expression relative to each other in the studied samples or groups of samples. Difficulties arise when it is necessary to compare expression in tissue cells with significantly different physiologies, since the expression of some genes including housekeeping genes can alter prominently among tissues. There will be a separate set of stable genes for separate sets of tissues. This gives rise to mosaicism in multiple comparisons; a similar mosaicism appears when comparing results from different experiments, when the set of studied genes is different. This creates natural difficulties in data processing, which are often solved individually in each analysis.

Aims: Create a unified, automated differential expression calculation tool capable of processing mosaic data and resistant to single outliers.

Methods and Algorithms: Expression data are calculated in logarithmic scale. First of all, the relative pairwise expression of all genes against all for each sample is calculated. For each pair separately, the mean and scatter are calculated between samples. Exponentially decaying statistical weight (statweight) is assigned to the values of the pair in the sample in accordance with the difference from the pair mean and the scatter for the pair. Taking into account the calculated statistical weight, we recursively recalculate the mean and the spread until the result stabilizes between iterations. A similar procedure is performed for the differences of a sample pair from the average pair across samples, but now the enumeration occurs for the sample gene over all its pairs – in fact, how stable is the deviation of the gene relative to the general pattern for each paired gene. The statweight is also used, which now takes into account the total instabilities of the gene relative to pairs in the sample, the instabilities of these pairs relative to each other in the samples, and the instability of the sample itself. For a gene, the differences of each pair from its average for samples are summed up, taking into account the statweight of this pair, so we get a matrix of corrections for gene expression by samples. We correct the initial differential expressions with the obtained corrections and recalculate recursively until stabilization in form of the corrections becomes less than a predetermined value. We sum up the final corrections accumulated in iterations, and this is the resulting differential expression data.

Results: An algorithm was developed and tested in Python v.3.7 for Microsoft Windows. The program allows processing incoming data as separate blocks in Excel list, scripts perform the primary sorting of incoming data. The output is in the form of an Excel table with matrices of values of relative expressions and estimated errors. The program allows assigning a deliberate statweight to the elements under consideration: a low statweight

can be set for obviously unstable objects or for a whole group of related genes that are stable relative to each other, but not relative to a reference gene. This allows introducing known assumptions about biological objects into calculations.

Conclusion: The program allows accelerating and automating the calculations of differential gene expression. Moreover, the algorithm as well as the program can also be applied in other areas requiring differential analysis and assuming the presence of characteristic profiles of the ratio of objects, such as, for example, in tasks on microbiota populations. The introduced ability to determine the preset statweight enables preferences in the calculations according to third-party information about the object. The use of the statweight makes the algorithm resistant to outliers and dynamically handles them. The algorithm assumes the presence of a distribution similar to a normal one and cases of multimodal distributions will probably be handled incorrectly.