massiveGenesetsTest: a web tool to run enrichment analysis.

Luigi Cerulo
Stefano Maria Pagnotta

Universitá degli Studi del Sannio,
Department of Science and Technology,
82100, Benevento, Italy

February 15, 2021

Abstract

Motivation: Inferring biological phenotypes from genomic data and sample clusters is a routinely task usually performed with Gene-Set Enrichment Analysis (GSEA), a tool that queries gene-profiles. In previous work, we scrutinized the approach based on Mann-Witney for Gene-Sets test. We highlighted the Mann-Witney test-statistics sensitivity to uncover weak signals and the drastic decreasing of time complexity.

Results: We propose web implementation of the Gene-sets testing based on the Mann-Witney procedure. The test-procedure has reshaped to decrease the computational expense, now about tens of seconds, even if a large collection of gene-sets queries the same gene-profile. The probabilistic interpretation of the normalized test-statistic has been investigated to a better understanding of the enrichment results. A novel prioritization method across enrichment-scores, gene-set dimensions, and p-values, draws attention to relevant gene-sets. The web tool provides both tabular and graphical enrichment results. A complimentary R function allows integrating the enrichment procedure in a complex context.

Contact: pagnotta (at) unisannio.it

Supplementary information: Example data and guidelines are included in the supporting material of the web-site.

Keywords: GSEA, MWW-GST, gene-sets enticements, Mann-Witney test, prioritization.

Availability: massiveGenesetsTest is freely available at http://www.massivegenesetstest.org/

1 Introduction

Gene-sets enrichment analysis is a routinely used technique to uncover phenotypes of a gene profile usually associated with the differential expression between two conditions [7]. The inputs of the enrichment analysis are a gene-profile and a gene-set. The gene-profile is a ranked list of genes, usually correlated to the differential level of expression between two sets of samples, such as treatment and control. A gene-set is a collection of genes cooperating to a specific phenotype derived from annotated databases, such as Gene Ontology [1]. From the work of [12], the statistical framework of enrichment analysis is just a significance test where a p-value is assigned to a set of genes considered as a unit: if the p-value is below the significance level of the test then the gene-set is associated with the treatment group, assuming the alternative hypothesis to upper tail. The enrichment score is a metric that measures how relevant is the association between the subsets of gene-sets and the treatment group.

GSEA [9] is the most used gene-set enrichment methodology adopting a modified version of the two-sample Kolmogorov-Smirnov test. Its main drawback is the heavy computational load that for N gene-sets is O(KN), with K high due the resampling strategy to compute the null distribution. Other methodologies, such as [6] and [12], concerning Mann-Whitney test [11], known as Mann-Witney-Wilcoxon (MWW) or rank-sum test as well, captured our attention for two main reasons: 1) the computational time can be reduced, 2) insufficient attention has been paid to the test-statics which is crucial to interpret the results. We explored this second point in [3] where we proposed a methodology named MWW-GST (Gene Set Test), based on the normalization of the MWW test-statistic defined as Normalized Enrichment Score (NES), that is an estimate of a probability.

In this note we propose an online implementation of MWW-GST in javascript that allowed us to implement a very fast enrichment analysis tool. Essentially, given a gene-profile, and a collection of gene-sets, to speed up the MWW-GST we precompute the rank of the gene-profile, and then we compute the MWW-statistic for each gene-set. As implemented in javascript, the tool does not run on server side, so the speed of the analysis reflects the speed limit of the client web-browser.

2 Methodological details

2.1 The Normalized Enrichment Score and the p-value

The Normalized Enrichment Score (NES) and the p-value come from the Mann-Whitney (MW) test [4]. The null hypothesis of this test states that there's no mutual dominance of the distribution functions, $F_{in}(x)$ and $F_{out}(x)$, that describe the intensities of the genes, respectively in and out the gene-set. The alternative hypothesis states that the distribution function $F_{out}(x)$ dominates $F_{in}(x)$, i.e. $\mathcal{H}_1$: $F_{out}(x) > F_{in}(x)$. Under the alternative hypothesis, the genes in the gene-set have intensities higher than those of the genes outside the gene-set. The test statistic U of the MW-test is the number of times that the relation $x_j^{out} < x_i^{in}$ is true $\forall\, i, j$, where x_j^{out} $(j = 1, 2, \ldots, m')$ is the intensity associated with the j^{th} gene outside the gene-set, while x_i^{in} $(i = 1, 2, \ldots, m'')$ is the intensity associated with the i^{th} gene in the gene-set. $m' + m''$ amounts to the total number of genes in the gene-profile. The computation of the U-statistic requires that the values x_j^{out} and x_i^{in} are combined and then rank-transformed. The rank-sum test statistics reported in [11] helps to speed up the computation of the U-statistic (then Mann-Whitney-Wilcoxon (MWW) test).

According to [2], the ratio $U/m'm''$ is an unbiased estimator of the probability $P\big[X_{in} > X_{out}\big]$, where $X_{in} \sim F_{in}(x)$ and $X_{out} \sim F_{out}(x)$. Given a gene-set, the event $X_{in} > X_{out}$ says that "*a gene randomly drawn from the gene-set has an intensity greater than the one of a second gene randomly sampled from outside the gene-set*". We define the Normalized Enrichment Score

(NES) as $P\left[X_{in} > X_{out}\right]$ and the associated p-values comes from the MWW-test. Assuming that a gene-profile recapitulates the differential expression of treatment samples versus control, a NES close to 1 means association of the gene-set with the treatment. Instead, a NES close to 0 suggests an association with the control group. This interpretation allows us to restate NES as

$$NES = P\left[\textit{the gene-set is associated with the treatment group}\right].$$

A different way to look at the NES is the odds = NES/(1-NES) that is the imbalance of the probability that the gene-set is associated with the treatment group to the probability that the gene-set has no association with it, or the gene-set is related to the control group.

$$odds = \frac{P[\textit{the gene-set is associated with the treatment group}]}{P[\textit{the gene-set is not associated with the treatment group}]}$$

The association with the treatment is as strong as the odds diverges to infinity; it is weak when the odds approaches to zero. In this last case, the association is with the control groups. An odds about 1.0 means no association with either the treatment nor with the control group.

A further transformation of NES is the $logit2NES = \log_2(\text{odds})$. In this version of the NES, a zero value means no association; a positive value is a measure of the association of the gene-set with the treatment group, while a negative value points at the control.

2.2 Enrichments prioritization

It is known that gene-sets with smaller dimension are inclined to get higher NES and lower p-value. The final table of the analysis is usually ranked according to the NES or the p-value, in this way, the attention focuses on marginal significant gene-sets instead of those with larger sizes that could provide a robust understanding of the treatment group. To balance among NES, p-value, and the gene-set size, we introduced the recap variable *relevance* (rel). Let assume we run a two sided enrichment test so that some gene-sets have $logit2NES > 0$, and some others $logit2NES < 0$. For the k'^{th} gene-set, $k' = 1, 2, \ldots$, in the collection having $logit2NES > 0$, then $rel^{+}_{k'} = \text{rank}\left(\text{actual-size}_{k'}\right) + \text{rank}\left(\text{logit2NES}_{k'}\right) + \text{rank}\left(1 - p\text{-value}_{k'}\right)$, where $\text{rank}\left(\cdot\right)$ is a function that associates the highest rank with the highest value of its argument, and actual-size is the gene-set size. Similarly, the relevance in the subsets of gene-sets (with index k'') such that logit2NES < 0, $rel^{-}_{k''} = \text{rank}\left(\text{actual-size}_{k''}\right) + \text{rank}\left(-\text{logit2NES}_{k''}\right) + \text{rank}\left(1 - p\text{-value}_{k''}\right)$. Finally, given the k^{th} gene-set, its rel_k value is rel^{+} when $\text{logit2NES}_k > 0$, and $-rel^{-}$ when $\text{logit2NES}_k < 0$. For the "greater" alternative hypothesis, $rel \equiv rel^{+}$, and for the "less" hypothesis $rel \equiv -rel^{-}$.

2.3 Enrichments visualization

We integrate the results with a network-graph of gene-sets. A node represents a significant gene-set. The sizes of the node are proportional to the size of gene-sets, while the intensity of the color is proportional to NES values. The connection between two gene-sets A and B is instead proportional to their similarity $S(A, B)$. The similarity is computed as a convex combination of the Jaccard, $\delta_0(A, B) = |A \cap B|/|A \cup B|$, and the overlap, $\delta_1(A, B) = |A \cap B|/\min\left(|A|, |B|\right)$, indexes. $S(A, B) = \epsilon \cdot \delta_1(A, B) + (1 - \epsilon) \cdot \delta_0(A, B)$, with $0 \leq \epsilon \leq 1$. When $\epsilon = 0$ we get $S(A, B) \equiv \delta_0(A, B)$, while $\epsilon = 1$ means $S(A, B) \equiv \delta_1(A, B)$.

3 Results

To run the analysis, the user needs to load two files: a gene-profile (as a two columns tab-separated text format, the gene-name and the associated value), and one or many collection of gene-sets (in .gmt format). The next steps are: 1) set the significance-level of the enrichments

Analysis results

Analysis from www.massivegenesetstest.org started at 4/23/2019, 4:29:58 PM (running time: 0.90 seconds)
Gene set collection c5hallmark.6.2.gmt (5967 gene sets found)
Gene-profile geneProfile_of_FGFR3-TACC3_fusion_positive_samples_in_GBM.txt (17814 symbols found)
Alternative: twosided (B-value < 0.01 and abs(logit2NES) > 1)

gene set	collection	size	actualSize	NES	odds	logit2NES	absLogit2NES	p-value	BH-value	B-value	relevance
HALLMARK OXIDATIVE PHOSPHORYLATION	c5hallmark.6.2.gmt	200	194	0.701	2.343	1.229	1.229	0.000e+0	0.000e+0	0.000e+0	8543.500
GO ENERGY DERIVATION BY OXIDATION OF ORGANIC COMPOUNDS	c5hallmark.6.2.gmt	217	204	0.669	2.019	1.014	1.014	1.110e-16	1.840e-14	6.625e-13	8506.500
GO CELLULAR RESPIRATION	c5hallmark.6.2.gmt	143	134	0.713	2.488	1.315	1.315	0.000e+0	0.000e+0	0.000e+0	8394.500
GO MITOCHONDRIAL MEMBRANE PART	c5hallmark.6.2.gmt	173	153	0.672	2.053	1.038	1.038	1.874e-13	2.110e-11	1.118e-9	8389.500
GO MITOCHONDRIAL PROTEIN COMPLEX	c5hallmark.6.2.gmt	136	121	0.698	2.306	1.205	1.205	6.362e-14	7.592e-12	3.796e-10	8320.000
GO INNER MITOCHONDRIAL MEMBRANE PROTEIN COMPLEX	c5hallmark.6.2.gmt	106	95	0.744	2.902	1.537	1.537	2.220e-16	3.487e-14	1.325e-12	8237.000
GO ELECTRON TRANSPORT CHAIN	c5hallmark.6.2.gmt	94	85	0.730	2.708	1.437	1.437	2.177e-13	2.362e-11	1.299e-9	8150.000
GO OXIDATIVE PHOSPHORYLATION	c5hallmark.6.2.gmt	84	76	0.725	2.632	1.396	1.396	1.288e-11	1.114e-9	7.688e-8	8082.000
GO RESPIRATORY CHAIN	c5hallmark.6.2.gmt	80	73	0.736	2.785	1.478	1.478	3.273e-12	3.150e-10	1.953e-8	8067.500
GO MITOCHONDRIAL RESPIRATORY CHAIN COMPLEX ASSEMBLY	c5hallmark.6.2.gmt	76	67	0.715	2.508	1.327	1.327	1.180e-9	7.493e-8	7.043e-6	7985.000
GO AEROBIC RESPIRATION	c5hallmark.6.2.gmt	53	51	0.734	2.759	1.464	1.464	7.466e-9	4.368e-7	4.455e-5	7819.000
GO MITOCHONDRIAL RESPIRATORY CHAIN COMPLEX I BIOGENESIS	c5hallmark.6.2.gmt	56	52	0.714	2.500	1.322	1.322	9.064e-8	4.362e-6	5.409e-4	7812.500
GO NADH DEHYDROGENASE COMPLEX	c5hallmark.6.2.gmt	42	42	0.721	2.588	1.372	1.372	6.986e-7	2.798e-5	4.169e-3	7655.500
GO TRICARBOXYLIC ACID METABOLIC PROCESS	c5hallmark.6.2.gmt	37	36	0.781	3.574	1.838	1.838	5.141e-9	3.099e-7	3.068e-5	7555.500
GO ACETYL COA METABOLIC PROCESS	c5hallmark.6.2.gmt	26	25	0.814	4.377	2.130	2.130	5.485e-8	2.773e-6	3.273e-4	7215.500

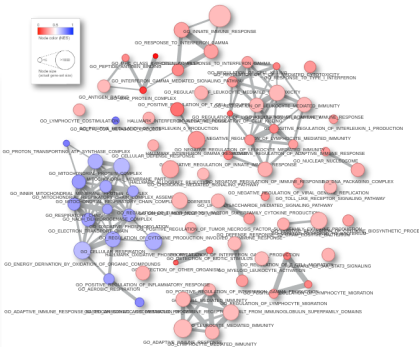

Figure 1: Example of the output of enrichment analysis. On the left, a tabular view of the significant gene-sets; on the right, the same significant gene-sets displayed as a network.

(the user can choose between the p-value, and two versions of adjusted p-values: Benjamini-Hochberk and Bonferroni), and 2) set the threshold value of the $logit2NES$. To require that the probability of association of the gene-set with the treatment group be twice the probability of non association, the $logit2NES$ threshold must be set to 0.9 (equivalent to $NES > 0.65$, or $odd > 1.5$).

Triggered the computation with the run button, a tabular version of the results is generated (see figure 1). This table respects the constrains given as input, while the full table of the enrichments associated with every gene-set can be downloaded as .csv or .tsv text format. The html version of the table can be downloaded as shown. Both the displayed table, and its .html version, allows the user to re-sort the results according to any of the columns.

To visualize the network-graph of the current results, the user can click on the network tab. Here, the similarity between any two of the gene-sets in the table is computed and the network of the gene-sets is shown. The user can chose between two similarity measures, the Jaccard or the overlap, or any convex combination of the twos by tuning the parameter ϵ with a slider box. A second slider-box allow to set the threshold value so that a segment joins two nodes when the similarity is above it. As the user operates the two sliders, the network is updated in real time. The plot of the network allows some editing actions and it can be downloaded as .png file.

In figure 1 we present an example of analysis results. We interrogated the gene-profile of the FGFR-TACC3 fusion positive samples in the glioblastoma multiforme study from the TCGA (see [3]) with the C5 and Hallmark collections (MsigDB v.6.1) of gene-sets from the Broad Insitute. On the left, there is a list of the significant gene-sets (alternative = two-sided, B.value < 0.01, and abs(logit2NES) > 1), while the corresponding network is shown on the right. This analysis can be reproduced with the gene-profile provided on the web-site, and gene-sets collections provided by the http://software.broadinstitute.org/gsea/msigdb/index.jspMSigDB.

Acknowledgements

SMP designed the statistical analysis, and LC engineered the web-site and implemented the statistical functions.

Funding

This work has been supported by 1) the Department of Science and Technology, Universitá degli Studi del Sannio, Benevento, 82100, Italy; 2) the AIRC under IG 2018 - ID. 21846 project P.I. Ceccarelli Michele, and 3) PRIN2017 id: 2017XJ38A4_004.

References

[1] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25 EP –, May 2000.

[2] D Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387 – 415, 1975.

[3] V Frattini, SM Pagnotta, Tala, JJ Fan, MV Russo, SB Lee, L Garofano, J Zhang, P Shi, G Lewis, H Sanson, V Frederick, AM Castano, L Cerulo, DCM Rolland, R Mall, K Mokhtari, KSJ Elenitoba-Johnson, M Sanson, X Huang, M Ceccarelli, A Lasorella, and A Iavarone. A metabolic function of fgfr3-tacc3 gene fusions in cancer. *Nature*, 553:222 EP –, Jan 2018.

[4] HB Mann and DR Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60, 03 1947.

[5] Huaiyu Mi, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. Large-scale gene function analysis with the panther classification system. *Nature Protocols*, 8:1551 EP –, Jul 2013.

[6] J Michaud, KM Simpson, R Escher, K Buchet-Poyau, T Beissbarth, C Carmichael, ME Ritchie, F Schütz, P Cannon, M Liu, X Shen, Y Ito, WH Raskind, MS Horwitz, M Osato, DR Turner, TP Speed, M Kavallaris, GK Smyth, and HS Scott. Integrative analysis of runx1 downstream pathways and target genes. *BMC Genomics*, 9(1):363, Jul 2008.

[7] VK Mootha, CM Lindgren, KF Eriksson, A Subramanian, S Sihag, J Lehar, P Puigserver, E Carlsson, M Ridderstråle, E Laurila, N Houstis, MJ Daly, N Patterson, JP Mesirov, TR Golub, P Tamayo, B Spiegelman, ES Lander, JN Hirschhorn, D Altshuler, and LC" Groop. Pgc1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267 EP –, Jun 2003. Article.

[8] R. G. O'Brien and J. M. Castelloe. Exploiting the link between the wilcoxonmannwhitney test and a simple odds statistic. In NC Cary, editor, *Proceedings of the Thirty-first Annual SAS Users Group International Conference, Paper 209-31*. San Francisco, CA, SAS Institute Inc.

[9] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, and JP Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[10] L Tian, SA Greenberg, SW Kong, J Altschuler, IS Kohane, and PJ Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

[11] F Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[12] D Wu and GK Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, 2012.