Review of: "Systematic comparison of ranking aggregation methods for gene lists in experimental results"

Karoly Heberger

Potential competing interests: The author(s) declared that no potential competing interests exist.

Review target:

Source: bioRxiv2022.01.09.475491v1.full

Title: Systematic comparison of ranking aggregation methods for gene lists in experimental results

Authors: Bo Wang*, Andy Law, Tim Regan, Nicholas Parkinson, Joby Cole, Clark D. Russell, David H. Dockrell, Michael U. Gutmann* and J. Kenneth Baillie*

Body

The authors have completed a meta-analysis of gene lists using simulated and real data. Indeed, many studies answer the same-similar questions and hence, they can be combined in a meta-analysis to find a consensus or a more reliable answer. The selected performance parameter (merit) was the coverage rate. The authors have correctly realized that "the result can be largely influenced by the way a measurement weights each position.", *i.e.*, no preference can be set honestly to data sets (which one is more reliable and on what extent).

However, they have made a simple univariate comparison (colored nicely in a heatmap form), though the methods to be compared work differently on the various data sets. Hence, only a multicriteria decision making (MCDM) analysis (or multiobject optimization, MOO) would provide proper solutions. Any MCDM method would do, *e.g.*, TOPSIS, Promethee II, but preferably sum of ranking differences (SRD) [1,2], as it realizes a consensus of eight independent MCDM techniques [3].

Transpose of table 1 is suitable to make a fair method comparison. The table contains coverage rates, *i.e.*, 'the higher the better rule' should be applied. Then, the gold standard should be chosen as row maximums. It is equivalent to define the hypothetically best method, which provides the best coverage rate for all examined data sets. Abundant data sets were investigated, *i.e.*, reliable conclusions might be drawn. The computer code can be downloaded from the link [3]:

http://aki.ttk.mta.hu/srd

Figure 1 shows the ordering and grouping the methods by their SRD values (Manhattan distances to the ideal rankings (gold standard)):
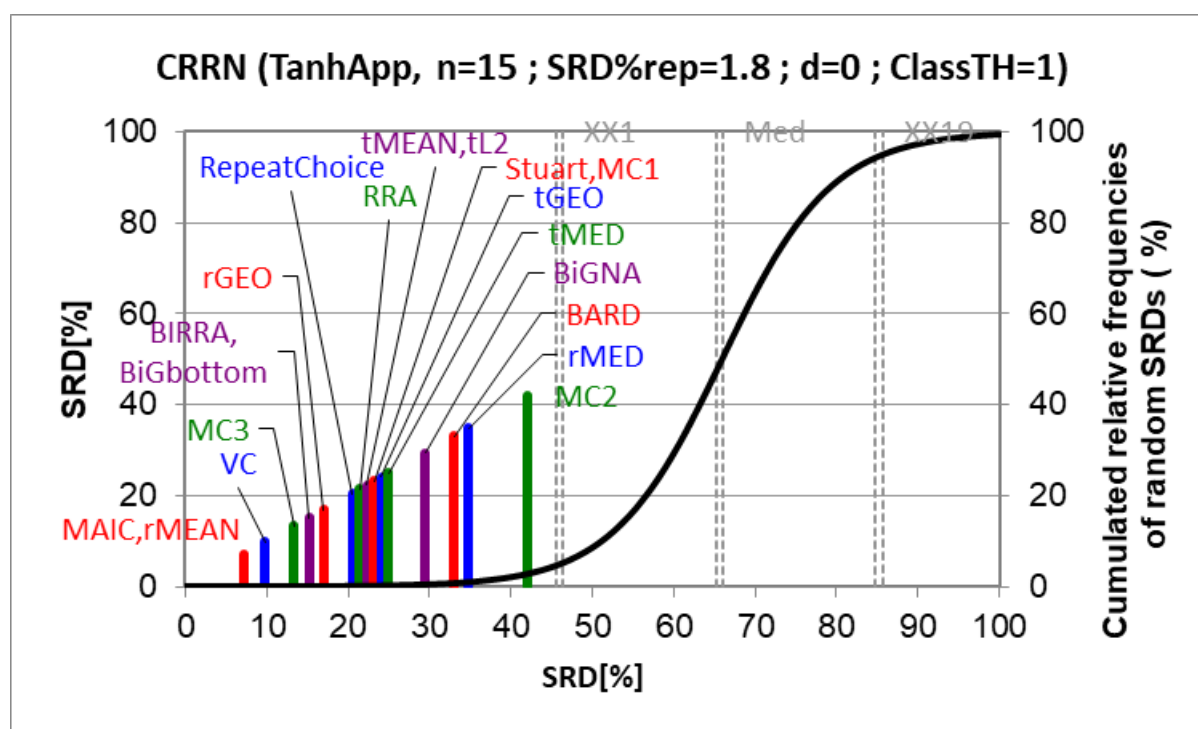
Figure 1 Comparison of ranking aggregation methods based on coverage rates with top-1000 cutoff. Scaled SRD values (Between 0 and 100) are plotted on the left y and x axes alike, *i.e.*, not the length of the bars carry the information, but their proximity to 0 (gold standard), to random ranking (black cumulated Gaussian curve), and their grouping. The abbreviations are the same as in bioRxiv2022.01.09.475491v1.full

To assign uncertainties the SRDvalues, sevenfold cross-validation was carried out by two different manners: Type A – sevenfold split blockwise; Type B – repeated random resampling with return (bootstrap).

Then, variance analysis (ANOVA) was carried out with two factors: F1 – resampling options, two levels: A and B, and F2 – rank aggregation methods, 19 levels. So, $n=14*19$.

F1 is a random factor and not significant at the 5 % level, as expected. Significant grouping can be observed in the rank aggregation methods (F2):
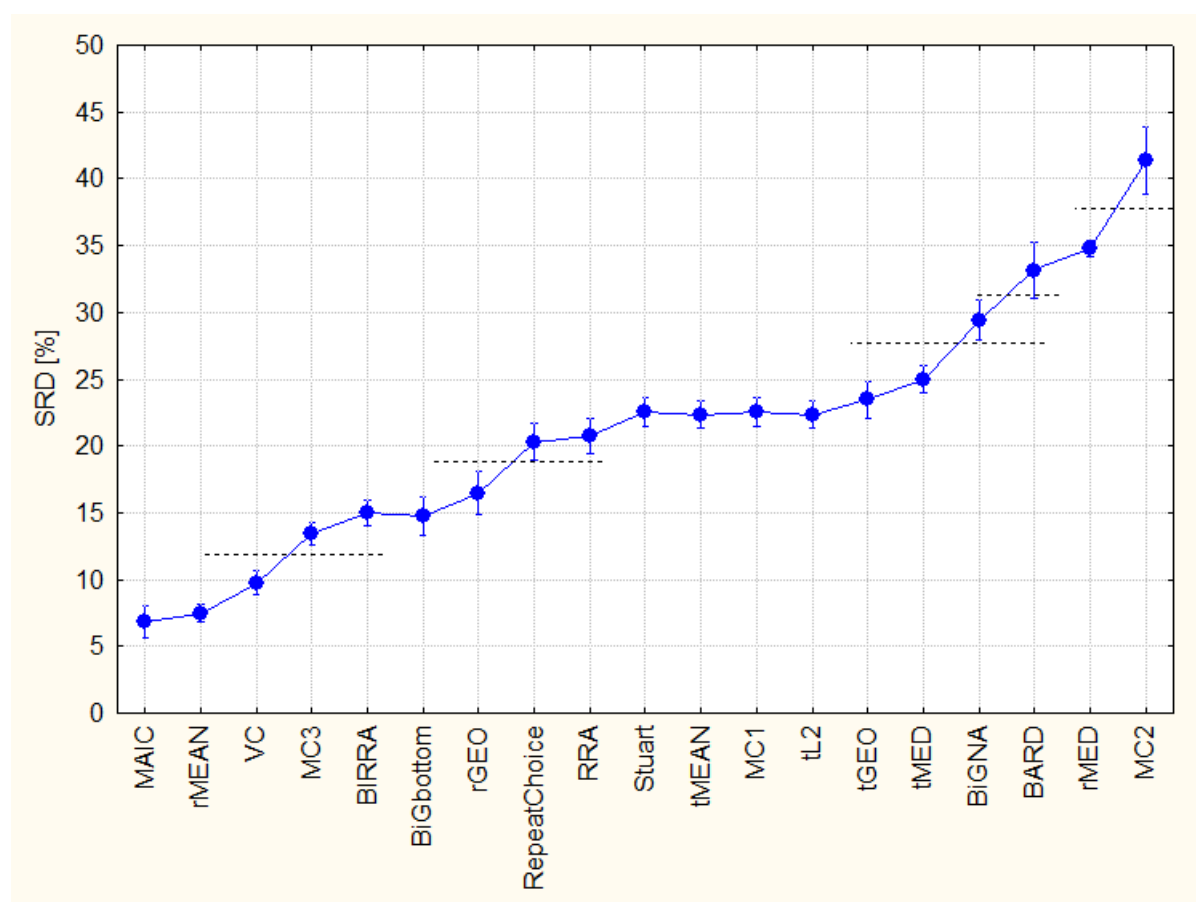
Figure 2 Scaled SRD values (between 0 and 100) for rank aggregation methods. The smaller the better (closer to the ideal ranking). Tukey's HSD (honestly significant differences) test indicates the groups separated by dotted lines. (The largest cluster in the middle is comprised of two overlapping ones. The same *post hoc* grouping is observed by Bonferroni's test.

Discussion

The figures are self-explanatory. Cealr ranking can be seen in **Figures 1** and **2**. The latter shows the significant groupings. On the one hand, sum of ranking differences confirms and, on the other hand, modulates previous findings. MAIC is the best method, but not significantly better than rMEAN and VC. MC3, BiGbottom, BIRRA, and rGEO compose the second group, which can be recommended and also recommended by the authors for ranked inputs. Stuart is placed in the middle; methods on the right-hand side are not recommended (BARD and rMED are not different significantly).

It is true the ranked or not ranked inputs are mixed up in this SRD analysis. However, the above methods should be preferred in unknown cases, for large and small datasets, evaluating many sources, and in case of large heterogeneity. Anyway, these ideas are subjective: when is a data set large or heterogeneous? There were many equal numbers (ties) in the input file; therefore, SRD with ties should be used only. Consequently, a method might have the same ranking difference for more than one data set.

Table 1 is not complete; four deficient lines can be involved in the study using multiple imputation or substituting averages to the missing places.

Some of the errors are minor, *e.g.*, DOI links are missing from the references or Figure 4 is redundant, but these can be corrected easily.

Declaration of potential competing interests

No competing interest declared

References

[1]   K. Héberger: Sum of ranking differences compares methods or models fairly. TRAC – Trends in Analytical Chemistry, 29 (2010) 101-109.
https://doi.org/10.1016/j.trac.2009.09.009 .

[2]   K. Kollár-Hunek, K. Héberger: Method and model comparison by sum of ranking differences in cases of repeated observations (ties) Chemometrics and Intelligent Laboratory Systems, 127 (2013) 139-146.
http://dx.doi.org/10.1016/j.chemolab.2013.06.007 .

[3]   J. M. Lourenço, and L. Lebensztajn: Post-Pareto Optimality Analysis with Sum of Ranking Differences. IEEE Transactions on Magnetics, 54, Issue: 8 (2018) Article Sequence Number: 8202810 pp. (99), 1–10.
https://doi.org/10.1109/TMAG.2018.2836327 .

Supplementary material

Table 1 from bioRxiv2022.01.09.475491v1.full

(Four deficient lines was omitted)

MAIC	0.14	0.27	0.46	0.11	0.21	0.36	0.39	0.11	0.11	0.22	0.37	0.11	0.19	0.34	0.19
BIRRA	0.10	0.23	0.42	0.10	0.19	0.34	0.32	0.07	0.10	0.23	0.42	0.10	0.19	0.34	0.14
BiGbottom	0.12	0.20	0.33	0.11	0.18	0.34	0.31	0.12	0.12	0.20	0.33	0.11	0.18	0.34	0.18
rGEO	0.12	0.19	0.29	0.11	0.18	0.29	0.30	0.05	0.12	0.19	0.29	0.11	0.18	0.29	0.18
VC	0.13	0.22	0.36	0.11	0.16	0.27	0.36	0.13	0.11	0.17	0.29	0.10	0.15	0.26	0.19
rMEAN	0.12	0.19	0.30	0.11	0.17	0.28	0.29	0.14	0.12	0.19	0.30	0.11	0.17	0.28	0.19
MC3	0.12	0.19	0.30	0.11	0.18	0.30	0.30	0.11	0.12	0.19	0.30	0.11	0.18	0.30	0.17
Stuart	0.11	0.18	0.27	0.11	0.18	0.29	0.30	0.05	0.11	0.18	0.27	0.11	0.18	0.29	0.17
tGEO	0.11	0.17	0.24	0.10	0.17	0.28	0.35	0.07	0.11	0.17	0.24	0.10	0.17	0.28	0.11
RRA	0.11	0.18	0.26	0.11	0.17	0.28	0.29	0.05	0.11	0.18	0.26	0.11	0.17	0.28	0.16
tMEAN	0.10	0.15	0.20	0.10	0.15	0.24	0.29	0.12	0.10	0.15	0.20	0.10	0.15	0.24	0.10
MC1	0.11	0.17	0.24	0.11	0.17	0.25	0.31	0.08	0.11	0.17	0.24	0.11	0.17	0.25	0.13
tMED	0.10	0.16	0.23	0.10	0.16	0.27	0.29	0.11	0.10	0.16	0.23	0.10	0.16	0.27	0.09
tL2	0.09	0.14	0.19	0.09	0.14	0.22	0.29	0.13	0.09	0.14	0.19	0.09	0.14	0.22	0.09
RepeatChoice	0.09	0.14	0.22	0.09	0.13	0.24	0.29	0.07	0.09	0.14	0.19	0.08	0.12	0.24	0.13
BiGNA	0.09	0.14	0.19	0.09	0.14	0.22	0.31	0.03	0.09	0.14	0.19	0.09	0.14	0.22	0.03
BARD	0.09	0.13	0.26	0.09	0.15	0.31	0.11	0.13	0.09	0.13	0.26	0.09	0.15	0.31	0.09
MC2	0.08	0.12	0.16	0.11	0.17	0.28	0.16	0.11	0.08	0.12	0.16	0.11	0.17	0.28	0.09
rMED	0.09	0.13	0.18	0.10	0.15	0.26	0.16	0.15	0.09	0.13	0.18	0.10	0.15	0.26	0.08