BIAS-AWARE CONSENSUS PERTURBATION PATHWAY ANALYSIS

NEVADA NASA SPACE GRANT CONSORTIUM

UNIVERSITY OF NEVADA, RENO

Hung Nguyen, Duc Tran, and Tin Nguyen*

Department of Computer Science and Engineering, University of Nevada, Reno

Contact: tinn@unr.edu. Website: https://bioinformatics.cse.unr.edu

05/08/2020

BACKGROUND

More than 70 pathway analysis techniques have been developed to understand the molecular mechanisms under certain conditions, especially with complex diseases [1]. However, most methods are sensitive to noise in expression data [2] and are bias toward certain pathways [3].

OBJECTIVES

Developing a powerful ensemble approach, Bias-Aware Consensus Perturbation Analysis (BACPA), that (i) takes advantage of each method's strength, (ii) is robust against noise, and (iii) performs unbiased analyses.

RESULTS

Data: 22 datasets containing 4 diseases with a total of 1,713 samples (742 control samples, 971 disease samples).

Metric: Target pathway's ranking (smaller the better).

Methods: ORA [6], KS, Wilcox, GSEA [7], GSA [8], and BACPA.

Results: BACPA produces the best rankings for target pathways.

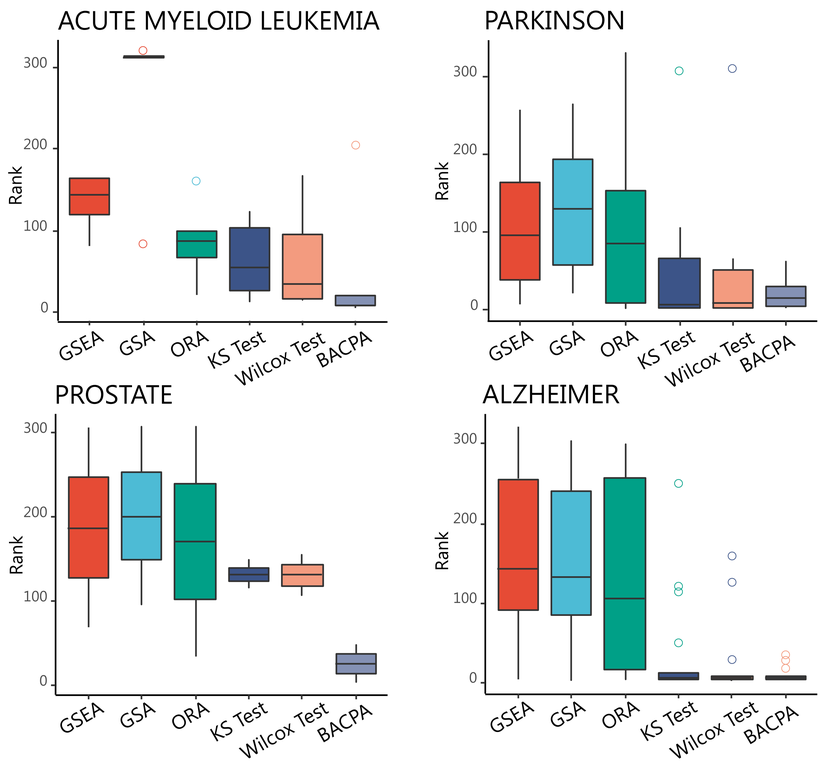

Fig. 3. Ranking of target pathways on 4 diseases

METHODS

Perturbation Pathway Analysis: Samples in the input data is repeatedly perturbed and filtered to mitigate the effect of the noisy nature of the expression data [4].

Consensus Statistical Testing: Three statistical hypothesis tests including Fisher's exact test (ORA), KS test, and Wilcox test are used to compute the significance values of impacted pathways, which are then combined using the **Additive Method or Central Limit Theorem** [5].

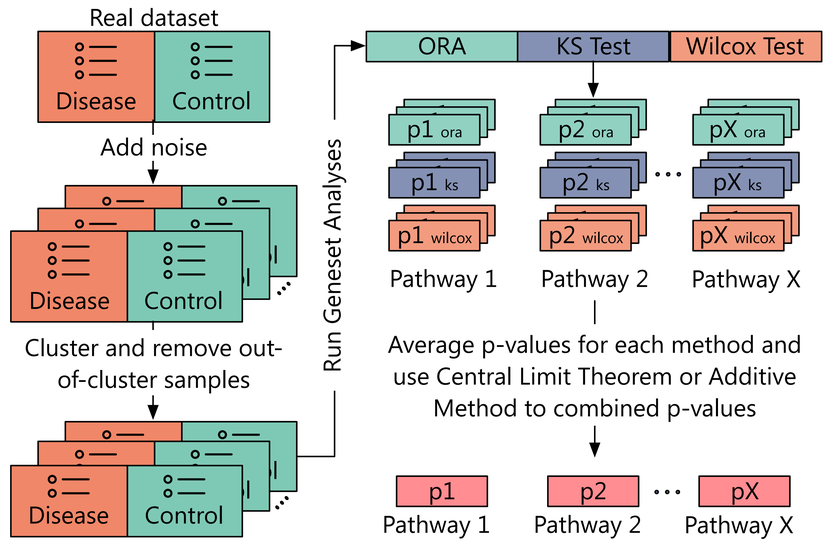

Fig. 1. Consensus perturbation pathway analysis

Bias Correction: A random dataset is repeatedly generated from a pool containing only control samples and is used as the input for the pathway analysis to obtain the empirical null distribution of p-values for each pathway. These distributions are used to correct the resulted p-values.

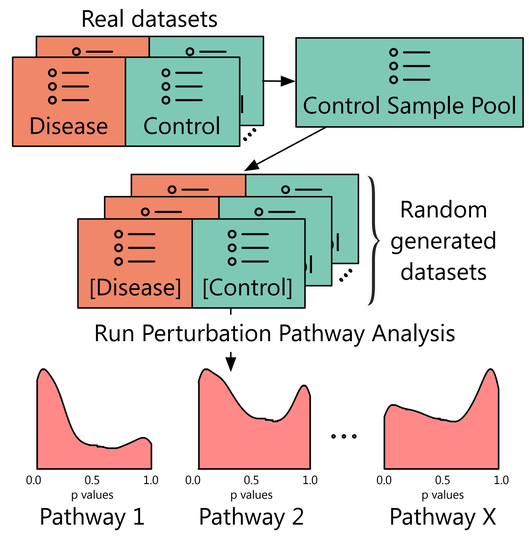

Fig. 2. Empirical null distribution generation

CONCLUSION

BACPA is an effective method for pathway analysis. It is fast, non-bias and robust against noise.

FUTURE WORK

- Apply BACPA to study the effects of microgravity on living organisms using data from NASA GeneLab.
- Build an interactive web interface to visualize and explore the analysis results [9].
- Apply the methodology to related fields using genomic data including meta-analysis [10], omics integration [11], subtyping [12], and single-cell analysis [13].

References

1. Nguyen et al. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20(1), 1-15.
2. Borisov et al. (2017). Data aggregation at the level of molecular pathways improves stability of experimental transcriptomic and proteomic data. *Cell Cycle*, 16(19), 1810-1823.
3. Nguyen et al. (2017). DANUBE: data-driven meta-ANalysis using UnBiased empirical distributions—applied to biological pathway analysis. *Proceedings of the IEEE*, *105*(3), 496-515.
4. Nguyen et al. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, 27(12), 2025-2039.
5. Nguyen et al. (2016). A novel bi-level meta-analysis approach: applied to biological pathway analysis. *Bioinformatics*, 32(3), 409-416.
6. Draghici et al. (2003). Global functional profiling of gene expression. *Genomics*, 81(2), 98-104.
7. Subramanian et al. (2005). 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.
8. Efron et al. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1), 107-129.
9. Cruz et al. (2019). Functional analysis tools for post-translational modification: a post-translational modification database for analysis of proteins and metabolic pathways. *The Plant Journal*, 99(5), 1003-1013.
10. Nguyen et al. (2020). NBIA: a network-based integrative analysis framework–applied to pathway analysis. *Scientific Reports*, 10(1), 1-11.
11. Shafi et al. (2019). A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Frontiers in Genetics*, 10, 159.
12. Nguyen et al. (2019). PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics, 35(16), 2843-2846.
13. Tran et al. (2019). Fast and precise single-cell data analysis using hierarchical autoencoder. *bioRxiv*, 799817.

This material is based upon work supported by the National Aeronautics and Space Administration under Grant No. 80NSSC19M0170