Pathway Expression Analysis

Nathan Mankovich[*,1], Eric Kehoe[*,2], Amy Peterson[*,3] and Michael Kirby[*,4]

Abstract— **This talk introduces a novel pathway expression framework for biological data analysis. Using this framework, we distinguish between shedding subjects post-infection and all subjects pre-infection in human blood transcriptomic samples challenged with various respiratory viruses: H1N1, H3N2, HRV (Human Rhinoviruses), and RSV (Respiratory Syncytial Virus). The classification results and selected discriminatory pathways from pathway expression data are benchmarked against standard gene expression based classification and pathway ranking methodologies. We find that using pathway expression data along with selected pathways, which have minimal overlap with high ranking pathways found by traditional methods, improves balanced success rates across experiments.**

I. Background

Analyses of respiratory illnesses aid in understanding the mechanisms of shedding and in developing methodologies that succeed across multiple infectious diseases. Understanding the imprint of viral shedding on human gene expression may uncover latent effects which are beyond disease symptoms. Previous work used machine learning (ML) models, e.g., neural networks, support vector machines (SVM), centroid encoders (CE), and spectral gene graph analysis, to identify discriminatory biomarkers within early shedders challenged with influenza and subsequently classify those subjects [1], [2], [3]. Most applications of ML models which learn on gene expression data do not utilize known biological relationships between genes. However, a biological pathway analysis uses those known relationships, usually by grouping related genes, to build a biologically informed model [4]. Given the numerous definitions of pathway membership and number of ways to relate genes within a pathway, many of these pathway analyses require an a priori set of "important" genes to determine significant pathways [5], [6]. Standard tools that rely on an a priori gene set include: over representation analysis (ORA) [7], gene set enrichment analysis (GSEA) [8] and Centrality-based pathway enrichment (CePa) [9].

II. Results

We demonstrate that novel pathway expression methods produce higher balanced success rates (BSRs) than gene expression methods on 3 out of 4 classification experiments on the GSE73072 data set (a human transcriptomics respiratory virus data set). We are able to select pathways using influenza training data then use these selected pathways to produce competitive, even improved, classification BSRs on HRV and RSV data sets when compared to gene based methods. We improve upon the classification results from Aminian et. al. [1] by 3 to 4 percent in BSR using selected pathways from pathway expression data rather than selected genes from gene expression data. In addition, we compare these selected pathways from pathway expression methods to two standard gene expression pathway analysis methods: CePa and ORA. We find that the pathways selected from pathway expression methods generally have little similarity to pathways from these gene expression methods.

III. Conclusion

We used pathway expression methods to produce improved classification results and select discriminatory pathways on the GSE73072 data set. In addition, the pathway expression methods appear to be more robust than gene expression methods to subject differences within a class. Our results also suggest that our pathway selection methods with pathway expression provide a unique pipeline that selects discriminatory pathways which are not detected by standard pathway analyses on gene expression data.

References

[1] M. Aminian, T. Ghosh, and A. e. a. Peterson, "Early Prognosis of Respiratory Virus Shedding in Humans," *Scientific Reports*, vol. 11, no. 1, pp. 1–15, 2021.

[2] M. Chaturvedi, T. Ghosh, M. Kirby, X. Liu, X. Ma, and S. Stiverson, "Explorations in Very Early Prognosis of the Human Immune Response to Influenza," pp. 562–570, 2016.

[3] N. Mankovich, "Methods for Network Generation and Spectral Feature Selection: Especially on Gene Expression Data," 2019.

[4] P. Khatri, M. Sirota, and A. J. Butte, "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges," *PLoS computational biology*, vol. 8, no. 2, p. e1002375, 2012.

[5] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork *et al.*, "The STRING Database in 2021: Customizable Protein–Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets," *Nucleic acids research*, vol. 49, no. D1, pp. D605–D612, 2021.

[6] M. Gillespie, B. Jassal, R. Stephan, and M. e. a. Milacic, "The Reactome Pathway Knowledgebase 2022," *Nucleic Acids Research*, vol. 50, no. D1, pp. D687–D692, 11 2021. [Online]. Available: https://doi.org/10.1093/nar/gkab1028

[7] G. Yu and Q.-Y. He, "ReactomePA: an R/Bioconductor Package for Reactome Pathway Analysis and Visualization," *Mol. BioSyst.*, vol. 12, pp. 477–479, 2016. [Online]. Available: doi.org/10.1039/C5MB00663E

[8] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[9] Z. Gu, J. Liu, and K. e. a. Cao, "Centrality-Based Pathway Enrichment: a Systematic Approach for Finding Significant Pathways Dominated by Key Genes," *BMC Systems Biology*, vol. 6, no. 1, pp. 1–13, 2012.

This work was partially funded by NSF-ATD 1830676.

* Department of Mathematics, Colorado State University

Emails: nmank@colostate.edu[1], two@ekehoe@colostate.edu[2], apete41lsu@gmail.com[3], michael.kirby@colostate.edu[4]