**Analysis of Interconnection Between Reactome Pathways**

Ahmed S Abouhashem†  
 Biotechnology Graduate Program  
 The American University in Cairo  
 New Cairo, Cairo, Egypt  
 a.safwat.a@aucegypt.edu

Mohammad Elkholy  
 Computer Science and Engineering  
 The American University in Cairo  
 New Cairo, Cairo, Egypt  
moelkholy@aucegypt.edu

**ABSTRACT**

Biological databases have distinct collections of information for genes, their relationship with each other, their functional classifications and their association with diseases. Researchers use those databases to infer meaningful information for results they obtain from high-throughput biological experiments. Overrepresentation analysis could infer information about pathway activity using altered sets of genes between different conditions. Although identification of a pathway to be upregulated or downregulated could be interpreted as a reflection of the cellular activity, the interpretation still lacks causal relationship. In this work, we aim to use the gene regulatory network from the Reactome database to characterize the connections between pathways and to model gene expression data. Characterization of the interconnections will identify the overlapping between pathways from the proportion of shared genes and will identify the regulations among pathways from the activity of genes. Next, edge weights will be inferred from modeling of gene expression data from different conditions, so we may identify the active paths where information flow occurs in each condition. The current work would consider the signaling transduction flow difference between distinct conditions.

1**Literature review**

The amount of big data in bioinformatics increases on a daily basis worldwide. Biological databases include information about physical interaction between proteins [1], signaling pathways [2], metabolic pathways [3], and collections of Gene Ontology (GO) sets [4]. Signaling and metabolic pathway datasets define categories where different genes/ proteins belong to such as development and DNA repair. Of note, each gene can belong to multiple pathways or not to belong to a defined pathway. GO includes collections of gene sets sharing similar localization, biological processes or molecular functionality regardless of their pathway such as glucose transport, angiogenesis (formation of new blood vessels) and catalase activity. GO gene sets, among others, were utilized in multiple research articles to gain more insights about the results at the level of genome, transcriptome or epigenome. For example, when comparing gene expression level (transcriptome; ~20,000 protein coding genes) between two conditions such as COVID patients and healthy individuals, results might include hundreds of differentially expressed genes (DEGs) between the two conditions. Enrichment analysis can then be performed to identify which pathways or biological processes are overrepresented by the DEGs [5,6]. Thus, alterations at pathway level whether upregulated or downregulated can be identified by comparing different cellular conditions.

Despite the distinct collections for gene annotations, it is still challenging to explain how a pathway is upregulated. It is very challenging to analyze the relationship between cellular genes/ proteins, between proteins and diseases or between drugs and proteins. Network analysis has been applied to better understand how genes behave collectively. Novel ways were developed to interpret interactions between diseases, drugs and proteins. For instance, identification of coexpressed interacting proteins within the protein-protein interaction network might point towards that those proteins regulate each other activity and might form complexes (machinery like). Other research involves integration of protein-protein interaction network, drug-protein network and disease association network (disease-protein) to identify drugs that might be effective for different diseases rather than what they were prescribed [7,8]. Another type of biological network considers the relationship between genes whether a gene activates (upregulates) another gene or deactivates (downregulates) another gene [9]. Multiple research was performed to infer gene regulatory networks from different high-throughput datasets in order to perform cell reprogramming, changing cell state through identification of altered transcription factors [10].

However, the interaction between signaling pathways is not fully understood. Different pathways/ subpathways are not working independently. Alteration of a pathway will cause alteration in multiple other pathways. Same gene can be involved in different pathways. Additionally, the same gene can have opposing effects on different pathways (upregulate a pathway and downregulate another one). Signal transduction involves flow of information through multiple genes (proteins). For example, cells can respond to environmental stimuli via binding of a stimulatory molecule to the extracellular domain of a cell membrane receptor. Upon binding, the intracellular domain of the membrane receptor will have conformational change allowing it to interact with another protein. The cascade continues and could reach the nucleus to activate a transcription factor which in turn could alter expression level of multiple genes [11]. How activation of a pathway affects other pathways? How does signal transduction differ between health and disease? Do cells favor a signaling path over others under different states? How to identify influential genes aiming to alter the information flow back to desired states? How to downregulate a pathway involved in disease progression instead of targeting a single molecule? Those questions need to be further investigated in order to move beyond gene set enrichment analysis and start to understand how to rewire signaling transduction pathways to reprogram the cell.

2 **Ideas**

Firstly, we will build a gene regulatory network from the gene-gene relationships in the Reactome database including 9,300 genes and > 100,000 interactions (activation or activation/ inhibition). The network files are publicly available on Reactome website (<https://reactome.org/download-data>) including gene-gene relationship, pathway and their included genes and pathway-pathway relationship (hierarchical structure). We will add some gene annotations such as the type (transcription factors, enzymes, …etc) and the pathway/s the gene belongs to from the ingenuity pathway analysis (IPA) software. Also, we will obtain gene expression data (transcriptome) from the Gene Expression Omnibus (GEO) database for different conditions (<https://www.ncbi.nlm.nih.gov/geo/>). The gene expression data will be analyzed using the Seurat package in R to perform data quality check and normalization. Afterwards, we will focus on a cell type and compare between different conditions. By comparing the different conditions (either the extreme cases or the change over time), we will use the resulting fold change values for each gene to model the difference using the generated gene regulatory network.

We aim to analyze the network at the level of labeled pathways and at gene level. At the level of pathways, nodes will represent main or sub-pathways and edges will represent either amount of overlapping between pathways or the collective effect of individual genes inter-pathways. Then, we can classify the interconnection between pathways into high and low. At the gene level, we aim to identify how to increase or decrease the level of a predefined ***k*** set of genes using an undefined ***I*** set of genes. Also, we aim to be able to infer edge weights from gene expression data in which highly weighted edges might represent the active signaling transduction “***active tracks***”. Also, given a gene, we aim to identify similarity of its upstream ***active tracks*** between different conditions. The current methods for enrichment analysis depend on predefined pathways or biological processes annotations and hence they are biased towards the current knowledge (if correct). Our analysis will use information from the gene regulatory network to identify ***active tracks*** where signal transduction occurs.

**REFERENCES**

[1] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., & von Mering, C. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research.

[2] Griss J, Viteri G, Sidiropoulos K, Nguyen V, Fabregat A, Hermjakob H. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. Mol Cell Proteomics. 2020.

[3] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.

[4] GO Enrichment Analysis: Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. Jan 2019;47.

[5] Abouhashem, A. S., Singh, K., Azzazy, H., & Sen, C. K. (2020). Is Low Alveolar Type II Cell SOD3 in the Lungs of Elderly Linked to the Observed Severity of COVID-19?. Antioxidants & redox signaling, 33(2), 59–65.

[6] Ximerakis, M., Lipnick, S. L., Innes, B. T., Simmons, S. K., Adiconis, X., Dionne, D., Mayweather, B. A., Nguyen, L., Niziolek, Z., Ozek, C., Butty, V. L., Isserlin, R., Buchanan, S. M., Levine, S. S., Regev, A., Bader, G. D., Levin, J. Z., & Rubin, L. L. (2019). Single-cell transcriptomic profiling of the aging mouse brain. Nature neuroscience, 22(10), 1696–1708.

[7] Morselli Gysi, D., do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S. D., Patten, J. J., Davey, R. A., Loscalzo, J., & Barabási, A. L. (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proceedings of the National Academy of Sciences of the United States of America, 118(19).

[8] Suratanee, A., & Plaimas, K. (2018). Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions. PloS one, 13(6).

[9] Levine, M., & Davidson, E. H. (2005). Gene regulatory networks for development. Proceedings of the National Academy of Sciences of the United States of America, 102(14).

[10] Tran, A., Yang, P., Yang, J., & Ormerod, J. (2022). Computational approaches for direct cell reprogramming: from the bulk omics era to the single cell era. Briefings in functional genomics, 21(4), 270–279.

[11] Esbenshade, T. A., & Duzic, E. (2006). Overview of signal transduction. Current protocols in pharmacology, Chapter 2, Unit2.1. https://doi.org/10.1002/0471141755.ph0201s31