Pathway Expression Analysis

Nathan Mankovich*, Eric Kehoe, Amy Peterson, and Michael Kirby

Department of Mathematics at Colorado State University

Abstract

This project introduces a pathway expression framework as an approach for constructing derived biomarkers. The pathway expression framework incorporates the biological connections of genes leading to a biologically relevant model. 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). Additionally, pathway expression data is used for selecting discriminatory pathways from these experiments. The classification results and selected pathways are benchmarked against standard gene expression based classification and pathway ranking methodologies. We find that using the 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.

Methods

Probe ID Networks

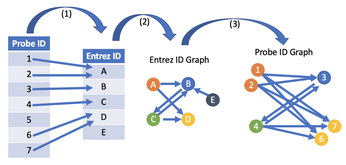


Figure: How to generate Probe ID pathway networks using the platform file and Entrez IDs pathway networks.

Pathway Expression

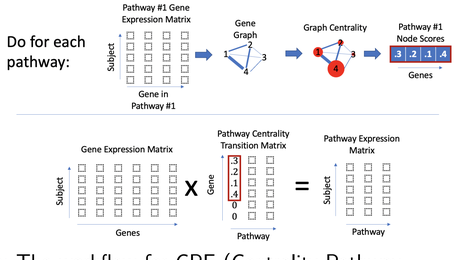


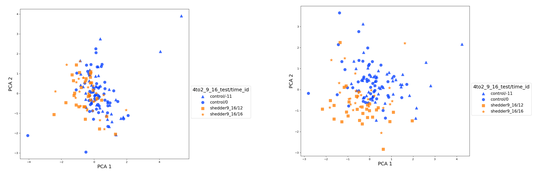
Figure: The workflow for CPE (Centrality Pathway Expression). For centrality in CPE we use either PageRank or out-degree. Linear pathway expression uses vectors of all 1s instead of using centrality scores.

SSVM Feature Selection: Take all m features ordered with respect to their corresponding magnitude of their SSVM (Sparse Support Vector Machine) weights calculated on the control/shedder classification experiment. We calculate the weight ratios and look for a "jump" (for our experiments we set our "jump" ratio at 5). We take the all features with the largest weights before the "jump". We then add in the features that are at least .9 correlated to these features using their training data.

Results

Time Bin	Method	4 to 2	4 to 3	6 to 1
1 to 8	CPE	**80.04**	**75.32**	**89.44**
1 to 8	LPE	71.80	73.43	80.35
1 to 8	GE	60.48	61.09	62.16
9 to 16	CPE	**87.19**	**89.15**	**95.45**
9 to 16	LPE	82.13	79.10	89.44
9 to 16	GE	73.51	73.60	90.90
17 to 24	CPE	**78.35**	**80.93**	**80.35**
17 to 24	LPE	69.74	69.28	74.33
17 to 24	GE	66.62	59.09	77.27
25 to 32	CPE	94.81	**90.52**	**90.91**
25 to 32	LPE	92.49	80.19	68.31
25 to 32	GE	**95.89**	80.19	75.80

Table: Classification BSR by method, experiment and time bin on the test studies. The highest BSR for each time bin is bold face.

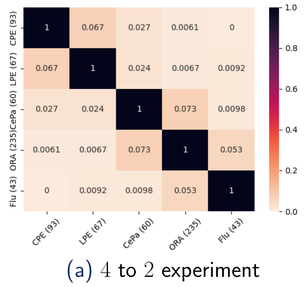


(a) GE test data with the selected the features.

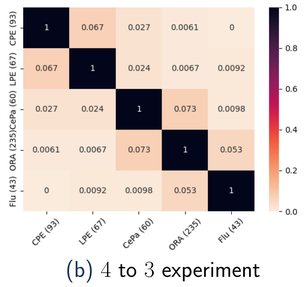
(b) CPE test data with the selected the features.

Figure: A PCA embedding of the 2 test studies in the time bin 9 to 16 hours after infection. Features are selected on the 4 training studies.

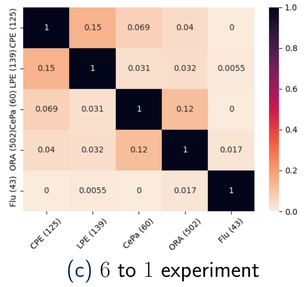
Top pathways from CPE include: R-HSA-74713 (IRS activation), R-HSA-3595177 (Defective CHSY1 causes TPBS), R-HSA-5603037 (IRAK4 deficiency (TLR5)), R-HSA-168330 (Viral RNP Complexes in the Host Cell Nucleus, R-HSA-451326 (Activation of kainate receptors upon glutamate binding), R-HSA-2485179 (Activation of the phototransduction cascade), R-HSA-5626978 (TNFR1-mediated ceramide production), R-HSA-9694631 (Maturation of nucleoprotein), R-HSA-1638091 (Heparan sulfate/heparin (HS-GAG) metabolism).



(a) 4 to 2 experiment



(b) 4 to 3 experiment



(c) 6 to 1 experiment

Figure: Jaccard overlap between the selected pathways for different methodologies at the 25 to 32 time bin.

Methods (Cont.)

Testing Features: Use the features from SSVM feature selection in a LOSO (Leave one Subject Out) SVM (Support Vector Machine) experiment on the test data.

The Bottom Line:

- We compare GE (Gene Expression) features to LPE (Linear Pathway Expression) and CPE (Centrality Pathway Expression) that were found on the training data in a classification experiment on the test data.
- We compare the pathways LPE and CPE that were found on the training data to pathways found using standard pathway analysis methods applied to the GE features found on the training data.

Human Influenza Data (GSE73072)

- From the NCBI GeneExpression Omnibus (GEO) Microarray gene expression data
- 7 studies by Duke, UVA and hVIVO
- Funded by the Defense Advanced Research Projects Agency (DARPA)
- 22277 probe identifiers and 148 human subjects
- Four different types of respiratory viruses: HRV, RSV, H1N1 and H3N2
- Data samples are collected at irregular time intervals from 38 hours before infection to 680 hours after infection

Data Partitioning

Control/ Shedder Experiment: *Controls* are subjects before infection. *Shedders* are subjects labeled as shedding after infection.

Time bins: We do 4 different time bins where shedders are taken from 1 to 8, 9 to 16, 17 to 24 or 25 to 32 after infection.

Data Partitioning

Partition Name	Train	Test
4 to 2	H3N2, H1N1	HRV
4 to 3	H3N2, H1N1	HRV, RSV
6 to 1	HRV, RSV, H1N1, H3N2	H3N2

Table: The train/test splits by disease for the 4 to 2, 4 to 3 and 6 to 1 experiments. The partition n to m means that we train on n studies and test on m different sequestered test studies. For each partition, one experiment was done with Limma on subject ID on the train and testing data.

Comparison to State of the Art Gene Expression

Time Bin	Paper	4 to 2	4 to 3	6 to 1
1 to 8	this paper	80.04	75.32	**89.44**
1 to 8	Aminian et al.	**84.74**	**82.06**	87.97
9 to 16	this paper	87.19	89.15	95.45
9 to 16	Aminian et al.	**93.21**	**90.37**	**100.00**
17 to 24	this paper	78.35	**80.93**	80.35
17 to 24	Aminian et al.	**81.58**	78.82	**86.36**
25 to 32	this paper	**94.81**	**90.52**	**90.91**
25 to 32	Aminian et al.	88.21	85.46	89.44

Table: Classifications rates of SVM in a LOSO experiment on test data across different experiments within 32 hours after infection. All experiments in this table use Limma normalization on subject identifier. The best results from this paper are using CPE with pre-computed undirected edges and PageRank centrality.

Acknowledgement: This work was partially supported by National Science Foundation award NSF-ATD 1830676.

***Email:** nmank@colostate.edu

MATHEMATICS
COLORADO STATE UNIVERSITY