**An R Shiny Application for Facilitation of Structure Equation Modeling to Reveal Regulatory Networks**

Day, Kevin1; Li, Jianying2; Wang, Tianyuan2; Bushel, Pierre3; Wu, Steve4; Li, Jian-Liang2

1Duke University; 2Integrative Bioinformatics Support Group, ESCBL, NIEHS; 3Biostatistics & Computational Biology Branch, NIEHS; 4Reproductive & Developmental Biology Laboratory

**Background:** Interaction of molecular activities among a genetic network can be examined by the path analysis with structural equation modeling (SEM) on gene signature data. Removing a subset of genes from the gene signature dataset allows further *in silico* investigation of the role of these genes in the given genetic network. However, the generation of the gene signatures and latent (hidden) variables are often tedious especially when verifying significance based on bootstrap resampling. As such, this creates a bottleneck in testable hypotheses when swapping in different exogenous variables in the SEM framework. Automation of the process will assist in efficiently revealing the biological functioning network, thus predicting which biological pathways are significant and should be primarily targeted in future investigations.

**Objective:** Create a user-friendly R Shiny app which will analyze genomic data to generate gene signatures for application to an SEM.

**Methods:** Open-source R and R Shiny were used for calculations and presentation. The microarray data file is normalized by keeping only the probe with the highest variation for replicate genes and centering to the median across the genes on the array. A T-score is calculated for each gene by first finding overlap of genes between the gene signature and microarray files, then by separating the induced (high) and repressed (low) genes to perform a two-sided T- test between the two groups. To examine the impact of removing a subset of genes on the genetic network, a T-score is calculated by removing these genes from the gene signature list, then proceeding with the aforementioned T-score method. To assess significance of the SEM, bootstrap resampling is performed in parallel by iteratively removing an equal random number of genes from the overall gene list as the number of genes in test, then proceeding with the T-score method. Using lavaan R package syntax, SEM model fits the T-scores of variables in the format: Y ~ X1 + X2, where Y is the endogenous variable and Xi is the exogenous variable.

**Results:** Bootstrap resampling of the data is run in parallel to assess significance of the modeling. Inferred regulatory networks are revealed and visualized through significant gene interactions. ~~We test the app on a real data set of gene expression signatures of recurrent implant failure following~~ *~~in vitro~~* ~~fertilization and show that the SEM frameworks extracts a regulatory network where~~ *~~in silico~~* ~~alteration of latent gene interactions disrupts the progesterone receptor pathway in the uterus of pregnant mice.~~

**Conclusions:** Overall, this app can be used by all scientists in efficiently creating SEM models and mathematically determining significant biological pathways from genomic data.

Last year’s academic level of presenting author: College Freshman