Structural Equation Modeling of In silico Perturbations

**Jianying Li1,2,3** †**, Pierre R. Bushel3,4** †**, Lin Lin5,6, Kevin Day7, Tianyuan Wang1,2, Francesco J. DeMayo6, San-Pin Wu6 \*, and Jian-Liang Li1 \***

1 Integrative Bioinformatics, Epigenetics and Stem Cell Biology Laboratory, Division of Intramural Research, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

2 Kelly Government Solutions, Research Triangle Park, NC 27709, USA

3 Massive Genome Informatics Group, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

4 Biostatistics and Computational Biology Branch, Division of Intramural Research, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

5 Department of Family Health Care Nursing, University of California at San Francisco, San Francisco, CA 94143, USA

6 Reproductive and Developmental Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

7 Duke University, Durham NC 27708

†These authors have contributed equally to this work and share first authorship

**\* Correspondence:**Jian-Liang Li ([jianliang.li@nih.gov](mailto:jianliang.li@nih.gov))and San-Pin Wu ([steve.wu@nih.gov](mailto:steve.wu@nih.gov))

Running Title: Structural Equation Modeling In silico

Keywords: Structural Equation Modeling, gene expression, *in silico* perturbation, molecular interaction, R, Shiny

Word count: 2249

Number of Figures: 2 Number of Supplementary Figures: 4

Number of Tables: 0 Number of Supplementary Tables: 2

# Abstract

Gene expression is controlled by multiple regulators and by the interaction among these factors’ activities. Data from genome-wide gene expression assays enable a mathetical estimation of molecular activities via a projection from the gene signature of a model system to gene expression profiles of human specimens using a T-score calculation. With the quatification of a given gene activity in each individual specimens, structural equation modeling (SEM) was able to determine the concurrent regulatory effects of two or more upstream regulators on levels or activities of a downstream reporter gene. We developed an R Shiny application, termed “Structural Equation Modeling of In silico Perturbations (SEMIPs)” to compute a two-sided t-statistic, or T-score as a surrogate gene activity in a given human specimens, which can be used in either correlation studies between outcome variables of interest or subsequent model fitting on multiple variables. This application implements a 3-node SEM model that consists of two upstream regualtors as input variables and one downstream reporter as an outcome variable to examine the significace of interactions among them. SEMIPs enables scientists of non-bioinformatic background to examine the genetic interactions among the three variables *in silico*. As a case example for SEMIPs, we showed that putative direct downstream genes of the GATA2 transcription factor are sufficient to infer GATA2’s activities *in silico* for the conserved PGR-GATA2-SOX17 genetic network in the human uterine endometrium.

# Introduction

While gene expression data in public repositories provide a valuable resource for investigators to infer regulatory processes {Edgar, 2002 #9}, the causal relationships among variables of interest are not always directly measurable in a system. Moreover, it is challenging to test the knowledge obtained from experimental model systems in humans due to undetermined clinical outcomes and ethical considerations. Genome-wide gene expression assays on human speciments allow observations of correlations among the gene expression levels as well as between RNA abundances and phenotypic outputs. Meanwhile, these assays can also determine the downstream targets of a factor of interest in model systems that are relevant to the particular type of human specimen via genetic or phamalogical perturbations. The resulting gene signature, manifested by the behavior of these downstream target genes in response to a perturbation, could unbiasly serve as a surrogate of the activity of the factor of interest in a given context. Assuming that gene functions are preserved between human tissues and relevant model systems, the degree of similarity between the gene signature of the factor of interest and the specimen’s gene expression profile could be quantitatively estimated by a T-score calculation to repsent activities of the factor of interest in the said specimen [PMID: 25295534, 19666588, 19490893, 18757322]. This scoring system have been employed to establish correlations between the prognosis outcome and manifestation of activities of the factor of interest in corresponding tumors [PMID: 25295534, 23201680, 19666588, 19490893, 18757322]. The T-score calculation has also been utilized to determine the association among activities of factors of interest or between the activities of an upstream regulator and levels of its downstream targets within a set of human specimens [PMID: 26356605, 27783953]. Results of these studies demonstrated applications of such a surrogate score of molecular activites in investigation of gene functions and inference of regulatory processes.

To determine the relations among multiple variables, structural equation modeling (SEM) is one of the statistical techniques to indicate the strength of influence among variables by getting an overall fit of model with existing data. The fit of the model can be assessed using various criteria, including the root mean square error of approximation (RMSEA), along with a 90% confidence interval, the Comparative Fit Index (CFI), the Tucker-Lewis Fit Index (TLI), and the standard root mean square residual (SRMR). For the RMSEA, the general rule of thumb is that values <.05 indicate close fit, values between .05 and .10 indicate marginal fit, and values >.10 indicate poor fit (MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods. 1996;1:130-149). For both the CFI and the TLI, a value of 1 indicates perfect fit, and the general rule of thumb is that values >.90 indicate adequate fit ([ Hu L, Bentler PM. Fit indices in covariance structure modeling: sensitive  to underparameterization model misspecification. Psychol Methods. 1998;3:424-453], [Hu L, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6:1-55]). Also, SRMR values <.08 indicate a very good fit between the model and the data. Therefore, SEM offers a statistical framework to make casual inferences about the causality of multiple variables in a system.

We were motivated to develop a Structural Equation Modeling of In silico Perturbations (SEMIPs) Shiny application to facilitate casual inference of gene regulatory processes, especially on multifactoral impacts on outcome variables concurrently. SEMIPs enables quantification of a projected activity metric (T-score) and allows users to fit desired SEM models using variables of interest. For hypothesis generation purpose, SEMIPs provides two different bootstrap random sampling procedures (elimination with or without replacement) to test the significance of a model after removing a subetset of downstream targets that are pertinent to pathways of interest in the gene signature (PMID: 18757322). Previously, the T-score system and SEM were applied to gene expression data to evaluate gene interactions that regulate the progesterone signaling pathway in the mouse uterus and inference of the gene regulation processes in a human uterine specimens {Rubel, 2016 #1}. SEMIPs streamlines this process and allows scientists of limited bioinformatic background to perform computations and analyses through a user-friendly interface.

# Materials and Methods

## Overview of SEMIPs

This SEMIPs RShiny App allows users to compute a two-sided t-statistic (T score) from gene expression data to infer the activities of genes of interest in a quantitative manner. Shown in Figure 1, SEMIPs App, which is highlighted in the orange dotted rectangle, facilitates the hypothesis generation and testing framework, this app also provides a 3-node model fitting function using structural equation modeling to test the joint regulation of a target gene by two upstream regulators *in silico*. In addition, for hypothesis generation purposes, a two-class bootstrap method, elimination with replacement or elimination without replacement, is included in the app to examine the impact of removing genes that belong to the same signaling cascade from the downstream targets of the gene of interest.

## T-score Calculation

The T-score calculation requires the input of two components, a normalized gene expression matrix of the human specimens and a gene signature of the factor of interest. To generate the normalized gene expression matrix of human tissues, such as microarray or RNAseq data, the expression values of each gene were centered to the median across all samples. If the gene had multiple probes, the probes with the highest variation (standard deviation) was chosen to represent that gene. The gene signature was first determined by identifying downstream target genes whose RNA abundance are associated with the levels of the upstream regulator in a given set of statistical criteria. These associated downstream targets were further subgrouped based on the positive (up-regulated signature) or negative (down-regulated signature) correlations on the RNA abundance between the upstream regulator and the downstream targets. The T-score was then calculated based on the following formula:

Tscore = d\*TINV(p, df);

Where as,

d =1, if the average expressions of homologous genes of up-regulated signature genes is larger than the average expressions of homologous genes of down-regulated signature genes). Otherwise, d = -1.

TINV: the function of inverting t statistic.

p: p value of 2 tailed t-test of the expressions of homologous genes of up-regulated signature genes and the expressions of homologous genes of down-regulated signature genes with equal variance.

df: degree of freedom; total number of the homologous genes of signature genes minus 2.

The biological hypothesis generation relies on results obtained from the model animal perturbation system, normally with experimental mice or rats, then projects into human or other animal systems when either direct perturbation is not possible or the variables of interest are not directly measurable. The SEMIPs RShiny App provides an automated route to calculate such T-score with a separated tab “T-Scores” shown in Figure 2. The application will conduct the analysis and produce inferred activity results that can be used in subsequent downstream analyses.

## Structural Equation Modeling

The second feature of SEMIPs App is the structural equation modeling (SEM).We implement the SEM with lavaan package {Rosseel, 2018 #6} to provide a 3-node model fitting function using structural equation modeling to test the joint regulation of a target gene by two upstream regulators *in silico*. T-scores or normalized RNA levels of two upstream regulators are the two input variables, while the outcome variable is the RNA abundance values of a chosen downstream reporter gene that are expected to be regulated by the two upstream regualtors. The app comes with a sample data file “app\_installation\_dir/dataSEM/sampleDAT.txt”. When the SEM tab is selected (Figure 2), this data will be loaded, users can select three variables from the drop-down list to test the SEM model. The SEMIPs App also provides a data file template “app\_installation\_dir/dataSEM/\_sampleDAT.txt” that users can modify and save as “sampleDAT.txt” to overwrite the default data. As a result, users’ data will be loaded when the App is launched next time. Users can save the modeling figure and all fitting statistics from the app.

## Bootstrap Simulation

The third feature (the bootstrap tab shown in Figure 2) assesses the potential impact from a perturbation on on the proposed genetic network such as removing a downstream molecular pathway or the gene signature of a downstream effector from the upstream regulator. We implemented a two-class (elimination with or without replacement) bootstrap resampling for statistical inference (Supplementary Figure 1), which eliminates unrelated signatures and provides statistical significance to the SEM fitting. For this feature, it is assumed that the users have successfully run a T-score analysis. The users also need to enter the signatures associated with the downstream system of interest to evaluate. To improve the rigor of the statistical test, it is recommended to run the bootstrap a minimum of 1,000 times. This feature involves bootstrapping simulation, it needs multicore hardware and can take more times depending on how many iterations users choose.

## Sample Data

The SEMIPs App comes with four sets of testing data and data templates for user to use the application and further modify their own data for any customized research projects. They are located at app\_installation\_dir/testData.

## Hardware and Software Requirement

SEMIPs was written in R with the Shiny package {Rstudio, 2014 #11} that is known for its light weight web development framework with shiny-related features. The lavaan package {Rosseel, 2018 #6} was used for the SEM, other depending packages will be checked at the installation and need to be installed if not already available. The application requires modern multicore CPUs for the backend parallel processes. SEMIPs was developed under Linux CentOS7 and has been successfully tested on MacOS (v. 10.14.6) and Windows10. To install and run this application, users can follow the detailed instructions provided in the README.txt file. The SEMIPs Shiny app andsource code are freely available at <https://github.com/NIEHS/SEMIPs> under the MIT license.

# Results

## An Integrated Hypothesis Generation and Testing Framework

As shown in Figure 1, the SEMIPs workflow depicts a genetic interaction among genes of interest that is initially revealed in an model systems and then tested for its manifestation in human specimens via model fitting. proposed or has been tested in model systems. SEMIPs is designed to test concurrently contributions of regulatory effects of two upstream regulators “Fac1” and “Fac2” on the expression of a downstream reporter gene “Endpoint”. Meanwhile, two-directional interactions between the two upstream regulators are also examined. Under this structure, users could test the relationships among the gene expression levels of all three variables. If a hypothesis is involved testing of molecular activities of two upstream regualtors, gene signatures of the upstream regualtors are first projected to a gene expression matrix of human specimens of interest (e.g. a dataset of microarrays that are derived from human biopsies) through the T-score calculation function. The resulting T-scores will serve as the surrogate molecular activities to test for the manifestation of the proposed genetic network in human specimens via model fitting. with SEM.

For the hypothesis generation purpose, a subet of genes that are associated with pathways of interest or downstream effectors could be removed from the upstream regulator’s gene signature as a in silico perturbation to infer the potential impact of losing the downstream signaling on the activities of the upstream regulator (PMID: 18757322). Based on the SEM model, a presumed relationship can be tested in humans by determining the significance of the inference via a non-parametric bootstrap resampling framework. The resulting perturbed pathways would help to prioritize experimentations in model systems. These workflow steps are shown within the dotted rectangle on the right side of Figure 1 with three major features implemented in the SEMPIPs App as the function tabs when the Shiny App is launched (Figure 2).

## T-Score Calculation Assisted Translational Research

The T-score was employed to project molecular activites of a gene of interest from a model system experiment to human specimens where a perturbation was not directly applicable [PMID: 25295534, 19666588, 19490893, 18757322]. In a model system, the biological replicates are randomly assigned into two groups, where one group will receive “placebo” and/or no treatment and another group will receive the perturbation treatment. Experimental measurements will be properly collected from both groups (i.e. gene expression profile from a genomewide gene expression experiment). Significantly changed genes/probes (signatures) will be obtained from this analysis according to pre-determined thresholds followed by a statistical analysis with directionality (up/down regulation). Such a list of genes/probes are deemed collectively as the “gene signature” of biological responses to a particular perturbation in a given context such as cell or tissue typs of interest. And these downstream target genes of the perturbed molecule are referred as “signature genes” of the molecule of interest. This gene signature information will be projected into the human specimen of interest bearing the assumption that the biological behavior of the gene of interest is conserved between the chosen model system and the human specimens.

In the gene expression dataset (i.e. human) that molecular activites of the factor of interest on individual samples are to be estimated, the counterparts of the signature genes were first identified and grouped based on the directionality of the signature genes. The T-scores of individual samples in the dataset were calculated by a normal t-statistics from these two groups of measurements to derive a single number as a quantitative surrogate of molecular activities of interest. Specimens with T-scores larger than 0, which share a similar signature gene expression profile from the model system, were classified as having gene signature activities and vice versa.

As an example in the SEMIPs, users can upload (1) A list of gene signature (in Entrez gene symbol format) obtained from a study of interest (e.g. Human Sig.xlsx) and (2) A gene expression data matrix that consists of gene expression profiles in a given context (e.g. HumanArray2Shiny.xlsx) located under “/app\_installation\_dir/testData/t-score/”. Once successfully uploaded, top few lines of data will be visible for preview (Figure 2). For illustration purposes, we provide both mouse signature (e.g. Mouse Sig.xlsx) and human signature files, the proper matched specie needs to be selected. T-scores will be calculated by clicking the green “Go!” button, the top 10 rows of the T-scores will be shown for preview. The users are encouraged to download the T-Scores for further analysis. Since the T-Scores are calculated from two-side T-test, the corresponding p-values are also reported (the second column in T-Scores results shown in Figure 2).

## Flexible Structural Equation Modeling

The impact of genetic interactions among regultors on downstream target genes is often tested by simultaneous manipulations on levels or activities of the regualtors in a model system. The SEMIPs app takes advantage of publicly available or existing gene expression information to examine such potential interactions *in silico* by SEM. SEMIPs supports the test of a hypothesis in which two upstream regulators (“Fac1” and “Fac2” in Figure 1) concurrently regulate the levels of one downstream reporter gene (Endpoint in Figure 1) in a 3-node model (Figure 1). The input variables for upstream regulators could be either the gene expression levels or the molecular activites in a T-score format. Our current SEM model tests both upstream regulators in a regression model on the “endpoint”, where γ11 and γ21 are the coefficients in the regression model and ε1 is the model residual (Figure 1). The model also assumes and tests the correlationshp between these two upstream regulators represented by the arc both-ended error pointing to each other. This model also examines the mutal influence between the two upstream regualtors’ activites or levels, which may serve as a predication on candidate genetic interactions between the two factors within the context of the gene expression data matrix. Operationally, once the SEM tab is selected, the default data (“app\_installation\_dir/dataSEM/sampleDAT.txt”) will be loaded, and all features are available for users to choose from the drop-down windows (Figure 2). Two variables are hypothesized as “causal variable” and one variable will be the “endpoint”. The tool reports model fitting statistics in a compressed (zipped) file that can be downloaded, the three-node SEM figure can also be downloaded. This feature also allows users to test a separate system by uploading their relevant dataset. The dataset requires the same format as the example data. Results derived from the SEMIPs could aid prioritizing wet lab experimentations and establishing clinical relevance.

## Two-class Bootstrap Simulation

Biological signaling is often transduced by a cascade of downsteam effectors in a hierachyal manner. The gene signature of an upstream regulator is usually a summary presentation of activities of multiple downstream effectors whose mRNA abundance may or may not be altered upon stimulations. In silico dissection of the contribution of effectors to the upstream regualtor’s effect has been utilized previously by removing genes that reflect the effector’s activities from the upstream regulator’s gene signature (PMID: 18757322). In SEMIPs, genes that are associated with biochemical pathways or belong to the downstream effector’s gene signature could be tested with this two-class (elimination with or without replacement) bootstrap resampling for statistical inference (Supplementary Figure 1). In the test data folder “/app\_installation\_dir/testData/bootstrap/”, four downstream gene sets are available. Under the “Bootstrap” tab, users can navigate to this location and run the bootstrap simulation analysis. The impact on the downstream system can be assessed by either elimination without replacement or with replacement. To ensure the rigor of the statistical test, it is recommended to run the bootstrap a minimum of 1,000 times. Depending on the hardware configuration, this analysis can take a considerable amount of time. Users can download the zipped results after the analysis is completed. The results derived from this function could serve as a rationale to further genetic or phamalogical experimentations.

## A User Case Application

Previously we demonstrated that the mouse gene signatures of GATA2 and PGR allow inference of the interaction between GATA2 and PGR for regulation of SOX17 expression in the human endometrial tissues {Rubel, 2016 #1}. The full GATA2 gene signature consists of both direct and indirect downstream genes of GATA2 in the uterus {Rubel, 2016 #1}. Since GATA2 is known as a transcription factor that occupies cis-acting elements and confers genomic actions, we hypothesize that expression levels of GATA2’s direct downstream targets reflect its activities *in silico.* Here, a GATA2 direct downstream target is defined as a GATA2 regulated gene with GATA2 genome occupancy within 2-kilobase vicinity of the said gene’s transcription start site in the uterus (Gene Expression Omnibus (GEO) accession: GSE40659, {Rubel, 2016 #1}). This stringent criterion led to the identification of 634 genes (Supplementary Table 1), which is termed “GATA2 direct signature”. The GATA2 activity, as represented by the GATA2 direct signature in a T-score, was quantified by the SEMIPs app from gene expression data of the endometrium tissue for each individual human subject (GEO accession: GSE58144, {Koot, 2016 #13}). T scores for the uterine GATA2 in all 115 patients were calculated by the app with the GATA2 direct signature and the data matrix of GEO accession: GSE58144 (Supplementary Table 2). Similarly, T scores for the uterine PGR (termed PGR signature) were obtained using the GEO accession: GSE39920 dataset {Rubel, 2016 #1} on the same data matrix via the application’s T score calculation function. To test whether the GATA2 direct signature fits the model of the 3-node PGR-GATA2-SOX17 genetic network, the application was fed with T scores of GATA2 direct signature and PGR signature as exogenous variables and the SOX17 expression levels as the endogenous variable under the “SEM” function. The output data shows that, with GATA2 direct signature in place of the full gene signature, the model significantly fits the GEO accession: GSE58144 dataset with all proposed paths (Supplementary Figure 2) and this model is considered not rejected by the human data. This finding suggests that the expression levels of GATA2 direct downstream targets, a subset of the full GATA2 regulated genes, can mathematically serve as surrogate reporters of the GATA2 activities in the human endometrium tissues, which supports our hypothesis. Results of this analysis not only reduce the number of reporter genes for GATA2 activities to 634, but also implicate possibilities of a further reduction with additional filtering criteria on the gene list. A small and manageable panel of markers for GATA2 activities could serve as a future diagnostic tool for pregnancy failure {Díaz-Gimeno, 2011 #16}.

# Discussion

The SEMIPs R Shiny app offers an easy to use *in silico* perturbation testing system with several advantages. First, it has capability to calculate response activities using large datasets representative of biological systems. Second, it leverages the power of SEM to test the relationship among end points in a study and provides users with the flexibility for testing new hypotheses. Lastly, it integrates a non-parametric testing procedure for assessing statistical significance.

This user-friendly app allows quick assessments on genetic interactions and subsequent hypothesis generation without the requirement of extensive knowledge on computation languages and statistical analyses. Due to its simplicity in design, this app is limited to a 3-node model fitting capability. Models of higher complexity can be tested using the R package MplusAutomation that focuses on automating the SEM modeling which was originally implemented in Mplus {Hallquist, 2018 #12}, a commercial software.

Currently, the two-class bootstrap analysis can only be conducted separately. Integration of these into the SEMIPs methodology for formulation into a single test will be investigated for future design, development and implementation. As noted in the manuscript and mentioned previously, the SEMIPs app has been adopted by wet lab researchers with a few papers published recently {Liu, 2019 #8;Wetendorf, 2020 #7}. We hope that it can serve a wider research community to address additional scientific questions.

**Author Contributions**

JYL and PRB designed the framework, performed the analyses and drafted the paper. LL provided the guidance on SEM and wrote part of the manuscript. KD developed and draft the Rshiny code. TW prepared gene signatures, processed gene expression matrix data, and wrote part of the manscript. SPW wrote part of the manuscript. JLL, SPW, and FJD conceived the idea, provided overall guidance and oversaw the project progression.

**Funding**

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences Z1AES103311 (FJD), Z99-ES999999 (SPW), Z01-ES102345 (PRB) and by the Gaine Research Foundation GRF-2018-01 (LL).

**Acknowledgments**

The authors would like to thank our colleagues Drs. Hamed Bostan, Eric Thomson, James Ward and Matt Wheeler for kindly testing SEMIPs and for providing valuable feedbacks to improve the application. We also thank for Drs. John House and Rong Li for their critique of the draft of this manuscript.

**Conflict of Interest**

The authors would declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Reference

# Figure Legends

**Figure 1**. The workflow and application of SEMIPs. The left four rectangles and arrows indicate our hypothesis testing and generation schema; the components bounded by dotted orange rectangle are features provided in the web-application. A biological hypothesis is tested in a model animal system (mouse) on relationship between two interacting factors (Fac1 & Fac2) and their endpoint. The hypothesis is translated to another species (i.e. human in our research) via T-score computation (represented by the upper blue arrow noted as “assisted by”) and verified with SEM model (represented by the lower blue arrow noted as “achieved through SEM”). This process is accomplished with our shinyapp indicated by two curved arrows. γ11 and γ21 are correlation efficient and ξ1 is the model residual. The two-class bootstrap analysis is shown in the red rectangle box. Hypothesis generating and exploring steps are explained by the bottom two rectangles.

**Figure 2**. The user interface is shown when it is launched. The main panel contains four tabs: “T-Scores”, “SEM”, “Bootstrap”, and “Instruction”. The right panel shows the screen when the “T-scores” is selected and generated. In the left panel, the application accepts two inputs, 1) a list of signatures (in Entrez gene symbol format) and 2) a data matrix of expression measurement with the top lines shown for viewing. The green “Go!” button is clicked to launch the T-score generation and grayed out to denote the process is running. The first 10 rows of the T-scores matrix are shown, which can be downloaded by clicking the “Download T-Scores” button.