SEMIPs: Structural Equation Modeling of In silico Perturbations

**Jianying Li1,2,3** †**, Pierre R. Bushel3,4** †**, Lin Lin5, Kevin Day6, Tianyuan Wang1,2, Francesco J. DeMayo7, San-Pin Wu7 \*, and Jian-Liang Li1 \***

1 Integrative Bioinformatics, Epigenetics and Stem Cell Biology Branch, 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 Duke University, Durham NC 27713

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

†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))

San-Pin Wu ([steve.wu@nih.gov](mailto:steve.wu@nih.gov))

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

**Abstract**

**Summary**: Structural Equation Modeling (SEM) is a statistical approach for studying complex cause-effect hypotheses in a “closed system” of latent (hidden) endogenous variables. SEM has been widely used in various fields involving perturbations and measurable outcomes. We developed an R Shiny application, termed “Structural Equation Modeling of In silico Perturbations (SEMIPs)” to aid in the transfer of perturbations in gene expression pathways from one system to another for determining casual inference of molecular interactions *in silico*. SEMIPs computes a two-sided t-statistic (T score) to rank signature gene activities for modeling. It implements a basic SEM model and then performs bootstrap random sampling for statistical significance. As a use 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

Although gene expression data in public repositories provide a valuable resource for investigators to infer regulatory processes (Edgar et al. 2002), the variables of interest are not always directly measurable in a causal response model system. Moreover, it is challenging to test the knowledge obtained from experimental model systems in humans due to undetermined clinical outcomes and ethical considerations. Structural equation modeling (SEM) offers a statistical framework to make casual inferences about the causality of latent (hidden) endogenous variables in a system (Grace 2006). We were motivated to develop a Structural Equation Modeling of In silico Perturbations (SEMIPs) Shiny application to facilitate casual inference from *in silico* alterations of gene expression pathways. SEMIPs enables quantification of a projected activity metric (two-sided t-statistic, i.e. T score) calculated from gene expression activity upon exposure to a perturbation (Wu et al. 2015), thus allowing users to fit desired SEM models using selected endogenous and exogenous variables. This application also provides two different bootstrap random sampling procedures (elimination with or without replacement) for testing the significance of a model based a non-parametric distribution.

Previously, SEM was applied to gene expression data to evaluate an alteration of latent gene interactions that disrupts the progesterone receptor pathway in the uterus of pregnant mice and the model was then transferred *in silico* to a human reproductive system (Rubel et al. 2016). SEMIPs streamlines this process and allows bench scientists to perform the computations and analysis through a user-friendly interface.

# Method

**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-way 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 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. Under an experimental perturbation, the genomics system response was measured and exemplified through the significantly changed gene set. Such an information was projected into another system via gene orthologues and the activities of genes of interest will be calculated by t-statistic (T score)

Structural Equation Modeling

Bootstrap simulation

Sample Data

**Hardware and software requirement** SEMIPs was written in R with the Shiny package (Rstudio 2014) that is known for its light weight web development framework with shiny-related features. The lavaan package (Rosseel 2018) 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

**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 et al. 2016). The full GATA2 gene signature consists of both direct and indirect downstream genes of GATA2 in the uterus (Rubel et al. 2016). 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 et al. 2016)). This stringent criterion led to the identification of 634 genes (Supplemental 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 et al. 2016)). 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 (Supplemental Table 2). Similarly, T scores for the uterine PGR (termed PGR signature) were obtained using the GEO accession: GSE39920 dataset (Rubel et al. 2016) 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 (Supplemental Figure 1) 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 et al. 2011).

# Discussion

# Author Contributions

~~The Author Contributions section is mandatory for all articles, including articles by sole authors. If an appropriate statement is not provided on submission, a standard one will be inserted during the production process. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work. Please see~~ [~~here~~](http://home.frontiersin.org/about/author-guidelines#AuthorandContributors) ~~for full authorship criteria.~~

# Data Availability Statement

The datasets for this study can be found in the Gene Expression Omnibus (GEO) accession: GSE40659, GSE58144, and GSE39920 (https://www.ncbi.nlm.nih.gov/geo/)

# 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.