OXFORD UNIVERSITY PRESS | Bioinformatics

SEMIPs: Structural Equation Modeling of In silico Perturbations

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2021-0974
Category:	Applications Note
Date Submitted by the Author:	07-May-2021
Complete List of Authors:	Li, Jianying; National Institute of Environmental Health Sciences, Integrative Bioinformatics Bushel, Pierre; National Institute of Environmental Health Sciences Lin, Lin; University of California San Francisco, Department of Family Health Care Nursing Day, Kevin; Duke University Wang, Tianyuan; National Institute of Environmental Health Sciences, Integrative Bioinformatics DeMayo, Francesco J; National Institute of Environmental Health Sciences Wu, San-Pin ; National Institute of Environmental Health Sciences Li, Jian-Liang; National Institute of Environmental Health Sciences, Integrative Bioinformatics
Keywords:	

SEMIPs: Structural Equation Modeling of In silico Perturbations

Jianying Li[1,2,3,8], Pierre R. Bushel[3,4,8], Lin Lin[5], Kevin Day[6], Tianyuan Wang[1,2], Francesco J. DeMayo[7], San-Pin Wu[7 *], and Jian-Liang Li[1 *]

[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

[8]These authors contributed equally

* Correspondence
Jian-Liang Li (jianliang.li@nih.gov)
San-Pin Wu (steve.wu@nih.gov)

Running Title: Structural Equation Modeling for In silico Perturbations

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

Word Count (excluding references): 1026

Total number of figures: 1

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.

Availability and implementation: The SEMIPs Shiny app and source code are freely available at https://github.com/NIEHS/SEMIPs under the MIT license. SEMIPs is developed in R.

Contact: **Jianying.li@nih.gov**

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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.

Implementation and usage

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. 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.

As shown in Figure 1A, the SEMIPs workflow depicts a biological question initially tested in an animal model and then applied to a human system. 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 can be eventually tested in the animal model. These workflow steps are shown within the dotted rectangle on the right side of Figure 1A.

The SEMIPs Shiny application has three main features. The first feature allows users to quantify the projected "regulator activity" of the gene of interest from a study in the form of T scores (Wu, S.P. et al 2015, Liu et al. 2019, Wetendorf et al. 2020). The "T Scores" tab (Figure. 1B) was designed in the main panel to conduct such an analysis. It requires two components: (1) A list of gene signature (in Entrez gene symbol format) obtained from a study of interest; and (2) A gene expression data matrix that consists of gene expression profiles in a given context. The application will conduct the analysis and produce inferred activity results reflected as T scores that can be used in subsequent downstream analyses. The second feature (the SEM tab) provides users a convenient SEM model fitting interface with the T scores. The users can hypothesize a 3-

node structural equation model by selecting the desired endogenous and exogenous variables. The tool reports model fitting statistics in a compressed (zipped) file. 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. The third feature (the bootstrap tab) assesses the potential impact from a perturbation on any downstream system. We implemented a two-class (elimination with or without replacement) bootstrap resampling for statistical inference, 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.

The application provides a user-friendly interface (Fig. 1B) with special features that are provided as separate tabs. The data needs to be in the same format as shown under the "Instructions" tab. Further details for running the application are provided under the "Instructions" tab.

Case study

With the 634 putative GATA2 direct downstream genes, SEMIPs provided evidence to infer GATA2's activities that permitted modeling fitting of the PGR-GATA2-SOX17 genetic network in the human endometrium (See the Supplemental information for details).

Conclusion

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.

Acknowledgements

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.

Funding

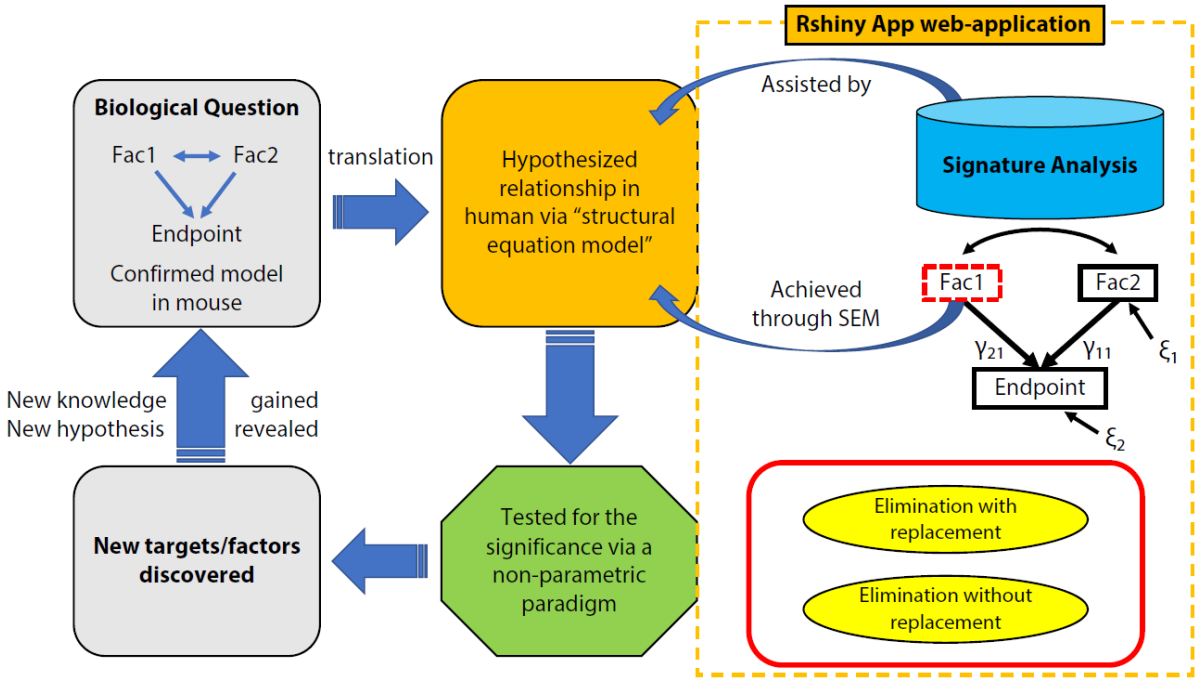
This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences Z1AES103311 (FJD), Z99-ES999999 (SPW), Z01-ES102345 (PRB) and by the Gaine Research Foundation GRF-2018-01 (LL).

References

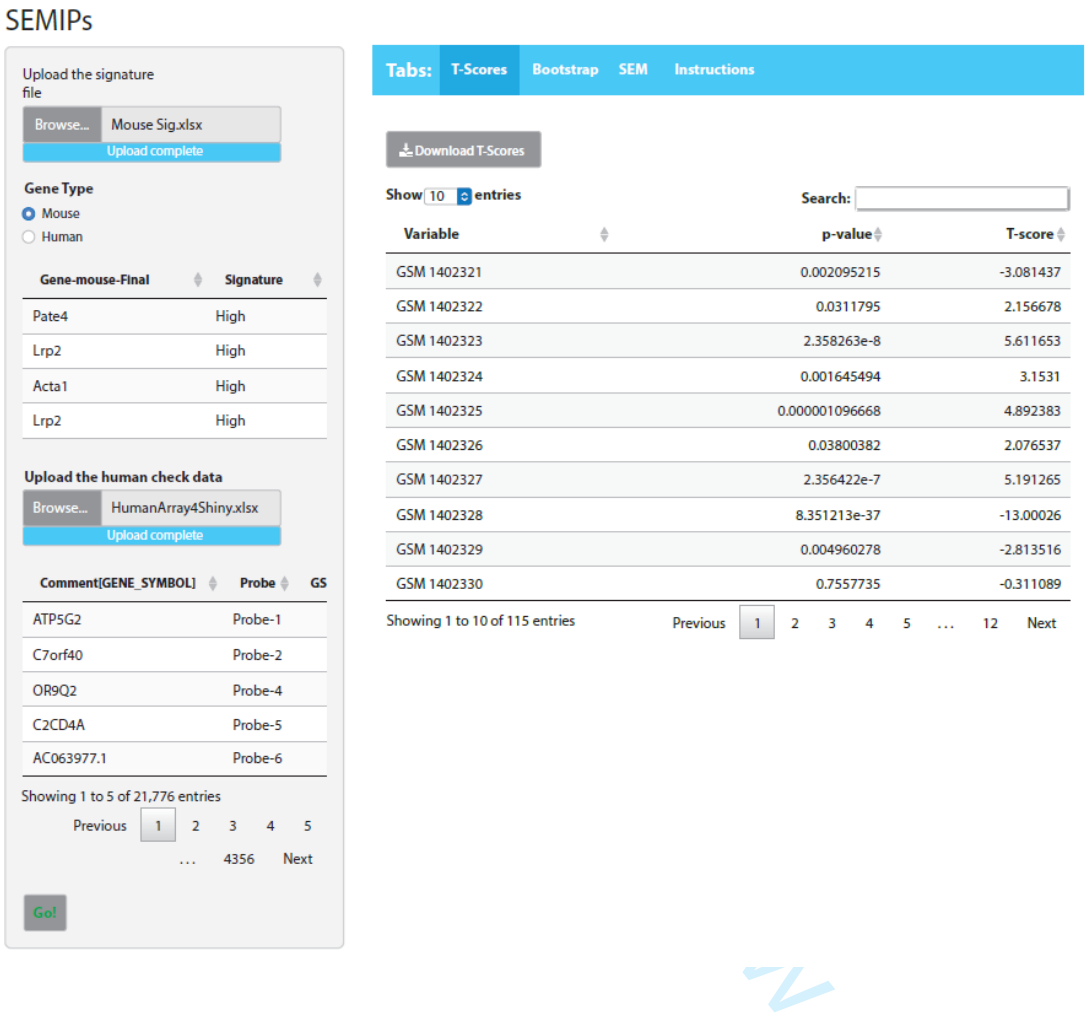
Edgar, R., et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.
Grace, B. J. (2006). Structural Equation Modeling and Natural Systems, Cambridge University Press.
Liu, J., et al. (2019). "JNK(1/2) represses Lkb(1)-deficiency-induced lung squamous cell carcinoma progression." Nat Commun **10**(1): 2148.
Rosseel, Y. (2018). "Latent Variable Analysis."
Rstudio, I. (2014). "Shinny: Easy web applications in R."
Rubel, C. A., et al. (2016). "A Gata2-Dependent Transcription Network Regulates Uterine Progesterone Responsiveness and Endometrial Function." Cell Rep **17**(5): 1414-1425.
Wetendorf, M., et al. (2020). "Constitutive expression of progesterone receptor isoforms promotes the development of hormone-dependent ovarian neoplasms." Sci Signal **13**(652).
Wu, S. P., et al. (2015). "Increased COUP-TFII expression in adult hearts induces mitochondrial dysfunction resulting in heart failure." Nat Commun **6**: 8245.

Figure 1. The SEMIPs web application. **A**. The workflow and application of SEMIPs. The left four rectangles and arrows indicate our hypothesis testing and generation schema. A biological hypothesis is tested in a model animal (mouse) on relationship between two interacting factors (Fac1 & Fac2) and their endpoints. The hypothesis is translated to another species (i.e. human in our research) via T-score computation and verified with SEM model. This process is accomplished with our shinyapp indicated by two curved arrows. γ_{11} and γ_{21} are correlation efficient and ξ are model residuals. The two-class bootstrap analysis is shown in the red rectangle box. Hypothesis generating and exploring steps are explained by the bottom two rectangles. **B.** 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. The left panel shows that 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 that 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.

A



B


SEMIPs
Upload the signature file
Browse... Mouse Sig.xlsx
Upload complete
Gene Type
Mouse
Human
Gene-mouse-Final
Signature
Pate4 High
Lrp2 High
Acta1 High
Lrp2 High
Upload the human check data
Browse... HumanArray4Shiny.xlsx
Upload complete
Comment[GENE_SYMBOL]
Probe
GS
ATP5G2 Probe-1
C7orf40 Probe-2
OR9Q2 Probe-4
C2CD4A Probe-5
AC063977.1 Probe-6
Showing 1 to 5 of 21,776 entries
Previous 1 2 3 4 5 ... 4356 Next
Go!
Tabs: T-Scores Bootstrap SEM Instructions
Download T-Scores
Show 10 entries
Search:
Variable
p-value
T-score
GSM 1402321 0.002095215 -3.081437
GSM 1402322 0.0311795 2.156678
GSM 1402323 2.358263e-8 5.611653
GSM 1402324 0.001645494 3.1531
GSM 1402325 0.000001096668 4.892383
GSM 1402326 0.03800382 2.076537
GSM 1402327 2.356422e-7 5.191265
GSM 1402328 8.351213e-37 -13.00026
GSM 1402329 0.004960278 -2.813516
GSM 1402330 0.7557735 -0.311089
Showing 1 to 10 of 115 entries
Previous 1 2 3 4 5 ... 12 Next