Revealing the Biological Network via a Statistical Framework

-- Structural Equation Modeling

Manuscript Overview

November 13th, 2019

Jianying Li1, Lin Lin2, Pierre Bushel3, Kevin Day5, Tianyuan Wang1, Francesco J. DeMayo4, San-Pin Wu4, and Jian-Liang Li1

**Abstract**

A previous application of Path Analysis confirmed an important biological model for the transcriptional network regulating uterine progesterone responsiveness and endometrial functions. The Structural Equation Model (SEM) is a statistical modeling approach for studying of complex cause-effect hypotheses in a closed system. SEM has been widely used in various fields with involving perturbations and measurable outcomes. Upon successfully establishing a framework in R (an open source program), we extended a previous model and leveraged the SEM features to examine broader systems involving regulators and their endpoints. In addition, a user-friendly R Shiny app was developed to infer the activity of human reproductive regulators based on the mouse gene expression profiles in an automated fashion. We started from an *in vitro* alteration of latent gene interactions that disrupted the progesterone receptor pathway in the uterus of pregnant mice and then projected to a human reproduction system via an *in silico* route. The results eventually revealed a regulatory network in human that has the potential to facilitate human infertility study. In conclusion, by leveraging the features in the SEM framework, we established a translational framework that revealed a functional network system in a human reproduction system from a laboratory established mouse-model. In the future, we plan to extend this framework to other complex systems for in-depth elucidation of biological networks and hidden regulatory relationships.

**Introduction**

Successful pregnancy requires the endometrium, the inner lining of the uterine, to prepare for receiving and supporting the embryos. This process is controlled by the pregnancy hormone progesterone and its cognate receptor PGR. Dysfunctional progesterone signaling often leads to pregnancy failure. *In vivo* research model established in mouse revealed that PGR and GATA2 transcription factors work together to mediate the progesterone signaling (Rubel, Wu et al. 2016), such a relationship was proved to be conserved between human and mouse via a SEM framework. The present study aims to implement this SEM framework in an open source R programming language so that we can streamline the process in an automated fashion. This will allow us to test hypotheses and reveal novel factors that could perturb or enhance the casual relationship. Additionally, we constructed a graphical user interface in R shiny. It allows bench scientists to leverage gene expression signature analysis and calculate “regulator activity” (Wu, S.P. et al 2015). Ingenuity Pathway Analysis (IPA) on the GATA2 significant gene lists reveals many upstream regulators whose downstream target genes comprise of an informative gene set. Our study suggests that those upstream regulators could potentially enhance or disturb the existing relationship, which themselves could be targets of interest for further scientific hypothesis testing. Finally, we use a non-parametric distribution to provide statistical significance of the SEM results. In summary, we have shown the applicability and potential of such a SEM framework to make inferences of cause and effect gene network perturbations by leveraging gene expression data from a different species model system. With the flexibility of open source environment, such a framework can be integrated into analytical pipelines for streamlining and automation.

**PGR, Progesterone and GATA2 signatures identification**

The Partek Genomics Suite 7.17 software (Partek Inc., St. Louis, MO) was utilized to

process raw data from CEL files. The Robust Multichip Analysis (RMA) algorithm with

quantile for normalization and log2 transformation was applied to preprocess the data. A one-way ANOVA model was used to compare expression profiles from different groups. Differentially expressed genes (DGE) were defined using an unadjusted p value < 0.01 and absolute fold change >1.3.

*DEGs identified: GATA2 (1,913), PGR (211), and Progesterone (129)*

**GATA2, PGR, and Progesterone activities in human dataset (GSE58144)**

The publicly available human recurrent implantation failure (RIF) array dataset

GSE58144 was scored for manifestation of the mouse model-derived GATA2, PGR,

and Progesterone signature, using published methods (Qin et al., 2013). For multiple probes mapping to the same gene, the probe with the highest variation was chosen to represent the gene. For each gene, the profiles were centered to the median across samples.

**Structural Equation Model (SEM) for biological network**

The hypothesized networks derived from the previous study (Rubel, C.A. 2016) on the functional relationship among endogenous/exogenous variables were tested for causal relationship in a SEM framework. All SEMs were implemented in R (Figure 1) with lavaan package (Oberski 2014). A SEM was fitted in each network and evaluated with a chi-square goodness of fit test. The fit was evaluated with different indices including Tucker-Lewis Index, Comparative Fit Index, RMSE, etc.

**Signature Analysis web development**

The web-based application was developed through integrating publicly available R packages and internally developed R functions by using the Shiny framework in R environment (Figure 2A). The Signature Analysis is for calculating the functioning activity (Wu, S.P. et al 2015) that is later used in the SEM. This application is accessible within NIEHS network and the source codes are deposited in GitHub and a docker image is also available.

**Mechanistic analysis**

Ingenuity Pathway Analysis (Qiagen) was performed on GATA2 significant DEGs. Significant upstream regulators with p-value < 0.01 were reported

**Significance assessment**

To assess the significance of the analysis, we implemented two bootstrap strategies (Figure 3A). An overall diagram of two bootstrap strategies. On the left hand, we randomly select and exclude the same number of genes (N) as downstream target of an upstream regulator of interest. The re-calculate the “GATA2 activity” with our R Shiny app, then rerun the SEM with an updated GATA2 activity. On the right, we create a universe candidate gene lists using all genes on the human array by removing the original GATA2 significant genes list, plus PGR significant gene list, and SOX17. We exclude the downstream target of an upstream regulator of interest, then the randomly select the same number of genes (N) from the universe and add it back to the shrunk list, eventually, “GATA2 activity” with our R Shiny app, then rerun the SEM with an updated GATA2 activity. With each strategy design, we iterate 1000 times. To test whether an upstream regulator plays a significant role, we use a non-parametric distribution-based test (Figure 4 diagram). The significance was indicated by the final index (I): the smaller the index, the greater the impact.

**Tables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regulator** | **Molecule Type** | **SEM p-value (GATA2)** | **SEM p-value (PGR)** | **p-value (IPA)** | **NumOfDownstreamGenes** | **Predicted Activation State (IPA)** |
| ATF4 | transcription regulator | 0.0709 | 1.24E-11 | 0.000000252 | 34 | Inhibited |
| beta-estradiol | chemical - endogenous mammalian | 0.2264 | 2.02E-12 | 6.02E-24 | 289 | Inhibited |
| EGFR | kinase | 0.1184 | 6.02E-12 | 1.1E-10 | 73 | Inhibited |
| ERN1 | kinase | 0.0456 | 1.93E-11 | 1.58E-08 | 31 | Inhibited |
| FOXO3 | transcription regulator | 0.0513 | 1.40E-11 | 3.87E-10 | 62 | Activated |
| HOXA10 | transcription regulator | 0.0353 | 4.27E-11 | 2.74E-08 | 39 | Inhibited |
| MITF | transcription regulator | 0.0483 | 2.29E-11 | 5.91E-02 | 27 | Activated |
| progesterone | chemical - endogenous mammalian | 0.0927 | 4.72E-12 | 1.03E-20 | 124 | Inhibited |
| PTEN | phosphatase | 0.0741 | 1.78E-11 | 4.30E-08 | 83 | Activated |
| RUNX3 | transcription regulator | 0.0399 | 2.96E-11 | 3.10E-03 | 20 | Inhibited |
| TGFBR1 | kinase | 0.0327 | 5.00E-11 | 8.22E-03 | 14 | Activated |
| Vegf | group | 0.0699 | 6.57E-12 | 5.22E-06 | 72 | Inhibited |
| XBP1 | transcription regulator | 0.0276 | 5.68E-11 | 2.88E-09 | 45 | Inhibited |

Table 1. **.** Function of 13 upstream regulators whose downstream target genes were parts of the GATA2 significant gene list and were tested in the SEM model. Most of them played a role in affecting the GATA2 activity which were further reflected by the SEM (p-value of GATA2).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Upstream regulator** | **Molecular type** | **Index(I)** | **Impact** | **NumOfDownstreamGenes** | **IPA (p- value)** |
| BCL6 | transcription regulator | **<0.001** | **Favorable** | 28 | 9.96E-03 |
| BRCA1 | transcription regulator | **<0.001** | **Favorable** | 32 | 2.07E-05 |
| CTNNB1 | transcription regulator | **<0.001** | **Favorable** | 92 | 2.50E-04 |
| EHMT2 | transcription regulator | **<0.001** | **Favorable** | 9 | 7.90E-03 |
| EP300 | transcription regulator | **<0.001** | **Favorable** | 65 | 3.44E-06 |
| MAX | transcription regulator | **<0.001** | **Favorable** | 20 | 2.82E-05 |
| NFYA | transcription regulator | **<0.001** | **Favorable** | 15 | 3.68E-03 |
| PML | transcription regulator | **<0.001** | **Favorable** | 22 | 1.42E-03 |
| SIN3A | transcription regulator | **<0.001** | **Favorable** | 13 | 1.53E-03 |
| SIN3B | transcription regulator | **<0.001** | **Favorable** | 7 | 1.11E-03 |
| SP1 | transcription regulator | **<0.001** | **Favorable** | 80 | 1.66E-06 |
| SP3 | transcription regulator | **<0.001** | **Favorable** | 39 | 3.25E-06 |
| TP73 | transcription regulator | **<0.001** | **Favorable** | 47 | 8.98E-03 |
| TWIST1 | transcription regulator | **<0.001** | **Favorable** | 23 | 5.50E-03 |
| YAP1 | transcription regulator | **<0.001** | **Favorable** | 17 | 5.31E-03 |
| ZNF282 | transcription regulator | **<0.001** | **Favorable** | 3 | 2.16E-03 |
| EZH2 | transcription regulator | **0.0023** | **Favorable** | 45 | 3.04E-03 |
| TCF3 | transcription regulator | **0.0024** | **Favorable** | 41 | 7.42E-03 |
| **EGFR** | **kinase** | **<0.001** | **Favorable** | **73** | **1.10E-10** |

**Table 2** A test analysis of 97 transcription regulators plus several pregnancy-related factors that were identified by an IPA upstream analysis. Only top 20 best index score results were shown. The initial 13 regulators used for the SEM model testing were also included and shown in blue.

**Figures and legends**

**Figure 1.** Overall workflow

**Figure 2.** Testing the biological hypothesis.

**Figure 3.** A screenshot of the Signature Analysis R Shiny application. It has four tab and the last tab shows the instruction of how to use this application. Two QR codes are for code repository on GitHub and a docker image.

**Figure 4A** A two-fold bootstrap strategies. On the left hand, we coin the name “elimination without replacement”; whereas the one on the right, “elimination with replacement”.

**Figure 4B.** Based on the two-fold bootstrap strategies, an index value is calculated to infer the impact significance

**Supplemental table**

**S1.** Differentially expressed genes (DEGs)/probes from GATA2 knockout study (Rubel, C.A. et al. 2016). The analysis was done with Partek and DEGs were obtained at different threshold cut offs (GATA2\_dff\_probes\_at\_thresholds.xlsx)

**S2.** IPA analysis was performed on GATA2 DEGs list, upstream regulator results were retained (all\_UpstreamRegulator\_from\_IPA.xlsx)

**S3.** Upstream regulator results from IPA were retained at p-value less than 0.01 and for “transcription regular” only. (UpstreamRegulator\_TranscriptionRegulator\_pValue\_lessthan\_01.xlsx)

**Supplemental figures**

**F1.** Validating the results from M-Plus with newly developed R-program

Our paper will focus on the following four main contribution

1. Successfully establish the R environment for SEM framework
2. Successfully leverage the SEM concept in genomic/biology research (SEM paper in GWAS)
3. R-shinyapp for SEM and t-score implementation for wider user’s scope
4. New targets were identified for biology
5. Extend the model to include more variables

Reference document

<https://www.kdnuggets.com/2017/03/structural-equation-modeling.html>

SEM variables

In the path diagram above, ovals represent factors, also known as latent variables, unobserved variables or unmeasured variables in SEM lingo. These are theoretical concepts which can be inferred but not directly measured

**Results**

1. Translational projection from mouse to human, automatically deriving “functional factor” activity (Rshinny application)
2. Screening platform for biological hypothesis generation, validation and testing (simple model to complex models)
3. Testing unseen and untestable relationship among important factors the play role in a systematical way
4. Biological hypothesis testing
5. Non-parametric testing scheme for upstream regulators function testing

**Abstract**

A previous application of Path Analysis confirmed an important biological model for the transcriptional network regulating uterine progesterone responsiveness and endometrial functions. The Structural Equation Model (SEM) is a statistical modeling approach for studying of complex cause-effect hypotheses in a closed system. SEM has been widely used in various fields with involving perturbations and measurable outcomes. Upon successfully establishing a framework in R (an open source program), we extended a previous model and leveraged the SEM features to examine broader systems involving regulators and their endpoints. In addition, a user-friendly R Shiny app was developed to infer the activity of human reproductive regulators based on the mouse gene expression profiles in an automated fashion. We started from an *in vitro* alteration of latent gene interactions that disrupted the progesterone receptor pathway in the uterus of pregnant mice and then projected to a human reproduction system via an *in silico* route. The results eventually revealed a regulatory network in human that has the potential to facilitate human infertility study. In conclusion, by leveraging the features in the SEM framework, we established a translational framework that revealed a functional network system in a human reproduction system from a laboratory established mouse-model. In the future, we plan to extend this framework to other complex systems for in-depth elucidation of biological networks and hidden regulatory relationships.

**Introduction/rationale**

Successful pregnancy requires the endometrium, the inner lining of the uterine, to prepare for receiving and supporting the embryos. This process is controlled by the pregnancy hormone progesterone and its cognate receptor PGR. Dysfunctional progesterone signaling often leads to pregnancy failure. *In vivo* research model established in mouse model revealed that PGR and GATA2 transcription factors work together to mediate the progesterone signaling (Rubel, Wu et al. 2016). This progesterone responsive genetic network is conserved between human and mouse as indicated by the Path Analysis using the M-Plus algorithm (Rubel, Wu et al. 2016). The present study aims to extend the Path Analysis framework to test additional candidate factors and potential endpoints to further investigate this causal-responsive network. We established such platform with an open source R language and lavaan package developed by Yves Rossel in Belgium (Oberski 2014) and tested it with two established gene candidate lists: one is a clinical testing panel for human infertility (Diaz-Gimeno, Horcajadas et al. 2011) and one intrinsic list (personal communication with Wu, S). We also established a non-parametric testing strategy to provide statistical significance metrics on the testing procedures. In summary, we have shown the applicability and great potential of such a well-established framework in the genetic and molecular biology research. With the flexibility of open source environment, such framework can be integrated into analytical pipelines for streamlining and automated processes.

**Methods**

**PGR, P4, and GATA2 signatures identification** The Partek Genomics Suite 7.17 software (Partek Inc., St. Louis, MO) was utilized to process raw data from CEL files. The Robust Multichip Analysis (RMA) algorithm with quantile for normalization and log2 transformation was applied to generate signal values of all samples. The one-way ANOVA model was used to compare expression profiles from different groups. Differentially expressed genes (DGE) were defined using the filters of ANOVA unadjusted p value < 0.01 and absolute fold change >1.3.

*DEGs identified: GATA2 (1,913), PGR (211), and P4 (129)*

**GATA2, PGR, and P4 activities in human dataset (GSE58144)** The publicly available human recurrent implantation failure (RIF) array dataset GSE58144 was scored for manifestation of the mouse model-derived GATA2, PGR, and P4signature, using published methods (Qin et al., 2013). When multiple probes in GSE58144 were used for the same gene, the probe with the highest variation was chosen to represent the gene. For each gene, the profiles were centered to the median across samples.

**Path analysis for biological network** The hypothesized network was derived from previous study on the functional relationship among GATA2, PGR, and other important genes in the system with various end points signature. These components were placed in the structural model as either exogenous or endogenous variable. The normalized gene expression or functional component activities were used in the SEM model. All SEM models were implemented in R with lavaan package (Oberski 2014), after initial validation with the commercial software M-Plus. SEM model was fitted in each proposed Model and tested with chi-square. The fit was evaluated with different indices including Tucker-Lewis Index, Comparative Fit Index, RMSE, etc.

**Significance assessment** In this research, we tested our framework on a genomic diagnostic transcriptomic panel – 238 genes (Diaz-Gimeno, Horcajadas et al. 2011). To test for potential significance, we implemented a non-parametric testing. Through repeating1000 processes, for each process, random select a same size of gene panel and fit the SEM model, a summary statistic was recorded. The empirical distribution of the summary statistics served as the null distribution to assess the significance for a predefined realization panel.

**Datesets** Mouse: Genome-wide profiling of progesterone receptor and GATA2 binding in the mouse uterus (GSE34902. Affymetrix Mouse Genome 430 2.0 Array). Mouse: Identification of murine uterine genes regulated in a ligand-dependent manner by the progesterone receptor (GSE39920. Affymetrix Murine Genome U74A Version 2 Array) .Human: The endometrial gene expression signature of recurrent implantation failure after IVF  (GSE58144. A-UMCU-HS44K-2.0)