**SEMIPs: online tool for data processing and analysis**

SEMIPs (Structure Equation Modeling for In-silico Perturbations) is a shinyapptool developed to perform data processing and analysis using the Structure Equation Model to facilitate an *in-silico* perturbation system in our mouse/human infertility research. Here are a few test cases for internal reviewers/users before we can proceed with our manuscript preparation.

**Installation**

To install and run this app, please follow the detail instruction in the README.txt. Please be noted, currently, we are NOT going to release the online portal version at shinnyapps.io until we get a server set up with proper security compliances. Please only test local installation and running.

There are three major features, as noted by the navigating tabs, need to be tested.

**T-score calculation**

A “t-score” that represents a gene signature score was defined and reported to use earlier (Wu, Kao et al. 2015) for each external profile. This feature has been extensively used and reported in a recent research on Gata2-dependent transcription network study (Rubel, Wu et al. 2016), however, it has been handled manually and limits the integration of such process in an automated pipeline analysis.

To calculate the T-score with the SEMIPs app, the researchers will need two files: (1). A significant gent list and (2). An expression data matrix. To test this feature, please use the data provided within the package located “/installationRoot/testData/” (figure 1).

We provide two signature gene list, one for human and one for mouse. After a user browses the directory and uploads the signature file, he needs to check one of the radio buttons that matches the specie. Once the file is successfully uploaded, the first few rows of data will be displayed.

The users can use their “own significant” gene list and the “expression file”. The gene list needs to use Entrez gene symbol as the “id”, and the expression file needs to be in the same format as the example.

There is only one “expression file” to test (HumanArray4Shinny.xlsx), it needs to be uploaded also. Once the file is successfully uploaded, the first 5 rows of data will be displayed.

Once both files are uploaded successfully, the user can click the green button “Go!” and the T-scores results will be shown under “T Scores” tab with the first 10 rows of data will be displayed. The user can click the “Download T-scores” to save the results locally.

Graphical user interface, website

Description automatically generated

Figure 1 The test data (for T-scores) location

**Bootstrap simulation**

In our research, we proposed for a two-category bootstrap (please see the graphically illustration under the “Instruction” tab) strategy. To test this feature, the user needs to run T-score calculation first.

The test data is provided under “/installationRoot/testData/bootstrapData/” in figure 1. We tested four different datasets, and tester is encouraged to use a small set “KEGG\_4\_boostrap.xlsx”. After uploading the data properly, the top rows will be shown. The user needs to select either of the two categories by clicking the corresponding ratio button, the user also needs to select bootstrap repetition. The default is 1000 bootstraps, but the testers are highly recommended to use “small number i.e. 10). Once these are done, the user can click green “Go” button and a processing status bar will be shown until it is finished.

**Structural Equation Model**

One of our effort in this research is to use an open source software R to implement the Structure Equation Model so that we can integrate this into our automated pipeline. Under the SEM tap, we provide a simple three variable SEM model as a prototype. By default, the SEM is implemented on a pre-loaded data and all three variables are selected. The users have the option to select different two “exogenous variables” and one endogenous variable from the drop-down list. To the right of the selection is the fitting with the lavaan package (Rosseel 2018), and a graphical display is shown below the text output. If the user is interested in saving the result, he can click the orange “Download Zip” button to save the result locally.

If the users want to use our SEM model to apply on their own data, they can provide the data and name it as “sampleDAT.txt” under “/installationRoot/dataSEM/”. Currently, we provide a dummy data called “\_sampleDAT.txt”, the user can rename this file to “sampleDAT.txt”. Then, relaunch the application as usual, the default SEM data will be replaced.

We need our testers to test this feature as well.

With all due respects, we would like to thank our helpers/colleagues to test the above-mentioned features. All questions and concerns can be submitted via email to [li11@niehs.nih.gov](mailto:li11@niehs.nih.gov). Your effort is greatly appreciated.

**References**

Rosseel, Y. (2018). "Latent Variable Analysis."

Rubel, C. A., S. P. Wu, L. Lin, T. Wang, R. B. Lanz, X. Li, R. Kommagani, H. L. Franco, S. A. Camper, Q. Tong, J. W. Jeong, J. P. Lydon and F. J. DeMayo (2016). "A Gata2-Dependent Transcription Network Regulates Uterine Progesterone Responsiveness and Endometrial Function." Cell Rep **17**(5): 1414-1425.

Wu, S. P., C. Y. Kao, L. Wang, C. J. Creighton, J. Yang, T. R. Donti, R. Harmancey, H. G. Vasquez, B. H. Graham, H. J. Bellen, H. Taegtmeyer, C. P. Chang, M. J. Tsai and S. Y. Tsai (2015). "Increased COUP-TFII expression in adult hearts induces mitochondrial dysfunction resulting in heart failure." Nat Commun **6**: 8245.