Debiased Sparse Partial Correlation Network Modeling and Visualization Using the Network Explorer Module

By: Le Chang, Jeff Xia

Date: 19/11/2020

This tutorial aims to demonstrate how the Network Explorer module of MetaboAnalyst can build and visualize partial correlation networks using a data-driven network approach. This feature implements the debiased sparse partial correlation (DSPC) algorithm from Basu et al. 2017 (https://doi.org/10.1093/bioinformatics/btx012). The example data used in this tutorial comes from the original Java implementation, which are amino acid concentrations of 240 plasma samples measured by GC-MS.

Table of Contents

Introduction to DSPC Network

Biological interpretations of metabolomics data depend heavily on knowledge-based tools that contain information about metabolic networks. However, these tools' applications are restricted by the insufficient coverage of metabolism and the lack of knowledge on non-canonical interactions between metabolites. Moreover, the presence of a significant number of unknown compounds also hindered the biological interpretations of the metabolomics data. Data-driven approaches that permit the inclusion of unknown compounds hold the promise to construct biologically relevant networks and aid in identifying unknown compounds. Therefore, to address concerns in incomplete knowledge of metabolic networks and infer the putative identity of unknown metabolites, we introduce a data-driven network feature.

This feature aims to support the inference of biologically relevant networks using both known and unknown metabolites data by adopting a data-driven approach. In particular, we aim to support building partial correlation networks by using the novel debiased sparse partial correlation algorithm (DSPC) based on a graphical LASSO model. DSPC can distinguish between direct and indirect associations and provide insights into the structure of dependencies between metabolites. The main assumption of this DSPC method is that the amount of real metabolites interactions is far smaller than the sample size (i.e., the real partial correlation network among the metabolites is sparse). Under this assumption, DSPC reconstructs a network and calculates partial correlation coefficients and P-values for each pair of metabolic features. Thus, DSPC makes it possible to uncover connectivity patterns among a large number of metabolic features by using fewer samples. The inferred results can be visualized as weighted networks, where nodes represent the metabolic features, and the edges depict the correlations among them. Due to its popularity, we have implemented the DSPC algorithm in the Network Explorer module. The main steps of DSPC network analysis are as follows:

I) Data upload and processing;

II) Compute DSPC network;

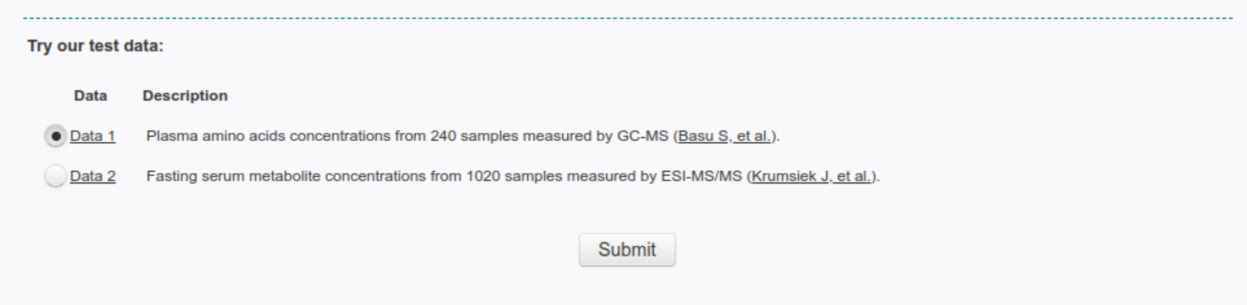
V) Network visualization and exploration.

DSPC Network Step-by-Step

Data upload and processing

1) Go to the MetaboAnalyst homepage (https://www.metaboanalyst.ca/home.xhtml). Click the "click here to start" to enter the "Module Overview" page. Click the "Network Explorer" button to enter the data upload page. Click on "A concentration table" to bring up the data uploading tab. The input data is a concentration table that contains measurements of metabolic features across multiple samples. Samples may be in rows or columns. This tutorial will use the first example dataset, which contains plasma amino acid concentrations from 240 samples measured by GC-MS (Basu S et al., 2017). Click the "Submit" button under "Try our test data" (data 1 is selected by default).

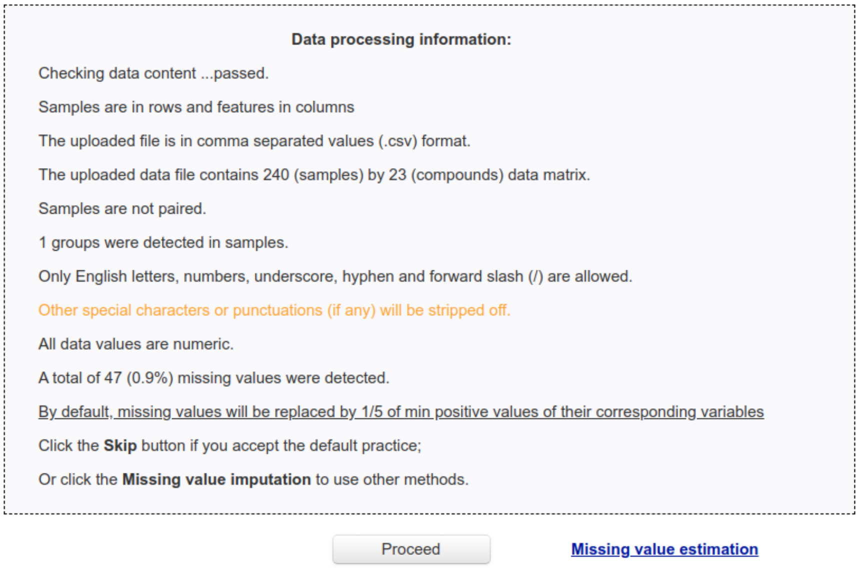
MetaboAnalyst 4.0 currently supports three types of metabolite identifiers (compound name, HMDB, and KEGG ID) as well as unknown metabolites.



Try our test data:

Data	Description
● Data 1	Plasma amino acids concentrations from 240 samples measured by GC-MS (Basu S, et al.).
○ Data 2	Fasting serum metabolite concentrations from 1020 samples measured by ESI-MS/MS (Krumsiek J, et al.).

2) The "Data Integrity Check" page should now be displayed. The data integrity check is carried out to ensure that the data meets the basic criteria for proper downstream analysis.



Data processing information:

Checking data content ...passed.

Samples are in rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 240 (samples) by 23 (compounds) data matrix.

Samples are not paired.

1 groups were detected in samples.

Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.

Other special characters or punctuations (if any) will be stripped off.

All data values are numeric.

A total of 47 (0.9%) missing values were detected.

By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables

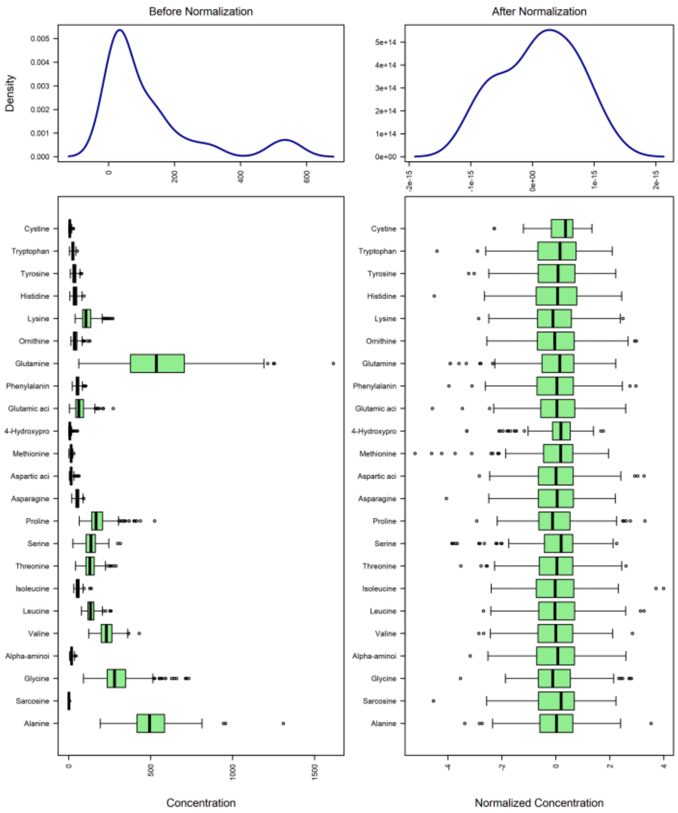
Click the **Skip** button if you accept the default practice;

Or click the **Missing value imputation** to use other methods.

3) The results indicate that 0.9% of the values in this dataset are missing. By default, missing values will be replaced by a very small value (half of the minimum positive value in the dataset), as they are presumed to be caused by signals below the detection limits. In this case, accept the default option by clicking the "Proceed" button to go to the next page. To learn more about the advanced procedures to deal with missing values, please refer to Basic Protocol 1 published in Chong et al., 2019.

4) The "Data Normalization" page is now displayed. In this case, select "None" for sample normalization, "Log transformation" for data transformation, and "Auto-scaling" for data scaling. Then click the "Normalize" button. To learn more about the different normalization strategies, please refer to Basic Protocol 1 published in Chong et al., 2019.

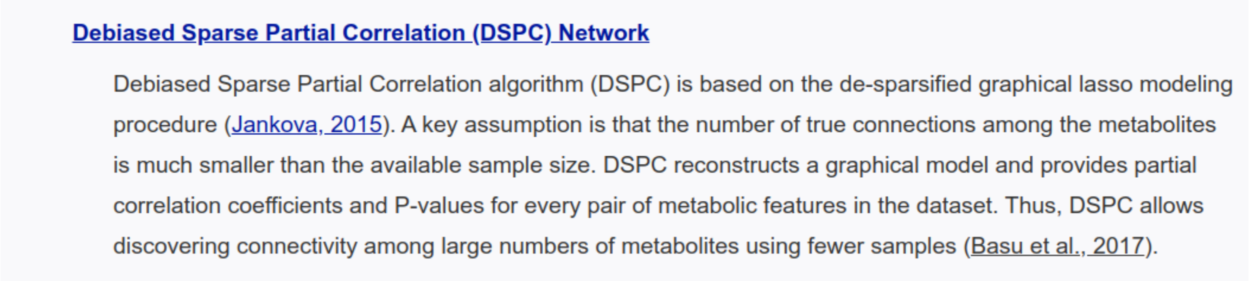
5) Click "View Result" to see a graphical summary of the normalization results. The density plots at the top show the overall data distribution based on kernel density estimation before (left) and after (right) normalization. In contrast, the box plots at the bottom display the distribution of individual metabolite concentrations before and after normalization. Users should compare the graphical summary before and after normalization to direct them to select the best methods for their data.



6) Once you are satisfied with the normalization results, click the "Proceed" button.

Compute DSPC network

7) Click on the "Debiased Sparse Partial Correlation (DSPC) Network" hyperlink to construct the DSPC network and calculates partial correlation coefficients and P-values for each pair of metabolic features.

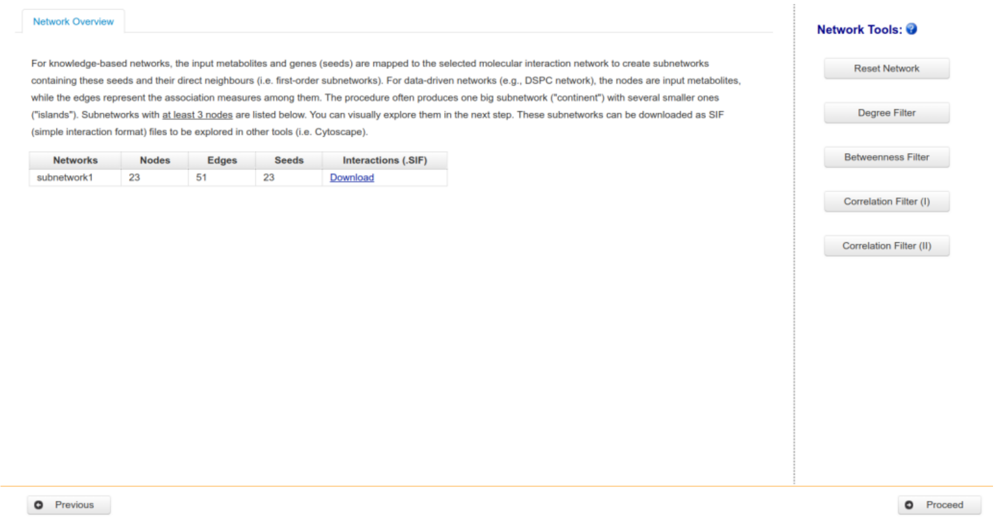


Debiased Sparse Partial Correlation (DSPC) Network

Debiased Sparse Partial Correlation algorithm (DSPC) is based on the de-sparsified graphical lasso modeling procedure (Jankova, 2015). A key assumption is that the number of true connections among the metabolites is much smaller than the available sample size. DSPC reconstructs a graphical model and provides partial correlation coefficients and P-values for every pair of metabolic features in the dataset. Thus, DSPC allows discovering connectivity among large numbers of metabolites using fewer samples (Basu et al., 2017).

8) The next page provides an overview of the constructed network. For data-driven networks (e.g., DSPC network), the nodes are input metabolites, while the edges represent the association measures among them. Subnetworks with at least three nodes are listed in the table, all of which can be visually explored in the next step or downloaded as SIF (Simple Interaction Format) files to be viewed in other tools, such as Cytoscape.

9) The correlation filters (I&II) may be used to filter metabolites for those with P-values below a specified threshold or correlation coefficients within a defined range. The default network only shows top edges based on P-value ranking (top 20% when the total number of edges is less than 1000; top 100 edges when the total number of edges is over 1000).



For knowledge-based networks, the input metabolites and genes (seeds) are mapped to the selected molecular interaction network to create subnetworks containing these seeds and their direct neighbours (i.e. first-order subnetworks). For data-driven networks (e.g., DSPC network), the nodes are input metabolites, while the edges represent the association measures among them. The procedure often produces one big subnetwork ("continent") with several smaller ones ("islands"). Subnetworks with at least 3 nodes are listed below. You can visually explore them in the next step. These subnetworks can be downloaded as SIF (simple interaction format) files to be explored in other tools (i.e. Cytoscape).

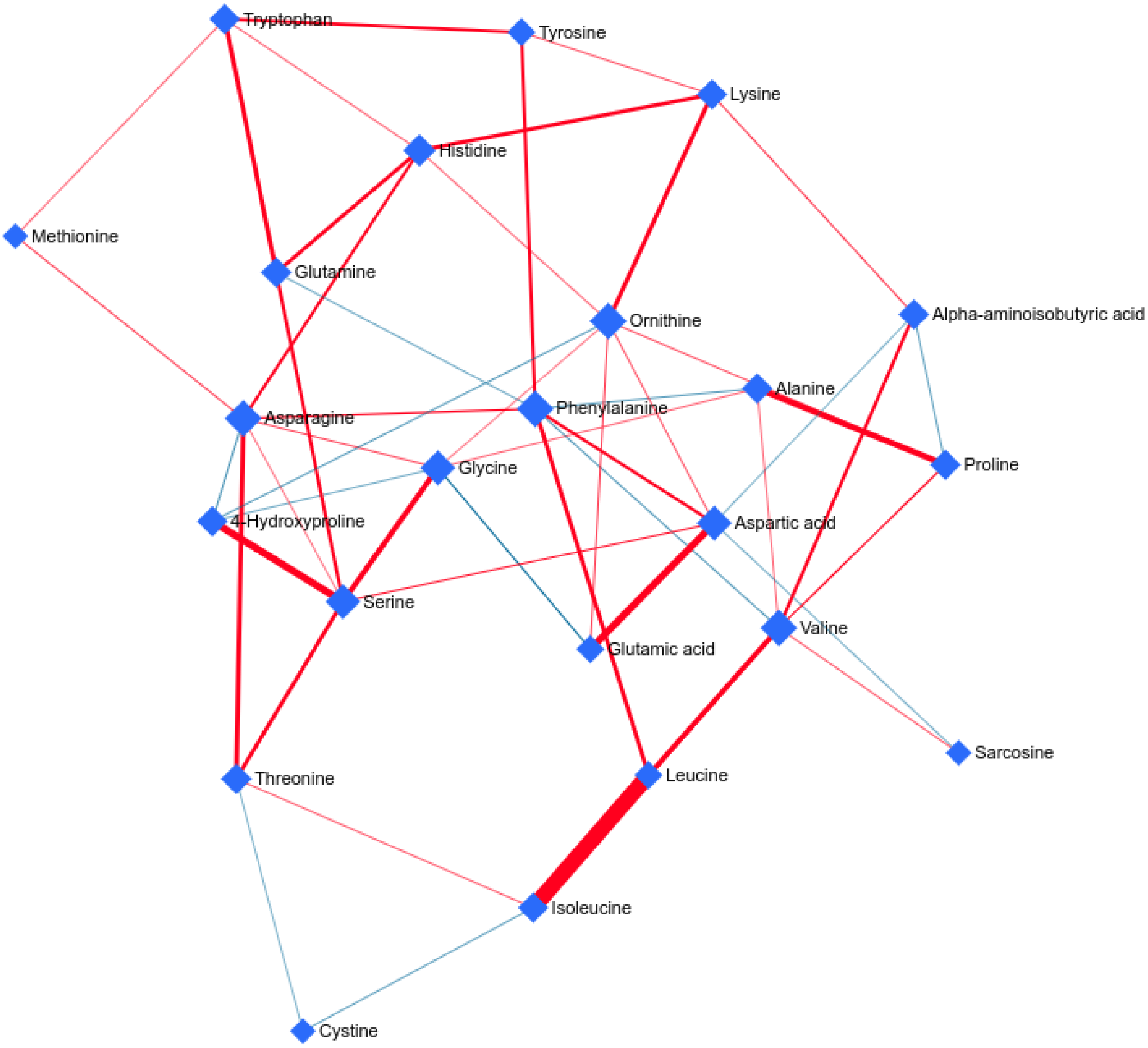
Networks	Nodes	Edges	Seeds	Interactions (.SIF)
subnetwork1	23	51	23	Download

10) Click the "Proceed" button to perform the interactive visualization and exploration of the network.

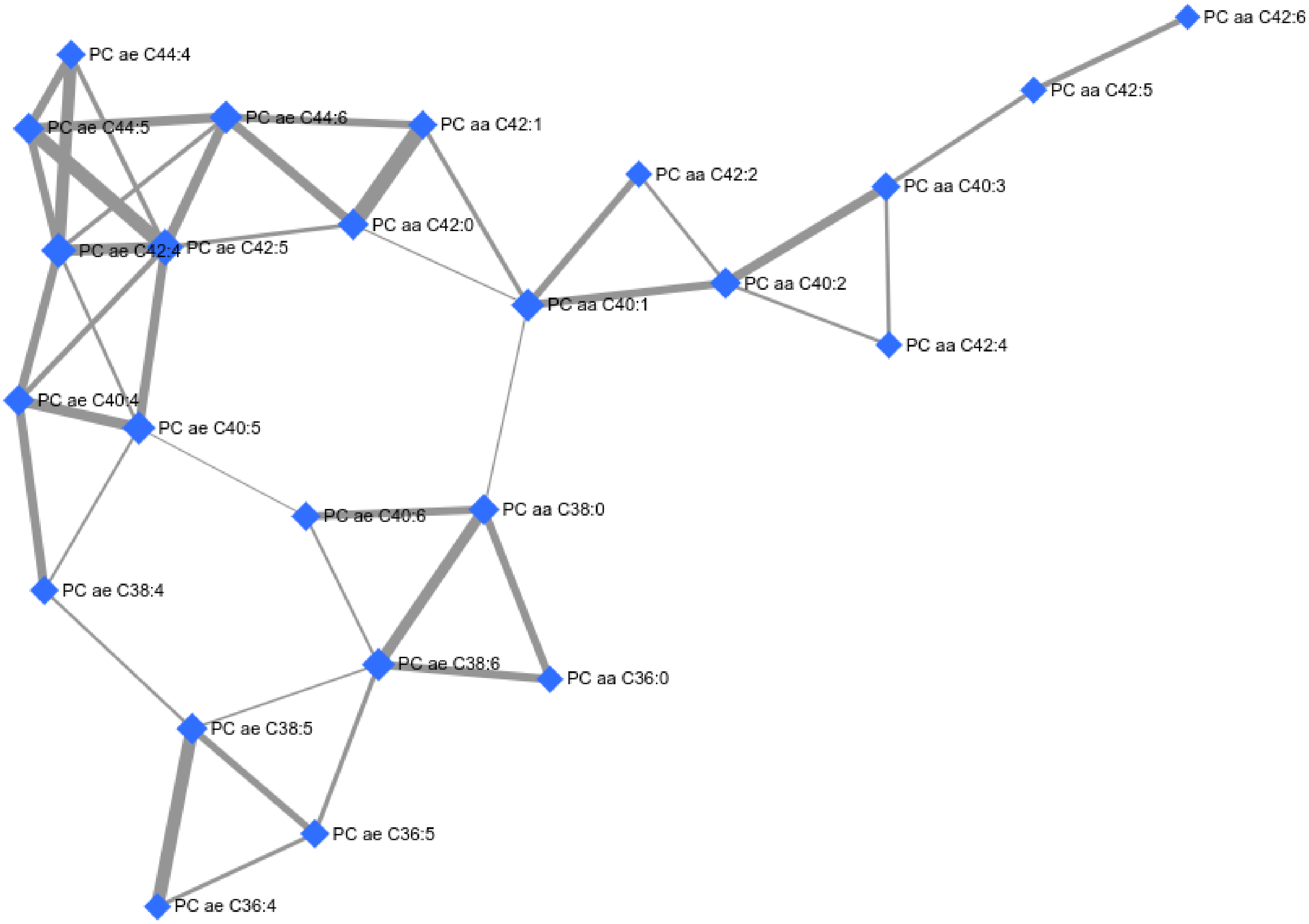
Network visualization and exploration

11) The next page shows the default view of the first subnetwork. You can see that the results are visualized as a weighted network. The thicker the edges, the stronger the correlation.

12) You may turn on the edge color by using the "Advanced View Options". An example network is shown below where the red edges represent positive correlations, while the blue edges represent negative correlations. Note, if a compound is not found in the system, the ID will display "unmapped" under "Node Explorer". To learn more about the functions of various network customization tools, please refer to Basic Protocol 11 published in Chong et al., 2019.



13) The figure below shows an example that DSPC can group the lipids belong to the same class - phosphatidylcholines (PC). The network is constructed using the second example dataset, which contains 151 metabolites measured in 1020 blood serum samples from the KORA study (Krumsiek et al., 2011).


PC aa C42:6
PC ae C44:4
PC aa C42:5
PC ae C44:5
PC ae C44:6
PC aa C42:1
PC aa C42:2
PC aa C40:3
PC aa C42:0
PC ae C42:4
PC ae C42:5
PC aa C40:2
PC aa C40:1
PC aa C42:4
PC ae C40:4
PC ae C40:5
PC ae C40:6
PC aa C38:0
PC ae C38:4
PC ae C38:6
PC aa C36:0
PC ae C38:5
PC ae C36:5
PC ae C36:4