Metabolic Network Inference.

Margriet M.W.B. Hendriks1, Age K. Smilde2

1. Leiden/Amsterdam Centre for Drug Research, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands
2. Biosystems Data Analysis, Faculty of Sciences, University of Amsterdam, Science Park 904, P.O. Box 94215, 1090 GE Amsterdam, The Netherlands

# Introduction

Metabolic networks can be studied at different levels. This is shown in Figure 1.

a) b)

A

B

C

D

E

F

2

A - 1B

1

B - 1C

1

B - 2D

1

D - 1E

A

B

C

D

E

F

2F - 1C

c)d) k1

A

B

C

D

E

F

A

B

C

D

E

F

2

A -> 1B

1

B -> 1C

1

B -> 2D

1

D -> 1E

2

A -> 1B

k

2

1

B -> 1C

k

3

1

B -> 2D

k

4

1

D -> 1E

k

5

k

1

k

2

k

3

k

4

k

5

2F -> 1C 2F -> 1C

Figure 1: The idea of metabolic network inference.

The metabolic network of yeast which is used in the exercise is shown in Figure 2. This is the glycolytic pathway in S. cerevisiae (Teusink et al. 2000). It contains 13 metabolites (including P); has 18 reactions (of which 2 are lumped : http://jjj.biochem.sun.ac.za) and 21 interactions.

**List of reactions**

1. GLCo ↔ GLCi
2. P + GLCi ↔ G6P
3. G6P ↔ F6P

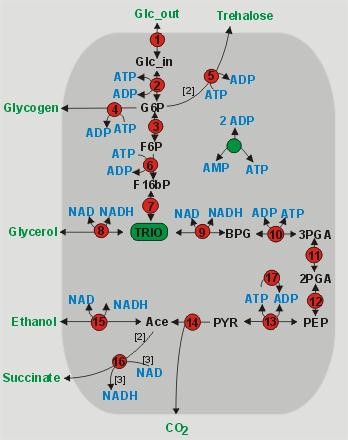


Figure 2: The network of yeast.

1. P + G6P ↔ Glyc
2. P + 2 G6P ↔ Trh
3. F6P + P ↔ F16bP
4. F16bP ↔ 2 TRIO
5. NADH + TRIO ↔ NAD + GLY
6. NAD + TRIO ↔ NADH + BPG
7. BPG ↔ 3PGA + P
8. 3PGA ↔ 2PGA
9. 2PGA ↔ PEP
10. PEP ↔ P + PYR
11. PYR ↔ ACE + CO2
12. ACE + NADH ↔ ETOH + NAD
13. 2 ACE + 3 NAD ↔ 3 NADH + SUCC
14. P ↔ X

**List of metabolites**

1. GLCi
2. G6P
3. F6P
4. F16bP
5. TRIO
6. BPG
7. 3PGA
8. 2PGA
9. PEP
10. PYR
11. ACE
12. P (ATP/ADP)
13. NADH

# Practical details

All datasets can be loaded using RStudio (under files right lower panel icon for files) by double-clicking the file networkdata.RData. The numbers in the object names indicate the number of repeats, or the level of measurement error. Each file contains 13 columns representing the 13 metabolites (see the list of metabolites, the ordering of the columns is the same as the number in the list). The rows represent the observations. The number of observations may differ for different datasets.

For visualization we need to install a package and make it available:

install.packages("igraph")

library(igraph)

Two scripts and one R functions are available:

The script with name network\_2022.R generates a Pearson correlation matrix and a network plot and the adjacency matrix given a certain significance level for data999.

The script with name partnetwork\_2022.R generates a Partial Pearson correlation matrix and a network plot and the adjacency matrix given a certain significance level for data999.

quality.R. This function calculates the true and false positives, the true and false negatives, the true and false positive ratios as well as the g-score of an inferred network. Input is the adjacency matrix for that network and the adjacency matrix of the true network. This function is used by the two following scripts. So in order to run the scripts please open this file in R-studio and source it.

# Exercises

## Exercise 1

Use the first part of the network\_2022.R script to make plots of the true network with and without the node for ATP/ADP denoted as P. Compare the two resulting figures with the network shown in figure 2 above.

**The problem is that, in the real network, reactions are unique but metabolites can be repeated, e.g. ADP and ATP, but in the graph we created, the metabolites (nodes) are unique but reactions (edges) can be repeated, so distributed metabolites (like P in this case) will have a lot of connections with other nodes, so we remove P to have a clear view of the graph.**

So distributed metabolites (as P in the exercise) have a lot of connections that makes the picture difficult to read (left figure). To make the picture clearer, we remove such metabolites from the graph (right figure).

## Exercise 2

Use dataset data999 to generate a network based on Pearson correlations.

1. Inspect the matrix with Pearson correlations.
2. Inspect the correlation between metabolites F16bP on the one hand with F6P, TRIO and BPG on the other hand using the plot function of R.

Pairs of metabolites F16bP - F6P and F16bP - TRIO are **directly connected on the network**. Comparative concentration plots of both pairs show a good correlation. As opposite to that, the concentration plot of F16bP and BPG shows the weakest correlation. Metabolites BPG and F16bP do not have the direct connection **but they have an indirect connection that is formed by metabolite TRIO**.

(**You should ALWAYS READ THE NETWORK GRAPH BEFORE ANSWERING QUESTIONS!**)

1. Use a correlation cutoff of e.g 0.1. Visualize the graph without P.
2. Calculate p-values using permutations. Use a p-value cutoff of 0.01. Visualize the graph without P.

(**WHY DO WE USE P-VALUE HERE?**)

The visualized network based on the Pearson correlations shows a very dense graph. The connections can be defined based on value of correlation or on the p-value. These are connected, but depend on the number of samples used to calculate the correlation.

Almost all metabolites are connected. Pearson correlations are not able to distinguish between direct and indirect links. Spurious edges are observed. BPG is connected with three metabolites on the real graph (TRIO and 3PGA, NADH), but on the visualization of the adjacency matrix based on Pearson correlations BPG has connections with PEP, F6P, 3PGA and 2PGA.

1. Calculate the fn, fp, tn, tp, tpr, tnr and g-scores for the # part

Comment on your findings and interpret the results.

**The problem is: Pearson correlation cannot distinguish between direct and indirect connections**

## Exercise 3

Use the partnetwork\_2022.R file. The same dataset is used as in exercise 2 to generate a network based on partial Pearson correlations with a cutoff value of 0.01.

The partial correlation was explained in the lecture by calculating the correlation between residuals after the effect of all other variables has been corrected for. E.g.

Y = ZBY + E\_Y

X = ZBZ + E\_X

Then the partial correlation between Y and X, corrected for Z is the correlation between E\_Y and E\_X. Another (much faster) approach to calculate the partial correlation is byusing the Graphical Gaussian Network approach. If we want to find the partial correlation between variables 1 and 2, we correct for the influence of all other metabolites in the data. In the wiki page <https://en.wikipedia.org/wiki/Partial_correlation> you can find how the partial correlation is calculated using **a matrix inversion**. This is the way it is calculated here. The solve function in R calculates the inverse of the correlation matrix.

Convince yourself that the slow and the approach by regression and the fast approach by the GGN approach provide the same results.

1. Visualize the network with the script for the partial correlation network. Concentrate on the without P network.
2. Calculate the tp, tn, fp, fn, tpr, tnr and g-scores.
3. Compare with the network of exercise 2.

Comment on your findings and interpret the results.

**Most indirect connections between metabolites disappear (although some are still there, such as TRIO - F6P)**

## Exercise 4

Calculate partial Pearson correlations for datasets data500, data250, data100, data50 with a cutoff value of 0.01. Focus only on the networks without P.

1. What is the difference between those datasets?

**They have different number of observations.**

1. Compare the partial pearson correlation networks for these four datasets with the PPC network of dataset data999.

**Data999:**

**tp tn fp fn tpr tnr g**

**15.0000000 45.0000000 12.0000000 6.0000000 0.7142857 0.7894737 0.7509393**

**Data500:**

**tp tn fp fn tpr tnr g**

**14.0000000 47.0000000 10.0000000 7.0000000 0.6666667 0.8245614 0.7414227**

**Data250:**

**tp tn fp fn tpr tnr g**

**11.0000000 54.0000000 3.0000000 10.0000000 0.5238095 0.9473684 0.7044435**

**Data100:**

**tp tn fp fn tpr tnr g**

**8.0000000 55.0000000 2.0000000 13.0000000 0.3809524 0.9649123 0.6062884**

**Data50:**

**tp tn fp fn tpr tnr g**

**5.0000000 57.0000000 0.0000000 16.0000000 0.2380952 1.0000000 0.4879500**

1. Calculate the fn, fp, tn, tp and g-scores of the networks.

**See above**

Comment on your findings and interpret the results.

**Dataset with larger observations have better performance, generating networks that are close to the true network, and have the largest G-score. When the observations are too small, there are not enough information to build a good network model, resulting a sparse network and small G-score.**

**From the pictures we conclude that partial correlation requires a high number of repeated experiments.**

**50 replicates for 13 metabolites is not enough and renders most connections not significant. The more**

**replicates are used the more true-direct links are obtained (compare the graphs for data50, data100,**

**data250, data500 and see Table 1). However, with growing number of replicates the number of false**

**positives or indirect links also grows.**

## Exercise 5

The dataset data999 contained no measurement error, the datasets data n1, data n5, data n10 contain the same data, but now with respectively 1, 5 and 10% measurement error added. Construct PPC networks for these 3 datasets with a cutoff value of 0.01. Compare them with the PPC network of the original noiseless dataset using the quality criteria fn, fp, tn, tp and g.

Comment on your findings and interpret the results.

**Data\_n1:**

**tp tn fp fn tpr tnr g**

**11.0000000 48.0000000 9.0000000 10.0000000 0.5238095 0.8421053 0.6641557**

**Data\_n5:**

**tp tn fp fn tpr tnr g**

**6.0000000 50.0000000 7.0000000 15.0000000 0.2857143 0.8771930 0.5006262**

**Data\_n10:**

**tp tn fp fn tpr tnr g**

**3.0000000 56.0000000 1.0000000 18.0000000 0.1428571 0.9824561 0.3746343**

**With larger error, the quality of network is lower, which matches our intuition.**