DATA-DRIVEN RELATIONAL GRAPHS

Marianyela Petrizzelli[1]

[1]Institut Curie, PSL Research University, INSERM U900, F-75005 Paris, France and CBIO-Centre for Computational Biology, MINES ParisTech, PSL Research University, F-75006 Paris, France.

Abstract

I provide a source code (written in R) that generates data-driven relational graphs based on the method implemented by Verbeke et al. (2015). It consists in the conversion of all available data about the entities under study into a single comprehensive network representation containing not only genes but also individual patient samples. Precisely, the method allows for missing data about genes but requires data-sets containing information for all samples.

1 Overview of the global network construction

As a first step of the method, data-sets (e.g. gene expression, methylation) must be converted into a binary form where 1's design molecular modifications (such as over- or under-expression, hyper- or hypo-methylation) and 0's correspond to normality. Secondly, to construct the global network, each binary input data-set is represented as an individual network, by converting the 1's in the binary data to an indirect link connecting a sample node with a gene node. Next, the individual networks are merged. In the merging process, sample nodes are joined, but gene nodes are not. Consequently, the network will contain multiple gene nodes with the same gene identifier for genes showing multiple molecular abnormalities.

In addition, all data are linked through a molecular interaction network derived from public repositories in order to connect heterogeneous and potentially sparse genetic aberrations with their downstream effects.

2 Case study of the relational graph construction from a multi-omics medulloblastoma dataset

The construction of the global network is demonstrated using the medulloblastoma multi-omics data-set (Forget et al., 2018) which included gene expression, methylation, proteomic and phosphoproteomic data (provided in folder "data").

In this study, I choose to use a trivial approach for data conversion into a binary matrix: I set a hard threshold to the (absolute) values for each data-set. The data-set dependent threshold is chosen in such a way that the fraction of entries equal to "1" in each data matrix is as close as possible to a predefined parameter, here set to 30%.

Individual graphs are constructed and merged as described above. In particular, the data-driven network is returned as a data-frame with "from-to-layer" attributes for network representation. Next, a gene-gene interaction (GGI) network is downloaded from Pathway Commons (https://www.pathwaycommons.org/) and merged with the data-driven network by linking genes in the data-sets to those present in the GGI. Naively, I included in the global network the GGI sub-graph containing only those genes present in at least one of the input data-sets.

The final graph consisted in 59102 nodes and 1819581 edges. A sub-representation of the data-driven relational graph is given in Figure 1. As can be seen, there are several types of nodes with labeled relations between them:

- Patients/tumors are denoted in pink (MB01, MB02, MB03);
- Genes are denoted in red (RPL4, RPL41, FAP, GFAP, DBNL, WAS, SLC10A3);
- Molecular modifications in tumors (nodes .RNASeq, .Protein, .Phospho, .Methylation) that always connect a patient and a gene.

3 Implementation of the approach to construct relational graphs from muti-omics datasets

The code is written in R. Scripts “packages.R” and “functions.R” provide all packages and functions needed for the construction of data-driven relational graphs. Scripts “multilayer.R” and “graphical_representation.R” perform construction and representation, respectively, of the case study in Section 2.

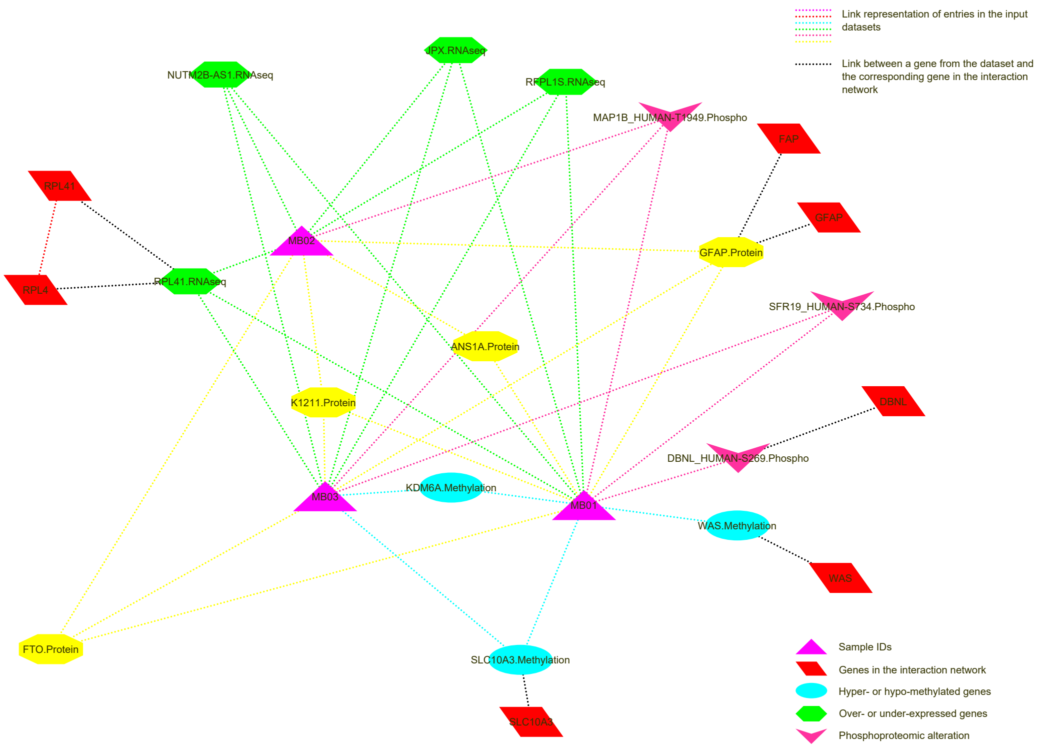
Link representation of entries in the input datasets
Link between a gene from the dataset and the corresponding gene in the interaction network
NUTM2B-AS1.RNAseq
JPX.RNAseq
RFPL1S.RNAseq
MAP1B_HUMAN-T1949.Phospho
FAP
RPL41
GFAP
MB02
GFAP.Protein
RPL4
RPL41.RNAseq
SFR19_HUMAN-S734.Phospho
ANS1A.Protein
K1211.Protein
DBNL
DBNL_HUMAN-S269.Phospho
KDM6A.Methylation
MB03
MB01
WAS.Methylation
WAS
FTO.Protein
SLC10A3.Methylation
SLC10A3
Sample IDs
Genes in the interaction network
Hyper- or hypo-methylated genes
Over- or under-expressed genes
Phosphoproteomic alteration
Proteomic alteration

Figure 1: Sub-graph representation of the global network.

References

Forget, A., Martignetti, L., Puget, S., Calzone, L., Brabetz, S., Picard, D., Montagud, A., Liva, S., Sta, A., Dingli, F., Arras, G., Rivera, J., Loew, D., Besnard, A., Lacombe, J., Pagès, M., Varlet, P., Dufour, C., Yu, H., Mercier, A. L., Indersie, E., Chivet, A., Leboucher, S., Sieber, L., Beccaria, K., Gombert, M., Meyer, F. D., Qin, N., Bartl, J., Chavez, L., Okonechnikov, K., Sharma, T., Thatikonda, V., Bourdeaut, F., Pouponnot, C., Ramaswamy, V., Korshunov, A., Borkhardt, A., Reifenberger, G., Poullet, P., Taylor, M. D., Kool, M., Pfister, S. M., Kawauchi, D., Barillot, E., Remke, M. and Ayrault, O. (2018) Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. *Cancer Cell*, **34**, 379–395.e7. URL: https://linkinghub.elsevier.com/retrieve/pii/S1535610818303568.

Verbeke, L. P. C., Van den Eynden, J., Fierro, A. C., Demeester, P., Fostier, J. and Marchal, K. (2015) Pathway relevance ranking for tumor samples through network-based data integration. *PLOS ONE*, **10**, 1–22. URL: https://doi.org/10.1371/journal.pone.0133503.