Network elements

One network per GO term. Thus predictions are independent for each GO term*
By predicting the function, we aim to report a possible association between the gene and a GO-term. More specifically, the Biological process (BP) category of GO.

- Target GO-term: the GO term for which we want to identify genes that are associated. Given the target GO term, we will investigate its possible association with each of the genes in the data set.

- Gene of interest: The gene whose association with the target GO term we are interested in at a specific moment.

- Nodes: genes or proteins. In this thesis we work with genes, but sometimes the literature refers to proteins

- label: a category that specifies whether a node is or not associated with the target GO term. In the network, the labels can be interpreted as the colors of the nodes. And in the script the labels are coded as follows:
  1: The gene has the function
  0: The gene is not known to have the function
  -1: The gene may or may not be known to have the function. If the function is known, we hide that information by placing a -1 instead of a 0 or a 1. Predictions are evaluated based on the fraction of genes that are -1 and whose label is known.

- Edges: Two nodes are linked by an edge if they are co-expressed. The magnitude of the connection is neglected

- Neighbors. Genes that are co-expressed or linked by an edge

sets of genes

(A) positive, non-validated-positives, negatives, reliable negatives, unlabeled, unkown

- positive: known-positive gene with validated proof, where positive means that its association with the target GO term has been validated with experimental evidence scores (EES). We only considered as positives those associations that regard the BP category.

- Non-validated-positives: the gene is associated with the target GO term but: (1) the association has not been validated with EES, or (2) the association does not regard the BP GO-category

- negatives: genes that are not associated with the GO terms. Since in biology the lack of evidence for an association does not imply that such an association does not exist, this set of genes cannot be knwon.

- reliable negatives: set of genes that stand for representative set of genes that are not associated with the target GO-term (negative genes). Note that although they stand for

representative set of negative genes, we cannot be certain of this. Therefore, more strictly speaking these are genes that are, to some extend, unlikely to be associated with the target GO-term

- unlabeled: gene that is not known to be associated with the target GO term. This excludes the genes Non-validated-positives genes. The set of unlabeled includes: (1) genes that are associated with the target GO term but whose association is not known yet. These are the ones that we want to identify; (2) The reliable negatives that we aim to extract in the first step of PU-BMRF

- unknown: a special case of unlabeled gene that is unlabeled for all the GO-terms in the database. Unknown genes are treated differently because we expect that these genes are not associated with any GO term in the database not because they are less functional than the other genes but because, for some reason, their function is more difficult to predict.

(B) training and test

- train set: a set of genes whose label is either known or unkowun, and that will enter the model with its label.

- Test: a set of genes whose function is known but hidden (see concept 'label')

folds and replicates

- folds, each fold consists of a set of test and a set of train. In total the test and the train set account for all the genes in the dataset. However the assignation of the genes to the train or test set changes with with k-fold in the crossvalidation. When all the folds are created, each gene has been assigned to the test set once and k-1 times to the train set. However, there are some genes that are never included in any test set, an remain always in the train set. These genes are the unknown genes and the Non-validated-positives genes (explained in the section 'sets of genes' in this appendix)

- replicates

Other concepts

GO-size filter

features

Model parameters

* Strictly speaking this is not true. Explained in Part 1 of Materia and methods

Model parameter used in the original code (https://github.com/jwbargsten/bmrf)	
name parameter	**Description**
minGOsize	Minimum # of labels per GO term in the train set
minDFsize	Minimum # of labels per domain term in the train set
maxGOsize	Maximum # of labels per GO term in the train set. (i.e. 0.9 means 90% of the #labels in network)
maxDFsize	Maximum # of labels per domain term in the train set. (i.e. 0.9 means 90% of the #labels in network)
k	Number of folds in the BMRF cross-validation
Additional parameters considered	
network size	Subsets of netowrk used: **coexpresion** (#conexions)*
only EES	Whether only associations of category "Biological process" and with Experimnetral-evidence-scores are considered

Table 1: Description of model parameters

Prediction peerformance: AUC…

Explain that depth is not such a good idea and that less depth is mor egeneral

edges from positives
te/tpe from positive genes (appenidx ii)