Data preparation

Data sources for the different species is shown in Appendix I [*include a head of the 3 files*]. In the GO-terms file, the associations were labelled as "valid" if, for at least on of the association available in data, they correspond to EES ('EXP', 'IDA', 'IEP', 'IMP', 'IPI', 'IGI') and to the category of BP, and as "non-valid" otherwise. In order to maintain this distinction throughout the analysis between "valid" and "non-valid" associations these two groups of associations were up-propagated independently and then the files were combined keeping the label. The code [...] was used for uppopagating.

Domain and GO-terms files were pruned to exclude genes that are not available in the network file, as a requirement for the BMRF code. Then, the GO-size filter was applied to exclude the GO terms that are too general or whose number of known associated genes is excessively low to allow for predictions with BMRF (details about which number of genes is excessively low are given in Appendix x...*this changes based on k... and explian the error*). The GO-size filter is based solely on the "valid" associations because the non-valid assoications are not used in the validation and the GO-size filter allows to make sure that there are enough number of henes n the validation. Analogous to the GO-szie filter, BMRF uses a DF-size filter to exclude form ythe analuysis the domains whose number of genes is below a certain threshold because levels with few observation give proiblems inthe sparse matrisx.

Part 1

BMRF [1] was used to do PFP and learn about the impact of different method and network parameters on the prediction performance. In this framework, predictions are made individually for each GO term that passes the GO-size filter (see Appendix I - Concepts). The predictions, however are not entirely independent of the other GO terms in the dataset because the genes that are not associated with any of the GO terms are treated differently. These genes are coded as "unkowns", and the number of unkowns depends on the number of GO terms in the database, and therefore on the values used un GO-size filters. Each gene in the network file will enter the BMRF code with one of three labels (1,0 or -1), and it will be predicted as 1 or 0, where "1" stands for posotive, "0" stands for unlabeleld (we cannot tell whether it has the function) and "-1" stands for unknown (we cannot tell whether it has the function and its label in the training set will be 0 or 1 based on Gibb sampling taking into account the label of its neighbours). Thus, although the genes enter the BMRF code with three labels, the training set consists only of two labels.

The advantage of treating the unknown genes (genes known to be associated with zero GO terms) differently is that genes that have never been predicted as positives are less likely to be non-associated for a given function than those genes that have been found as positives for some function but not for the function of interest. This has to do with the fact that some function are more difficult to predict than others. Thus, if based on data, a gene has never been identified as positive for any function, it is more fair to assume that the function is particularly difficult to predict using experimental approaches, that assuming that a very low number of genes have the function. In other words, a large portion of unlabbeld genes ("0") implies that the function is rare, (almost never observed), wheras a large portion of -1s implies that the function is difficult to predict. This distinction between unlabeled and unknowns genes, additionally, allows to do predictions in a fairer fasshion. By treating the gene of interst as unkwon (-1) instead of as unlabeleld (0), the prediction of the gene-GO association we are interetsted in, will be more free from the prediction of the GO term. If the portion of "-1" however becomes very large with respect to the portion of "0", however, Gibb sampling will fail in the relabeling, because it will expect that the portion of genes that have the function is very large.

The labelling is as follows ["make table"]:

- genes assoicaited with the GO term labelled as "non-valid": In all folds enter the model as 1s and are not tested
- unwns genes always take label -1, and they do not enter the test set
- positive gens in the test sset take label -1, and are said to be succesfully classifies if the take label 1 in the predictions, or unsuccesful otherwise
-negative genes in the test set,
...

The validation was as follows: k-fold 10, was used and only the associations coded as "valid" entered the test set. Genes classified as unkowns wre also excluded form the test set. The k-folds were made independetly for the positive set and the unkowns but with a similar value of k (k:10 in part 1 and k:2 in part 2). Because the allocation of genes to folds is a random process, 20 replicates we carried.

In order to investigate the impact that the quality of the data has on the predictions, a portion of associations were randomly removed. For this, we distinguished between 2 typeas of associations: association of the GO term of interest and other association sof other GO terms. The prediction performance was computed when the portion was
Analogous o the removal of association analysis were carried also when a portion of the edges wsa removed from the anlaysis. For this, we distinguishes four types of edges: edges positive -positive, epn, enn, te. Portions subtracted were.
Finally we carried analysis adding random associations between genes and the GO-term of interest. We did not sample genes from the set of genes that are positives for the GO term but not validated. Portions of noisy associations added were...

In order to compute the correlation between different parameters, the following information was extracted:

For each gene-GO association:
-number of edges
-number of edges positives-positives
-number of edges positives-negatives
-number of edges negatives-negatives

For each GO term:
-depth, using the function 'getAllBPChildren" from the R-packe GO.db
-number of genes
-number of validated labels
-sum of the number of edges, of the genes associated with the GO term
-sum of the number of edges positives-positives, of the genes associated with the GO term
-sum of the number of edges positives-negatives, of the genes associated with the GO term
-sum of the number of edges negatives-negatives, of the genes associated with the GO term

For each gene:
-number of GO-terms

Part 2

features GO-specific/not... num,ber of RN.

scaled

Explian that you can extract RN in 3 ways: AUC, # RN, value of toleerance as in [,]

7.BHARDWAJ, N., GERSTEIN, M. & LU, H. 2010. Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique. Bmc Bioinformatics, 11.
The folloing information was extracted