Appendix III – Additional results

Part 1- Choice of the network data and tune the model parameters for prediction performance using BMRF.

- **Section (a) Impact of the co-expression threshold on the prediction performance**

Conventionally a Pearson Correlation of 0.7 is used as a threshold for co-expression analysis. However, in the case of Chickens, using a Pearson correlation of 0.7, leads to an excessively low number of validated associations. Subsequently, only 9 GO term pass the GO-size filter used by BMRF (see appendix I-Concepts). Furthermore, we observed that the number of GO terms that passed the filter did not improve much when the the GO-size filter was adjusted to include in the analysis GO terms with a low number of associated genes (for minGOsize=0.8, 52 GO passed the filters). We therefore investigated whether by lowering the Pearson correlation threshold we obtained better prediction performance. We choose different co-expression threshold values and we computed AUC, standard deviation across replicates for a given GO -term and number of gene sin network and GO term sin analysis.

Illustration 1 shows the prediction performance (AUC) for chickens using different Pearson correlation thresholds. The distribution of the AUC is hyperbolic, showing its maximum at a Pearson correlation of 0.35.

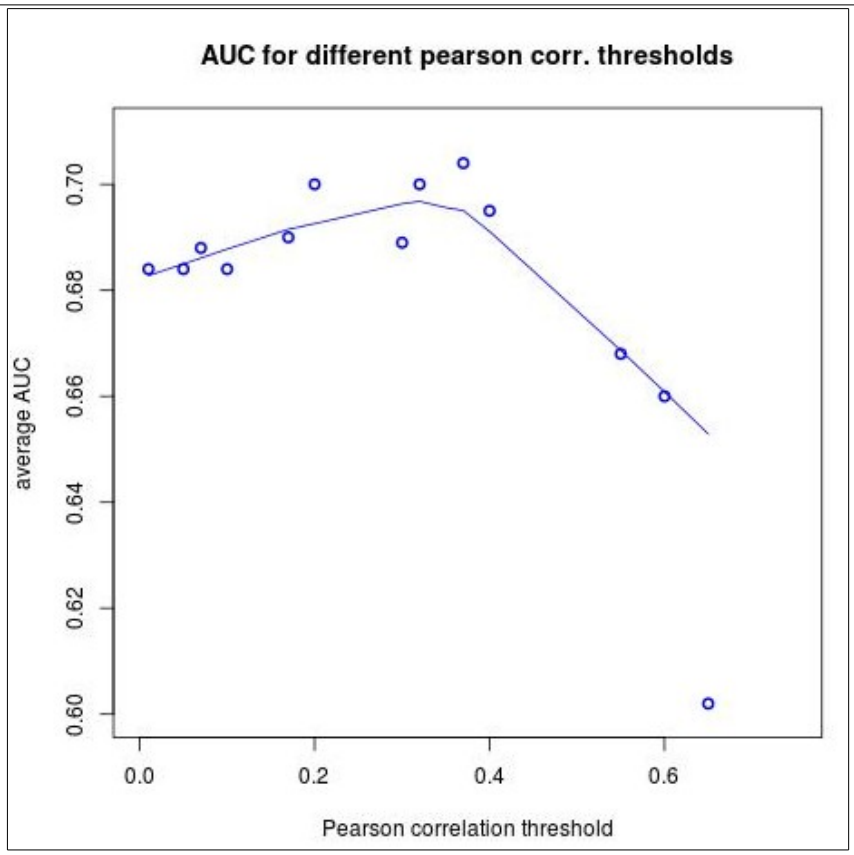


Illustration 1: AUC for different Pearson correlation thresholds for the chickens' co-expression network.

The Y-axis corresponds to the average AUC for all GO-terms considered in the analysis. The line corresponds to the lowness line. Values corresponding to the plot are in table 1.

The network changes considerably base don the co-expresion threshold. Two important parameters that change based on this are the number of genes and the standard deviation across replicates for any given GO-term, as shown in Illustrations 2 and 3, respectivelly.

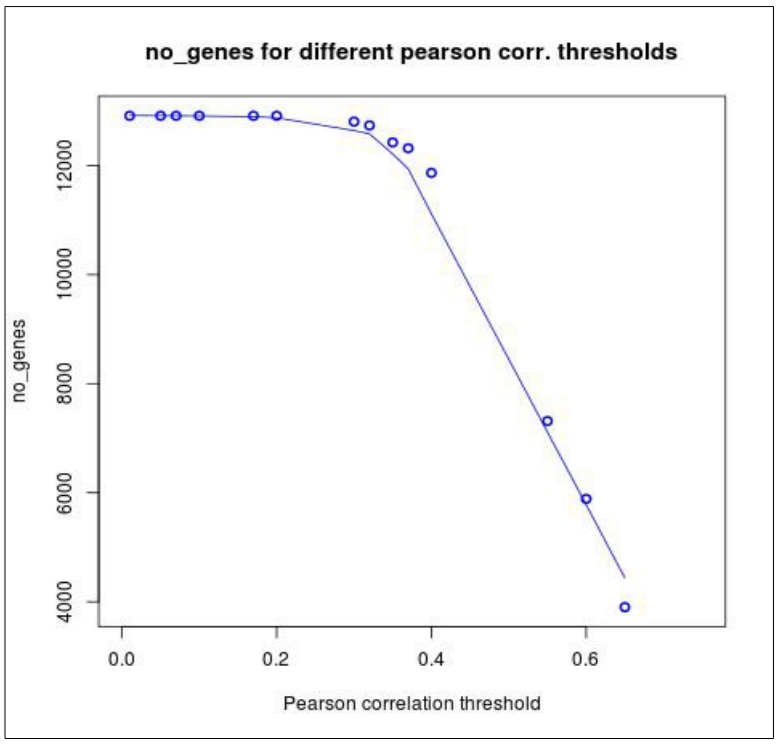


Illustration 2: Number of genes in the network for different co-expression thresholds.

Values corresponding to the plot are in table 1.

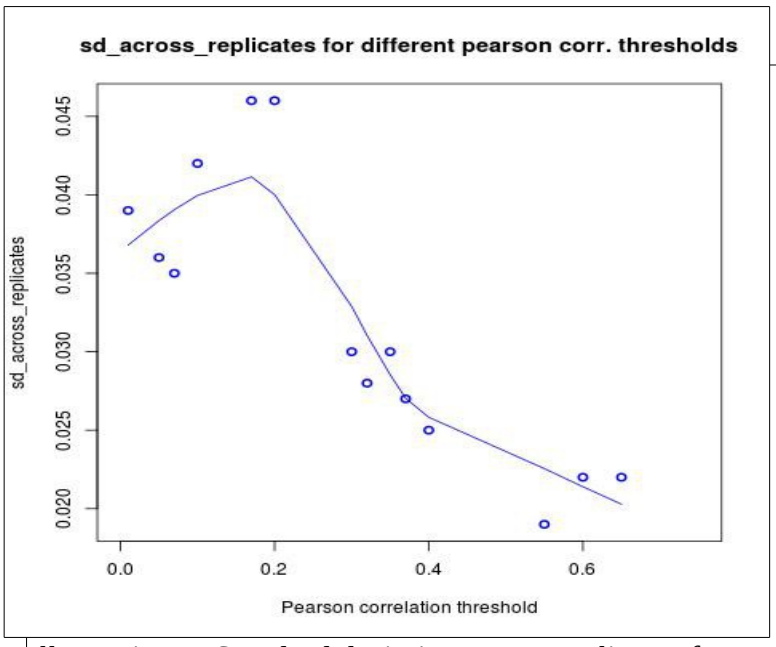


Illustration 3: Standard deviation across replicates for differnet co-expression thresholds.

Values corresponding to the plot are in table 1.

Illustrations 2 show a decrease in the number of genes as the threshold for co-expression becomes more strict. The decrease becomes more sharp when the threshold is above 0.4. However, this may depend on the characteristics of the co-expression analysis.

In illustration 3 we observe that the standard deviation across replicated decreases also as we become more strict in the co-expression threshold. Thus, choosing a high co-expression threshold allows for better reproducibility. On the left hand side of the illustration, we observe that the standard deviation was lower for Pearson correlation thresholds (0.01, 0.02 and 0.015), than for a Person correlation of 0.17 and 0.2. This effect is counter-intuitive, but does not seem to be significant and may be an artifact of the analysis.

Illustrations 1-3 are based on the results in Table 1.

Pearson corr. Threshold	#GO terms pass normal filter	#genes in network	AUC	sd AUC	Mean of SD across replicates
0.7	9	2,784	0.714	0.065	0.022
0.65	21	3,901	0.603	0.064	0.022
0.6	33	5,887	0.660	0.055	0.019
0.55	54	7,314	0.668	0.063	0.022
0.4	134	11,866	0.695	0.056	0.025
0.37	140	12,317	0.704	0.056	0.027
0.35	**138**	**12,424**	**0.726**	**0.080**	**0.030**
0.32	148	12,735	0.700	0.052	0.028
0.3	148	12,805	0.689	0.054	0.030
0.2	150	12,912	0.700	0.068	0.046
0.1	150	12,912	0.684	0.068	0.042
0.07	150	12,912	0.688	0.075	0.035
0.05	150	12,912	0.684	0.074	0.036
0.01	150	12,912	0.684	0.078	0.039

Table 1: Prediction performance choosing different Pearson correlation thresholds, for chicken data. In bold the threshold for which a highest AUC was achieved.

Predictions were highest for a Pearson correlation threshold of 0.35. In this study we used a Conditional independent network to obtain reproducible results. However, It should be considered that this value depends on the characteristics of the co-expression analysis. Furthermore, this value may depend on the species. For instance, in species with a large number of validated data, it is expected that the best possible threshold is higher, given that overall quality of the data is higher; whereas in a species in which the portion of validated data is low (as is the case for chickens), including data of relative quality may be helpful. In other words, the quality of the data may become more important once the criteria for minimum of data required has been satisfied.

- **Section (b). Impact of the different model parameters on the prediction performance**

In this section we investigate the impact in the prediction performance of the number of replicates of analysis, the GO-size filters, the number of k in the k-fold validation and the number of iteration in the Gibbs-sampling. Also we investigated the effect of adding non-validated data and/or domain information.

- Number of replicates

We considered as reproducible, results with less than 0.02 standard deviations in AUC across replicates. Note that the standard deviation that is given in most of the tables, for instance in tables 3 and 4, refer to the standard deviation across GO-terms rather than across replicates of the same GO term.

We investigated how many replicates were required to achieve an average standard deviation of 0.02 across replicates. For this, we carried different runs, of 10 or 20 replicates each and we computed the standard deviation across runs. We observed that the standard deviation across replicates was 0.006 lower when 20 replicates were used instead of 10. Furthermore, we observed that 20 replicates were required to achieve an average stabndrad deviation across replicates below 0.02. We therefore decided to use 20 replicates for each analysis, except for part 3, that we used only 4 replicates due to time constrains.

In another analysis we showed that the standard deviation across 5 runs of 20 replicates was slightly lower for a GO-size filter of (20,0.1) than for (5,0.9): 0.008 vs 0.01, respectively. This makes sense since the standard deviation is slightly larger for those GO terms with fewer genes and more of these GO terms were considered in the analysis when the GO-size filter was (5,0.9) (less strict).

- GO-size filter

Default value for maxGOsize was 0.9, however, for this thesis, we are not interested in predictions for the most general GO terms and we chose value 0.1. A preliminary analysis carried on yeast co-expression data showed that there is not significant increase in the accuracy of prediction when 0.1 was used instead of 0.9. In this analysis, AUC was 0.779 with a minGOsize of 0.1 and 0.775 for 0.9. The standard deviations being 0.08 in both cases. The same analysis showed how the data changes with the GO-size filter (table 2)

scenarios			**data**			
scenario name	**Min GO-size**	**Max GO-size**	**Network size (#conn)**	**#unkown genes***	**#assoc.**	**#GO-terms**
normal	20	0.1	598,174	655	132,249	1,104
default**	20	0.9	598,174	4	264,279	1,187
more GO-terms	10	0.9	598,174	4	273,977	1,738
Only large GO-terms	30	0.07	598,174	688	104,582	832

Table 2: Effect of the GO-size filter on the data

We then investigated the effect of the GO-size filter in the 3 species considered and for yeast ppi.

For this, we carried the analysis in three scenarios with different GO-size filters. We called “normal scenario” to the analysis in which minGOsize was 0.1 and minGOsize:20. We considered that by changing the minGOsize to 9, we are adding to the analysis more specific GO-terms and by removing the filter on maxGOsize (maxGOsize=1) we are allowing for more general GO terms. The lowest value used for minGOsize was 9 because values below this give problems in the computation of the sparse matrices.

	# GO-terms		
scenario	**MinGOsize:9 maxGOsize:0.1**	**MinGOsize:20 maxGOsize=0.1**	**MinGOsize:20 maxGOsize=1**
Description scenario	**Adding more specific GO-terms**	**Normal**	**Adding more general GO-terms**
humans	3328	1982	2069
Chickens0.35	307	138	138
yeast	1772	1104	1187
yeast PPI	1734	1057	1153

Table 3: Impact of GO-size filter in the prediction performance

From table 3, we learn that the number of GO-terms after passing the filter was still low for chickens (307 GO-terms) when minGOsize was set to 9. Due to time constrains we will carry the analysis for the 138 GO terms in chickens when the GO-size filter is (20,0.1). The analysis, however, could be extended to 307 GO terms if the filter was changed to (9,0.1).

	Average AUC (sd)		
scenario	**MinGOsize:9 maxGOsize:0.1**	**MinGOsize:20 maxGOsize=0.1**	**MinGOsize:20 maxGOsize=1**
Description scenario	**Adding more specific GO-terms**	**Normal**	**Adding more general GO-terms**
humans	0.701 (0.017)	0.701 (0.017)	0.703 (0.017)
Chickens0.35	0.726 (0.08)	0.726(0.08)	0.726 (0.08)
yeast	0.757 (0.023)	0.764 (0.016)	0.761 (0.015)
yeast PPI	0.707 (0.028)	0.714 (0.02)	0.712 (0.018)

Table 4: Impact of GO-size filter in the prediction performance

From table 3, we learn that in the case of humans and chickens, the effect of the GO-size filter on the prediction performance was almost null (for the GO-size filters considered). For yeast and yeast PPI, the impact was very small, leading to slightly better performance when the filter was “normal”. Intuitively we would expect a better performance in the scenario “Including more General GO-terms”. However, we did not observe so. This could be regarded as an indication that the relationship between the specificity of the GO-term and the prediction performance is not linear.

This increase in AUC may come because predictions are made for a different subsets of GO terms. For instance, more strict filter leads less GO terms passing the filter. An additional analysis, however, showed that the prediction of individual GO terms are not affected by the GO-size filter. We carried applied BMRF on human data with minGOsize:20, 150 and 400, and we computed AUC only on those GO terms that were predicted with the three filters. The mean AUC (and sd) were:: 0.693 (0.05), 0.693 (0.05) and 0.692(0.05), respectively, indicating that the prediction performance of a GO term is independent on how many GO terms are considered. Table xx in this appendix, however we observe dthat the predictions are, howver, not independt on the number of labelled

genes of other GO terms in the data differnet form th etarget GO terms.

- Number of folds in k-validation

The number of folds in the k-validation is an important model parameter because the association data is highly unbalanced (for each GO term, the number of unlabeled genes is much larger than the number of labeled genes), and therefore low values of k may result in a inadequate use of the train set (using less labeled genes than are actually available), and high values of k may result in an inaccurate prediction performance because AUC may be estimated based on a excessively low number of labeled genes. Table 4 shows the results when we carried analysis using different values of k.

	Average AUC (sd)			
K-fold	**2**	**5**	**10**	**20**
humans	0.667 (0.034)	0.694 (0.021)	0.701 (0.017)	0.705 (0.01)
Chickens0.35	0.7 (0.083)	0.718 (0.082)	0.726(0.08)	0.75 (0.066)

Table 5: Impact of the number of folds in the prediction performance.

Table 4 shows that in human data k:10 is sufficient to achieve the highest possible prediction performance, whereas in chickens, the prediction performance using k:20 instead of k:10 is slightly larger. This not surprising since in chickens the number of labeled genes per GO term is much lower than in humans and a higher value of k is translated in train set with more labeled genes and better prediction performance. In humans, however, the train set seems to have already a large number of positive cases when k is equal to 10. We will choose the value 10 for the model parameter 'k', because larger values imply that the number of positive cases in the training set may not be sufficient (at least for some of the folds), and the estimates of accuracy of prediction become less reproducible.

- Number of iteration in Gibb-sampling

In BMRF, Gibb sampling is used to estimate the label of the unknown genes (a definition of unkown genes is given in Appendix I-concepts. We investigated whether more iterations are required when the number of unknown genes was large (default value of GS is 30 iterations). We carried analysis when the number of unknown genes was above 3000 (this was achieved with a minGOsize of 400), and we estimated AUC when GS was 30 and 500. We observed that the mean AUC and the mean standard deviation across replicates were near identical in both analysis. We therefore concluded that increasing the number of GS iterations is not helpful when the number of unknown genes becomes very large. We therefore chose value for the number of GS iterations 30 in all analysis.

- Non-validated data and domain information

We investigated the effect of adding domain information and non-validated GOterm-gene associations in the analysis. Table 6 shoes the results in three scenarios defined based on the information used.

	Average AUC (sd)		
scenario	**Not including information**	**Including domain information**	**Including both, domain info. and non-valid info. (normal approach)**
humans	0.654 (0.027)	0.701 (0.017)	0.705 (0.017)
Chickens0.35	0.57 (0.068)	0.724 (0.08)	0.726(0.08)
yeast	0.773 (0.093)	0.792 (0.089)	0.764 (0.016)
yeast PPI	0.73 (0.103)	0.747 (0.0992)	0.714 (0.02)

Table 6: Impact of domain info. and non-valid data on the prediction performance.

In table 6, we would intuitively expect an increasing AUC from the left of the table to the right (as more information was included in the analysis). However, in the case of yeast and yeast PPI, we observed the best prediction performance when the domain information was included but not the non-valid info. Furthermore, predictions were better when non of the sources of information was included (left) than when both sources were considered. This is a clear indicator that in the case of yeast, and yeast_ppi, the non-valid information worsens the prediction performance. A possible explanation for this is that in the case of yeast, a very large portion of the gene-function associations are validated and therefore including the non validated information increases the noise without a corresponding increase in the accuracy of prediction.

In the case of chickens and humans, the domain information is more relevant than for yeast and yeast_ppi, accounting for roughly 5% higher AUC. Whereas the non-valid information slightly helps in chickens and humans. An explanation could be that in the absence of enough validated data for humans and chickens, adding non-validated information may add noise but it also improves the resources in the network method.

Part 2 – Choice of the improvement strategy

Part 2 consists of 4 sections:

Section 2a: Differences in prediction performance across species
Section 2b: Impact of the GO-term-properties on the prediction performance
Section 2c: Impact of the quality of the data on the prediction performance
Section 2d: Impact of the nature of the network on the prediction performance

All analysis in part 3 were carried with the model parameters described in tableX1 of Results.

Section 2a. Differences in prediction performance across species

By comparing the characteristics of the network in the different species and the prediction performance, we can gain some understanding on which network characteristics are more relevant for protein function prediction via BMRF. In order to have more cases and species to compare, in this section, we carried the analysis with the chicken data when the co-expression threshold was 0.7 (common threshold), in addition to 0.35 (chosen threshold).

It is expected that the total number of edges of the network may be of limited importance for PFP because it may be that most of these edges are connecting genes that are not known to have the function, or genes that are known to have a given function with genes that are not known to have the same function. A more important network parameter therefore may be, for instance, '#epp' (edges of positive-positive). We compared the #te, #epp, #epn and #enn (Appendix I-Concepts) for the different species and we studied the relationship between these parameters and the prediction performance (AUC - Area under the curve). Tables 6 and 7 summarize the main differences in the characteristics of the network of the different species and the prediction performance. Note that here, #epp, for instance, refer to the sum of all the epp of all the GO terms in that particular species. And the same applies to #epn, #enn and #te.

	#te	#epp (epp*100/te)	#epn (epn*100/te)	#enn (enn*100/te)	AUC
yeast ppi	401,820	264,347 (65.79)	123,152 (30.65)	14,321 (3.56)	0.734
yeast	598,174	382,450 (63.94)	186,722 (31.22)	29,002 (4.85)	0.775
humans	1,548,622	481,792 (31.11)	754,276 (48.71)	312,554 (20.18)	0.712
Chicken_07	100,764	24 (0.02)	2,232 (2.22)	98,508 (97.76)	0.728
Chicken_035	2,094,870	576 (0.03)	51,610 (2.46)	2,042,684 (97.51)	0.762

Table 7: #edges and AUC

#te: total number of edges, epp: edges positive-positive. epn: edges positive-negative. enn: edges negative-negative

Based on table 6, we concluded that #te, #epp, #epn and #enn may be of limited importance for the prediction performance. We would expect that epp*100/te would be more directly related to the prediction performance than #epp because #te is the sum of #epp, #enn and #epn and, in principle, prediction will be more difficult when #epn and #enn are larger. This is because #enn and #epn may be connecting positives and negatives genes, and BMRF will fail to classify the positives and negatives genes. In table 1 we observed that epp*100/te may be playing a role in the prediction accuracy, because in yeast, epp*100/te is higher than for humans and so it is the prediction performance. However, in the case of chickens the epp*100/te is very low and predictions performance is still high. We therefore concluded that the epp100/te is also not informative of the prediction performance.

We then studied the degree of connections between the genes of a given GO terms, in the different species. One way to do this is by comparing the portion of epp with respect to the total possible number of epp (tpepp). Tpepp is a constant different for each GO term and refers to the total number of edges if all the genes associated with the GO term were interconnected. Tpepp is calcuated as: n*(n-1)/2, where n is the number of genes that are associated with the GO Term.

	epp*1000/tpepp	epp/tpepp*1000 corrected by epp and standarized	AUC
yeast ppi	47.88	-0.449	0.734
yeast	63.37	-0.449	0.775
humans	38.63	-0.449	0.712
Chickens_07	210.56	1.789	0.728
Chickens_035	28.15	-0.442	0.762

Table 8: epp by tpepp

epp: edges positive-positive; tpepp: total possible epp

AUC: area under the curve. Mean AUC of all GO terms that pass the filter considering only validated associations between the GO term and genes.

From table 3, we learn that with the exception of chickens data, there seem to be a favorable relation between epp/tpepp and AUC. For chickens_07, epp/tpepp is considerably larger than for the other cases. A possible explanation is that for chickens_07, the quality of the network dat ais very high and epn are less common. AUC, nevertheless, is not larger for chickens_07. Thus, we concluded that epp/tpepp is not a direct indicator of the prediction performance.

Then, we investigated the relationship between the level of annotation for one species (this includes the average of genes per GO-term and average GO-terms per gene), the #edges per gene and the #epp per GO-term (Appendix I-concepts), with the prediction performance.

	mean (sd)				AUC
	labels/go	go/labels	edges/label	Epp/GO	
yeast_ppi	11.31 (46.29)	17.05 (22.67)	156.39 (179.3)	960.12 (10844.65)	0.734
yeast	12.01 (48.78)	18.12 (22.77)	213.7 (146.61)	1311.15 (15209.25)	0.775
humans	11.24(59.85)	25.63 (42.51)	310.36 (80.50)	954.16 (11417.71)	0.712
Chickens_07	0.28 (0.97)	0.12(0.85)	43.02(49.12)	0.14642(1.199437)	0.728
Chickens_035	24.55 (26.14)	0.87 (6.23)	811.29 (1097.3)	6778.91 (110825.9)	0.762

Table 9: Differences between the network data of the different species

Co-expression data for yeast has higher #labels per GO and epp/GO, whereas human co-expression data has higher in #GO/labels and #edges/label. Since we achieve a higher overall AUC for yeast, we can expect that #labels per GO and #epp/GO are more directly affecting the AUC. Further, we could expect that in order to achieve higher AUC (>0.75), data should have a large #labels/GO (~12) and ~1000 epp/GO. We observe that co-expression data for chickens_07 is currently far from these numbers. However, when we use chicken_035 data “#labels per GO”, #edges/label and “Epp/GO” increase to levels even higher than for the other species (#go/labels reminds much lower than for the other species). It is therefore not surprising that the mean AUC is higher for chickens_035 than for yeast_ppi and humans.

From section-A, we concluded that the network characteristics of the different species (#te, #epp, #epn, #enn, #epp/te and epp/tpepp) are not direct indicators of the prediction performance. However, when we looked at these parameters at the level of individual GO-terms, and then doing the average for all the GO-terms, (instead of summing these values), we observed that some of these properties (#labels-per-GO and #epp/GO) are potentially good indicators of the prediction performance. We therefore, followed the analysis, in section B, investigating the effect of GO-term-properties on the prediction performance rather than the differences in network characteristics between the species.

Note that some of the GO-term-properties defined in section B may have a similar name than the network-characteristics that we used in section A to describe the differences between the networks of the different species. However, in section A, the network characteristics refer to the sum of all the GO terms of the species, whereas in section B, the GO-term-properties refer to the same parameter at the level of individual GO-terms. Also, in section A we adress the question 'Why predictions are overall more accurate for one species than for others?', whereas in section B we adress the questions 'Why predictions are more accurate for some GO terms than for others?'

In this part we get a better insight on the prediction performance for the species considered using BMRF. Illustration4 and 5 show the distribution of AUC for the different GO terms, and the standard deviation, respectively.

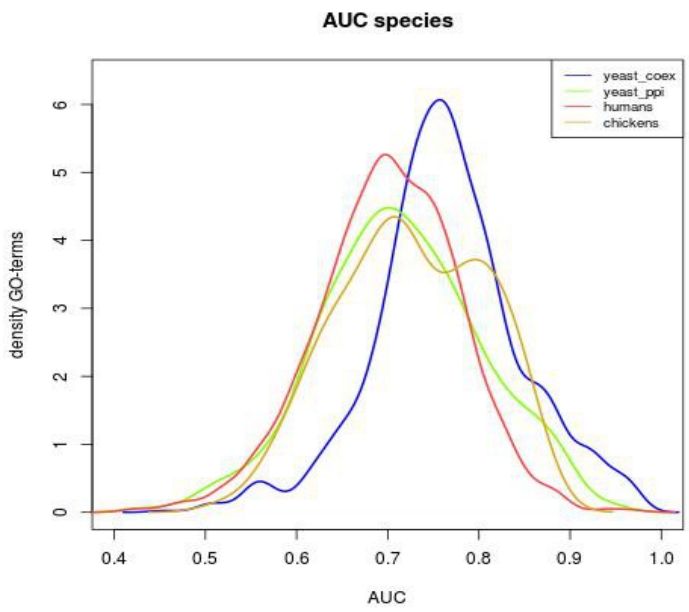


Illustration 4: Distribution of AUC for the different GO-terms in the different species.

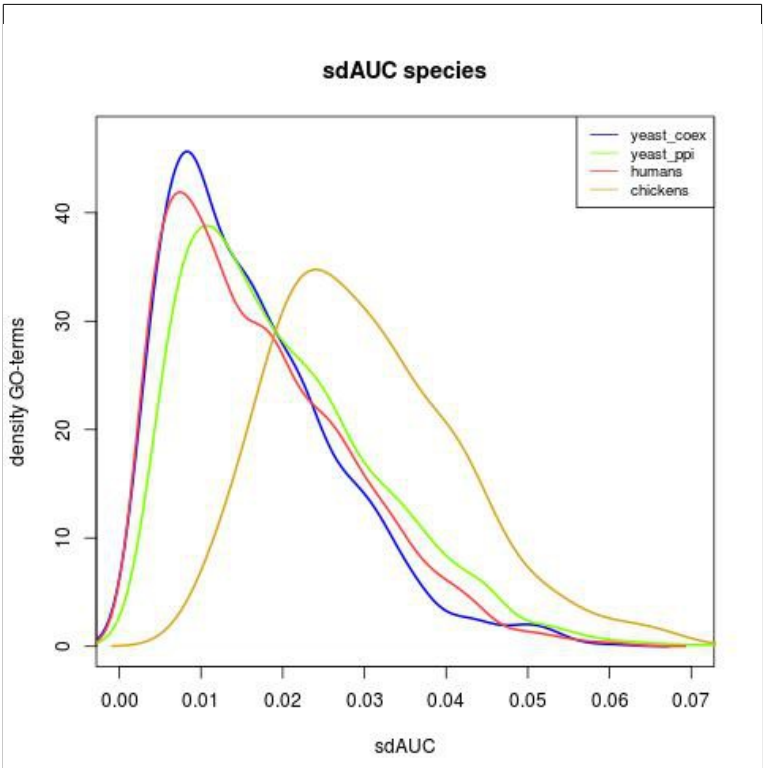


Illustration 5: Distribution of the standard deviation of AUC across replicates for the different GO-terms in the different species.

From illustration 4, we observe that AUC is marginally larger for yeast and that the highest value of AUC is translated in an overall incerase of AUC in the individual GO terms. Prediction perfomance for yeast_ppi, humnas anc hcikens is more similar, and seem sto be larer for chichekns than fo rhumans and for humans than for yeast_ppi. We also observe that the right side of the curce decreases more sharply for choickens than in the other 3 cases.

Illustration 5 shows a similar distribution of standtad deviation across replicates for the different GO terms in humans, yeast amnd yeast_ppi, whereas for chckens the standr deviation is considerably larger. This was expeted since the number of genes that are associated with the GO terms is much lower fo this species (Illustration 2 -Appendix II-Overview of data) and a lower number of positive scases is expected to be associated wit ha higher standrad deviation, because the training set is more unblanced and the predictionwill depend on which of the postive genes enter the trainng roi the test set.

To get an overview of how the AUC of specific GO terms change depending on thespecies considerd, we ploted the AUC of 20 randomly chose GO terms that were predicted in the four cases (illustration 7).

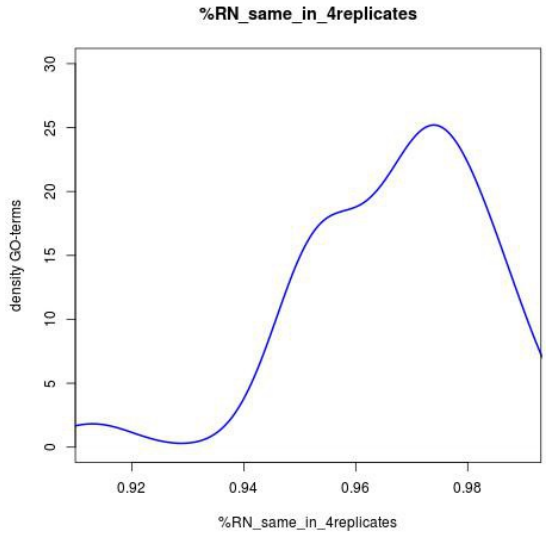


Illustration 6: Reproducibility in the process of extraction of RN when a maximum of 3000 RN were exctracted

Illustration 6 shows a more visual rep[resentation of the prediction performance in the species considered.

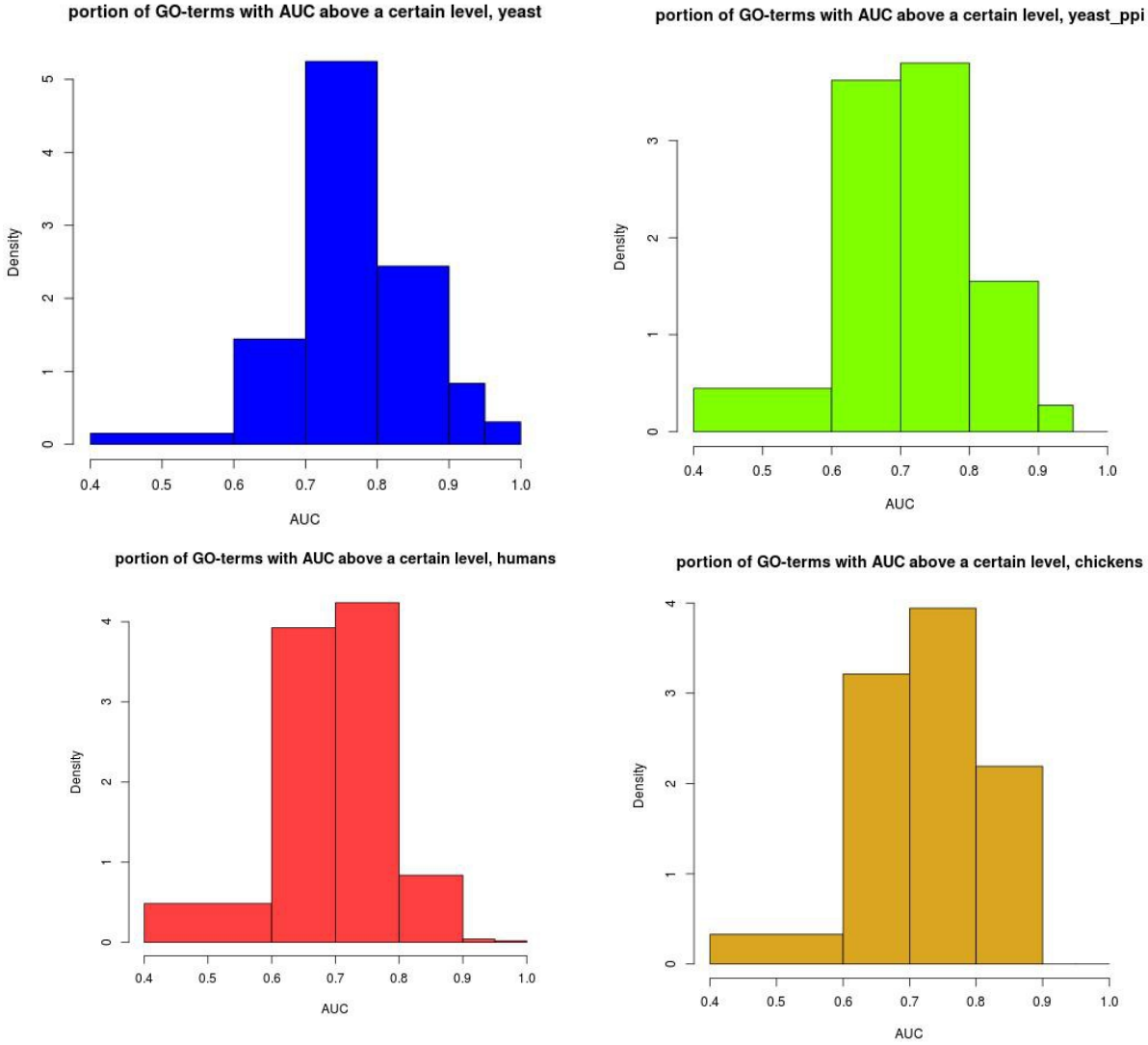


Illustration 6, shows once more that predictions are better for yeast, the nchieksn and yeast_ppi and finally humans. Howeverm, we observe that whereas for yeas_ppi there are a few eGO terms whose accuracy is above 0.9, this is not the case for any of the Go terms predictens with chicken data. It is interesting to investigate how the histogram will change when PU-BMRF is used.

		AUC(sd)	
#RN	BMRF	BMRF random extract	BMRF-PU
1000		0.643 (0.094)	0.723 (0.08)
2000		0.65 (0.09)	0.75 (0.072)
3000		**0.64 (0.1)**	**0.758(0.084)**
4000		0.636 (0.095)	0.751(0.086)
5000		0.652 (0.087)	0.725 (0.094)
6000		0.644 (0.087)	0.728 (0.089)
7000		0.638 (0.096)	0.716 (0.092)
8000		0.636 (0.095)	0.701 (0.095)
all	0.706 (0.0793)		

Table 10: Comparison accuracy of prediction BMRF vs PU-BMRF

Section 3b. Impact of the GO-term properties in the prediction performance

This section should call to the netwrok propertie sin appendix II

We defined a total fo 9 GO-term-properties including epp/tpEpp, eppV/tpEppV, #genesV, sepc, teV/tpeV, depth, AUC and sdAUC. Definitions of these are in Appendix I-Concepts. Tables 5-9 show the correlation between these parameters for yeast, yeast_PPI, humans and chickens, respectivelly. Only Significant correlations (pvalue<0.05) are shown. Note that tables 5-9 are a detailed version of the correlation shown in table XX of results. Although in results only the correlations with AUC are shown and tables 5-9 include correlations between these parameters.

Var1	Var2	corr	p_value
epp/tpEpp	epp_V/tpEppV	0.968	0
#genesV	spec	0.875	0
epp_V/tpEppV	teV/tpeV	0.834	0
epp/tpEpp	teV/tpeV	0.83	0
AUC	epp_V/tpEppV	0.62	0
#genesV	sdAUC	-0.591	0
AUC	epp/tpEpp	0.582	0
sdAUC	spec	-0.497	0
AUC	sdAUC	-0.408	0
AUC	teV/tpeV	0.394	0
depth	spec	-0.341	0
depth	#genesV	-0.303	0
AUC	depth	0.265	0
depth	epp_V/tpEppV	0.264	0
depth	epp/tpEpp	0.262	0
epp_V/tpEppV	spec	-0.192	0
epp/tpEpp	spec	-0.186	0
depth	sdAUC	0.176	0
epp_V/tpEppV	#genesV	-0.145	0
epp/tpEpp	#genesV	-0.135	0
epp_V/tpEppV	sdAUC	-0.104	0.0006
depth	teV/tpeV	0.099	0.001
AUC	spec	-0.095	0.0015
sdAUC	teV/tpeV	-0.093	0.002
epp/tpEpp	sdAUC	-0.092	0.0021
spec	teV/tpeV	-0.085	0.0046

Table 11: correlations_yeast

Table 12: Correlations yeast_PPI

Var1	Var2	corr	p_value
epp/tpEpp	epp_V/tpEppV	0.959	0
#genesV	spec	0.754	0
epp_V/tpEppV	teV/tpeV	0.575	0
#genesV	sdAUC	-0.533	0
epp/tpEpp	teV/tpeV	0.479	0
AUC	epp_V/tpEppV	0.373	0
sdAUC	spec	-0.359	0
AUC	epp/tpEpp	0.34	0
depth	spec	-0.268	0
AUC	teV/tpeV	0.241	0
depth	epp_V/tpEppV	0.234	0
depth	epp/tpEpp	0.23	0
depth	#genesV	-0.226	0
epp_V/tpEppV	spec	-0.199	0
epp/tpEpp	spec	-0.196	0
epp/tpEpp	#genesV	-0.163	0
epp_V/tpEppV	#genesV	-0.158	0
AUC	depth	0.104	0.0007
depth	sdAUC	0.102	0.0009

Var1	Var2	corr	p_value
#genesV	spec	1	0
epp_V/tpEppV	teV/tpeV	0.723	0
#genesV	sdAUC	-0.562	0
sdAUC	spec	-0.562	0
AUC	#genesV	-0.518	0
AUC	spec	-0.518	0
depth	#genesV	-0.512	0
depth	spec	-0.512	0
AUC	depth	0.478	0
epp/tpEpp	teV/tpeV	0.45	0
epp/tpEpp	sdAUC	-0.355	0
AUC	teV/tpeV	-0.263	0.0017
depth	teV/tpeV	-0.254	0.0025

Table 13: Correlations GO-term-properties chickens

Table 14:

Var1	Var2	corr	p_value
#genesV	spec	0.999	0
epp/tpEpp	epp_V/tpEppV	0.728	0
#genesV	sdAUC	-0.539	0
sdAUC	spec	-0.538	0
AUC	epp_V/tpEppV	0.462	0
AUC	epp/tpEpp	0.378	0
AUC	sdAUC	-0.337	0
epp/tpEpp	#genesV	-0.248	0
epp/tpEpp	spec	-0.247	0
epp_V/tpEppV	#genesV	-0.236	0
epp_V/tpEppV	spec	-0.235	0
depth	spec	-0.134	0
epp/tpEpp	sdAUC	0.13	0
AUC	teV/tpeV	-0.13	0
sdAUC	teV/tpeV	0.129	0
spec	teV/tpeV	-0.129	0
#genesV	teV/tpeV	-0.128	0
epp_V/tpEppV	sdAUC	0.122	0
epp/tpEpp	teV/tpeV	-0.104	0
depth	#genesV	-0.09	0.0001
depth	sdAUC	0.073	0.0011

Correlations humans

From Table 5-9, we observed:

- The strong correlation between epp_V/tpeppV (or epp/tpepp) and teV/tpeV could be regraded as an indicator of the positive genes are more interconnected than the genes that do not have the functions. Thus, to some extend, a higher correlation between epp_V/tpeppV and teV/tpeV is related to a most completed state of annotation. We observe that the correlation is higher for yeast, yeast_ppi (0.83, 0.57, respectively), than for humans and chickens (not significant in both cases), which is in line with what we observed in Illustration 1-3 of Appendix II-Data overview.

- The strong negative correlation between #genesV and sdAUC, was observed in the four cases, and suggests that reproducibility is higher for GO terms with a large number of validated labeled genes. Which makes sense, because when the number of labeled genes is low. The prediction will depend on which of the few labeled genes enter the train or the test set. This is in line with what we observed din Illustration 5 of this appendix, when we observed that the standard deviation was much large for chickens and an explanation for this could be that for this specie the number of associated genes is much lower than for the rest of the species.

- The correlation between specificity and depth was strong and negative (-0.3), indicating that GO terms that are associated with more GO terms have a lower depth, which is in line with the definition of depth.

- Corr between epp/tpEpp and epp_V/tpeppV is stronger for yeast than for yeast_ppi and

stronger in yeast_ppi than in humans, and is considerably lower for chickens. This is because a large portion of the gene-GO-term associations are validated in the first three cases whereas in chickens the portion is much lower.

- The correlation between specificity and #genesV is slightly lower for yeast and yeastPPI than for humans and chickens. An explanation for this is that spec by definition is the sum of all associated genes, whereas in the case of yeast or yeast_ppi, most of the annotations found are also found in the other three cases, whereas in chickens and humans some associations are found only in these species.

Additional analysis on yeast data showed that the correlation between AUC and the number of edges becomes stronger as the number of annotations in the data becomes smaller. Table XX shows the value of the correlation AUC-#edges for different sizes of association files.

#assoc.	CorrAUC_ #edges
264,279	0.025
132,249	0.124
111,480	0.252
110,682	0.255
104,582	0.174
104,303	0.133
58,358	0.376
32,336	0.356
8,862	0.219

Table 15: Impact of the number of annotations available in data on the correlations AUC-#edges and AUC.

From table 8, we learn that the highest correlation between AUC and t#edges (0.376) was reached when the number of association in data was around 1/4 of the total of associations available for yeast. Then, with this number of associations (58,358), we compared AUC in different subsets of GO-terms defined based on the number of edges. Table 9 shows the prediction performance for 5 bins of GO terms. The first bin corresponds to the 1th/5th of the GO terms with lower number of edges, the escond bin corresponds to the 2th/5th, and so on.

Bin of GO terms	AUC
1th/5	0.648
2th/5	0.654
3th/5	0.674
4^{th}/5	0.706
5^{th}/5	0.730

Table 16: Differences on AUC between groups of GO terms with different # of edges when data had 58.358 associations. The first bin (1th/5) refers to the 1th/5th of the GO terms with a lowest number of edges, and so on.

Results in table 9 show to which extend the number of edges of the GO term can affect the prediction performance when the data available for the analysis is limited. We therefore conclude that the correlations between the GO-term-properties and AUC may differ for species with a different portion of annotations gene-Goterm known.

Additional analysis on humans data showed the correlations between epn and epp is strong (0.897).

This coud be regarded as an indicator that either, the annotation is incomplete or the quality of the co-expression data is low. We would expect that as the quality of the data increases, this correlation would weaken. The same analysis also showed a significant negative correlation between epn and Epp/tpepp (-0.25), which was expected, since the tpepp is the sum of epn and epp.

Section 3 C. Impact of the quality of data on the prediction performance

In this section we investigate how the predictions vary when the data becomes more incomplete. This information can be used to get some insights on to which extend BMRF can be used for poorly annotated species. For this, we investigate how performance is altered in different situations. Using yeast data we investigated the effect of the network size on the prediction performance and then with humans data we investigated the effect of decreasing the quality of the data by removing edges of a specific class (i.e. epp, epn and enn) and associations between GO-terms and genes.

Using yeast data

We randomly extracted from the network a known percentage of the edges and calculated the prediction performance. Table X shows the results of this analysis.

Portion of edges extracted from data	Mean AUC
0% (all network data used)	0.744
10%	0.738
30%	0.733
50%	0.738
90%	0.719
95%	0.719

Table 7: Impact of number of edges in the prediction performance, using yeast data.

From table X, we observed that **r**emoving random edges from the data did not seem to affect much the prediction performance. We observed a larger impact after removing 10% of the edges and after removing more than 50% of the edges. Also, this decrease in the AUC can be caused by the fact that less GO terms were considered in the anlaysis. This is because in BMRF the GO-term annotation is pruned based on the network. However, when we looked at individual GO terms (Table X), we observed a similar pattern of decrease of AUC for all GO terms.

GO-term	total_labels	valid_labels	portion of edges substracted					
			0% (all data used)	**10%**	**30%**	**50%**	**90%**	**95%**
GO:0042981	30	30	0.741	0.726	0.734	0.778	0.685	0.699
GO:0014068	30	30	0.48	0.479	0.514	0.495	0.504	0.493
GO:0045931	31	25	0.649	0.628	0.632	0.63	0.665	0.682
GO:0000209	36	23	0.862	0.872	0.773	0.837	0.857	0.837
GO:0006664	39	32	0.844	0.853	0.837	0.821	0.77	0.775
GO:0031670	61*	50*	0.811	0.789	0.796	0.789	0.796	0.79
GO:0036503	62*	49*	0.855	0.844	0.85	0.827	0.803	0.819
GO:0006414	65*	40	0.756	0.752	0.755	0.757	0.733	0.714
GO:0006417	144	100*	0.728	0.732	0.741	0.745	0.73	0.731
GO:0044270	195*	166*	0.714	0.703	0.701	0.705	0.644	0.649

Illustration 8: Impact of the extraction of edges in prediction performance for individual GO terms

We are also interetsed in knowng how the annotation would worsen if less network data was available., We acrried ananalysis with yeast data and we observed that, for yeast dat, the number of validated proteins decreases almost linearlt with the number of edges (tables X and Illustratin). Therefore we should consider that for species for which network dat ais limited, the annotation will be consequently limited as well.

Table 17: Relationship between the

Scenario	**mean(#edges)**	**mean(#validated labels)**
“stress”	4200.727	86.05
only validated associations	18845.56	94.477
“normal”	24007.55	94.477
focus on top	25333.17	98.704
more goes	31496.92	123.806
default	44175.61	173.613

number of edges and the number of validated genes, for yeast data.

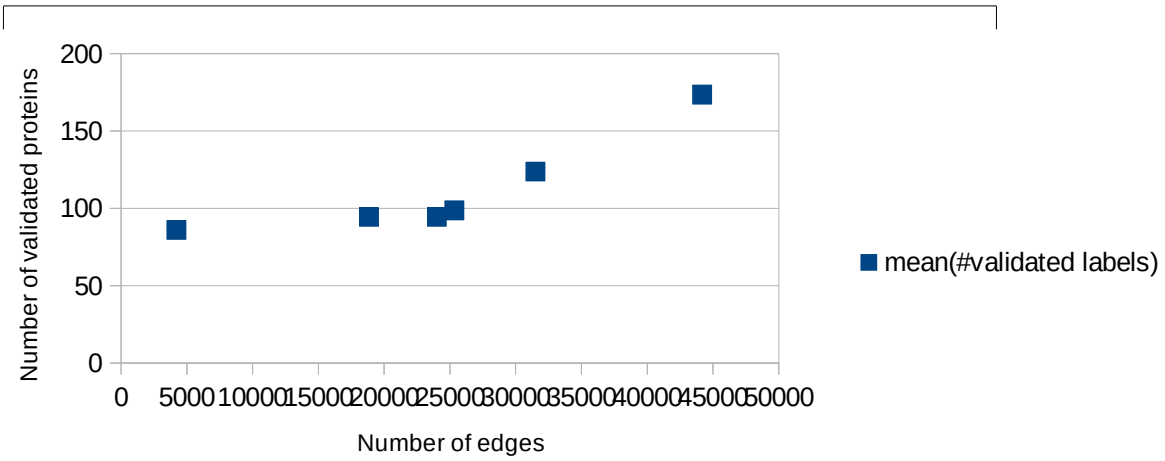


Illustration 9: Relationship between the number of edges in the network and the number of validated proteins in the analysis, for yeast data

Humans data

With human data, we investigated the effect of randomly decreasing the size of the network on the prediction performance.

- If several epp, epn or enn were removed from the data ('reduceEpp', 'reduceEpn', 'reuduceEnn', respectively)

- If several associations GO-gene are missing in the data. We defined two factors: 'reduceAmg' refers to removing from the data associations between the target GO term and the validated genes, and 'reduceOa' refers to removing associations between other GO terms and their validated-labeled genes. Removing associations of other GO terms is expected to have an effect on the prediction performance because more genes will appear as unknown and the relabeling will be initialized through Gibb-sampling.

	Correlation
AUC_reduceEpn	0.98
AUC_reduceEnn	0.95
AUC_reduceAmg	0.67
AUC_reduceOa	-0.47
AUC_reduceEpp	-0.26

Table 18: Correlation between AUC and data quality, for human data

From table 11 we learn:

- AUC will increase linearly as we remove epn from the data. This is in line with two of the observation from previous analysis. In table 3, we learn that the GO-term-property epp/tpepp is a good indicator of AUC, and then an additional experiment showed that epn is negatively correlated with epp/tpepp. Thus, as epn decreases, epp/tpepp increases and therefore AUC increases as well.

- AUC will also increase almost linearly as we remove enn. This also makes sense since we would expect that some of the enn are atually connecting negative genes with positive genes.

- AUC will increase if we remove associations from the target GO term. A possible explanation for this is that α will be lower in Equation 1, and therefore less genes will be classified as positive. Since a very low portion of the genes are true positives, AUC increases

$$\alpha \sum_{i=1}^{N} x_i + \beta^1 N_1 + \beta^0 N_0$$ Equation 1

- AUC however will be reduced if we remove annotations between genes and other GO terms. A possible explanation of this is that some of these genes will enter the category of unknown and their labeling will be initialized through Gibb-sampling, increasing the margin error.

- Lastly, as expected, removing epp leads to lower AUC.

•

We also investigated how the correlations in table 12 vary among GO terms.

AUC	AUC_reduceEpp	-0.38
AUC	AUC_reduceEnn	0.36
AUC	AUC_reduceEpn	0.20
AUC	AUC_reduceOa	0.18

Table 19: Correlation between AUC and the effect of data quality on AUC

From table 12 we learn that GO terms whose AUC increases when removing Epp, have high AUC. It is counter-intuitive that by removing Epp, for some GO terms we achieve higher AUC, It can nevertheless be the case because we will identify more labels as negatives, and consequentially it will be more easy to identify the negative cases, which are much more frequent. Thus AUC may increases by increasing the specificity at the cost of a lower sensibility.

Section 3D. Impact of the nature of the networks on the prediction performance

By nature of the network here we refer to the characteristics of the co-expression analysis. It is important to investigate whether, for instance, a co-expression analysis addressed to one specific tissue allows to make more accurate predictions for those GO terms whose function is more relevant in that tissue. Note that, from a biological perspective, and considering that network analysis exploit the principle of guilt-by-association, we would expect that the nature of the network has an impact on the prediction performance.

Using yeast data

Using yeast data, we investigated whether the nature of the co-expression network (characteristics of the experiment) have any impact on the prediction performance. For this, we choose 5 different subsets of the network. In the first subset, we included the co-expression data from all experiments refered to stress, in the second subset we considered all co-expression analysis involving oxidation. Then, since the size of the network differs between these subsets, we choose other subsets of experiments of a controlled size. Subsets 3 and 4 are of similar size and refer the names of the experiments are “Sodium arsenate response of wild type and slt2 deficient cells” and “Integration of the general amino acid control and nitrogen regulatory pathways in yeast nitrogen assimilation”,

respectively. The last subset of network corresponds to an experiment with very low number of co-expressed genes. The name of the experiment is “The metabolic response to iron deficiency in Saccharomyces cerevisiae”, under the dta source indicated in table X of appendix.

scenarios	data					
scenario name	Network size (#edges)	#unknown genes*	#proteins.	#assoc.	# GO-terms	AUC mean(sd) [median]
“stress” co-expression	98,479	471	4,879	110,682	1,021	0.727 (0.089) [0.723]
“oxidiation” co-expression	64,167	499	4,923	111,480	1,022	0.72 (0.086) [0.714]
similar_size_ network Experiment1	28,800	298	1,865	32,336	426	0.684 (0.101) [0.677]
similar_size_ network Experiment1	27,488	255	2,899	58,358	681	0.682 (0.089) [0.687]
very small network	7,073	112	661	8,862	203	0.635 (0.113) [0.614]

Table 20: Impact of the nature of the network on the prediction performance using yeast data.

From table 11 we concluded that the nature of the network does not seem to have a strong impact on the prediction performance, even if the network size and other parameters ()#unknown genes, #proteins, #edges and #GOs) differed between the different subsets of data.

Using human data

In order to investigate whether there is biological support in the data, we have identified the GO terms for which a highest AUC was achieved using network data from different tissues. For a fairer analysis this we normalized the networks of the different tissues based on epp/tpepp. Table 13 shows a list of the GO terms that were more accurately predicted using networks based on co-expression analysis in different tissues, and table 14 shows a list of the Tissues for which the co-expression analysis led to better and worse accuracy of prediction, for 10 randomly chose GO terms

tissue	top1_GOterm	top2_GOterm
Stomach	post-Golgi vesicle-mediated transport	positive regulation of lipid transport
Esophagus-Muscularis	anoikis	intrinsic apoptotic signaling pathway in response to oxidative stress
Thyroid	erythrocyte differentiation	cell aging
Whole_Blood	negative regulation of epithelial cell migration	keratinocyte proliferation
Brain-Amygdala	histone H4 acetylation	protein destabilization
Adrenal_Gland	regulation of protein oligomerization	negative regulation of response to biotic stimulus
Brain-Putamen(basal_ganglia)	regulation of protein complex disassembly	negative regulation of protein binding
Brain-Cortex	receptor internalization	regulation of heart rate
Skin-Not_Sun_Exposed(Suprapubic)	regulation of toll-like receptor signaling pathway	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
Testis	positive regulation of viral genome replication	negative regulation of telomere maintenance
Brain-Anterior_cingulate_cortex(BA24)	positive regulation of viral genome replication	peroxisome organization
Pancreas	regulation of receptor internalization	TOR signaling
Brain-Spinal_cord(cervical_c-1)	regulation of receptor internalization	regulation of microtubule polymerization
Brain-Hypothalamus	negative regulation of DNA binding	positive regulation of telomere maintenance
Brain-Caudate(basal_ganglia)	negative regulation of dephosphorylation	cellular extravasation
Artery-Tibial	regulation of cell adhesion mediated by integrin	negative regulation of telomere maintenance
Pituitary	negative regulation of blood vessel endothelial cell migration	protein localization to cytoskeleton
Esophagus-Mucosa	negative regulation of cell projection organization	response to temperature stimulus
Lung	intrinsic apoptotic signaling pathway in response to oxidative stress	histone deacetylation
Skin-Sun_Exposed(Lower_leg)	regulation of interferon-beta production	myeloid cell homeostasis
Nerve-Tibial	negative regulation of cell-substrate adhesion	anoikis
Muscle-Skeletal	homotypic cell-cell adhesion	regulation of cell adhesion mediated by integrin
Breast-Mammary_Tissue	receptor internalization	regulation of protein complex disassembly
Brain-Nucleus_accumbens(basal_ganglia)	negative regulation of epithelial cell migration	positive regulation of DNA binding
Adipose-Subcutaneous	regulation of protein oligomerization	negative regulation of blood vessel endothelial cell migration
Heart-Atrial_Appendage	positive regulation of macroautophagy	negative regulation of blood vessel endothelial cell migration
Adipose-Visceral(Omentum)	regulation of cell adhesion mediated by integrin	regulation of smooth muscle cell migration
Artery-Aorta	positive regulation of actin filament bundle assembly	cellular response to amino acid starvation
Brain-Substantia_nigra	homotypic cell-cell adhesion	regulation of epithelial to mesenchymal transition
Heart-Left_Ventricle	regulation of DNA recombination	regulation of sodium ion transport
Brain-Hippocampus	interleukin-10 production	histone ubiquitination
Brain-Cerebellar_Hemisphere	lipid storage	smooth muscle cell migration
Colon-Transverse	regulation of cell adhesion mediated by integrin	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
Brain-Cerebellum	peroxisome organization	ATP-dependent chromatin remodeling
Brain-Frontal_Cortex(BA9)	regulation of phosphatase activity	cell aging

Table 21: List of the GO terms that were more accurately predicted using networks based on co-expression analysis in different tissues.

GOterm	tissue_highest_AUC	highest_AUC	tissue_loest_AUC	lowest_AUC
regulation of receptor internalization	Pituitary	0.586	Pancreas	0.459
peroxisome organization	Brain-Caudate(basal_ganglia)	0.734	Brain-Anterior_cingulate_cortex(BA24)	0.612
mitotic cytokinesis	Adipose-Subcutaneous	0.783	Testis	0.665
post-Golgi vesicle-mediated transport	Testis	0.758	Stomach	0.644
regulation of DNA recombination	Brain-Hippocampus	0.806	Heart-Left_Ventricle	0.693
histone ubiquitination	Nerve-Tibial	0.74	Brain-Hippocampus	0.628
negative regulation of response to biotic stimulus	Brain-Anterior_cingulate_cortex(BA24)	0.695	Brain-Hippocampus	0.585
negative regulation of epithelial cell migration	Brain-Frontal_Cortex(BA9)	0.626	Brain-Nucleus_accumbens(basal_ganglia)	0.517
erythrocyte differentiation	Muscle-Skeletal	0.683	Thyroid	0.577

Table 22: Tissues for which the co-expression analysis led to better and worse accuracy of prediction. Results for 10 randomly chose GO terms

From table 13, we observe that there is some biological support in the network data. For instance, for the 'Stomach' co-expression experiment the GO term “post-Golgi vesicle-mediated transport” was the most accurately predicted and “positive regulation of lipid transport“ was the second. In another example, for brain-cortex, “regulation of heart rate“ was accurately predicted and we know that in the cortex there is a circuitry of the medulla oblongata, which serves critical functions such as regulation of heart and respiration rates.

Similarly, for each GO term we identified the tissues for which predictions were best, and worst. Table 14 also shows biological support. For instance, it is known that Pituitary is related to regulation of receptor internalization and that the hippocampus can regulate DNA recombination [1]. Thus it is not surprising that the GO term “regulation of receptor internalization” is more accurately predicted using a network from a Pituitary expression experiment than, for instance, using pancreas data.

Illustration 2, however, shows that for most GO terms the difference in AUC from one tissue network to another was close to 0. In order to investigate this further, we plotted the distribution of difference in AUC between the two most different-prediction-tissues for each GO term. Illustration 2 shows that for most of the GO terms, perfections were almost independent of the tissues co-expression networks.

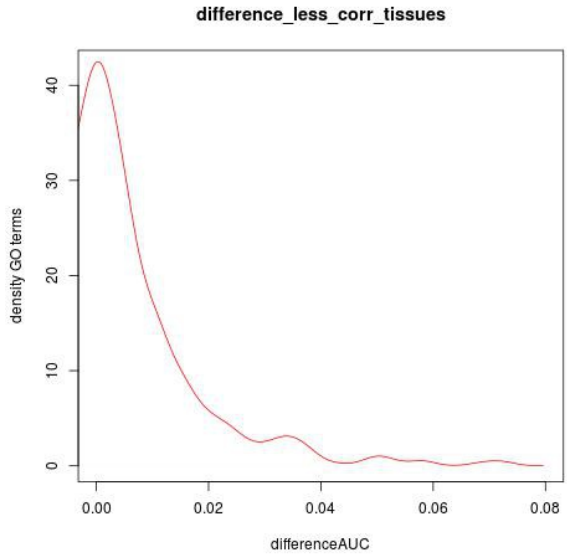


Illustration 10: Difference in AUC for the different GO terms

In Illustration 2, the y-axes represent the density of GO terms, and the x-axes represents the difference in AUC form the tissue with highest AUC for a particular GO term and the tissue with lowest AUC. Results imply that although there is biological difference between the networks, it does not seem to have a strong impact in the accuracy of PFP which network is used. This could be interpreted as “as long as there is enough data, the accuracy of prediction will depend more on the GO-term-properties than on the quality of the data”.

In order to have a more direct insight on what is the difference in PFP when using different tissues' co-expression networks, for each pair of tissues, we calculated the correlation between the AUC values for all GO terms. The minimum correlation between a pair of tissues was 0.977 (for Colon-Transverse and Brain-Frontal_Cortex). This implies that the effect of which network is used is very small, which is in line with the conclusion from Illustration 2.

Another prove of the low role that the nature of the network plays on th predictin perfomance is the fact that the overall AUC was very similar using the different subsets of network (sd across tissues=0/0005).

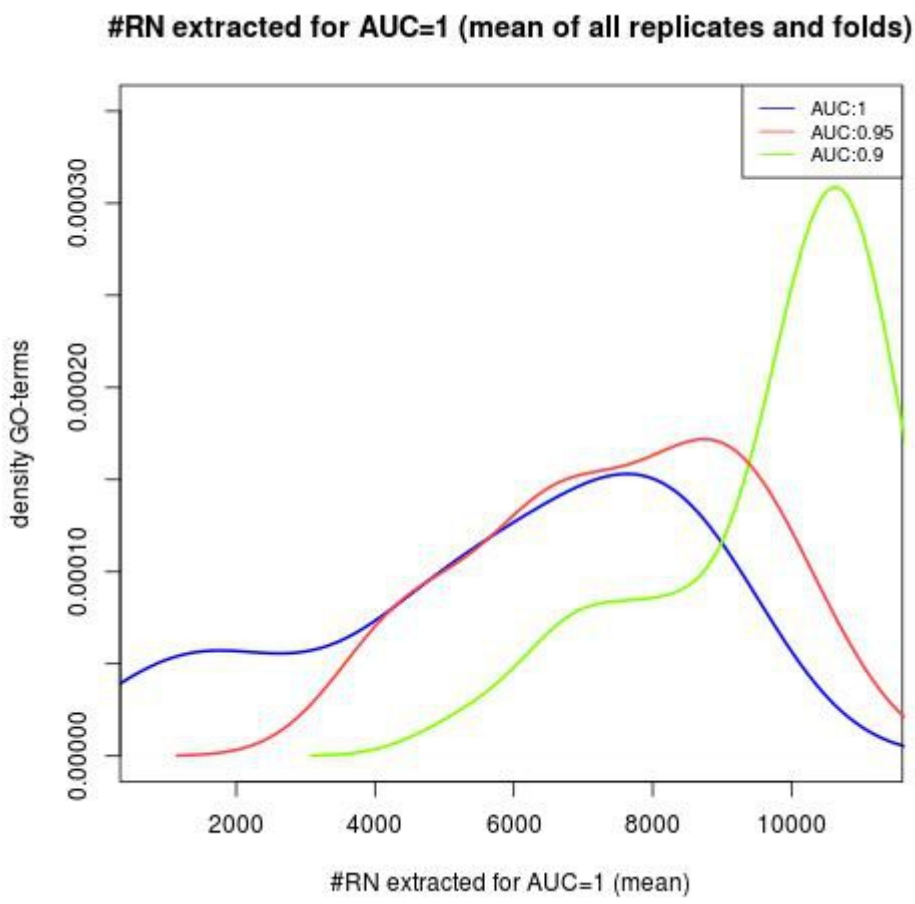
We, therefore, concluded that, as long as there is enough data, the accuracy of prediction will depend more on the properties of the GO term, rather than on the quality of the data. Therefore, in order to improve PFP, the development of method that can improve the GO-term-properties, like PU, that can improve the epp/tpepp ratio, may be more useful than achieving more accurate co-expression experiments.

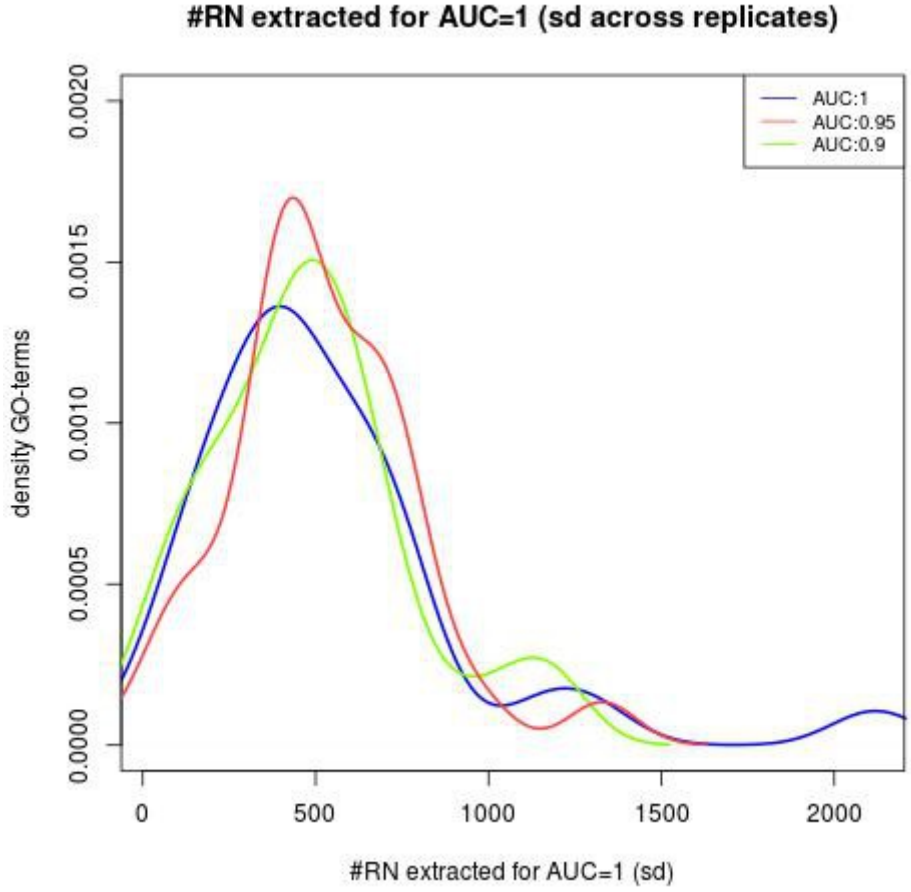
Part 3- PU-BMRF evaluation performance

		mean sd across replicates	
#RN	BMRF	BMRF random extact	BMRF-PU
1000		0.1436	0.129
2000		0.14	0.1268
3000		0.15	0.128
4000		0.135	0.121
5000		0.13	0.127
6000		0.129	0.1244
7000		0.123	0.116
8000		0.137	0.123
all	0.033		

Var1	Var2	corr	p_value
AUC_increase	epp_V/tpEppV	0.251	0.2058
AUC_increase	te/tp_e	0.222	0.2662
AUC_increase	epp/tpEpp	0.18	0.3689
AUC_increase	epp	0.153	0.4468
AUC_increase	#genes	0.148	0.4608
AUC_increase	eppV	0.134	0.5038
AUC_increase	#genesV	0.126	0.5319
AUC_increase	spec	0.126	0.5319
AUC_increase	teV	0.124	0.5361
AUC_increase	te	0.101	0.617
AUC_increase	depth	0.085	0.673
AUC_increase	teV/tpeV	0.04	0.8425

- **Extraction of RN**


#RN extracted for AUC=1 (mean of all replicates and folds)
AUC:1
AUC:0.95
AUC:0.9
density GO-terms
0.00000
0.00010
0.00020
0.00030
2000
4000
6000
8000
10000
#RN extracted for AUC=1 (mean)


#RN extracted for AUC=1 (sd across replicates)
AUC:1
AUC:0.95
AUC:0.9
density GO-terms
0.0000
0.0005
0.0010
0.0015
0.0020
0
500
1000
1500
2000
#RN extracted for AUC=1 (sd)

Results with k:2

#RN	AUC PU.BMRF	sd AUC PU.BMRF	AUC BMRF	sd AUC BMRF
2000	0.677	0.033	0.642	0.032
4000	**0.677**	**0.034**	**0.63**	**0.032**
6000	0.655	0.033	0.624	0.028
8000	0.641	0.033	0.617	0.03

Table 23: Comparison accuracy of prediction BMRF vs PU-BMRF

Accuracy in the extraction of RN:

Average value of tolerance (sd)			
tol_AUC85	**tol_AUC90**	**tol_AUC95**	**tol_AUC1**
0.885(0.088)	0.96(0.15)	1.2(0.346)	1.42(0.438)

Table 24: Values of "tolerance" that were required for diffrenet values of AUC in the process of extraction.

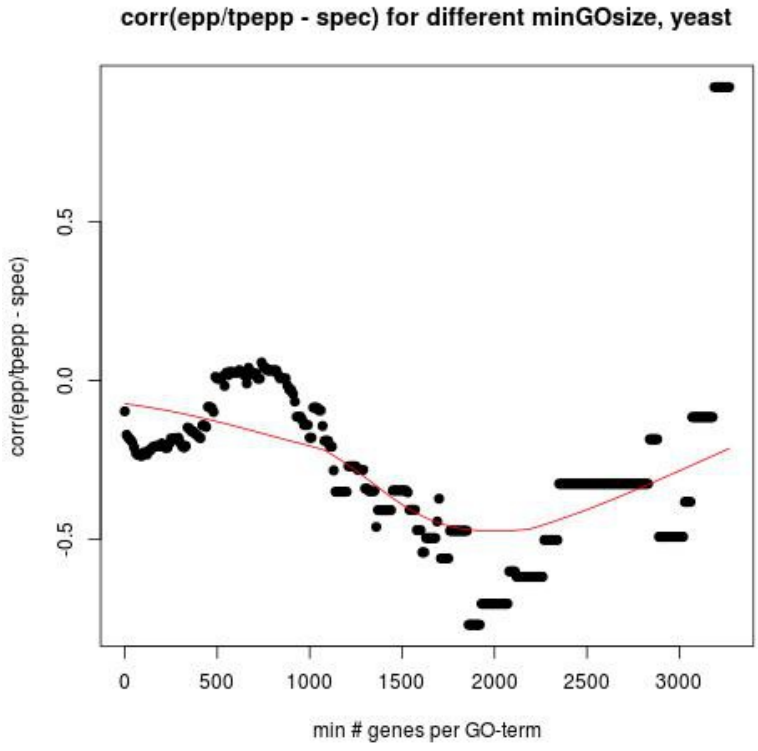
- **Computational time**

Step	Description	computational time in minutes for 1 GO-terms (% of total time)
1	Similarity Matrix	40 (18.2%)
2	Creating the folds	10 (4.5%)
3	Network features	40 (18.2%)
4	non-GO-specific features	20 (9.1%)
5	GO-specific features	60 (27.27%)
6	extraction of RN	50 (22.73%)

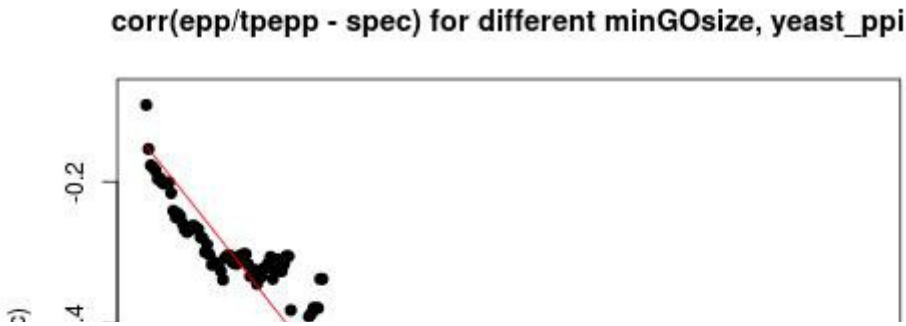
Table 25: Approximate computational time for steps of PU-BMRF for k:10 and 4 replicates

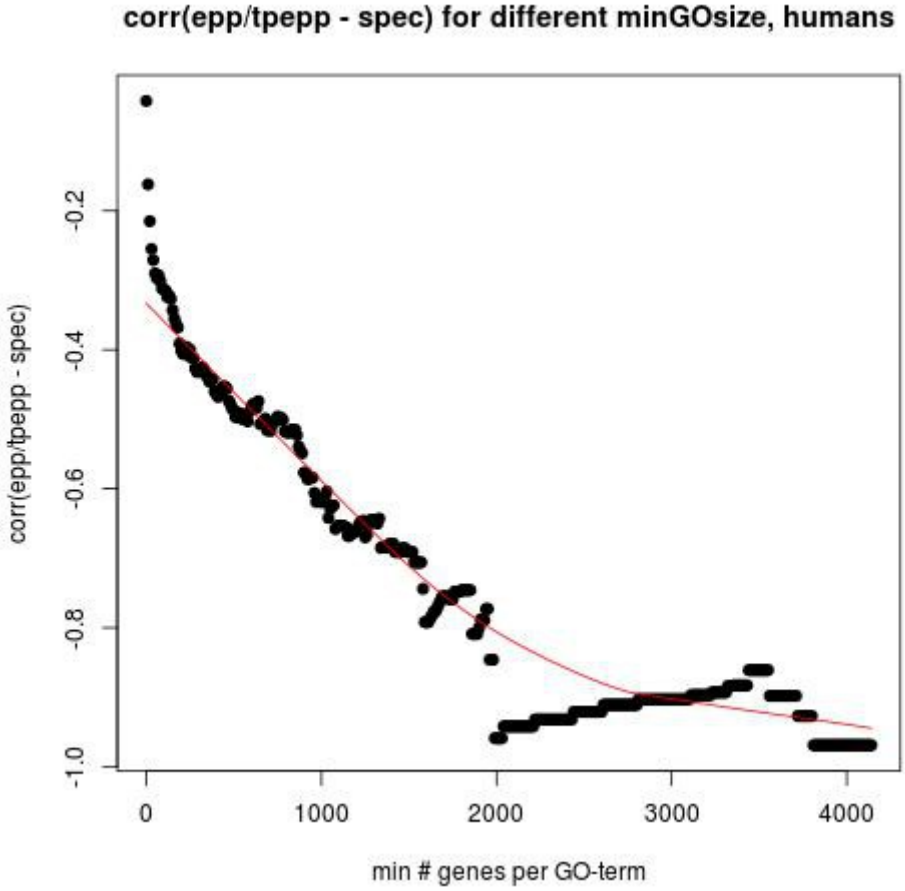
Computtaional time of BMRF for k:10 and 4 replicates is ~5 minutes. The time only increases because more analyses have to be done

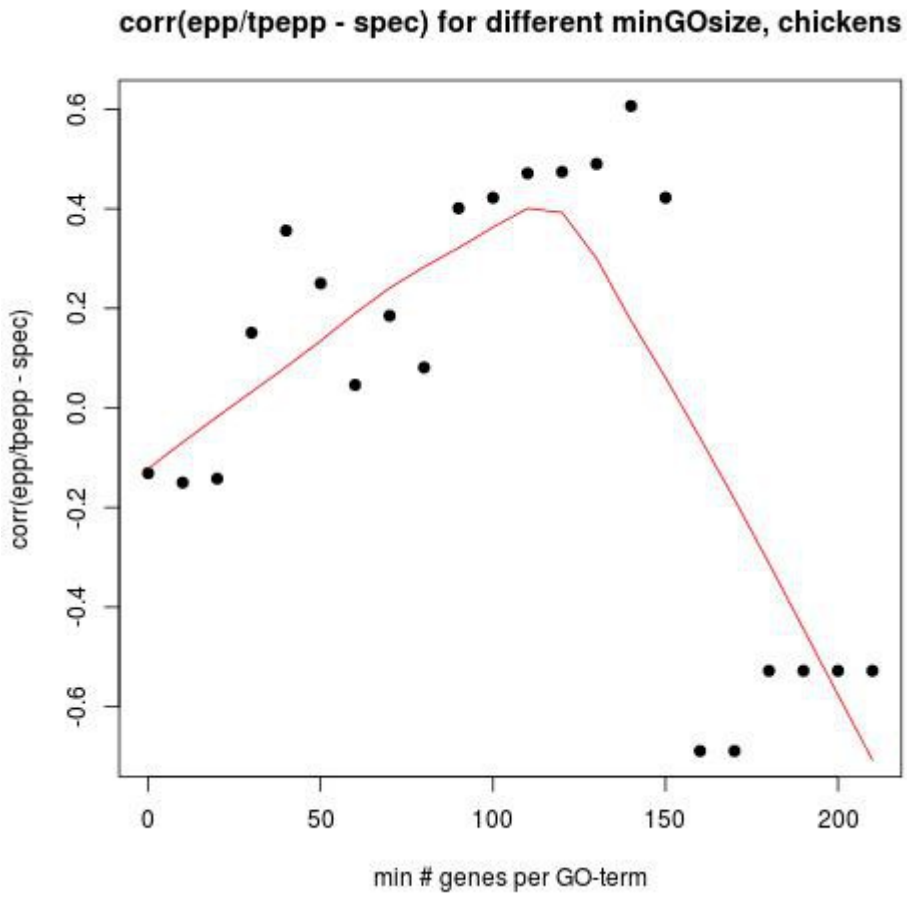
W observed a correlation of 0.67 between the number of labels of the GO-term and th eincrease in accuracy when PU-bMRF was used instead of BMRF, indicating that PU is more effective when the number of known association sis large.



Biology


corr(epp/tpepp - spec) for different minGOsize, yeast_ppi
-0.2


corr(epp/tpepp - spec) for different minGOsize, humans
corr(epp/tpepp - spec)
-0.2
-0.4
-0.6
-0.8
-1.0
0
1000
2000
3000
4000
min # genes per GO-term



In the 3 plots the maximum correlation was when the minGOsize was around 2000.