Appendix I – Additional results

Part 1	a) Model parameters: Choose values for model parameters		**Choose parameter values**	
		Analysis yeast data	k, GO-size filter, use of non-vsalidated data	
		Analysis human data	Number of replicates, k, GO-size filter, use of non-vsalidated data	
		Analysis chicken data	Pearson correlation threshold for co-expression network	
		Other analysis	GO-size filter	
	b) Network characteristics: Choose PFP improvement strategy		**Choose method**	
		Overview data species	epp/tpepp seems to be the most important network characteristic	PU-learning could be a good strategy
		Impact of network characteristics	Decreasing the number of edges negative-negative or positives-negatives improves AUC	
		Correlation between parameters	Increasing the quality of the experiment does not seem to help	
		Nature of the network (experiments)		
Part 2	BMRF with chicken data	**Current state of predictions**		
		Due to poor annotations, few GO terms can be predicted. The use of computational methods can improve this		
		Although the average prediction is high, accuracy of prediction is never larger than 85%. BMRF may improve this		
Part3	PU-BMRF with chicken data			
Part4	Biological support of the approach			

Illustration 1: Diagram of results

Part 1- Impact of the different model parameters and network characteristics in the prediction performance using BMRF. Choose parameter values.

1a. **Impact of the model parameters in the prediction performance with BMRF**

name parameter	Description
	Model parameter used in the original code (https://github.com/jwbargsten/bmrf)
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

* The following network sizes were considered for yeast co-expession data (# of associations): 10,973; 26,879; 26,774; 64,519; 111,390; 242,504; 598,194

Note on data used:
We used yeast and human data to choose the value of the model parameters. Some analysis were carried on yeast data because it was easily accessible, whereas some other analysis were carried on human as we though that chicken data would not become available and it resembles more the situation in chickens. Finally, some analysis were carried on chicken data, after this was made available.

Note on parameter values used:
Unless specified, the analysis were carried with the following values:
-k:10
-20 replicates
-30 iterations for the Gibbs-sampling
-GO-size filter of 20 and 0.1, for minGOsize and maxGOsize, respectively
-Using domain information as well as non-validated associations.

1a(i): Analysis with yeast-coexpression data.

Using yeast co-exopresion data we studied the impact of the GO-size filter, the addition of non-validated data and the number of k-folds on the prediction performance.

scenarios				data				AUC mean(sd) [median]
scenario name	Min GO-size	Max GO-size	only EES	Network size (#conn)	#unkown genes*	#assoc.	#GO-terms	
normal	20	0.1	F	598,174	655	132,249	1,104	0.779 (0.08) [0.778]
only validated associations			**T**		1,307	104,303	1,104	0.762 (0.083) [0.762]
default**		**0.9**			4	264,279	1,187	0.775 (0.08) [0.775]
more GO-terms	**10**	**0.9**			4	273,977	1,738	0.769 (0.1) [0.771]
Only large GO-terms	**30**	**0.07**			688	104,582	832	0.783 (0.075) [0.779]

Table 2: Impact of GO-szie in the data and the prediction performance, with yeast data

In blue, the parameter that were changed with respect to the "normal" scenario

*** default value in original BMRF code: https://github.com/jwbargsten/bmrf*

The number of protein was 5760 in all five scenarios.

From table 2, we conclude that using non-experimental evidence scores helps to achieve higher performance (AUC increased from 0.762 to 0.779). 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 (see scenario "normal").

Then, the effect of the GO-size filter was investigated at the level of individual GO terms and we observed a slight increase in the prediction perfomance as more GO-terms were considered. Table 3 illustrates this with 10 randomly chosen GO terms.

GO-term	#labels/#validated labels of the GO term	AUC filters 20,0.1	AUC filters 5,0.9
GO:0006417	100/144	0.738	0.747
GO:0031670	50/61	0.800	0.802
GO:0006414	40/65	0.752	0.766
GO:0051054	30/30	0.508	0.510
GO:0045931	25/31	0.642	0.641
GO:0007533	30/30	0.758	0.760
GO:0000209	36/23	0.863	0.869

Table 3: Impact of GO-size filter on individual GO-terms, with yeast data

In principle, we would expect that the prediction performance of one GO term would be independent of the other GO terms considered, however, we observed that this is not exactly the case. We observed a slight increase in AUC as the filter became less strict. This is due to the fact that when the filter is less strict, less genes will enter the category of unknown (see Appendix I - definitions).

Other analysis with yeast data showed

- The standard deviation across 5 runs of 20 replicates each was slightly lower for a GO-size filter of (20,0.1) than for (5,0.9): 0.008 vs 0.01, respectively. This is logical since the standard deviation is slighly larger for those GO terms with fewer genes and those GO terms were only considered in the analysis when the GO-size filter was 2,0.9 (less strict).

- There is no significant correlation between the standard deviation and the number of genes of the GO -terms(correlation: -0.05, pvalue:0.0826). Howvere, there is a strong corelation between the standard deviation and AUC (correlation: -0.4, pvalue:2.2e-16)

- Increasing the number of iterations from 10 to 20 did not improve the prediction performance, even though the training samples was slightly larger. AUC was 0.75 vs 0.751, respectively. Further, increasing the number o iterations, is not recommendable since the test-sample may become excessively small and this can cause problems in the computation of AUC.

1a(i): Analysis with human data.

Using human co-exopresion data, we studied the impact of the GO-size filter, the addition of non-validated data and domain information, the number of k-folds on the prediction performance and the number of replicates required for reproducible results.

Approach	AUC
AUC domains and nonValid (normal approach)	0.705
AUC domains but nonValid	0.701
AUC not domains and nonvalid	0.657

Table 4: Impact of domain information and non-validated gene-GO associations in the prediction performance with BMRF, using human data

AUC: area under the curve. Mean AUC of all GO terms that pass the filter.

Table 1 shows a significant increase in the prediction performance when the domain information was added. There was no increases, however, when the non-validated were added to the model.

Filter of GO terms for BMRF	#GO terms	average AUC
MinGOsize:9,maxGOsize=0.1	3328	0.701
MinGOsize:20,maxGOsize=0.1	1982	0.705
MinGOsize:20,maxGOsize=1	2069	0.704

Table 5: Ompact of the filters of "GO-term-size" in the prediction swith BMRF, with human data.

From table 2, we learn that the size of the GO terms does not seem to have an impact on the prediction performance. Results, however, may not apply in species for which the number of GO terms using different filters differ more.

We investigated the effect of the number of k-folds in the cross-validation.

AUC k:2	AUC k:5	AUC k:10	AUC k:20
0.668	0.695	0.705	0.705

Table 6: Impact of the number of folds in BMRF, with human data

From Table 3, we learn that the prediction performance increases with the number of iterations up to a point. This makes sense since the size of the training set increases for higher k and may be, insufficient for lower values of k. Passed a certain value of k the AUC does not increase further, which is in line with what we observed on yeast data.

The effect of the standard deviation across runs of 10 or 20 replicates was used as an indicator of the number of replicates that is required to achieve reproducible results. Here we considered as reproducible, results with less below 0.002 standard deviations in AUC

GO-term (#genes,#validated genes)	sd across runs (i.e. one run is the average of all replicates considered)		Difference
	10 replicates	20 reaplicates	
GO:0006417 (100/144)	0.007	0.002	0.005
GO:0006417 (100/144)	0.007	0.005	0.002
GO:0006414 (40/65)	0.013	0.007	0.006
GO:0051054 (30/30)	0.020	0.008	0.012
GO:0045931 (25/31)	0.024	0.010	0.014
GO:0007533 (30/30)	0.011	0.006	0.005
GO:0000209 (36/23)	0.012	0.014	-0.002

Table 7: Choosing the number of replicates, with humand data

1a(iii): Analysis with chicken-coexpression data.

With a pearson corr =0.7, the number of validated assoc is too low. AUC is still ~0.8 but only for 9 GO terms. Lowering the minGOsize does not help much (for minGOsize=0.8, we predict for 52 GOs). A better option could be to lower down the pearson corr. The number of assoc (181,735) seem to be high enough (9 times less than in humans; 1,213,376: ~1/4 those in yeast)

With 0.7, the number of edges is 1/15 those in humans (¼ those in yeast), which could be enough. The problem is when it comes to validated.

Of the 107,334 association savailable for chicken in “goa_chicken.gaf” I only used 18% due to network contsrains (19,000). Thus, the # of assoc can still increase a lot with lower corr threshold for coexpression.

In addition:
chickens is slighly more specific, so a bit more difficult to get very high AUC
An interesting think to test is which portion of all goes of chicken (without limiting based on network and after uppropagating), is common to all goes of humans

The worst parameter is probably the mena epp per GO: 0.146 vs 954 in humans

1a(iv) Other analysis;

Filter of GO terms for BMRF	#GO terms humans	Chickens_0.35	yeast	yeast_ppi
MinGOsize:9,maxGOsize=0.1	3328	307	1772	1734
MinGOsize:20,maxGOsize=0.1	1982	138	1104	1057
MinGOsize:20,maxGOsize=1	2069	138	1187	1153

Table 8: Number of GO terms with differnet GO-term-size filters, for the differnet species.

From table 7, we learn that the number of GO-terms after passing the filter was still low for chickens when minGOsize was set to 9. Due to time constrains we will carry the anlaysis for the 128 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).

1b) Network characteristics: Choose PFP improvement strategy

In this section, we first overview the differences in data between the three species considered and yeast ppI

1b(i). Differences in data between chickens, yeast and humans.

This subsection covers:

- The data sources,
- th eeffect of difffernet GO-size filters on the data and illustrations
- the distributions of goes per gene, gene p[er goes and number of edges per go, respectivelly, in the 4 cases.
- the number of genes available in data changes with te number of edges.
- the differences and similarities between the GO terms of the differnet species as well as information about the depth of the GO-terms.
- The domain information in the differnet cases

Yeast
Network file: http://www.inetbio.org/yeastnet/downloadnetwork.php
GO file: http://www.yeastgenome.org/download-data/curation
Domains file: http://www.uniprot.org/docs/yeast

yeast_ppi
Network file: /mnt/scratch/dijk097/Fernando/BMRF-R/
GO file: http://www.yeastgenome.org/download-data
Domains file: http://www.uniprot.org/docs/yeast

Humans
Network file: http://mostafavilab.stat.ubc.ca/gnat/
GO file: http://www.geneontology.org/page/download-annotations

Domains file: http://www.uniprot.org/help/homo_sapiens

Chickens
Network file: http://coxpresdb.jp/download.shtml
GO file: http://www.geneontology.org/page/download-annotation
Domains file: http://www.uniprot.org/help/homo_sapiens

Table 4: Data sources for the different species

Network data is from co-expression analysis, unless specified.

yeast_ppi: yeast protein-protein-interaction data

		total data	validated	validated after filter	Portion of data that is validated and passes the filter
#GO	yeast ppi	8,680	4,723	1,073	12.36
	yeast	8,680	4,723	1,104	12.72
	humans	19,549	10,271	1,982	10.14
	Chickens_07	9,247	877	9	0.10
	Chickens_035	16,205	2,350	142	0.88
#labels	yeast ppi	5,757	4,488	4,168	72.40
	yeast	5,757	4,488	4,453	77.35
	humans	8,574	5,582	5,535	64.56
	Chickens_07	2,152	53	53	2.46
	Chickens_035	9,038	300	296	3.28
#assoc	yeast ppi	474,389	227,420	98,192	20.70
	yeast	474,389	227,420	104,303	21.99
	humans	1,213,376	410,215	219,796	18.11
	Chickens_07	181,735	2,253	263	0.14
	Chickens_035	734,840	14,733	7,892	1.07

Table 5: Data available for the different species

ppi: protein-protein-interaction
#assoc: # of associations between GO terms and labels;

Chickens_07 and Chickens_05: Network data for Chicken when the pearson correlation was 0.7 and 0.5, respectively.

From Table 5, we observe:

- The network is considerably smaller for chickens, it should be investigated whether predictions are still accurate for this species.

- Validated data for chickens_0.7 and chickens_0.35 is very poor in comparison to yeast and humans. In the case of chickens_0.35 1% of the #associations is validated and passes the filter (vs 18% in humans), which stresses the difficulty of using BMRF in chickens. **It is thus recommended to investigate the results also when person correlation is lower (chicken_0.35).**

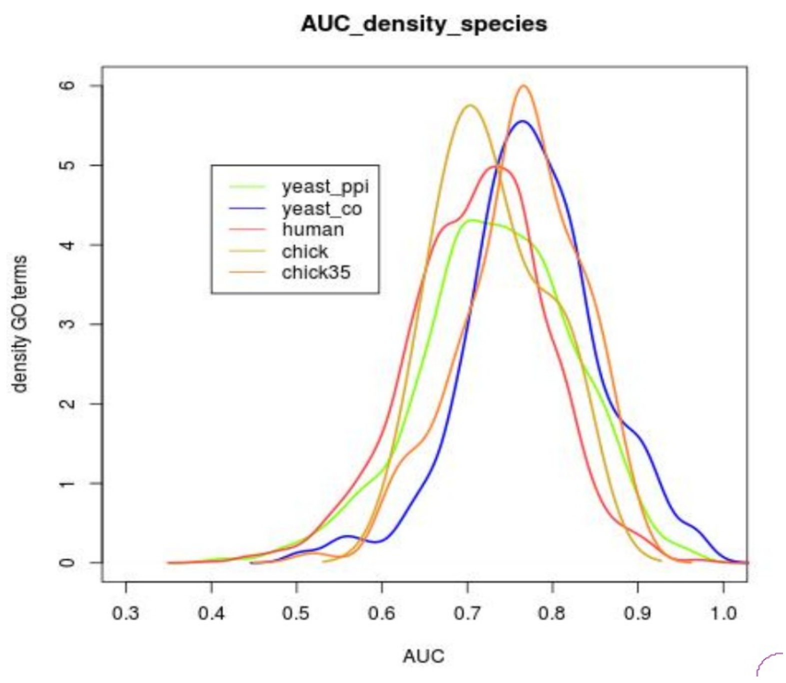
- For yeast, co-expression data is slightly more complete than ppi data.

- For chickens, predictions can only be made for 138 GO terms. **It should be tested whether with the current data, a lower value of minGOsize allows to get more results** (i.e. increasing the number of GO terms for which we make predictions at the cost of lowering the accuracy).
- **It should be investigated whether the BP of the 138 GO terms that can be predicted is known already as well as the depth of these BP GO terms.** This will determine how useful the method is with the current data.

- The proportion of validated data in chickens is very low with respect to the other two species. It is expected that if this proportion increases we will be able to PFP in more GO terms.

- Total data for humans is larger than for yeast but the proportion of validated data that passes the filter is lower (18% in humans vs 22% in yeast-coexpression in the case of #associations). **This offers an opportunity to investigate what is more important to achieve accurate predictions, network size (higher in humans) or proportion of data that is validated (higher in yeast)**.

- Validated data for chickens_0.7 and chickens_0.35 is very poor in comparison to yeast and humans. In the case of chickens_0.35 1% of the #associations is validated and passes the filter (vs 18% in humans), which stresses the difficulty of using BMRF in chickens. **We should, therefore, investigate the results also when person correlation is lower (chicken_0.35).**



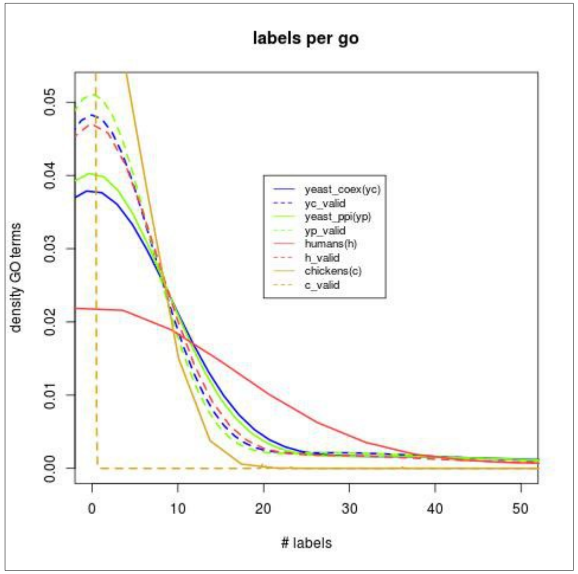


Illustration 3: Number of labels per GO-term in the different species.

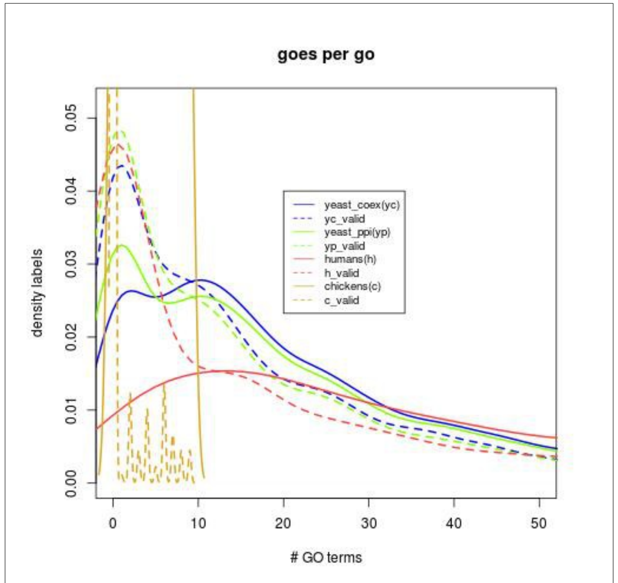


Illustration 2: Number of goes per gene

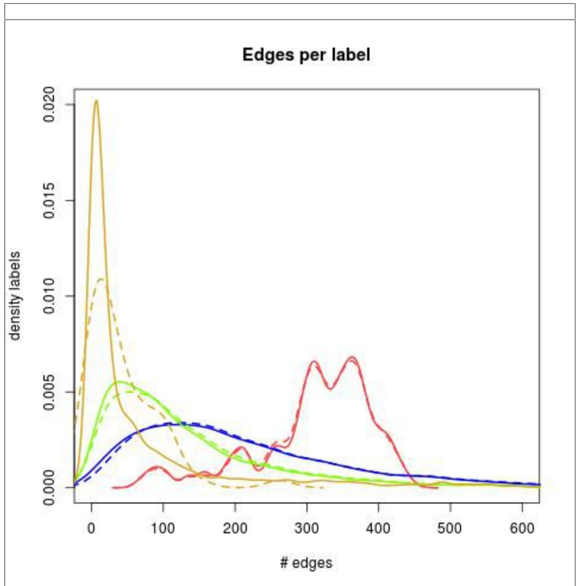


Illustration 5: Number of edges per GO-term

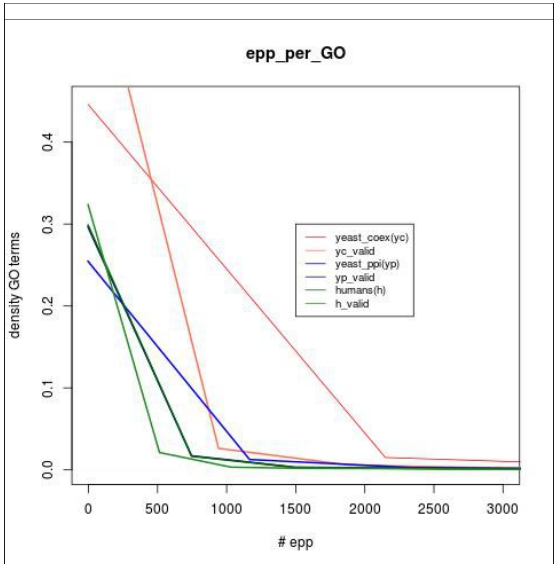


Illustration 4: Number of edges positive-positive er GO-term

From Illustration 1 and 2, we learn that the portion of labels per GO and number of GO terms per gene are much lower in chickens than in the other species and the differences becomes larger as we compare the validated data. Also we observe that for humans, since it is a more complex organism, the annotations are larger than for yeast but that the portion of data that is validated is considerably less for humans than for yeast.

The number of edges per gene is very similar for the validated genes and for the non-validated genes in all four cases: humans, chickens, yeast and yeast-ppi. In humans the number of edges per gene is considerably higher. We expect that this is the case, since humans is a more complex organism than yeast and a large portion of the data is available. In chickens the number of edges per gene is very low due to scarce annotation. As in Iluustrations 1 and 2, the coexpression data for yeast is more completed than the protein-protein interaction data.

We also observed that, for yeast dat, the number of validated proteins decreases almost linearlt with the number of edges

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

Table 9: Relationship between the number of edges and the number of validated genes, for yeast data.

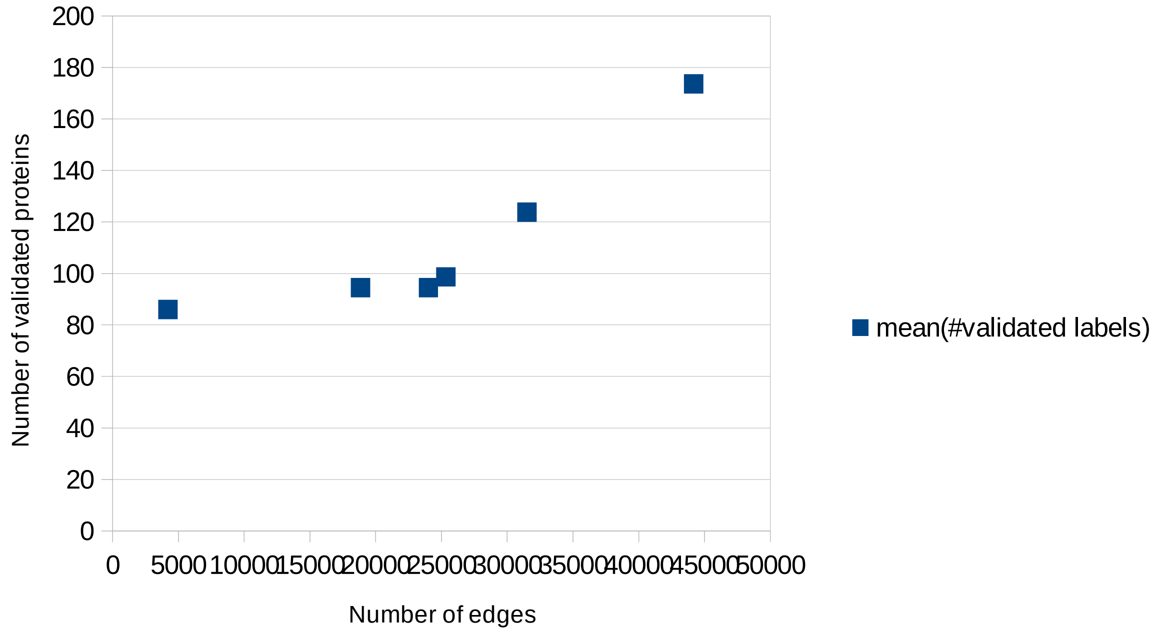


Illustration 6: Relationship between the number of edges in the network and the number of validated proteins in the analysis, for yeast data

	humans	**yeast**	**Common GO terms(6190) (71.31% of the GO terms of yeast)**	**GO terms exclusive in humans (11305) (64.6%)**	**GO terms exclusive in yeast (1948) (23.9%)**
Depth (mean(sd)[mean])	6.77(1.73)[7]	6.59(1.65)[7]	6.48(1.65)[7]	6.92 (1.76) [7]	6.93 (1.61) [7]

Table 10: Comparison GO terms and their depth, human vs yeast

From Tables 1 and 2, we learn that the depth of the GO terms is very similar for humnas, chickens and yeast

	humans	**chickens**	**Common GO terms(8397) (97% of the GO terms in chickens)**	**GO terms exclusive in humans (9098) (52%)****	**GO terms exclusive in chicken (254) (3%)**
Depth (mean(sd)[mean])	6.77(1.73)[7]	6.49(1.75)[6]	6.48(1.75)[6]	7.03 (1.67)[7]	6.89(1.78)[7]

Table 11: Comparison GO terms and their depth, humans vs chickens

	#domains	**#associations gene-domain**	**#genes with domain info.**
yeast	5,436	15,277	5,077
humans	12,193	60,733	17,282
chicken	12,193	60,733	17,282

Table 12: Domain information for the different species

1b(ii). **Impact of network characteristics in the prediction performance using BMRF**

By comparing the characteristics of the network of the different species and the prediction p[performance in each case, we can gain some understanding on the network properties that are more relevant for protein function prediction via BMRF. Prediction performance was as follows:

Tables 6 and 7 summarize the main differences in the characteristics network of the different species.

	#te	#epp (% te)	#epn (% te)	#enn (% 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 13: #edges and AUC

#te: total number of edges

epp: edges positive-positive. epn: edges positive-negative. enn: edges negative-negative

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 linking 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 the epp (edges of positive-positive). These are edges that are linking genes that are known to have a common function. We compare the epp, epn and enn for the differnet species and we study a relationship between these parameters and the prediction performance (AUC-Area under the curve)

In table 2, we observe that the #epp may not be as important as the ratio epp/te, as AUC is higher for yeast (higher ratio epp/te) than for humans (higher epp). This makes sense since, in principle, epn and enn make more difficult the task of PFP. Note that enn may be inafct edges between positives and negatives. Results also suggest that the ratio epp/te may be related to the portion of associations that are validated, as both quantities are higher in yeast co-expression.

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 epps (tpepp). Tpepp is a constant different for each GO term that 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/tpepp*1000	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 14: 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 higher than expected. A possible explanation is that for chickens_07, the number of tpepp is very low and, since the pearson correlation is large, t is more likely that a large portion of the genes associated with the same GO term are interconnected. AUC, nevertheless, is not larger for chickens_07; the relationshupo between epp/tpepp asn AUC is not straignt. Thus we investigated the relationship between Epp/GO and other GO-specific network parameter swith AUC. We considered: the number of labels per GO, the number of GO terms per label, the number of edges per label and the number of epp per label.

The numbetr of labels per GO and Goes per gene is larger for humans. (Appendix I)

Note: The curved corresponding to yeast valid ppi (noit visible) falls just under the curve of humans-non-validated.

From illustration 2, we learn that the postion of epp per GO is much larger for yeast coexpression than in the other 3 cases. If we consider that the number of genes per labe, however was much larger for humasn than for yeast (Innlustraions 1 in Appendix), we observe that the portion of data that is annotaed is much large for yeast than for humans.

and therefore we expect that prediction swill be better in this case. The differneces between the validated and non validtaed data is also larger for yeast coexpression. This may be due to the fact that under the name “non-validated” we include not onlt y the non-validated data frpom the Biological processGO category but also the validated nad non-validated data from Molecular function and Cell componets GO categories

	mean (sd)[median]				AUC
	labels/go	go/labels	edges/label	Epp/GO	
yeast_ppi	11.31 (46.29)	17.05 (22.67)	156.39 (179.3) [104]	960.12 (10844.65) [1]	0.734
yeast	12.01 (48.78)	18.12 (22.77)	213.7 (146.61) [178]	1311.15 (15209.25)[1]	0.775
humans	11.24(59.85)	25.63 (42.51)	310.36 (80.50) [323]	954.16 (11417.71) [0]	0.712
Chickens_07	0.28 (0.97)	0.12(0.85)	43.02(49.12)[24]	0.14642(1.199437)[0]	0.728
Chickens_035	24.55 (26.14)	0.87 (6.23)	811.29 (1097.3) [364]	6778.91 (110825.9) [0]	0.762

Table 15: 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 the first two parameters are more related to AUC. Further, we 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 still way far from this numbers. However, when we use chicken_035 data “#labels per GO”, #edges/label and “Epp/GO” increase to levels higher than in 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.

1b2. **Correlations**

With this purpose of identifying which parameters make predictions more accurate in some GO terms than in others, we have computed the correlation between AUC and different GO term parameters. Parameters considered are:

Epp/tpepp: Number of edges of positive positive of a GO term divided by the total number of possible epp that the GO term could have (if all the genes associated with the GO term were coexpressed).

Sd: standrad deviation of the AUC of the GO term across replicates. Different replicates do not get exact results because of the way the folds are made in the crossvalidation

depth: the depth of the GO term. For humans it ranges from 1 to 16, being 16 the most general GO term

#labels: # genes associated with the GO term

#epp (edges positive-positive), #epn (edges positive-negative), #enn (edges negative-negative)

Var2	Var1	correlation
epp/tpepp	AUC	0.431
sd	AUC	-0.287
depth	AUC	0.087
#enn	AUC	0.031
#epn	AUC	-0.028
#enn+#epn	AUC	-0.027
#labels	AUC	-0.024
#epp	AUC	-0.019

Table 16: Correlation between AUC and different GO term parameters

AUC: Area under the curve ffor a given GO term when BMRF attempted to predict whether the GO term is associated with each of the genes in data.

epp/tpepp: Portion of edges positive-positive

sd: standard deviation

epp: #edges positive-positive

epn: #edges positive negative

enn: #edges negative negative-negative

Details about how these correlations where computed are given in Appendix I.
Only epp/tpepp and sd seem to affect the prediction performance. Epp/tpepp shows a favorable correlation with AUC, meaning that for GO terms whose associated genes are interconnected in the network (coexpressed), the method has more chances to distinguish genes that are associated ssociated with the GO term from genes that are not. This makes sense, since identifying positive cases (associated genes)is a difficult task considering that among a very large set of genes (~8500 in humans), only a few may be associated with the GO term. The more interconnected the associated genes are, the easier will be to identify them.

Sd shows a negative correlation with AUC, meaning that for those GO terms whose AUC fluctuates more from replicate to replicate are overall worse predicted. A possible explanation for this is that the sd is high when epp/tpepp is low. Thus, indirectly, high sd means low overall AUC. This is because if only a few of the associated genes are interconnected with each other, the results will depend on whether the associated that are interconnected enter the trainning or the test set in the crossvalidation.

Table 9 confirms that some parameters that we may have considered important, such as #epp, depth or #lables are in fact not related to AUC.

We conclude that BMRF works better for those GO terms whose associated genes are coexpressed. In order to achieve high PFP accuracy, epp/tpepp should be high. Epp depends on data available and cannot be increased with methods, however tpepp could be reduced by PU. PU improves two ratios:
1) reduces the portion of epn within enn
2)(here) reduces the epn, as some unknows are droped from the anlaysis. [5th conlcussiojn in table 11).

We have also computed the correlation between these parameters. Correlations with magnitude larger than 0.2 are given in the following table:

Var1	Var2	correlation
epp+epn	#labels	0.996
#labels	epp	0.918
epn	epp	0.897
epn	sd	-0.509
#labels	sd	-0.491
#labels	epn	-0.436
#labels	depth	-0.344
epp	sd	-0.342
epn	depth	-0.332
depth	epp	-0.284
epn	epp.tpepp	-0.250
epp.tpepp	#labels	-0.238

Table 17: Correlation between GOP term parameters

epp/tpepp: Portion of edges positive-positive

sd: standard deviation

epp: #edges positive-positive

epn: #edges positive negative

enn: #edges negative negative-negative

From table 10 we learn:

- When the #labels is high the number of epp+epn will also be higher. Also the sd of AUC across replicates seems to reduce as epn and epp increase. Thus, as the number of labels per GO increase, the number of epp and epn will also increase and AUC results will be more consistent (less sd).

- The depth of the GO term decreases as epn, epp and # labels increase, which, in principle, is counterintuitive, as more general GO terms have higher depth.

- If we decrease epn, epp/tpepp will increase, which makes sense, since tpepp is the sum of two componets: epp and epn.

3b) Quality of data and PFP

We are interested in knowing how the predictions vary when the data becomes more incomplete. This information can be used to get an idea on in which species BMRF would be a successful method for PFP. We investigate how performance is altered in the different situations:

- If several epp, epn or enn are missing in the data
- If several associations GO-gene are missing in the data
- If the data has false associations GO-gene (noisy data).

	Correlation
AUC_reduceEpn	0.98
AUC_reduceEnn	0.95
AUC_reduceAmg	0.67
AUC_addNoise	0.60
AUC_reduceOa	-0.47
AUC_reduceEpp	-0.26

Table 18: Correlation between AUC and data quality

From table 11 we learn:

- AUC will increase linearly as we remove from the data Epn. This links together two points from previous analysis: In table 3 we learn that epp/tpepp is a good indicator of AUC; also in table 10 we saw that epn shows a negative correlation with epp/tpepp. Thus, as epn decrease, epp/tpepp increase and therefore AUC increases as well.
- AUC will also increase almost linearly as we remove enn. This is because some of the enn are actually epn (false negatives).
- AUC will increase if we remove associations. 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.
- A counter-intuitive results is that if we add fake associations between genes and a target GO term, AUC increases for that GO term.
- AUC however will be reduced if we remove from the network genes that do not have the function. This makes sense, as the epp/tpepp will increase.

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

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

We can also investigate 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.

However, we observed that,to some extend, the correlation between th enumber of egdges-AUC increases as the size of the network decreases. We did not observet any pattern on ho wthe size of the netwok alters the correlaion between AU Cnad the #labls, exept for the fact that the value of these correlation sis larger when the netwok is very small. the scenario “similar size network_1” the correlation betwee n#edges and AUC reached 0.376. We observed the differneces in AUC in 5 bins with differenet # of edges:

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

Table 20: Differncenes on

scenario name	#assoc.	CorrAUC_ #conn	CorrAUC_ #val_labels	CorrAUC_ #labels
very small network	8,862	0.219	-0.223	-0.125
similar_size network_1	32,336	0.356	0.072	0.124
similar_size network_2	58,358	0.376	0.223	0.272
only validated associations	104,303	0.133	0.05	0.05
Only large GO-terms	104,582	0.174	0.015	0.071
“stress” co-expression	110,682	0.255	0.113	0.161
“oxidiation” co-expression	111,480	0.252	0.113	0.162
normal*	132,249	0.124	0.014	0.053
default**	264,279	0.025	-0.008	0.002
more GO-terms	273,977	0.042	NA	NA

Table 21: Impact of the network size on the correlations AUC-#edges and AUC-#l(validated)able, using yeast data.

Also, in this scenario the correlation between AUC and the #genes is high (0.272). The differences in AUC between 5 bins of GO terms with different number of proteins was also significant.

Bin of GO terms	AUC
1th/5	0.649
2th/5	0.656
3th/5	0.689
4th/5	0.696
5th/5	0.720

Table 22: Table 11: Differncenes on AUC between groups of GO terms with different # of edges in scenario “similar size network_1”. The first fith refers to the 1/5th of the GO terms with a lowest numbver of edges, and so on

1b(iii). **Impact of quality of data in the prediction performance.**

- **Using yeast data**

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

Illustration 7: Impact of number of edges in th eprediction performance, using yeast data.

Removing random edges from the data dd not seem to affect much the prediction performance. We observed a large impact after removing 10% of the edges and after removing more than 50% of the edges.
When we looked at inividual GO terms, we did not observe differences in the effect between differenyt GO terms.

GO-term	total_labels	valid_labels	0% (all data used)	portion of edges substracted 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 predicion performance for individuyaol GO terms

- **Using human data**

1b(iv). Nature of the networks

By nature of the network here we refer to the characteristrics of the co-expression analuysis. It is imprtant to invetsigate whether for instance a coexpression analysis address to one specific tissue allows to make more accurate predictiosn for those GO terms whose function is more relevant in that tissue. Note that from a biological perspective we woul expect that this is the case, specially considering that netwokr analysis exploit the principle of guil-by-association.

- **Using yeast data**

Using yeast data, we investigated whether the nature of the co-expressionnetwork (cahracteristics of the experimnet) have any impact on th eprediction performance.

scenarios	data					
scenario name	Network size (#conn)	#unkown genes*	#proteins.	#assoc.	AUC mean(sd) [median]	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_1	28,800	298	1,865	32,336	426	**0.684** (0.101) [0.677]
similar_size network_2	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 23: Impact of the nature of the network on th eprediction performance using yeast data.

- **Using human data**

One important aspect to consider is the nature of the network 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 faiirer analysis this we normalized the netwoks of th differnet tissues based on epp/tpepp,

tissue	top1_GOterm	top2_GOterm	top3_GOterm
Stomach	post-Golgi vesicle-mediated transport	positive regulation of lipid transport	positive regulation of epithelial to mesenchymal transition
Esophagus-Muscularis	anoikis	intrinsic apoptotic signaling pathway in response to oxidative stress	ceramide metabolic process
Thyroid	erythrocyte differentiation	cell aging	regulation of histone acetylation
Whole_Blood	negative regulation of epithelial cell migration	keratinocyte proliferation	RNA-dependent DNA biosynthetic process
Brain-Amygdala	histone H4 acetylation	protein destabilization	regulation of membrane depolarization
Adrenal_Gland	regulation of protein oligomerization	negative regulation of response to biotic stimulus	sensory perception of sound
Brain-Putamen(basal_ganglia)	regulation of protein complex disassembly	negative regulation of protein binding	positive regulation of cell morphogenesis involved in differentiation
Brain-Cortex	receptor internalization	regulation of heart rate	mitotic DNA integrity checkpoint
Skin-Not_Sun_Exposed(Suprapubic)	regulation of toll-like receptor signaling pathway	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	regulation of cytokinesis
Testis	positive regulation of viral genome replication	negative regulation of telomere maintenance	negative regulation of cell projection organization
Brain-Anterior_cingulate_cortex(BA24)	positive regulation of viral genome replication	peroxisome organization	regulation of protein oligomerization
Pancreas	regulation of receptor internalization	TOR signaling	response to monosaccharide
Brain-Spinal_cord(cervical_c-1)	regulation of receptor internalization	regulation of microtubule polymerization	positive regulation of myeloid cell differentiation
Brain-Hypothalamus	negative regulation of DNA binding	positive regulation of telomere maintenance	regulation of membrane depolarization
Brain-Caudate(basal_ganglia)	negative regulation of dephosphorylation	cellular extravasation	histone H4 acetylation
Artery-Tibial	regulation of cell adhesion mediated by integrin	negative regulation of telomere maintenance	regulation of telomere maintenance via telomerase
Pituitary	negative regulation of blood vessel endothelial cell migration	protein localization to cytoskeleton	regulation of histone acetylation
Esophagus-Mucosa	negative regulation of cell projection organization	response to temperature stimulus	lipid storage
Lung	intrinsic apoptotic signaling pathway in response to oxidative stress	histone deacetylation	cellular response to amino acid starvation
Skin-Sun_Exposed(Lower_leg)	regulation of interferon-beta production	myeloid cell homeostasis	positive regulation of calcium ion transport into cytosol
Nerve-Tibial	negative regulation of cell-substrate adhesion	anoikis	regulation of striated muscle contraction
Muscle-Skeletal	homotypic cell-cell adhesion	regulation of cell adhesion mediated by integrin	membrane protein ectodomain proteolysis
Breast-Mammary_Tissue	receptor internalization	regulation of protein complex disassembly	intrinsic apoptotic signaling pathway in response to oxidative stress
Brain-Nucleus_accumbens(basal_ganglia)	negative regulation of epithelial cell migration	positive regulation of DNA binding	positive regulation of cell morphogenesis involved in differentiation
Adipose-Subcutaneous	regulation of protein oligomerization	negative regulation of blood vessel endothelial cell migration	endosome to lysosome transport
Heart-Atrial_Appendage	positive regulation of macroautophagy	negative regulation of blood vessel endothelial cell migration	zymogen activation
Adipose-Visceral(Omentum)	regulation of cell adhesion mediated by integrin	regulation of smooth muscle cell migration	positive regulation of lipid transport
Artery-Aorta	positive regulation of actin filament bundle assembly	cellular response to amino acid starvation	platelet activation
Brain-Substantia_nigra	homotypic cell-cell adhesion	regulation of epithelial to mesenchymal transition	positive regulation of myeloid cell differentiation
Heart-Left_Ventricle	regulation of DNA recombination	regulation of sodium ion transport	intracellular protein transmembrane import
Brain-Hippocampus	interleukin-10 production	histone ubiquitination	positive regulation of actin filament bundle assembly
Brain-Cerebellar_Hemisphere	lipid storage	smooth muscle cell migration	erythrocyte differentiation
Colon-Transverse	regulation of cell adhesion mediated by integrin	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	regulation of protein complex disassembly
Brain-Cerebellum	peroxisome organization	ATP-dependent chromatin remodeling	sensory perception of sound
Brain-Frontal_Cortex(BA9)	regulation of phosphatase activity	cell aging	negative regulation of autophagy

Table 24: Goes terms for which a highest AUC was achieved using network data from different tissues

From table 13 we observe that there is strong biological support in the network data. For instance, for Stomach data the GO term “post-Golgi vesicle-mediated transport” was the most accurately predicted and “positive regulation of lipid transport“ is the second. Also, for pituitary data, “protein localization to cytoskeleton“ is accurately predicted, which makes sense as the pituitary is related to bone development. In another example, for brain-cortex, “regulation of heart rate“ is 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 have identified the tissues for which predictions were best, and worst.

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 25: The 10 GO terms for which a higehst AUC was found between tissues

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.

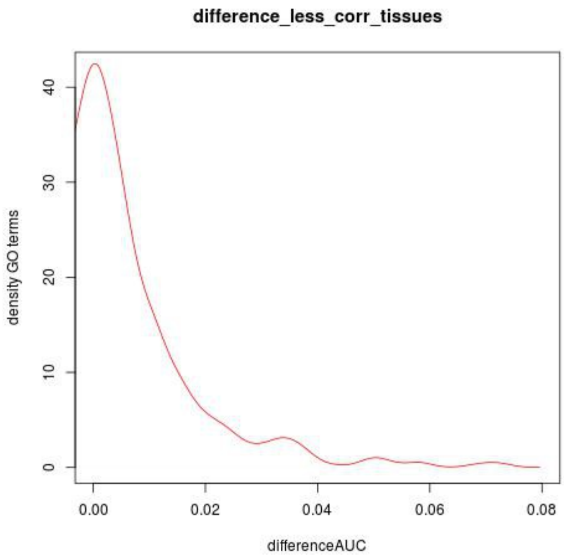


Illustration 9: Difference in AUC for the different GO terms

In Ilustration 2, the Y axes represent the density of GO terms, and the X-axes 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 much 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 nature of the GO term than on the quality of the data.**

Also in line with this, The overall AUC was very similar using the different subsets of network (sd across tissues=0/0005). A low value of sd across tissues is, in addition, not surprising considering that we report the mean of ~1800 GO terms, which is rather stable.

However, in order to have a more direct insight on what is the difference in PFP when using one tissues' network or another, for each pair of tissues, we have 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.

We, therefore, conclude 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**.

Part 2- Prediction with BMRF

	yeast_ppi	yeast_co	humans	Chicken_07	Chicken_035
number of GOS with AUC>0.6	1013 (92%)	1155 (97.3%)	1819 (92%)	8 (89%)	140 (99.29)
mean depth	6.2	6.0	6.1	2.4	4.0
sd depth	1.5	1.6	1.5	0.5	1.3
>0.7	723 (66%)	1016 (85.6%)	1200 (60.06%)	5 (56%)	113(80.14%)
mean depth	6.5	6.0	6.2	2.4	4.1
sd depth	1.5	1.6	1.6	0.5	1.3
>0.8	281 (25,6%)	428 (36%)	384 (19.8%)	2	44(31.2)
mean depth	7.0	6.5	6.4	2.5	4.3
sd depth	1.4	1.4	1.7	0.7	1.4
>0.9	54 (5%)	83 (7%)	158 (8%)	0	0
mean depth	8.2	7.3	7.4	NaN	NaN
sd depth	1.0	1.4	2.1	NA	NA
>0.95	37 (3.4%)	21 (1.77%)	143 (7.2%)	0	0
mean depth	8.0	8.1	6.3	NaN	NaN
sd depth	0.9	1.7	1.5	NA	NA

Table 26: Portion of GO terms above different AUC thresholds and their depth

Part 3- Prediction sin chickens with BMRF

GO term	#genes	AUC PU.BMRF	sd AUC PU.BMRF	AUC BMRF	sd AUC BMRF	difference
GO:0006928	24	0.614	0.043	0.578	0.048	0.571
GO:0006950	39	0.584	0.025	0.536	0.023	0.559
GO:0007423	24	0.657	0.054	0.644	0.041	0.603
GO:0008283	31	0.677	0.038	0.571	0.038	0.639
GO:0009605	37	0.716	0.037	0.612	0.03	0.679
GO:0009653	82	0.774	0.02	0.694	0.023	0.754
GO:0009719	33	0.691	0.061	0.577	0.052	0.63
GO:0009887	48	0.738	0.029	0.667	0.034	0.709
GO:0009987	224	0.728	0.013	0.704	0.015	0.715
GO:0032879	21	0.522	0.025	0.499	0.018	0.497
GO:0033554	21	0.539	0.046	0.534	0.037	0.493
GO:0040011	22	0.611	0.062	0.582	0.058	0.549
GO:0044699	219	0.785	0.011	0.75	0.016	0.774
GO:0044700	39	0.725	0.038	0.599	0.038	0.687
GO:0044707	133	0.788	0.016	0.719	0.016	0.772
GO:0044710	24	0.593	0.048	0.641	0.051	0.545
GO:0044763	173	0.772	0.012	0.711	0.017	0.76
GO:0044767	133	0.802	0.019	0.736	0.019	0.783
GO:0048646	37	0.62	0.03	0.589	0.03	0.59
GO:0051128	23	0.562	0.036	0.528	0.037	0.526
GO:0051240	26	0.63	0.034	0.587	0.03	0.596
GO:0060485	21	0.598	0.049	0.567	0.047	0.549
GO:0065008	31	0.574	0.039	0.518	0.029	0.535
GO:0097659	38	0.746	0.026	0.736	0.022	0.72
GO:1901360	56	0.717	0.027	0.666	0.022	0.69
GO:1901362	43	0.755	0.037	0.727	0.039	0.718
GO:1901576	56	0.746	0.027	0.672	0.022	0.719
GO:1901700	28	0.624	0.052	0.587	0.037	0.572
GO:1903506	36	0.755	0.029	0.731	0.029	0.726

Table 27: Results for 20 GO terms

Additinal results:

- The average difference between the set of positives and the set of unkoqns for the features considered was 0.105. with a standard dication of 0.141.
- We did not observe any correlation between the prediction performance of RN and the differnces between the positiveand the unkown set.
- We did not observe also a correlation between the differnces in feature sand the number of genens that are known to beassoicated with the diffrenet GO trerms
- The average AUC for a value of tolerance eual to 1 was 0.974 wiuth a standrad deviatio of 0.057