Part 1. Evaluate the performance of BMRF for PFP

The results and comments hereby aim to help in answering four questions:
(1) Can we annotate accurately function prediction in chickens with BMRF?,
(2) How chicken data should look like in order to achieve better results?
(3) Is PU useful for (1).
(4) Which type of data is required to achieve accurate PFP via BMRF. Thus for which species BMRF could be used for PFP

BMRF was used for PFP in yeast, humans and chickens to predict associations between genes and gene ontology (GO) terms. The PFP performance was evaluated in order to get an overview of which type of data is suits the method best. Area under the curve (AUC) was used to evaluate the PFP performance. The results also aim to investigate whether PU is may enable to make the best possible use of the biological data provided.

Network data was based on gene co-expression except for yeast, that we used both co-expression and protein-protein-interaction data. Predictions was on the Biological process (BP) category of GO ontology because co-expression networks are particularly promising for this category. This is because whereas Molecular Function (MF) and Cell Component (CC) can be predicted with sequence-similarity-based-methods, these methods do no enable to find associations between genes and BP.

Part 1 consists of 3 parts:
A) General overview of the factors that affect PFP with BMRF.
B) Investigate how BMRF depends on the input data, considering the expression data as a whole.
C) Investigate how BMRF depends on the input data, considering the expression data as a set of small subsetes each of which corresponding to an experiment (i.e. different tissues, different networks).

PART 1A- Overview factors BMRF

The BMRF method sets the parameter “minGOsize” to exclude from the analysis data from GO terms with less than “minGOsize” genes. Default value is 20, minimum value is 8. Values below 8, lead to problems when computing sparse matrices. Higher values of minGOsize will, in principle, be associate with higher mean AUC, but less GO terms will be predicted. Here we use the default value of minGOsize (20). BMRF also sets the parameter maxGOsize that is analogous to “minGOsize”, but that excludes for the analysis the GO terms tar are excessively general. We can also adjust this parameter depending on the results. Now we use maxGOsize = 0.1, meaning that GO terms that are with more than 10% of the genes are excluded from the analysis. MinGOsize and maxGOsize define thus a BMRF filter for the GO terms. We will indicate which data is available for the different species after applying such filter.

We performed different analysis in yeast co-expression data. The conclusion were as follows:

- Including GO-gene associations in the training set that are not used in the validation, such as non Experimental evidence scores (EES) or other categories than BP, increase the overall PFP prediction by ~0.02. (from 0.76 to 0.78). **We will therefore use non-EES and non-BP data to train the model.**

- When the maxGOsize was set to 0.9 (default) instead of 0.1 (chosen here), the overall AUC slightly decreased. This could be an artifact of the Gibbs sampling, that has much less unlabeled genes if maxGOszie is 0.9 than when maxGOszie was 0.1 (4 unlabeled genes vs 655 when minGOsize changed to 0.1). By unlabeled genes here we mean genes that are not associated with any GO term but that are present in the network. Note that the GO terms that are included if minGOsize is 0.9 are very general GOs and therefore PFP is less interesting for these. **We will therefore use minGOsize <0.1** (even though we will predict less number of GO terms). Other aspect to be considered is the overall depth of the GO terms for the species of interest. If the depth is large (Go terms are overall general), we may be interested in setting a lower value for minGOsize.

- When minGOszie was set to 10 instead of 20, overall AUC decreased because some very specific GO terms with less than 20 genes associated will also be considered. However, the decrease was very low ~0.01. On the contrary, when minGOsize was increased to 30, overall AUC increased by 0.004. We conclude, that depending on whether the overall AUC is high, and deepening ion the number of GO terms that pass the filter, **we may be interested in setting minGOsize to 10 instead of 20 in order to do PFP is a larger number of GO terms, or we may me interested in increasing maxGOszie in order to achieve accurate AUC in (at least) a small selection of GO terms.**

- The number of iteration in the gibbs sampling does not affect the AUC as long as it is above 20 iterations. We will keep a margin and use default value (30), as computational time barely increase form 20 to 30. The # iterations is independent of how many unlabeled genes there are. **We will therefore use default number of iterations for gibb sampling (30)**.

- The number of replicates that are required in order to achieve stable AUC results for the GO terms depends on the nature of the GO term. We conclude that in order to achieve a sd across replicates <0.05, 10 replicates are required if the GO term has AUC>0.7 and 20 replicates if AUC<0.7. **For simplicity, we will run the analysis with 20 replicates**.

- **When the filter was less strict, AUC was slightly higher for individual GO terms. We conclude that we should take this into account when it comes to decide the filter (second point).**

GO term	#labels/#validated labels	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

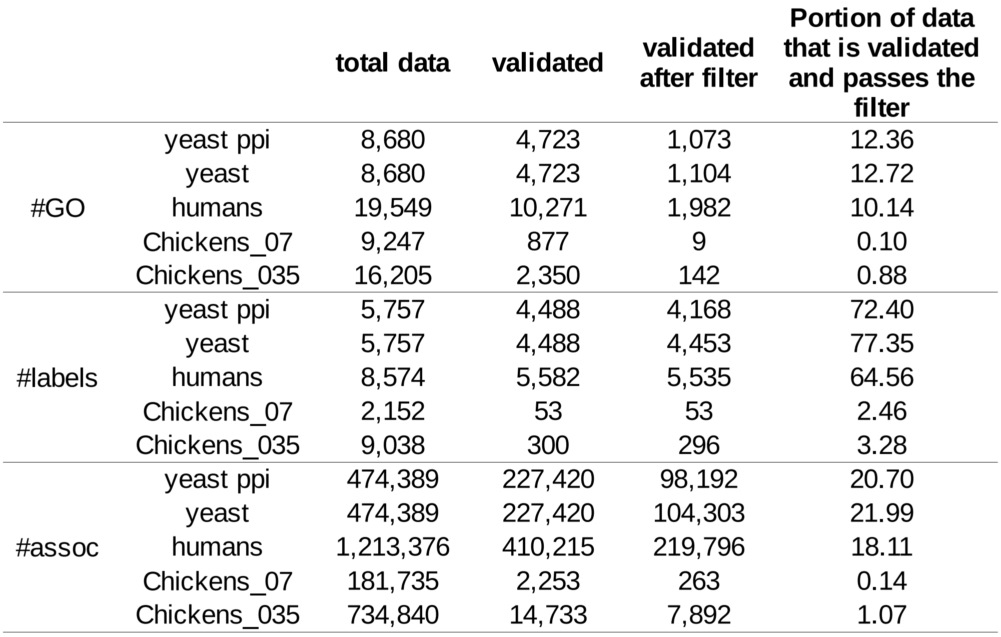
Table 1: Relationship between AUC and BMRF filter

PART 1B

In Part 1B we provide some insights on to which extend BMRF will allow PFP in chickens and how the method could benefit from PU. First, we compare the data available for yeast, humans and chickens, then we compare the networks and finally we investigate which network parameters are more important for PFP via BMRF.

1) Data available

Table 1 summarizes the data available for the different species.
In the case of chicken two different threshold of co-expression were used: person correlation 0.7 (standard threshold in co-expression analyses) and person correlation 0.35 (in order to get some insights on the results when the data is more and more unreliable). Note that the data has been trimmed, so genes and GO terms that were available but are not in the networks were discarded as BMRF cannot handle missing values in network file.



		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 1: Data available for the different species

ppi: protein-protein-interaction

#assoc: # of associations between GO terms and labels

From Table 1 we observe:

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

- 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, 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).**

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

- It seems that currently, for chickens we can make predictions only for 142 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 142 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.

We will continue the analysis with other networks parameters that refer in all cases to data after validation that passed the filter.

2) Compare networks.

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 fucntion. A more important network parameter may be the epp (edges of positive-positive). These are edges that are linking genes that are known to have a common function.

	#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 2: #edges and AUC

#te: total number of edges

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

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

From table 2, we learn that AUC for chickens_07 and chickens_035 is already quite high with the existing data. However, we would like to know whether we can achieve higher accuracy or whether we can do predictions for a larger number of GO terms. It is, therefore, important to know what type of data leads to highest AUC using BMRF. Thus, enabling to re-adapt the chicken data (i.e. with PU) and achieve better results. For this purpose it seems more proimisng to use the chickens_035 data rather than chickens_07.

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. 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 conclude that we should investigate the correlation between epp, epp/te, epn/te and enn/te

with AUC, as well as to which extend a larger portion of validated associations is related to a large te.

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 3: 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 the pearson correlation is large and therefore it is more likely that a large portion of the genes associated with the same GO Term are interconnected; and that the co-expressed genes are involved in the same function. AUC, nevertheless, is not larger for chickens_07. Thus, epp/tpepp is only a rough indicator of the accuracy of PFP.

We conclude that we investigate how epp/tpepp relates to AUC.

At the level of individual GO terms, some parameters are 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.

	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 4: 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 parametersare 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.

Other factors that differ between species but do not seem to affect AUC are the depth of the GO terms and the number of domains.

	yeast_ppi	yeast_co	humans	Chicken_07	Chicken_035
mean depth	6.14	5.99	6.13	2.38	3.98
median depth	6	6	6	2	4
sd depth	1.54	1.60	1.53	0.52	1.25

Table 5: Overview depth of the GO terms for the different species

	#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 6: Domain data for the different species

We now show in more detail the overall AUC for the different species. The sd (standard deviation) between replicates for a given GO term was roughly 0.015 (thus reproducibility is high):

	yeast_ppi	yeast_co	humans	Chicken_07	Chicken_035	Chicken_035_mgs_8
# GO terms	1,073	1,104	1,982	9	142	347
mean AUC	0.734	0.775	0.712	0.728	0.765	0.754
median AUC	0.736	0.775	0.717	0.718	0.771	0.765
sd AUC	0.090	0.080	0.083	0.062	0.07	0.093

Table 7: Overall AUC in the different species

Most accurate PFP was achieved with the yeast_coexpression data, where the genes associated with 1104 GO terms were identified with an AUC of 0.775. The standard deviation of AUC across GO terms was 0.08. The second best results where with the chicken_035 data. However, here only 142 GO terms were predicted. When we change the minGOsize (mgs) from 20 (default value) to 8 (lowest possible value), results were still good (~0.01 less AUC), but the number of GO terms reminded low (347).

To get a better insight on how the AUC differs between the different species, we can chechk for which portion of the GO term AUC was above a certain value.

	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 8: Portion of GO terms above different AUC thresholds

From table 8, we observe that predictions are overall better for yeast_co and chickens_035 (for 85.6% of the GO terms BMRF identified the positive and negative genes with an AUC >70%, and 80% in chicken_035 vs 66% in yeast_ppi and 60%% in humans). Similarly, for 36% of the GO terms predicted with yeast co-expression data AUC was above 80%, vs 25.6% in yeast_ppi and 19.8% in humans. This is in line with what we observed from tables 1-3, for yeast coexpression epp/et and epp/tpepp are higher and therefore predictions are better.

Interestingly, the portion of GO terms for which we achieved AUC>95% was higher for humans. (7.2% of the GO terms) than for yeast (1.77%). A possible explanation for this is that the standard deviation around the mean of label/GO is higher for this species, thus we can expect that for some GO terms the number of labels is considerably higher than for others, and predictions are very accurate for these GO terms with a high number of labels.

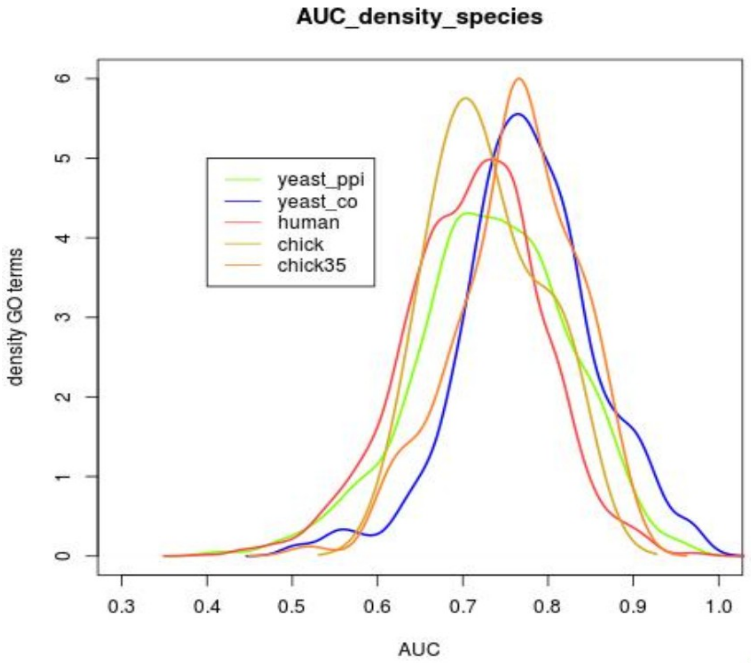


Illustration 2: AUC distribution for the differnet species

Illustration 2 confirms the founding of Table 7 and Table 8, AUC is highest for yeast co-expression data, then for chickens_035 (although few GO terms are predicted), then for yeast_ppi, and then for humans. Nevertheless, it is with humans data that we achieve a good proportion of GO ter,m with very high accuracy (>90%). In chickens_035, however,, for non of the GO terms we achieve 90%.

3) Identify important parameters.

With the purpose of knowing what data is required for PFP via BMRF and for which data the method performs best, we have performed two types of analysis:

3a) Investigate which GO parameters are more related with high AUC
3b) Investigate how the quality of the data affect the predictions

3a) GO parameters and PFP

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 9: 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 associated from genes non-associated with the GO term. This makes sense, since identifying the genes associated is a difficult task considering that among a very large connection of genes very few 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, then result will depend on whether the associated that are interconnected fall within the same training and test sets n the crossvalidation. Therefore, in this cases, the sd is high.

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 genes associated are coexpressed. In order to achieve high PFP accuracy, epp should be increased as much as possible and tpepp should be reduced. Epp depends on data available and cannot be increased with methods, but tpepp could be reduced by PU.

Results in table 9 suggest that overall AUC could be increased if the ratio epp/tpepp was increased, for instance by reducing tpepp with PU.

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 10: 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, if for a given GO term we somehow manage to increase the number of labels, 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,a s more epn means less epp.tpepp.

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 11: 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 12 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.
- 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.

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

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 12: Correlation betwen 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 much more frequent, Thus AUC may increases by increasing the specificity at the cost of a lower sensibility.

PART 1C – Data subsets – Tissues

One important aspect to consider is the nature of the network data. Thus, among networks with similar epp/tpepp, we investigate whether it makes any difference to use one network or another. In other words, we want to know whether in addition to the network size, the biological sport behind the edges should be considered. If there is strong biological support, some network parameters such as betweeness, closeness… will improve and it will be easier to achieve high AUC. To study the biological support of the network, co-expression data from different tissues can be used. We can, for instance, expect that for a given GO term we may be able to achieve higher AUC if we use a co-expression network from a tissue for which the genes associated with the GO term play an important role. Thus, there will be high closeness between these genes.

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.

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 13: 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 14: 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.

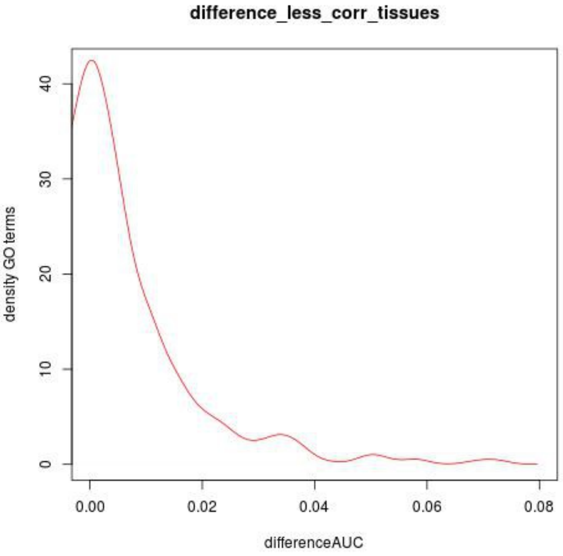


Illustration 3: Differece in AUC for the different GO terms

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 nature of the GO term than on the quality of the data**.

In yeast data, we performed a similar analysis, using subsets of networks from different experiments instead of from different tissues, and we reached to similar conditions (Appendix II).

We used different subsets of network data and we observed that there is a direct relationship between the network size and the AUC. Using all data available, (~600.000 edges) AUC was 0.78; with a subset of ~98000 edges, AUC was 0.73; and with a subset of ~27000 edges, AUC was 0.68.

Network	Network size (#edges)	Mean AUC (SE)
all data available	598,174	0.779 (0.0024)
experimenst yeast is stressed	98,479	0.727 (0.0028)
experimenst yeast is oxidated	64,167	0.72 (0.0027)
Experiments with ~28000 #edges	28,800	0.684 (0.0049)
Experiments with ~28000 #edges	27,488	0.682 (0.0034)

Table 15: Relationship between AUC and network size

When we investigated the correlation between AUC and the number of edges, we found:

Network size (#conn)	corrAUC_#edges
598,174	0.124
98,479	0.255
64,167	0.252
28,800	0.356
27,488	0.376

Table 16: correlation between AUC and #edges

We also observed that there is a strong correlation between the #edges of a GO term and AUC and that this correlation is higher in small. A possible explanation for this is that when the network is small there will be less genes per GO term and these genes can only be found via neighbors (so edges).

Thus, in equation 1, the α parameter is lower in small networks, and therefore it is less likely that we identify true positives unless there are positive cases in the neighbours (β).

We also investigate how the AUC of a given GO terms changes when some random edges are removed. We observed that decrease of AUC with network size is linear and that when we removed

95% of the edges, AUC was still high (AUC was only ~0.012 lower). This is because with 5% of the edges from the complete network, we still have ~30000 edges. Thus, the decrease of AUC with size is to similar decrease than the decreased we observed observed in Table1 when we used different substes of expression data. This indicates that although network size matters, the nature of the network does not seem to matter much. In other words, **it does not seem to matter much whether we use a network derived form a co-expression analysis or some random edges from a collection of experiments, as long as the size of the network is same.**

Further analysis could be:

- Check the counterintuitiveness in GO depth
- Consider also parameters betweeness, closeness, netwrok coeffciient… and relate to literature
- Report sd of correlation coeffcients
- pvalues and significance
- Make final conclussions
- In vestigate whetehr a target GO term has higher AUC in chickens than in humans…
- Include data regarding which portion of the chicken GO terms are also in human database
- Report how much domain information help

-Add the AUC info of chicken035

- Check that all porints covered in “key_points_part1.odt” are here also
- Check results in “/previous_to_week11”
- Is anything from the presentation or Evernote to be added
- Say how k affects
- The four analogous plots

References

[1] Wang Y, et al. Targeted DNA recombination in vivo using an adenovirus carrying the cre recombinase gene. Proceedings of the National Academy of Sciences of the United States of America. 1996;93:3932–3936.