Appendix II – Overview of data used

This appendix aims to show the differences in data available for the species considered. The appendix consists of two sections:

Section A) Current state of annotation in the species considered
Section B) Differences in network properties in the species considered

- **Section A) Current state of annotation in the species considered**

This section covers:

- The data sources
- An overview of the number of GO-terms, genes and number of edges in the different species and how many of these are validated and passes the BMRF filter
- A detailed plot of how the number of genes associated with the different GO term differs for the species.
- A comparison the current state of annotation
- The distributions of goes per gene and gene per go
- An overview about which portion of the GO terms are common in the different species or exclusive, as well as information about the depth of the GO-terms.
- The domain information that is available in the different cases

Data used in this study is freely available.

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 1: 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
#genes	yeast ppi	5,437	4,488	4,168	76.66
	yeast	5,760	4,488	4,453	77.31
	humans	9,998	5,582	5,535	55.36
	Chickens_07	2,152	53	53	2.46
	Chickens_035	12,424	300	296	2.38
#edges	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. Therefore, 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) (Appendix III-Additional results; part 1; section A).

- 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). (Appendix III-Additional results; part 2).

- 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-co-expression 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). (Appendix III-Additional results; part 2).

- 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). (Appendix III-Additional results; part 1).

- The number of assoc. is still large for chickens (734,840), although it is around half than for humans (1,213,376)

- For Pearson correlation equal to 0.07, the number of association was around 1/5 smaller than for Pearson correlation 0.05.

To gain a better insight on to which extend the annotations of chickens are poorer than for other species, we plotted #genes per GO term and #GO-terms per gene Illustrations 1-3.

We then compared the number of validated genes for a group of GO terms that were present in the four cases. Illustrations 5 to 7 show the differences in number of genes in the different species for 30 randomly chosen GO terms.

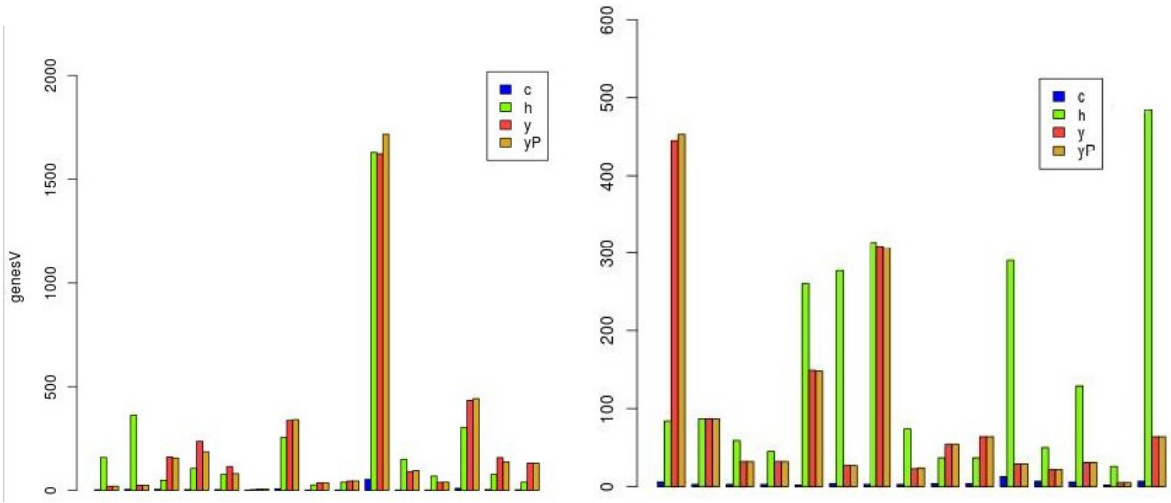


Illustration 1: Number of validated genes

We observe that for some of these GO terms the number of genes is larger for humans and in others for yeast or yeast_PPI. This may depend on whether the GO term is important for medical research. In all cases, the number is much lower for chickens, as expected. Also, that annotation is more completed in yeast, but since humans is a more complex species, the GO terms have a tendency towards being associated with more genes. The number of genes in human could be considered as an upper bound of what is achievable in chickens.

Another way to compare the level of annotation in the different cases is the ratio #assoc/total possible associations, considering that all GO-terms could potentially be associated with all genes Table 1

	#assoc*1000/total possible assoc	
	valid	**including non-valid**
yeast	4.82	10.05
yeast_PPI	4.46	9.83
humans	1.12	2.21
chickens	0.07	3.6

Table 1: Current state of annotations. Considering that all GO-terns could potentially be associated with all genes

From table 2 we observe that this ratio is considerably larger for yeast and yeast_ppi, both when we look only at validated associations and when we look at all the associations. This was expected because a although the annotation ofr humans is larger than for yeast. Yeast is proportionally better annotated than yeast.

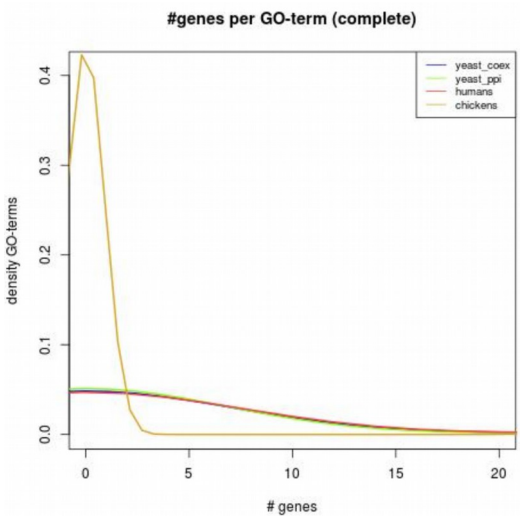


Illustration 3: Number of genes per GO-term in the different species. Current state of annotations.

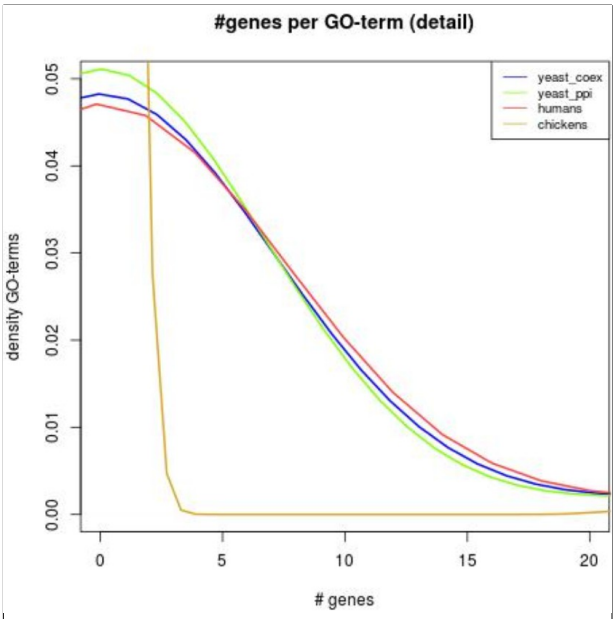


Illustration 2: Number of genes per GO-term in the different species. Detail form illustration 4. Current state of annotations.

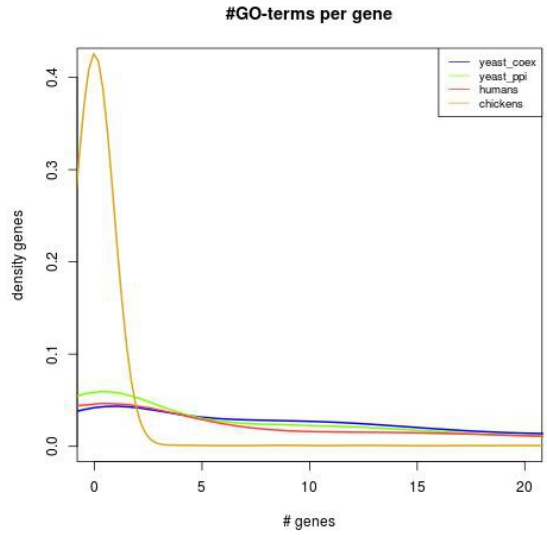


Illustration 4: Number of edges per GO-term. Current state of annotations.

In illustrations 1-3 we observe that annotations much lower for chickens, and the number of GO terms with more than 5 validated genes is almost 0. Whereas for the other two cases there is a very large number of GO terms with #genes between 5 and 2. We also observe that for humans the number of genes per GO is lower than or yeast. And so it is the number of GO per gene. Furthermore, yeast has more genes per Go and Go terms per gene than yeast_ppi.

An important aspect to consider is how general the GO terms are, since it is more desirable to do PFP in the most specific GO terms because for these it is more difficult to predict the function using more conventional approaches, like experimental approaches.

	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 2: Comparison GO terms and their depth, human vs 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 3: Comparison GO terms and their depth, humans vs chickens

From Tables 1 and 2, we learn that the depth of the GO terms is very similar for humans, chickens and yeast, independently on whether the GO terms are exclusive for these species.

Then, we looked in the domain annotation, which is more complete in ythe case of humans than for yeast. Which is not surprising since the identification of these modules is important for medical research.

	#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 4: Domain information for the different species

- **Section B) Differences in network properties in the species considered**

In this section, we study the differences between:

#edges,

Number of edges for the labelled genes,

Total edges from positive genes divided by all possible edges that could potentially link labeled genes (including epn and enn)

epp/tpepp. (EppV/tpeppV is shown figure XX in results).

These are shown in Illustrations 4-8, respectively.

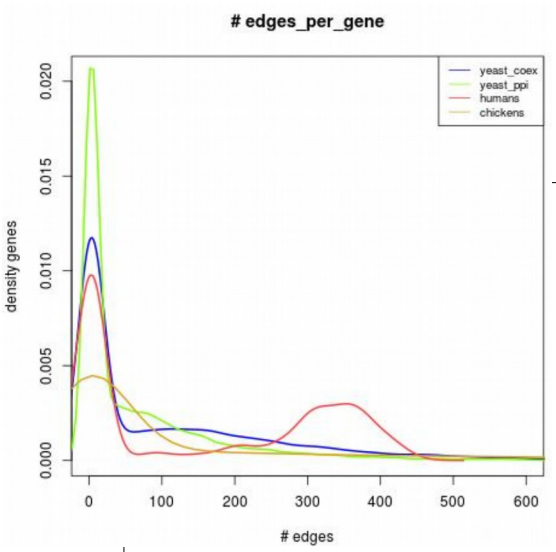
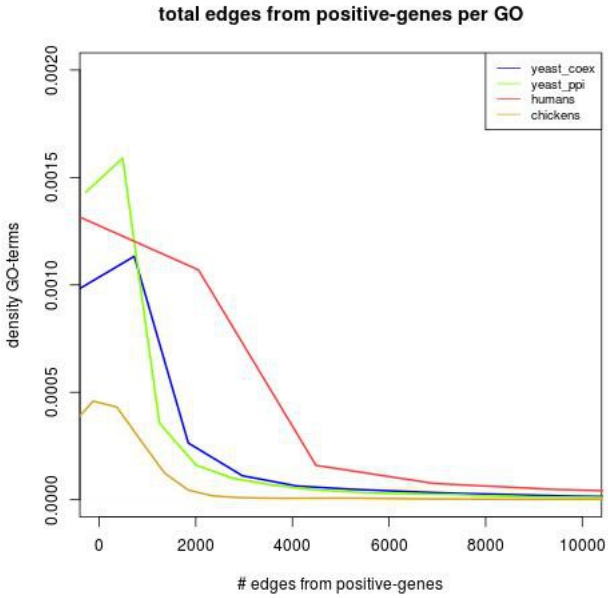


Illustration 7: Number of edges per gene



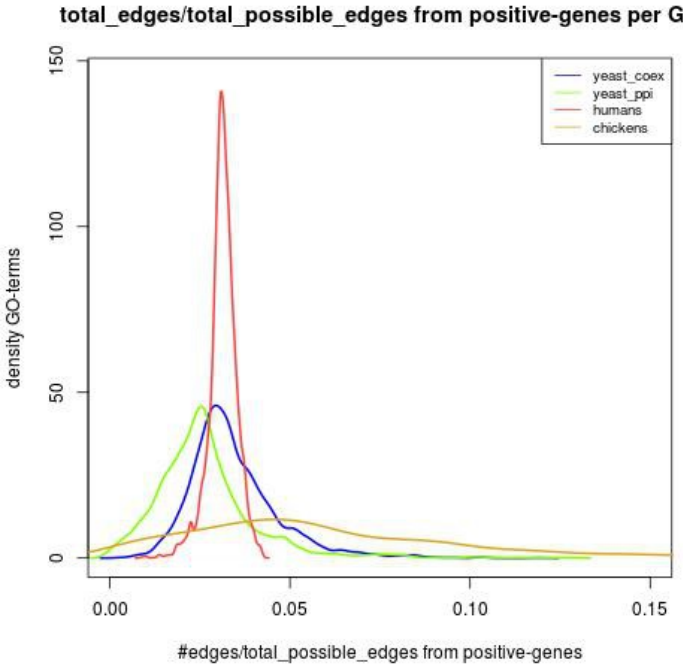


Illustration 6: te/tpe in positiv egenes .

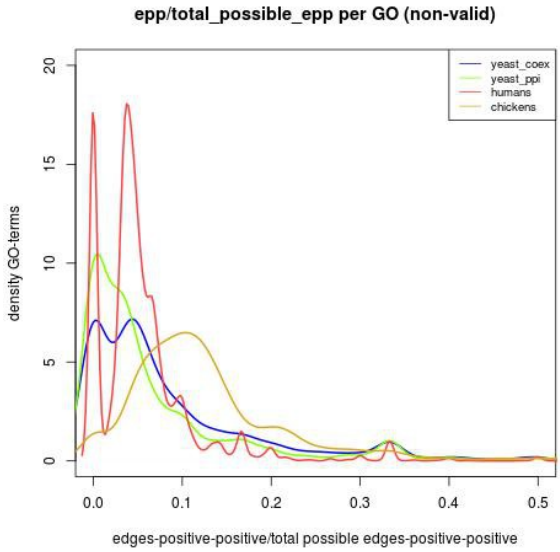


Illustration 5:

In illustration A, we observe that the portion of genes with a large number of edges is large in humans than in yeast, which is not surprising since the number of genes considered is much large than for yeast and yeast_ppi. It is surprising, however, that it is higher in humans than in chickens, since in chickens we are using pc0.35 and in humans 0.7. However, we do observe a larger number of genes with between 50 and 100 edges in chickens.

In illustration B we observe that ….i larger for humans and lower for chickens, which was expected, since the number of positive association is larger for humans than for yeast and for yeast than for chickens.

In illustration C, however, we observe that te/tpe from positive genes is larger in chickens, given the lower co-expression threshold. Also in humans the number of go terms with a very large ratio larger than 0.04 is very low. This also makes sense since the network is larger for this species, and therefore we expect a larger proportion of epn and enn edges.

In illustration D we observe tat opposite to what occurs for eppV/tpeppV, (Table XX results) the ratio is better for chickens when the non-validated data is considered, due to the low Pearson correlation.

We then compare the ratio epp/tpepp at the level of individual GO terms. (Illustration X)for 30 randomly chose GO terms that are predicted in the four cases

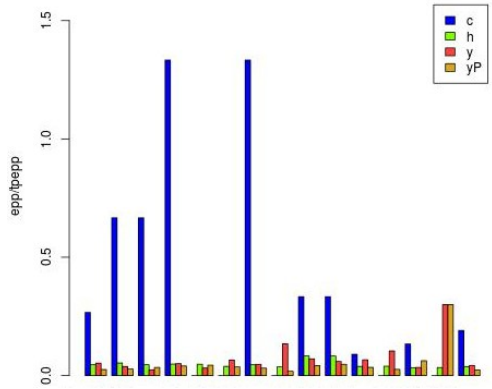


Illustration 9: GO_0000003,GO_0000075,GO_0000077,GO_0000086,GO_0000122,GO_0000165,GO_0000278,GO_0000302,GO_0000724,GO_0000725,GO_0000902,GO_0001101,GO_0001558,GO_0001676,GO_0001932

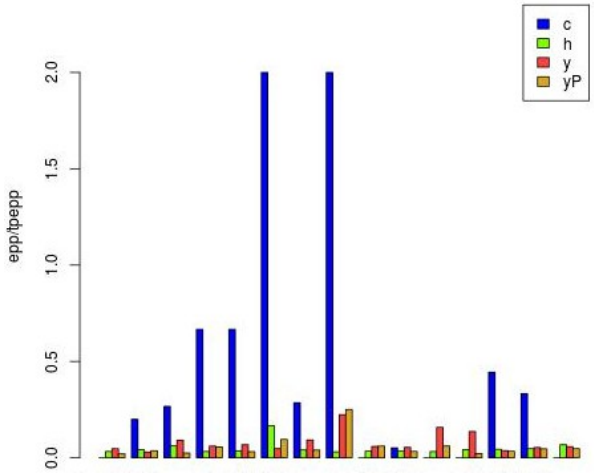


Illustration 8: GO_0001933,GO_0001934,GO_0003006,GO_0005975,GO_0006066,GO_0006071,GO_0006082,GO_0006090,GO_0006109,GO_0006139,GO_0006163,GO_0006164,GO_0006259, GO_0006260,GO_0006261

The ratio is considerably larger for chickens, although this could be an artifact of the scarce data that is available, as explained in Table XX of results. After that, the GO terms seem to have a higher ratio for yeast than for the other species, which is in line with the reasoning that epp/tpepp is an indicator of the prediction accuracy.