UNIPred: Unbalance-aware Network Integration and Prediction of protein functions: Supplementary Data

Marco Frasca,[*] Alberto Bertoni[*] and Giorgio Valentini[*] [†]

July 9, 2015

1 Mouse networks

In the MouseFunc setting, 21603 mouse proteins and 2815 GO terms with a number of annotations ranging from 3 to 300 have been considered, excluding GO annotations based solely on the "inferred from electronic annotation" (IEA) evidence code. A randomly selected set of 1718 proteins is held-out and their annotations have to be predicted using the annotations of the remaining proteins. We collected 17 types of protein profiles from MouseFunc, including protein family profiles, expression data, protein-protein interactions, phenotypes, phylogenetic profiles. These data are briefly described below, while the correspondences indices-networks are shown in Table S1. **The indices correspond to those of Figure 5, 8, 9 and Table 4 of the main paper.** The names in parenthesis correspond to the "Network" column of Table S1.

> *Expression data - 4 networks.* Expression data from oligonucleotide arrays for 13,566 genes across 55 mouse tissues (Zhang *et al.*, 2004) (*Zhang*); expression data from Affymetrix arrays for 18,208 genes across 61 mouse tissues (Su *et al.*, 2004) (*Su*); tag counts at quality 0.99 cut-off from 139 SAGE libraries for 16,726 genes (Siddiqui *et al.*, 2005) (*Sageavg, Sagesum*).

[*] DI - Department of Computer Science, University of Milan, Italy
[†] Corresponding author

Sequence patterns - 2 networks. Protein sequence pattern annotations from Pfam-A (release 19) for 15,569 genes with 3,133 protein families (Finn *et al.*, 2006) (*Pfam*); protein sequence pattern annotations from InterPro (release 12.1) for 16,965 genes with 5,404 sequence patterns (Mulder *et al.*, 2005) (*Interpro*).

Protein interactions - 2 networks. Protein-protein interactions from OPHID for 7,125 genes, downloaded on 20 April 2006 (Brown and Jurisica, 2005), in adjacency and distance format (*PPIbin, PPIdist*).

Phenotypes - 1 network. Phenotype annotations from MGI for 3,439 genes with 33 phenotypes, downloaded on 21 February 2006[a] (Eppig *et al.*, 2007) (*Pheno*).

Conservation profile - 4 networks. Conservation pattern from Ensembl (v38) for 15,939 genes across 18 species (Kasprzyk *et al.*, 2004) in binary and score format (*Phylobin, Phyloscore*); conservation pattern from Inparanoid (v4.0) for 15,703 genes across 21 species (O'Brien *et al.*, 2005), binary and score format (*Inpbin, Inpscore*).

Disease associations - 1 network. Disease associations from OMIM for 1,938 genes to 2,488 diseases/phenotypes, downloaded on 6 June 2006[b] (Wheeler *et al.*, 2007; Hamosh *et al.*, 2005) (*Omim*).

GO Annotation networks - 3 networks. In addition to genomics and proteomics association networks, we constructed three networks by using the annotations to a given domain of GO and computing the Pearson's correlation of binary vectors associated to each couple of genes. We obtain a network for each GO domain: $GO.XX$, where XX is one of BP (Biological Process), MF (Molecular Function) and CC (Cellular Component). When predicting GO term c in domain XX, we exclude the network $GO.XX$ from the set of networks to be integrated.

[a] Phenotype Annotations from MGI [ftp.informatics.jax.org/pub/reports]

[b] Disease Associations from OMIM [ftp.ncbi.nih.gov/repository/OMIM/]

[Table 1 about here.]

Then we constructed 17 functional association networks from the collected profiles, using different pre-processing procedures for respectively binary and real-valued profiles and PPI data (see next section for more details). To integrate the 17 networks we considered the union of their nodes, thus resulting in an integrated network with 21603 nodes. Finally, among the 2815 GO classes, we selected those with at least one annotation in the test set, obtaining 1847 GO terms.

2 Preprocessing of mouse networks

We constructed 17 functional association networks from the collected profiles, using different pre-processing procedures for respectively binary and real-valued profiles and PPI data. For binary data, if β is the proportion of ones (proteins for which a given feature is present), then all ones were replaced with $-\log(\beta)$ and zeros with $\log(1-\beta)$. In this way the "weight" of very uncommon features is emphasized (Mostafavi *et al.*, 2008). Finally the score for each gene pair has been set to the Pearson's correlation coefficient of the corresponding feature vectors. For continuous data we directly adopted the pairwise Pearson's correlation coefficient, and for gene expression data the squared correlation, in order to take in account both negative and positive correlation.

Finally for PPI interaction data we constructed pairwise interaction scores using the approach proposed in (Chua *et al.*, 2006), where the similarity score for genes i and j is

$$S_{ij} = \frac{2|N_i \cap N_j|}{|N_i \setminus N_j| + 2|N_i \cap N_j| + 1} \times \frac{2|N_i \cap N_j|}{|N_j \setminus N_i| + 2|N_i \cap N_j| + 1}$$

where N_k is the set of the neighbors of gene k (k is included).

To maintain sparse the resulting association networks, we set to 0 the negative correlations, and the edge threshold to a value such that each node has at least one neighbour. Finally, each network $\boldsymbol{W}$ has been normalized as follows:

$$\hat{\boldsymbol{W}} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{1}$$

where $\boldsymbol{D}$ is a diagonal matrix and $d_{ii} = \sum_j w_{ij}$ its diagonal elements.

3 MouseFunc methods compared with *UNIPred*

Table S2 contains the description of the best eight methods of the MouseFunc challenge. These methods are rankers, that is they provide for each gene i only a real score s_i. Accordingly, to compute the F-score we first scale these scores in the interval [0,1] by using the following equation:

$$s_i^* = \frac{s_i - min(s)}{max(s) - min(s)}$$

where $min(\boldsymbol{s}) = \min_i s_i$ and $max(\boldsymbol{s}) = \max_i s_i$. In this way the lowest score in $\boldsymbol{s}$ corresponds to 0 and the highest score in $\boldsymbol{s}$ corresponds to 1. Then, we set a threshold for scores at 0.5, i.e. genes corresponding to scores greater than 0.5 are predicted as positive and the remaining genes are predicted in the negative class. We outline that this technique for computing binary labels might be suboptimal for some of the compared algorithms; method-specific techniques to set the thresholds may lead to better results.

[Table 2 about here.]

4 *COSNet* and *UNIPred* correlation by GO ontology

Figure S1 reports the correlations averaged by GO ontology between F-scores achieved by the supervised linear classifier constructed at step 1.2 of *UNIPred* (see Section 2.3.2 of the main paper) on each single-source network and the corresponding F-scores computed by *COSNet*. The correlation is much higher for MF terms on networks 11, 14 and 15, whereas, on networks 11, 12 and 13, CC terms show a lower correlation than those of other two ontologies. For the remaining networks, CC terms in general achieve the highest correlations. Moreover, these correlations show also a higher variance w.r.t. the correlations relative to BP and MF ontologies.

[Figure 1 about here.]

5 Mouse updated GO annotations

Regarding GO annotation release (15 August 2012), we excluded all the annotations with IEA evidence, obtaining annotations for 18996 genes and 2712 GO terms with 3-300 annotations. Among these genes, 1255 are labeled genes belonging to the MouseFunc test set. The 2607 missing genes with respect to the GO 2006 release include pseudogenes (151), genes with solely IEA annotations (887), not classified genes (781), merged genes (419) and others (DNA segment, gene segment, etc., 520). Among the 2712 GO terms, 1782 have at least one gene annotated in the test set, with 1147 BP, 418 MF and 217 CC terms. We also re-computed the GO networks GO.BP, GO.MF and GO.CC using the new release of GO annotations.

6 Yeast and fly networks

The considered yeast and fly networks are briefly summarized in Tables S3 and S4 respectively. We integrated respectively 19 and 13 networks. To avoid biases, we excluded e.g. GO.XX association networks when we predicted GO XX terms, where XX is one of BP, MF, CC.

[Table 3 about here.]

[Table 4 about here.]

7 Results on integrated mouse networks

In Table S5 we show the overall average results of the MouseFunc I participants and *UNIPred* integration strategies WA, WAC WAP.

[Table 5 about here.]

[Table 6 about here.]

In addition to F-score results, we also report in Figure S2 the results in terms of AUC and P20R averaged by GO category.

[Figure 2 about here.]

[Figure 3 about here.]

Table S6 shows the statistically significant differences in average performance according to the Wilcoxon signed-ranks test (Wilcoxon, 1945) at $\alpha = 0.01$ significance level. The improvements of our method in terms of F-score are statistically significant w.r.t. all the methods and all the domains up to method G (Funckenstein) in domain MF. Moreover, even in terms of P20R the differences are significant in favour of *UNIPred*, except for method G and C (GeneMANIA), where there is no statistically significant difference. For the AUC results, *UNIPred* performs significantly worse than other methods only in BP domain when compared with method C, and hierarchical methods D and G.

Importantly, all the weighted strategies of *UNIPred* perform better than the unweighted sum average (UA) and among the weighted strategies, WA seems performing better than the others. These results are to some extent expected, since similar results have been achieved in the literature (Mostafavi and Morris, 2010). Nevertheless, in Figure S3 the results of *UNIPred* integration strategies and UA integration averaged by GO category show that per class strategy (WAP) tends to perform worse on more unbalanced categories (BP, MF, CC 3-10 categories) and better in the other categories (see e.g. BP categories for P20R results). For instance, the WAP strategy considerably improves performances on CC 11-30 category w.r.t. all the other strategies. These results are likely due to the excessive unbalance of classes with a very low number of annotations, which may affect the effectiveness of the cost-sensitive strategy of *UNIPred* for such extreme cases, and may introduce over-fitting problems.

8 Results on integrated yeast and fly networks

Fig. S4 shows the results in terms of AUC averaged by GO categories for yeast and fly organisms, and the fly P20R and F-score results. Confirming results obtained with mouse, *GeneMANIA* outperforms *UNIPred* in terms of AUC, and *UNIPred* performs significantly better than *GeneMANIA* in terms of both F-score and P20R. *MS-kNN* achieves competitive performance in terms of AUC, especially on yeast data, where it is the best method in BP 31/101, MF 101 and CC 11 categories.

[Figure 4 about here.]

9 Proof of Theorem 1

If $H = < \boldsymbol{W}, k, \rho >$ is the parametric Hopfield network constructed by *COSNet* on network $G = \langle V, \boldsymbol{W} \rangle$ (Frasca *et al.*, 2013), $f_{\hat{\alpha},\hat{\gamma}}$ is the optimum line computed by *UNIPred* with respect to the labeling function L_c, and F_c (see Section 2.3.2 in the main paper) the corresponding F-score value, then the following theorem holds:

Theorem 1 *If $\rho = \hat{\alpha}$ and $k = \hat{\gamma}$, then $F_c = 1$ iff $L_c(S)$ is an equilibrium state of H restricted to neurons in S.*

Proof. The prof follows from Fact 3 in Frasca *et al.* (2013) by setting $S^+ = S_+$, $S^- = S_-$, $U^P = U_+$, $U^N = U_-$, $I^+ = I_+$, $I^- = I_-$, $I^+_{\alpha,\gamma} = I_{+,\rho,k}$, $I^-_{\alpha,\gamma} = I_{-,\rho,k}$, $x(U^P, U^N) = \underline{0}$, $x(S^+, S^-) = L_c(S)$ and $Fscore(\alpha, \gamma) = F_c$. Remind that, in the learning phase, when $x(U^P, U^N)$ is set to $\underline{0}$ (*UNIPred* case), it means that to learn the optimal parameters just nodes in S are considered, i.e. both U^P and U^N are the empty set.

References

Aguilar, P. S., Frohlich, F., Rehman, M., *et al.* (2010). A plasma-membrane e-map reveals links of the eisosome with sphingolipid metabolism and endosomal trafficking. *Nat Struct Mol Biol*, **17**(7), 901–8.

Alamgir, M., Erukova, V., Jessulat, M., *et al.* (2010). Chemical-genetic profile analysis of five inhibitory compounds in yeast. *BMC Chemical Biology*, **10**(1), 1–15.

Apweiler, R., Attwood, T. K., Bairoch, A., *et al.* (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, **29**(1), 37–40.

Baradaran-Heravi, A., Cho, K. S., Tolhuis, B., *et al.* (2012). Penetrance of biallelic SMARCAL1 mutations is associated with environmental and genetic disturbances of gene expression. *Human Molecular Genetics*, **21**(11), 2572–2587.

Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*, **22**(7), 830–836.

Breitkreutz, A., Choi2, H., Sharom, J. R., *et al.* (2010). A Global Protein Kinase and Phosphatase Interaction Network in Yeast. *Science*, **328**(5981), 1043–1046.

Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, **21**(9), 2076–2082.

Busser, B. W., Shokri, L., Jeager, S. A., *et al.* (2012). Molecular mechanism underlying the regulatory specificity of a Drosophila homeodomain protein that specifies myoblast identity. *Development (Cambridge, England)*, **139**(6), 1164–1174.

Busti, S., Gotti, L., Balestrieri, C., *et al.* (2012). Overexpression of far1, a cyclin dependent kinase inhibitor, induces a large transcriptional reprogramming in which rna synthesis senses far1 in a sfp1-mediated way. *Biotechnology Advances*, **30**(1), 185–201.

Chen, Y. and Xu, D. (2004). Global protein function annotation through mining genome-scale data in yeast saccharomyces cerevisiae. *Nucleic Acids Res*, **32**(21), 6414–6424.

Chin, S. L., Marcus, I. M., Klevecz, R. R., *et al.* (2012). Dynamics of oscillatory phenotypes in saccharomyces cerevisiae reveal a network of genome-wide transcriptional oscillators. *FEBS Journal*, **279**(6), 1119–1130.

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from proteinprotein interactions. *Bioinformatics*, **22**(13), 1623–1630.

Colombani, J., Andersen, D. S., and Lopold, P. (2012). Secreted peptide dilp8 coordinates drosophila tissue growth with developmental timing. *Science*, **336**(6081), 582–585.

Costanzo, M., Baryshnikova, A., Bellay, J., *et al.* (2010). The Genetic Landscape of a Cell. *Science*, **327**(5964), 425–431.

Eppig, J. T., Blake, J. A., Bult, C. J., *et al.* (2007). The mouse genome database (mgd): new features facilitating a model system. *Nucleic Acids Research*, **35**, 630–637.

Finn, R. D., Mistry, J., Schuster-Bckler, B., *et al.* (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, **34**, 247–251.

Frasca, M., Bertoni, A., Re, M., *et al.* (2013). A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks*, **43**(0), 84 – 98.

Guruharsha, K. G., Rual, J., Zhai, B., *et al.* (2011). A Protein Complex Network of Drosophila melanogaster. *Cell*, **147**(3), 690–703.

Hamosh, A., Scott, A. F., Amberger, J. S., *et al.* (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**(Database-Issue), 514–517.

Joshi, T., Chen, Y., Becker, J. M., *et al.* (2004). Genome-scale gene function prediction using multiple sources of high-throughput data in yeast saccharomyces cerevisiae. *OMICS*, **8**(4), 322–333.

Kaake, R. M., Milenkovi, T., Przulj, N., *et al.* (2010). Characterization of cell cycle specific protein interaction networks of the yeast 26s proteasome complex by the qtax strategy. *J Proteome Res*, **9**(4), 2016–29.

Kasprzyk, A., Keefe, D., Smedley, D., *et al.* (2004). EnsMart: A generic system for fast and flexible access to biological data. *Genome Research*, **14**(1), 160–169.

Kim, W., Krumpelman, C., and Marcotte, E. (2008). Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biology*, **9**(Suppl 1), S5+.

Kovacs, L. A. S., Mayhew, M. B., Orlando, D. A., *et al.* (2012). Cyclin-dependent kinases are regulators and effectors of oscillations driven by a transcription factor network. *Molecular Cell*, **45**(5), 669 – 679.

Lee, H., Tu, Z., Deng, M., *et al.* (2006). Diffusion kernel-based logistic regression models for protein function prediction. *Omics : a journal of integrative biology*, **10**(1), 40–55.

Libuda, D. E. and Winston, F. (2010). Alterations in dna replication and histone levels promote histone gene amplification in saccharomyces cerevisiae. *Genetics*, **184**(4), 985–97.

Lundberg, L. E., Fiqueiredo, M., Stenberg, P., *et al.* (2012). Buffering and proteolysis are induced by segmental monosomy in Drosophila melanogaster. *Nucleic Acids Research*.

Mostafavi, S. and Morris, Q. (2010). Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, **26**(14), 1759–1765.

Mostafavi, S., Ray, D., Farley, D. W., *et al.* (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, **9**(Suppl 1), S4+.

Mulder, N. J., Apweiler, R., Attwood, T. K., *et al.* (2005). Interpro, progress and status in 2005. *Nucleic Acids Res*, **33**, 201–205.

Muller, P., Park, S., Shor, E., *et al.* (2010). The conserved bromo-adjacent homology domain of yeast orc1 functions in the selection of dna replication origins within chromatin. *Genes Dev*, **24**(13), 1418–33.

Obozinski, G., Lanckriet, G., Grant, C., *et al.* (2008). Consistent probabilistic outputs for protein function prediction. *Genome Biol*, **9 Suppl 1**, S6.

O'Brien, K. P., Remm, M., and Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, **33**(Database issue).

Ossareh-Nazari, B., Bonizec, M., Cohen, M., *et al.* (2010). Cdc48 and Ufd3, new partners of the ubiquitin protease Ubp3, are required for ribophagy. *Embo Reports*, **11**, 548–554.

Qi, Y., Seetharaman, J. K., and Joseph, Z. B. (2007). A mixture of feature experts approach for protein-protein interaction prediction. *BMC Bioinformatics*, **8**(Suppl 10), S6+.

Sanz, A. B., Garcia, R., Rodriguez-Pena, J. M., *et al.* (2012). Chromatin remodeling by swi/snf complex is essential for transcription mediated by the yeast cell wall integrity mapk pathway. *Molecular Biology of the Cell*.

Siddiqui, A. S., Khattra, J., Delaney, A. D., *et al.* (2005). A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc Natl Acad Sci U S A*, **102**(51), 18485–18490.

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**(3), 405–420.

Stark, C., joe Breitkreutz, B., Reguly, T., *et al.* (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, (Database-Issue), 535–539.

Su, A. I., Wiltshire, T., Batalov, S., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, **101**, 6062–6067.

Tian, W., Zhang, L., Tasan, M., *et al.* (2008). Combining guilt-by-association and guilt-by-profiling to predict saccharomyces cerevisiae gene function. *Genome biology*, **9 Suppl 1**(Suppl 1), S7+.

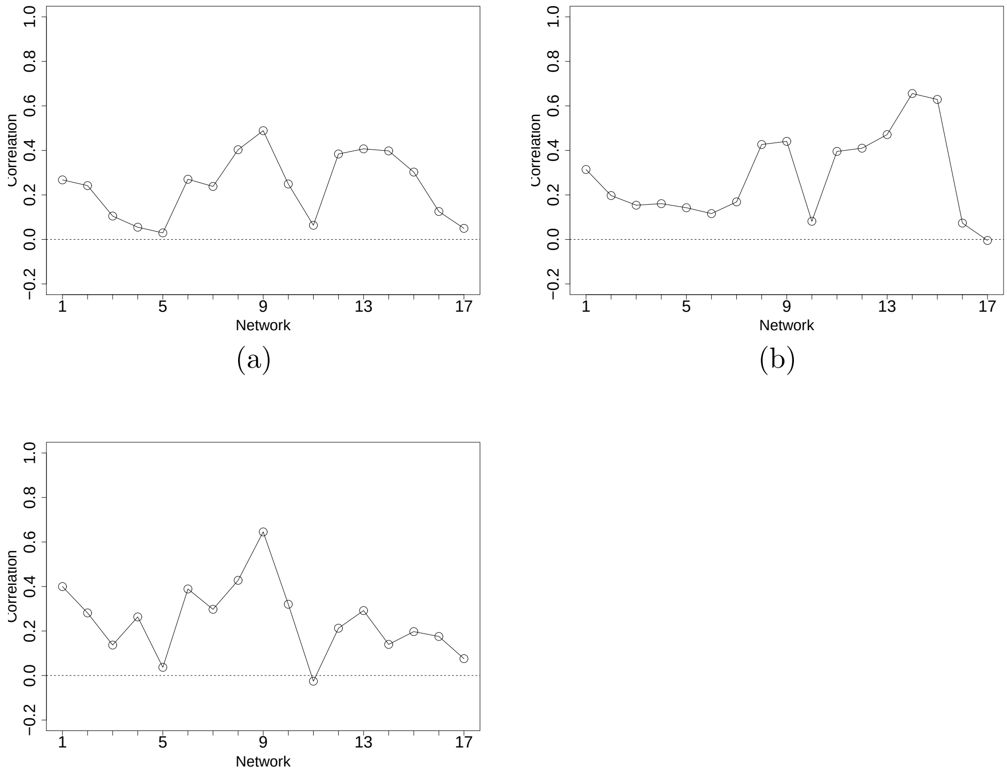
Wheeler, D. L., Barrett, T., Benson, D. A., *et al.* (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **35**(Database issue).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Journal of Computational Biology*, **1**(6), 80–83.

Yu, J., Pacifico, S., Liu, G., *et al.* (2008). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, **9**(1), 461+.

Zhang, W., Morris, Q., Chang, R., *et al.* (2004). The functional landscape of mouse gene expression. *Journal of Biology*, **3**(5), 21+.

List of Figures

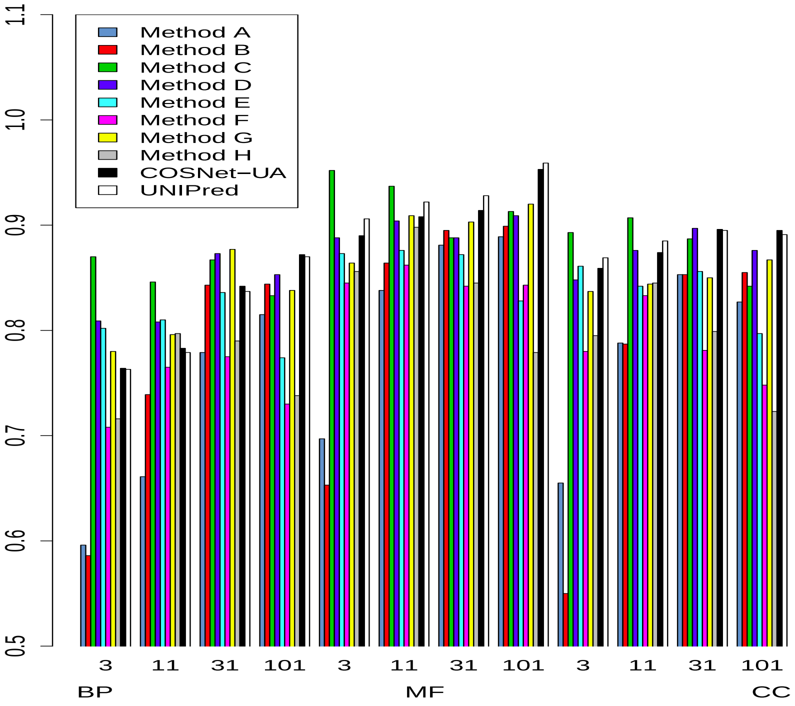


(a)

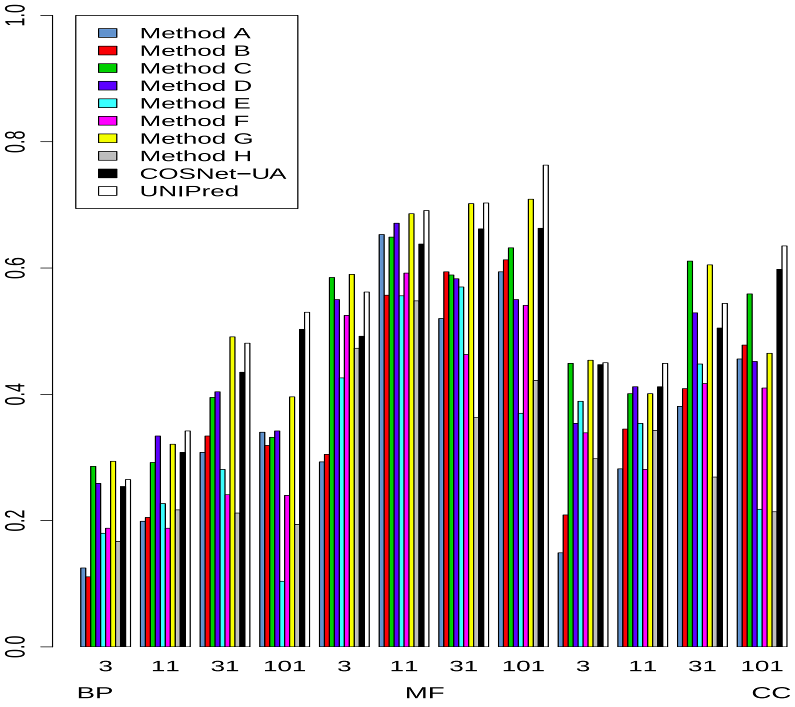
(b)

(c)

Figure S1: (a) Correlation of *COSNet* prediction per class F-score on 17 single mouse networks and the corresponding per class weights assigned by *UNIPred* by considering term belonging solely to (a) Biological Process, (b) Molecular Function and (c) Cellular Component ontologies.

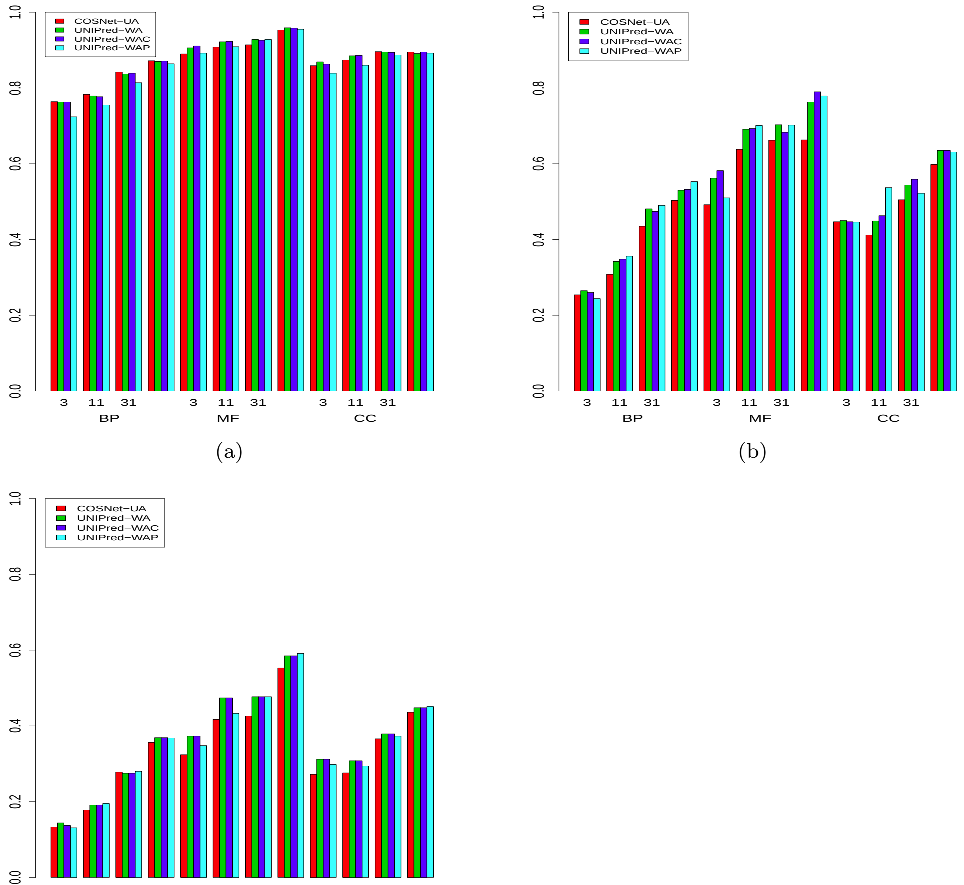


(a)



(b)

Figure S2: Comparison of the MouseFunc methods and *UNIPred* in terms of AUC (a) and P20R (b) averaged across the twelve considered GO categories.

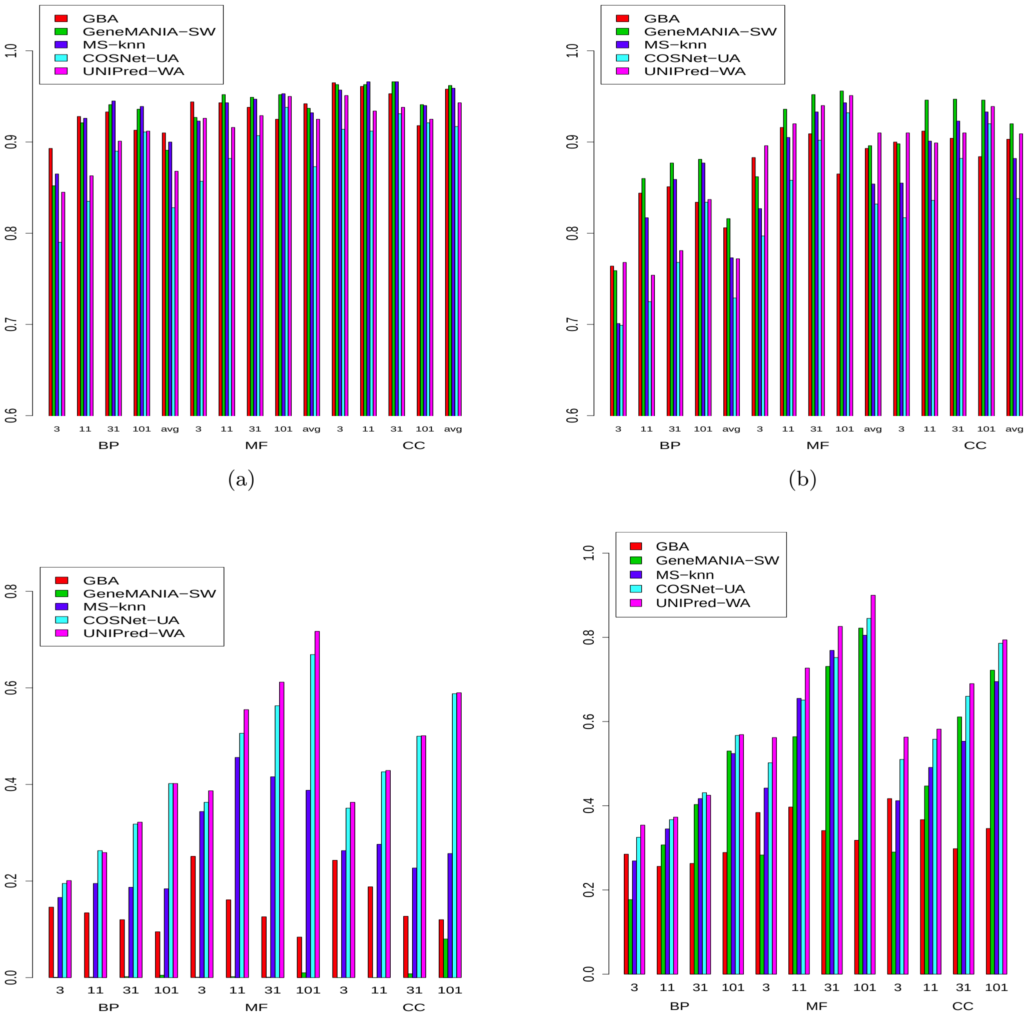


(a)

(b)

(c)

Figure S3: Comparison of *UNIPred* integration strategies on mouse data in terms of AUC (a), P20R (b) and F-score (c).



(a)

(b)

(c)

(d)

Figure S4: Performance on yeast and fly data averaged by ontology and cardinality of annotations of GO terms. AUC for yeast (a) and fly (b), F-score (c) and P20R (d) for fly. ”Avg” corresponds to the ontology average results irrespective of the cardinality of the annotations.

List of Tables

Index	**Network**	**Reference**
1	PPIbin	Brown and Jurisica (2005)
2	PPIdist	Brown and Jurisica (2005)
3	GO.BP	MouseFunc GO annotations
4	Su	Su *et al.* (2004)
5	Zhang	Zhang *et al.* (2004)
6	Sageavg	Siddiqui *et al.* (2005)
7	Sagesum	Siddiqui *et al.* (2005)
8	Pfam	Finn *et al.* (2006)
9	Interpro	Mulder *et al.* (2005)
10	Pheno	Eppig *et al.* (2007)
11	Omim	Wheeler *et al.* (2007); Hamosh *et al.* (2005)
12	Phylobin	Kasprzyk *et al.* (2004)
13	Inpscore	O'Brien *et al.* (2005)
14	Inpbin	O'Brien *et al.* (2005)
15	Phyloscore	Kasprzyk *et al.* (2004)
16	GO.MF	MouseFunc GO annotations
17	GO.CC	MouseFunc GO annotations

Table S1: Adopted indexes and references of Mouse networks.

Method	**Authors**	**Algorithm**
Method A	G. Obozinski, C. Grant, J. Qiu, G. Lanckriet, M. I. Jordan and W. S. Noble	Calibrated ensembles of SVMs (Obozinski *et al.*, 2008)
Method B	H. Lee, M. Deng, T. Chen, F. Sun	An Integrated Kernel-Logistic Regression Method for Protein Function Prediction (Lee *et al.*, 2006)
Method C	S. Mostafavi, D. W. Farley, C. Grouios, D. Ray and Q. Morris	GeneMANIA (Mostafavi *et al.*, 2008)
Method D	Y. Guan, C. L. Myers, O. G. Troyanskaya	Multi-label hierarchical classification (Barutcuoglu *et al.*, 2006) and Bayesian integration of diverse data sources (Aguilar *et al.*, 2010)
Method E	W. K. Kim, C. Krumpelman, E. Marcotte	Combination of classifier ensemble and gene network (Kim *et al.*, 2008)
Method F	T. Joshi, C. Zhang, G. N. Lin, D. Xu	GeneFAS (Chen and Xu, 2004; Joshi *et al.*, 2004)
Method G	W. Tian, M. Tasan, F. D. Gibbons, F. P. Roth	Funckenstein (Tian *et al.*, 2008)
Method H	Y. Qi, J. K. Seetharaman and Z. B. Joseph	Protein Function Prediction Using 'Query Retrieval' Methods (Qi *et al.*, 2007)

Table S2: MouseFunc I participants.

Type	**Source**	**Genes**
Co-expression	Busti *et al.* (2012)	5436
Co-expression	Chin *et al.* (2012)	5585
Co-expression	Sanz *et al.* (2012)	5585
Co-expression	Kovacs *et al.* (2012)	5585
Genetic interactions	Aguilar *et al.* (2010)	321
Genetic interactions	Alamgir *et al.* (2010)	90
Genetic interactions	Costanzo *et al.* (2010)	4346
Genetic interactions	Libuda and Winston (2010)	143
Genetic interactions	BioGRID (Stark *et al.*, 2006)	4280
Physical interactions	Breitkreutz *et al.* (2010)	887
Physical interactions	Kaake *et al.* (2010)	332
Physical interactions	Muller *et al.* (2010)	266
Physical interactions	Ossareh-Nazari *et al.* (2010)	406
Physical interactions	BioGRID (Stark *et al.*, 2006)	4752
Shared protein domains	InterPro (Apweiler *et al.*, 2001)	3964
Shared protein domains	Pfam (Sonnhammer *et al.*, 1997)	3541
GO association network	GO BP annotations	5775
GO association network	GO MF annotations	5775
GO association network	GO CC annotations	5775

Table S3: Yeast networks description.

Type	**Source**	**Genes**
Co-expression	Baradaran-Heravi *et al.* (2012)	8857
Co-expression	Busser *et al.* (2012)	8857
Co-expression	Colombani *et al.* (2012)	8857
Co-expression	Lundberg *et al.* (2012)	8857
Genetic interactions	BioGRID (Stark *et al.*, 2006)	929
Genetic interactions	Yu *et al.* (2008)	1414
Physical interactions	Guruharsha *et al.* (2011) A	1866
Physical interactions	Guruharsha *et al.* (2011) B	3833
Physical interactions	BioGRID (Stark *et al.*, 2006)	558
Shared protein domains	InterPro (Apweiler *et al.*, 2001)	5627
GO association network	GO BP annotations	9631
GO association network	GO MF annotations	9631
GO association network	GO CC annotations	9631

Table S4: Fly networks description.

Method	**AUC**			**P20R**			**F-score**		
	BP	**MF**	**CC**	**BP**	**MF**	**CC**	**BP**	**MF**	**CC**
Method A	0.672	0.796	0.766	0.204	0.470	0.284	0.113	0.340	0.163
Method B	0.709	0.789	0.737	0.204	0.469	0.334	0.113	0.328	0.197
Method C	**0.859**	**0.929**	**0.890**	0.314	0.607	0.479	0.175	0.406	0.281
Method D	0.825	0.894	0.872	0.320	0.591	0.423	0.140	0.346	0.229
Method E	0.809	0.870	0.845	0.209	0.492	0.366	0.028	0.170	0.208
Method F	0.742	0.848	0.795	0.203	0.529	0.343	0.104	0.340	0.198
Method G	0.810	0.890	0.846	0.351	**0.653**	0.467	0.188	0.434	0.231
Method H	0.759	0.859	0.805	0.194	0.462	0.297	0.091	0.322	0.143
COSNet-UA	0.795	0.906	0.887	0.329	0.587	0.465	0.196	0.392	0.314
UNIPred-WA	0.792	0.920	0.883	**0.356**	0.648	0.494	**0.205**	**0.443**	0.342
UNIPred-WAP	0.764	0.911	0.863	**0.356**	0.630	**0.519**	0.202	0.422	0.331
UNIPred-WAC	0.781	0.900	0.870	0.316	0.558	0.454	0.202	0.434	**0.350**

Table S5: Prediction performance in terms of AUC, P20R and F-score for the MouseFunc methods and *UNIPred* with integration strategies WA, WAP and WAC. UA is the unweighted average sum integration. The values are averaged across the three GO domains BP, MF, CC. The best results are reported in boldface.

Method	**AUC**			**P20R**			**F-score**		
	BP	**MF**	**CC**	**BP**	**MF**	**CC**	**BP**	**MF**	**CC**
Method A	+	+	+	+	+	+	+	+	+
Method B	+	+	+	+	+	+	+	+	+
Method C	−	=	=	=	=	=	+	+	+
Method D	−	+	+	+	+	+	+	+	+
Method E	=	+	+	+	+	+	+	+	+
Method F	+	+	+	+	+	+	+	+	+
Method G	−	+	+	=	=	=	+	=	+
Method H	+	+	+	+	+	+	+	+	+

Table S6: Statistically significant differences at $\alpha = 0.01$ significance level between *UNIPred* and the methods participating to the MouseFunc challenge. The symbol "+" means a difference statistically significant in favour of *UNIPred*, "=" means no statistically significant difference, "-" means a difference statistically significant in favour of the other method.