BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–3

Fast Integration of Multiple Networks for Predicting Protein Function with Limited Annotation

Sara Mostafavi [1,2] and Quaid Morris [1,2,*]

[1]Department of Computer Science,
[2]Center for Cellular and Biomolecular Research, University of Toronto.

1 SUPPLEMENT

Regularization for the GeneMANIA algorithm

We investigate the effect of four different forms of regularizations on the GeneMANIA algorithm: a) ridge with uniform prior, b) ridge with mean prior, c) LASSO [7] and d) elastic net [8]. Standard ridge regression with regularization parameter α_2 corresponds to the maximum a posteriori (MAP) estimation using a zero-mean Gaussian $p(\vec{\mu}) \sim N(0, \alpha_2^{-1} I)$ on the weights. We can also use a non-zero mean Gaussian $N(\vec{v}, S)$ as the prior on the weights, where $\vec{v}, S$ are the prior mean and diagonal covariance matrix; in this setting the network weights are encouraged to be close to a vector $\vec{v}$. In ridge with mean prior we set $\vec{v}$ to the average weights assigned by unregularized linear regression in predicting a large number of GO categories. In ridge with uniform, we set $v_d = 1$ for all networks.

As shown in the main text, ridge with mean and uniform prior perform better than LASSO and unregularized linear regression. One explanation for this observation is regularization methods that lead to many weights of zero are too selective and often identify only a few relevant networks. Figure S1 shows the proportion of categories for which different networks were assigned a positive weight. As shown, with LASSO, a few selected networks are assigned a non-zero weights in a large number of the GO categories.

Grouping GO categories for SW

SW simultaneously optimizes the network weights to a group of GO categories. We have investigated four different methods for grouping GO categories: Tree^0, Tree^1, Size, and Clust (see Figure S2). In Tree^0 we fit SW to all GO categories in the same GO hierarchy (e.g. BP) with 3-300 annotations, in contrast, in Tree^1 we fit the weights to all GO categories in the same hierarchy that have the same parent category which has 300 or less annotations (where each category is considered an ancestor of itself). In Size, we group GO categories based on their number of annotation and hierarchy; for example, we fit one set of weights to all BP categories which have 3-10 annotations. In Clust, we use hierarchical agglomerative clustering (single linkage) with Pearson Correlation Coefficient (PCC), of binary vectors which represent the gene annotated to categories, as the similarity metric to cluster GO categories. We investigate three different clusterings with increasing number of clusters $n = \{3, 10, 20\}$. Note that we only consider GO categories with 3-300 annotations; this is because GO categories with fewer annotations have too few examples for training and larger GO categories are too general. Once we compute the network weights based on a group of categories, we construct one composite network and use it to predict all categories in the given group.

We have compared the performance of composite networks constructed by SW when using the above four groupings of GO categories (see Figure S3). As shown, the various versions of SW perform similarly, however, SW-Tree^0 slightly out-performs the rest. In addition, Figure S3 shows that as the grouping of GO categories becomes more specific (for example with Tree^1 and Clust$^{\#n}$), the generalizability of SW decreases. In Tree^1, each group consists of the *ancestor* at 300-annotation level with all of its descendants; within these group 10 of 1188 categories where singletons (did not have any descendants or ancestors at 300-annotation level) and 414 of 1188 categories where placed in a group with 10 or more categories. If we remove these singleton categories the performance of Tree^1 is still lower than that of Tree^0 (average area under the ROC curve of 0.8067 for Tree^1 compared to 0.8273 for Tree^0).

Regularizing SW We have also investigated the performance of regularized versions of SW. Figure S4 shows the performance of SW-Tree^0 with 1) ridge, 2) ridge with uniform prior (ridge-uniform), and 3) ridge with mean prior (ridge-mean). For ridge, we set the regularization parameter using cross-validation. As shown, and expected, ridge-uniform degrades the performance. SW with ridge performs slightly better (but not significant) than the unregularized version.

Data sources for constructing networks

We constructed a large number of networks for yeast, fly, mouse, human, and E. coli. For yeast, majority of our networks are constructed from protein interactions (PPI) and genetic interactions (GI) that we downloaded from the BIOGRID database [6] (downloaded Nov. 2008). We note that we don't include BioGRID interactions that are derived from small scale experiments (those studies that reported less than 1000 interaction), as such interactions are often used to derive annotations. In addition, we have included networks constructed from gene expression and protein localization (see Table 1). For interaction based datasets, we have include both a direct and a correlation based network. For genetic interaction data, we have constructed networks for both *positive* and *negative* genetic interaction when possible.

In particular, for yeast, we constructed 44 networks which include interactions derived from gene expression, protein and genetic interaction (downloaded from BIOGRD [6]), and protein localization. For mouse, we construct 10 networks using the data

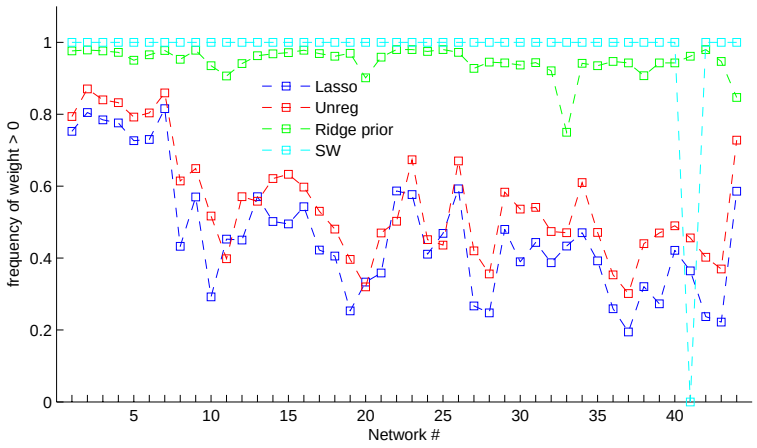

Fig. S1. Figure shows the proportion of times that a given network was assigned a positive weights with LASSO, unregularized linear regression, and ridge with mean prior.

provided by MouseFunc [4][1] (See Table 2). For human, we have constructed 8 networks based on 7 data sources provided by HPRD [5] (downloaded Sep. 2007), in addition we also included a co-expression network derived from a tissue expression dataset (See Table 1). For fly, we have constructed 38 networks; 32 of these are constructed from expression data downloaded from GEO [1], 4 networks are constructed from protein interaction data, and 2 networks are constructed from domain composition data (see Table 5). For E.coli, we use 7 networks provided by [2] (see Table 3) which include protein interaction, co-expression, co-inheritance, and shared sequence features.

Constructing functional association networks

For network-based data (e.g. protein interaction), we use both a direct interaction network and a correlation based network using the PCC on the frequency-corrected data (as done in [3]).

Co-expression networks Before constructing co-expression networks from the gene expression datasets we standardize each feature (condition) by subtracting the mean and dividing by the standard deviation. To construct co-expression networks we use Pearson correlation coefficient (PCC) to measure similarities between gene expression profiles. In particular, in the co-expression network W_d, the link (edge) between gene i and i is set to their pairwise correlation coefficient r_{ij}. We set to zero all negative r_{ij}. Correlation networks tend to be dense; a gene may have a positive PCC with a large number of other genes. However, efficient classification requires that we sparsify the correlation-derived networks. This is because the time and space complexity of our algorithm grows with the number of non-zero elements in the networks. In practice, we sparsified each network in yeast, fly, and human by keeping the top S=100 interactions for each gene and setting the rest to zero; for mouse we use S=50 interactions (as the mouse networks cover more genes and thus tend to be denser).

[1] available from http://hugheslab.med.utoronto.ca/supplementary-data/mouseFunc_I/

We note that previous studies have shown that sparsification of functional association networks does not degrade accuracy of gene function prediction [3]. For E. coli, we sparsify the co-expression network using a z-score cutoff ≥ 2.58 (p=0.001) (as done in [2]). We then normalized all our networks by: $\tilde{W}_d = D_d^{-1/2} W_d D_d^{-1/2}$ where D_d is the diagonal row sum matrix of W_d. After combining the networks, we also normalize the composite network W^* by left and right multiplying with $(D^*)^{-1/2}$.

Predicting cellular component and molecular function

In addition to predicting BP categories, we also report the performance in predicting cellular component (CC) and molecular function (MF) categories. Figure S5 and S6 shows the performance of SW, Uniform, and unregularized linear regression in predicting CC and MF categories in yeast, mouse, human, fly, and E. coli. Similar to our previous observations, this figure shows that SW improves the performance in categories with small number of annotations ([3-10]) and overall performance as these small categories make up the majority of GO functions.

REFERENCES

[1]R. Edgar, M. Domrachev, and AE. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, pages 207–210, 2002.

[2]P. Hu, SC. Janga, M. Babu, JJ. Daz-Meja, G. Butland, and W. Yang et al. Global functional atlas of escherichia coli encompassing previously uncharacterized proteins. *PLoS Biology*, 7:e96, 2009.

[3]S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(Suppl 1):S4, 2008.

[4]L. Pena-Castillo, T. Murat, C.L. Myers, H. Lee, T. Joshi, C. Zhang, and et al. A critical assessment of *Mus musculus* gene

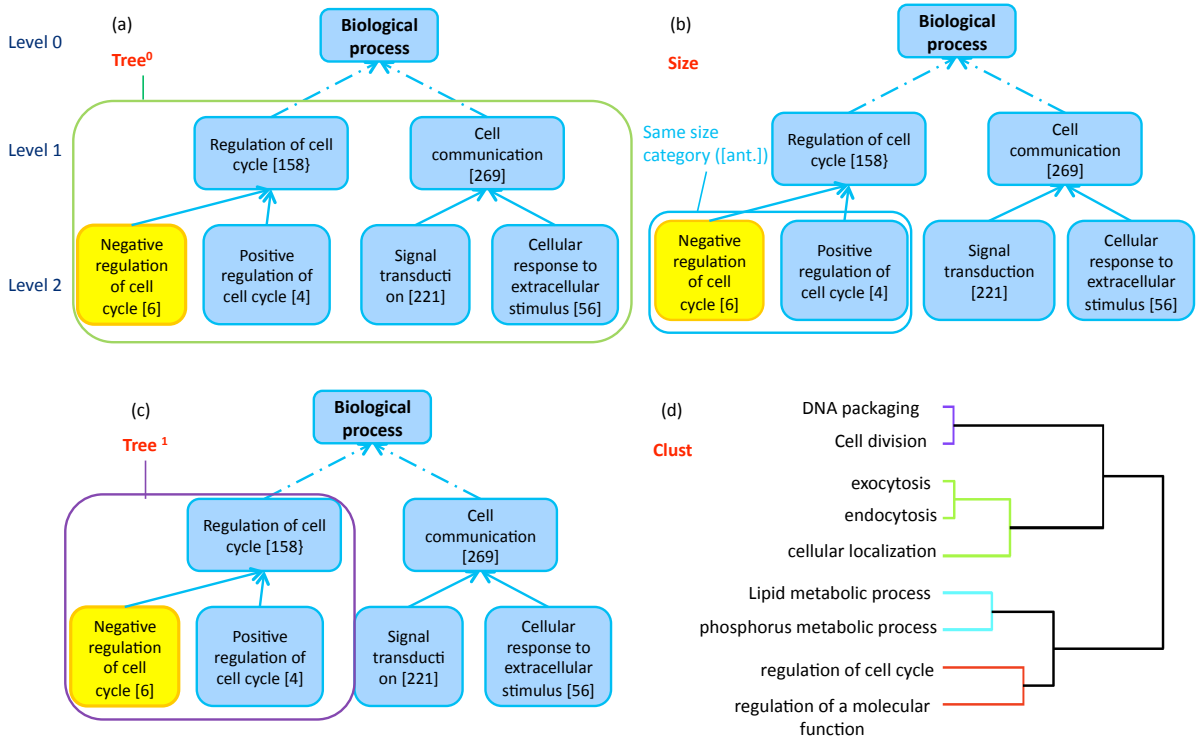

Fig. S2. We define four different methods for grouping GO categories: (a) Tree^0: all categories in the same hierarchy in GO, (b) Size: all categories in the same GO hierarchy with the similar annotation level where we define 4 annotations levels: [3-10], [11-30], [31-100], and [101-300], (c) Tree^1: all categories in the same hierarchy with the same ancestor which has no more than 300 annotations (each term is considered an ancestor of itself), and (d) Clust$^{\#n}$: all categories in the same hierarchy which are clustered together using hierarchical clustering with n clusters (we vary the number of clusters $n = \{3, 10, 20\}$)

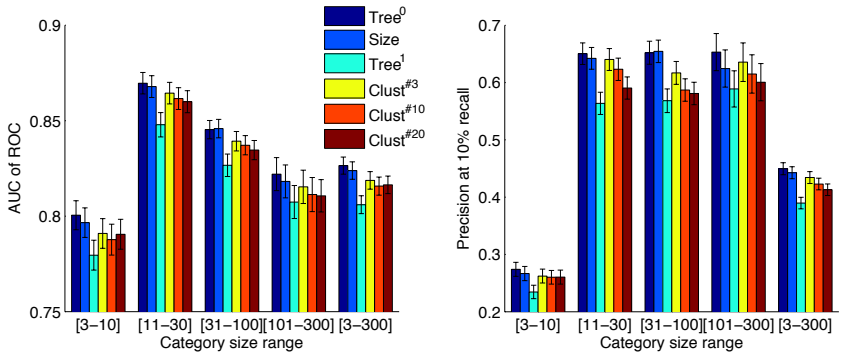

Fig. S3. Performance of SW with four different groupings of GO categories in predicting BP gene function with 44 yeast netowrks. The performance is shown in terms of AUC of ROC and precision at 10% recall using 3-fold CV.

function prediction using integrated genomic evidence. *Genome Biology*, 9(Suppl 1):S2, 2008.

[5]TS Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, and et al. Human protein reference database – 2006 update. *Nucleic Acids Research*, 34(Database Issue):D411–D414, 2006.

[6]C. Stark, BJ. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 1(Database issue):D539–D539, 2006.

[7]R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistics Society B.*, 58(1):267–288, 1996.

[8]H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of Royal Statistics Society*, 67(2):301–320, 2005.

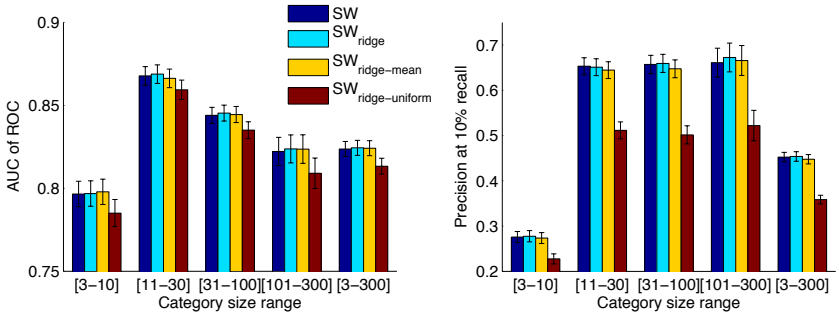

Fig. S4. Performance of SW with 1) ridge, 2) ridge with uniform and 3) ridge with mean prior. The performance is shown in terms of mean AUC of ROC and precision at 10% recall in predicting 1,188 GO BP categories with 44 yeast networks.

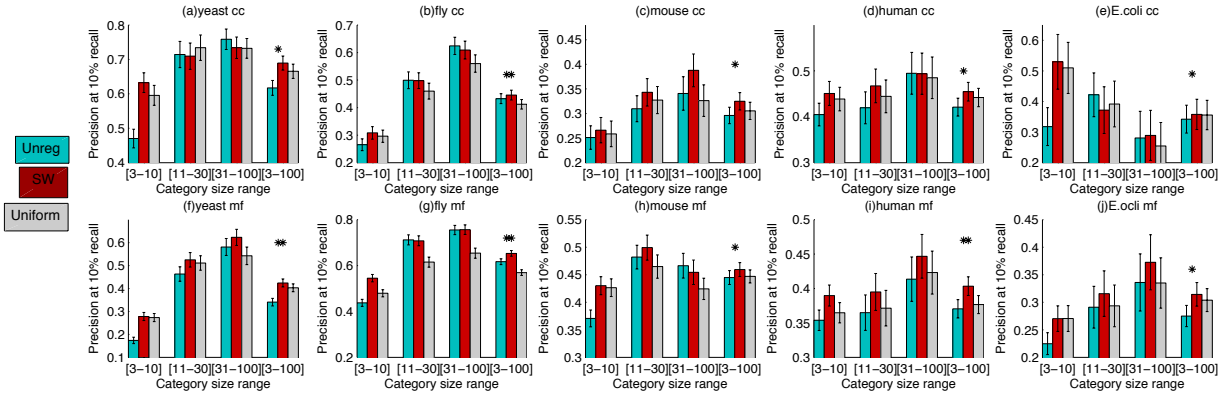

Fig. S5. Performance of SW, Uniform network combination, and unregularized linear regression in predicting CC (top row (a) through (e)) and MF (bottom row (f) through (j)) GO categories in yeast, fly, mouse, human, and E.coli. Performance is measured in terms of precision at 10% recall.

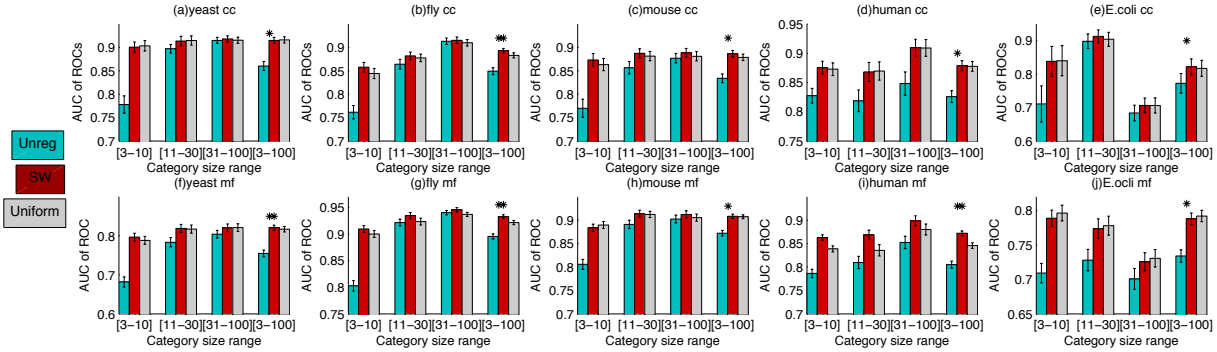

Fig. S6. Performance of SW, Uniform network combination, and unregularized linear regression in predicting CC (top row (a) through (e)) and MF (bottom row (f) through (j)) GO categories in yeast, fly, mouse, human, and E.coli. Performance is measured in terms of area under the ROC curve (AUC of ROC)

Table 1. Data sources used for constructing 44 networks for yeast.

PMID	Publication	Number of networks	Network # in the main text
10657304	Roberts et al. Science 2000	1	4
10929718	Hughes et al. Cell 2000	1	7
11102521	Gasch et al. Mol Biol Cell. 2000	1	2
11805826	Gavin et al. Nature 2002	2	8,17
11805837	Ho et al. Nature 2002	2	9,18
14562095	Huh et al. Nature 2003	1	1
14718668	Giaever et al, PNAS, 2004	1	44
14764870	Tong et al. Science 2004.	2	26,35
16093310	Miller et al. PNAS 2005	2	10,19
16269340	Schuldiner et al. Cell 2005 (PS)	4	27,28,36,37
16319894	Ptacek et al. Sciene 2005	2	11,20
16429126	Gavin et al. Nature 2006	2	12,21
16487579	Pan et al. Cell 2006 (SL)	4	29,30,38,39
16554755	Krogan et al. Nature 2006	2	13,22
16880382	Chua et al. PNAS 2006	2	5,6
17200106	Collins et al. Mol Cell Proteomics 2007	2	14,23
17314980	Collins et al. Nature 2007 (PS)	4	31,32,40,41
17923092	McClellan Cell 2007	2	33,42
18467557	Tarassov et al. Sciene 2008	2	15,24
18676811	Lin et al. Gens Dev 2008	2	34,43
18719252	Yu et al. Science 2008	2	16,25
9843569	Spellman et al. Mol. Biol. Cell. 1998	1	3

Table 2. Data sources used for constructing 10 networks in Mouse

PMID	Publication	Number of networks
1558831	Zhang et al. Journal of Biology 2004	1
15075390	Su et al. PNAS 2004	1
SAGE Lib.	Tag counts	1
OPHID	Protein interactions	1
Pfam	Domain composition	1
InterPro	Domain composition	1
MGI	Phenotype	1
bioMART	Phylogenetic profile	1
Inparanoid	Phylogenetic profile	1
OMIM	Disease genes	1

Table 3. Data sources used for constructing 7 networks in E.coli

PMID	Publication	Type	Number of networks
19402753	Hu et al. Plos Biol, 2009	Protein interaction	1
M3D database	M3D database v4Build5 affy	Co-expression	1
19402753	Hu et al. Plos Biol, 2009	Shared Operons	1
19402753	Hu et al. Plos Biol, 2009	Gene fusion	1
19402753	Hu et al. Plos Biol, 2009 (updated version)	Co-inheritance	1
19402753	Hu et al. Plos Biol, 2009	Distance in chromosome 1	1
19402753	Hu et al. Plos Biol, 2009	Distance in chromosome features 2	1

Table 4. Data sources used for constructing 8 networks in Human

Source	Type	Number of networks
HPRD	domain composition	1
HPRD	complexes	1
HPRD	protein-dna/rna-interaction	1
HPRD	post transcriptional modification	1
HPRD	tissue expression	1
HPRD	protein interaction	1
HPRD	OMIM disease	1
PMID:15075390	tissue expressionn	1

Table 5. Data sources used for constructing 38 networks for fly.

Accession	Source	Number of networks
GDS2674	Gene Expression Omnibus	1
GDS2399	Gene Expression Omnibus	1
GDS2485	Gene Expression Omnibus	1
GDS516	Gene Expression Omnibus	1
GDS1937	Gene Expression Omnibus	1
GDS2675	Gene Expression Omnibus	1
GDS1842	Gene Expression Omnibus	1
GDS2504	Gene Expression Omnibus	1
GDS2272	Gene Expression Omnibus	1
GDS1526	Gene Expression Omnibus	1
GDS23	Gene Expression Omnibus	1
GDS444	Gene Expression Omnibus	1
GDS732	Gene Expression Omnibus	1
GDS443	Gene Expression Omnibus	1
GDS667	Gene Expression Omnibus	1
GDS664	Gene Expression Omnibus	1
GDS653	Gene Expression Omnibus	1
GDS1690	Gene Expression Omnibus	1
GDS2665	Gene Expression Omnibus	1
GDS2071	Gene Expression Omnibus	1
GDS2479	Gene Expression Omnibus	1
GDS1911	Gene Expression Omnibus	1
GDS1739	Gene Expression Omnibus	1
GDS2673	Gene Expression Omnibus	1
GDS1977	Gene Expression Omnibus	1
GDS1877	Gene Expression Omnibus	1
GDS1686	Gene Expression Omnibus	1
GDS2830	Gene Expression Omnibus	1
GDS2784	Gene Expression Omnibus	1
GDS2228	Gene Expression Omnibus	1
GDS1395	Gene Expression Omnibus	1
GDS602	Gene Expression Omnibus	1
PMID14605208	BIOGRID	2
PMID15575970	BIOGRID	2
Interpro domains	Interpro	1
Pfam domains	Pfam	1